Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination

Awopeju, O. A. , PhD)
Afolabi, E. R. I. , Prof.
Obafemi Awolowo University, Ile-Ife, Nigeria
Department of Educational Foundations and Counselling

doi: 10.19044/esj.2016.v12n28p263 URL:http://dx.doi.org/10.19044/esj.2016.v12n28p263

Abstract

The study compared Classical Test Theory (CTT) and Item Response Theory (IRT)-estimated item difficulty and item discrimination indices in relation to the ability of examinees in Senior School Certificate Examination (SSCE) in Mathematics with a view to providing empirical basis for informed decisions on the appropriateness of statistical and psychometric tests.

The study adopted ex-post-facto design. A sample of 6,000 students was selected from the population of 35,262 students who sat for the NECO SSCE Mathematics Paper 1 in 2008 in Osun State, Nigeria. An instrument consisting of 60-multiple-choice items, May/June 2008 NECO SSCE Mathematics Paper 1 was used. Three sampling plans: random, gender and ability sampling plans were employed to study the behaviours of the examinees scores under the CTT and IRT measurement frameworks. BILOG-MG 3 was used to estimate the indices of item parameters and SPSS 20 was used to compare CTT- and IRT-based item parameters.

The results showed that CTT-based item difficulty estimates and one-parameter IRT item difficulty estimates were comparable (the correlations were generally in the -0.702 to -0.988 range in large sample and -0.622 to -0.989 range in small sample). Results also indicated that CTT-based and two-parameter IRT-based item discrimination estimates were comparable (the correlations were in the 0.430 to 0.880 ranges in large sample and 0.531 to 0.950 range in small sample).

The study concluded that CTT and IRT were comparable in estimating item characteristics of statistical and psychometric tests and thus could be used as complementary procedures in the development of national examinations

Keywords: Item Response Theory, Classical Test Theory, Item Difficulty, Item Discrimination

Introduction

Students' poor performance in Mathematics over the years has been attributed to the fact that the subject is difficult and that the teaching methodology has not been appropriate. However, some authors and researchers (Ashikhia, 2010; Adebule, 2004; Aremu & Sokan, 2003) have identified various factors that affect students' performances in Mathematics especially at the secondary school level. Prominent among these factors are the nature of the test items and the learners' characteristics. The performance of an examinee on a test item can be predicted (or explained) by the ability of the examinee and characteristics of the item.

A test can be studied from different perspectives and the items in the test can be evaluated according to different theories. Two of such theories are the classical test theory (CTT) and the item response theory (IRT). These theories are the two major frameworks that are used in educational measurement to develop, evaluate and study test items. These frameworks are based on different assumptions and use different statistical approaches. They are concerned not only to develop, evaluate, or determine the reliability and validity of tests but also to holistically improve the quality of test items. CTT was originally the leading framework for developing and analyzing standardized tests. Later, IRT was developed to compliment the role of CTT.

CTT is based on the assumption that an examinee has an observed score and a true score. The observed score of a test-taker is usually seen as a combination of an estimate of the true scores of that test-taker, plus/minus some unobservable error. The true score reflects what the test-taker actually knows, but it is always contaminated by different sources of errors. CTT utilizes measures of item characteristics, item difficulty and item discrimination, the values of which are dependent upon the distribution of examinee proficiency within a sample. Although the assumptions upon which classical test theory is based allow it to be applied to an assortment of test construction situations, these same assumptions appear to create weaknesses in the classical test theory model. The CTT based statistical indices are easy to compute, manipulate and understand by lay persons, but they vary from sample to sample. The major advantage of CTT is its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). While CTT has proven very useful in test development, the two statistics that form its cornerstones, item difficulty and item discrimination are both sample dependent. In particular, because the classical test theory model lacks information

regarding how an examinee is predicted to perform on a particular item, it cannot accommodate tests that target an examinee's proficiency level (Hambleton, Swaminattham & Rogers, 1991).

On the other hand, item response theory has become an important complement to CTT in development, interpretation and evaluation of tests and test items. The interest in IRT grew out of a combination of the concerns on the limitations inherent in CTT and the availability of computing systems. IRT has strong mathematical basis and depends on complex algorithms that are more efficiently solved via computer. It describes the relationship between an examinee's test performance and the traits assumed to underlie such performance on an achievement tests as a mathematical function called item characteristics curve (ICC) (Hambleton & Swaminathan, 1985; Harris, 1989). IRT primarily focuses on the item-level information in contrast to the CTT's primary focus on test level information. The relationship between examinee ability and performance on an item is described by one or more parameters depending on which IRT model is used. Despite the theoretical differences between IRT and CTT, little has been done to demonstrate empirically these differences in the measurement community.

The basic concern of test developers when constructing a test is the nature and quality of test items and how examinees respond to these items. The validity and the reliability of any test depend ultimately on the characteristics of its items. These characteristics are item difficulty and item discrimination. Test theories enable the prediction of outcomes of tests by identifying parameters of item difficulty, item discrimination and the ability of test takers.

Item difficulty is defined in both CTT and IRT in terms of the likelihood of correct response, not in terms of the perceived difficulty or amount of effort required. In CTT, the difficulty index, p, is the proportion of examinees who answer the item correctly. Discrimination of an item is the ability of a specific item to differentiate between high and low ability individuals on a test.

Some studies linking CTT and IRT item characteristics have shown signs of positive indications of a relationship that exist between them (Adedoyin, Nenty & Chilisa, 2008; Nukhet, 2002; Fan, 1998). Royce (2009) discovered that 2-parameter IRT model closely resembles CTT of the verbal and non-verbal test in terms of item characteristics. Paul and Sampo (2002) examined and compared item characteristics from the two measurement models and concluded that findings from past empirical investigations comparing IRT- and CTT-based item and person statistics should not be generalized to all educational and psychological tests.

Fan (1998), MacDonald and Paunonen (2002) among others have studied the empirical difference between these two models. They noted that

"because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between the IRT and CTT-based item and person statistics". However, findings by (Lawson, 1991; Skaggs & Lissitz, 1988; Stage, 1999) have all found differences between IRT and CTT theory estimates. On the other hand, some other researchers noted that the agreement between results from item-analyses performed within the two different frameworks, IRT and CTT, was very good. Also, Ojerinde (2013) found that the person and item statistics derived from the two frameworks are quite comparable.

Over the years, all research examining the empirical properties of CTT and IRT have failed to reveal consistent and demonstrable differences. Thus, findings from past empirical investigations comparing CTT- and IRT-based item and person statistics are not generalizable to all educational and psychological tests. Hence this study.

The objective of the study was to use the estimates of item difficulty and discrimination indices generated by IRT models to compare the estimates of item difficulty and discrimination indices generated by CTT.

Specific Objectives

The specific objectives of the study were to:

1. compare CTT and IRT estimated item difficulty values in relation to the abilities of examinees in Senior School Certificate Examination (SSCE) in Mathematics;
2. compare the CTT and IRT estimated item discrimination indices in SSCE in Mathematics;

Research Questions

In order to carry out this study, the following research questions were raised;

1. How comparable are the CTT-based and IRT-based item difficulty estimates?
2. Is the CTT-based and IRT-based item discrimination estimates comparable?

Method

The research design used was ex-post-facto. This design is relevant to this study because it allows analysis to be performed on existing data.

The population comprised all the students that sat for NECO senior school certificate Mathematics examination paper 1 (May/June, 2008) in Osun state, Nigeria. A computer-based simple random sample of responses of six thousand students (6,000 students), 3,000 male and 3,000 female, from

a total population of 35, 262 students who took the examination were selected.

Three sampling plans were employed to estimate item difficulty and item discrimination of the test scores under the CTT and IRT measurement frameworks. The sampling plans were random samples, gender group sampling and truncated group sampling. The sampling plans allow for the comparability of each framework across progressively less comparable samples.

According to Chang, Hanson and Harris (2001), stable estimates of CTT item difficulty and discrimination can be found with a sample size of 100 to 200. Wright and Stone (1979) found that sufficient sample sizes for CTT stability would allow for stable estimates of one-parameter IRT item indices. To investigate the functionality of CTT and IRT estimate under different conditions, two different sample size conditions were employed. In large scale measurement situations, one set of samples was randomly selected with n=1,000. And clinical situations were often constructed with small sample sizes; a second set of sample was randomly selected n=100 (Skaggs & Lissitz, 1986). The second set of random sample was drawn to look at the effect of small sample.

One set of random samples consisting of 1,000 examinees, were drawn from the 6,000 examinees. The second set of random samples, consisting of 100 examinees, was also drawn from the 6,000 examinees. 1000 random samples of each gender group were drawn. The same process was employed to generate the small sample replicates, 100 samples were randomly drawn from both the female and the male group. Fan (1998) noted that because the gender samples are subpopulations of the total population, theoretically, disparity between statistics calculated from different samples will be larger than that found in random sampling plan.

A third sampling involved truncated high-ability and low ability group samples. For this sampling plan, 1,000 samples were randomly drawn from both the low-ability and high-ability groups. For small samples, 100 samples were randomly drawn from both the low and high-ability groups. The low-ability sample was comprised of students whose total test score fell in the 0 to 21 mark out of 60 while the high-ability group fell in the 39 to 60 mark out of 60. One-hundred samples were randomly drawn from both the low and high ability group. These truncated high-ability and low-ability group samples should theoretically display the greatest dissimilarity between the CTT and IRT statistics, because “these two groups were defined in terms of test performance, not in terms of a demographic variable” (Fan, 1998).

The instrument for this study was the May/June 2008 NECO Senior School Certificate Examination Mathematics Paper 1. It was a dichotomous multiple choice examination consisting 60 items and based on the senior

secondary school mathematics curriculum in Nigeria. The Nigeria Senior School Certificate examination is administered at the end of the third year of senior school certificate course to measure the achievement level of candidates at that point. The examination is used as a tool to qualify students who are to proceed to the next level of education, which is tertiary institutions and also as an assessment mechanism that measures the extent to which basic competencies and skills have been acquired. The instrument was assumed to have been moderated and validated by NECO before it was administered on the students. The 60 multiple-choice Mathematics questions covered a wide range of topics in the Senior Secondary School (SSS) syllabus, showing that it had content validity. The reliability coefficients of the students' responses to the 60 multiple-choice Mathematics questions using Cronbach's Alpha coefficient was found to be 0.853, (n = 6000).

The data used in this study were responses of candidates who wrote May/June 2008 NECO SSCE Mathematics in Osun State. These responses were on marked Optical Recorder Mark (OMR) sheets and OMR sheets containing the responses of these candidates were collected from NECO office, Minna. NECO is an examination body in Nigerian that involves in conducting senior school certificate examinations and award certificates to candidates based on the individual candidate results. Senior school certificate examination in May/June is typically taken by school-bound students in senior secondary school 3.

Analysis of Data

Classical Test Theory

Classical test theory (CTT) analysis was obtained from the BILOG-MG 3 outputs. BILOG-MG presented its output in three phases. Phase 1 results were found in a file with the same name as the command file, but with the extension "PH1" (e. g, MATH.PH1), this phase contained information concerning the job setup, the reading of the data and classical item statistics. Appendix III contains the BILOG-MG output for Phase 1. The phase 1 output contains echoes of commands and the item statistics section which contains traditional item difficulty and discrimination statistics. The ratio of the number right (#RIGHT) to #TRIED is presented in the percent column (labelled PCT). This column indicates the percentage of examinees that correctly responded to an item. These percentages were divided by 100% to yield the measure of item difficulty.

The last two columns were collectively labelled ITEM*TEST CORRELATION and contain two traditional measures of item discrimination. The second to the last column (labelled PEARSON) contains the point-biserial, whereas the last column (labelled BISEARIAL) contains the corresponding biserial correlations. Each item was examined using the

proportion who answered the item correctly, p-values (item difficulty), and point-biserial correlation, rpbis (item discrimination). The point-biserial correlation is the correlation between the test-takers' performance on one item compared to the test-takers' performances on the total score.

Item Response Theory

For Item response theory Parameter Logistic models statistical method; the three known IRT models for binary response were used; one parameter (1PL), two parameter (2PL) and three parameter (3PL) logistic IRT models. Unidimensionality of the subject which is the major assumption of IRT models was investigated using SPSS version 20 through the eigenvalues in a factor analysis.

The second assumption, local independence, was examined through personal communication with NECO state coordinator and she gave assurance that no item gave a clue to any other item's answer.

The BILOG-MG 3 was used to estimate the item parameters. Outputs phase 2 of BILOG-MG contains the IRT calibration results. The beginning of this output contains information about the execution; the maximum number of EM cycles, the convergence criterion, the assumption of a Gaussian person prior and the quadrature point and corresponding weights.

The -2 LOG likelihood values showed the expected progressively decreasing pattern of a well-behaved solution. The marginal maximum log likelihood function value (-2 LOG LIKELIHOOD) after the last cycle was used for comparing model fit. The columns labelled SLOPE and THRESHOLD contain the IRT-based item discrimination parameter estimates and item parameter (item location) estimate respectively. While the column labelled ASYMPTON contains the guessing parameter estimates. This asympton is associated with three parameter model.

Comparability of irt and ctt statistics

Two item Statistics

The comparability of item characteristics for both methods was obtained by correlating (a) the item difficulty and (b) the item discrimination parameters. For each sampling plan, both the CTT- and IRT-based (one-, two- and three-parameter) item difficulty and discrimination estimates were obtained using BILOG-MG's marginal-maximum likelihood method.

The CTT-based item difficulty estimates were correlated with the 1PL, 2PL and 3PL IRT-based item difficulty parameter estimates, denoted by p in IRT models but referred to threshold parameter in BILOG-MG. Also, the CTT-based item discrimination parameter, both the item-test point-biserial and the transformed item-test point-biserial correlation, were

correlated with the 2PL and 3PL IRT-based item discrimination parameter estimates. 1PL IRT-based item discrimination parameter estimates were not available. All the correlation analysis was achieved using SPSS version 20.

Transformations for CTT p Value and Item-Test Correlations

In CTT, the item difficulty statistic is expressed on an ordinal scale. In an ordinal measurement scale, one is able to discern whether one item is more difficult than other item. However, it cannot tell us whether the differences in various item difficulties are the same across the different comparisons. For instance, if items 1, 2 and 3 have an item difficulty of .25, .20, and .15, just because the difference between 1 and 2 and 2 and 3 equals .05 does not indicate that the difference in difficulty is the same in these two comparisons.

However, if the trait being measured is normally distributed, the CTT item difficulty statistic can be expressed as equal interval normal curve units (Joseph, Jason & Ron, 2014). The transformation is achieved by finding the z score that corresponds to the proportion of examinees who answer an item correctly. The present study correlated both the CTT item difficulty estimates and the normalized CTT item difficulty estimates with IRT item difficulty estimates.

An item-test point biserial correlation, identified as the CTT item discrimination estimate, is not linearly scaled. As Hinkle, Wiserma and Jurs (1998) explained, "the sampling distribution of the correlation coefficient changes its shape as a function of both the magnitude and the sign of the coefficients. R.A. Fisher developed a transformation that in large samples allows the transformed correlation coefficient to be distributed approximately normal". Therefore, the assessment of the invariance of CTT item discrimination estimates was based on the correlation analysis between both the original and the Fisher z transformed point biserial for the different samples of examinees. For each sample plan, the individual point-biserial correlation coefficient was transformed to Fisher Zs.

Correcting for the Bias in Sample Correlation Coefficients

Because the sample correlation coefficient, r, is a ratio, it is a biased estimator of the population correlation coefficient. Zimmerman, Zumbo, and Williams (2003) noted that r can be biased as much as 0.03 or 0.04, which, as Zimmerman et al. (2003) indicated, may be vital when investigating the accuracy of the magnitude of r in measurement studies.

To correct for the bias in the sample correlation coefficient, R.A. Fisher developed a procedure to approximate the population correlation coefficient:

$$E[r] = r[1+\{(1-r2)/2n\}] \quad \text{.....Eq 1}$$

Later, Olkin and Pratt (1985) indicated that the following approximation is a more nearly unbiased estimator of r:

$E[r] = r[1+\{(1-r2)/2(n-3\}]$ ……Eq 2

As the sample size decreases, the effect of bias increases. The present study used both the Fisher and the Olkin and Pratt corrections to compare model parameters across CTT and IRT procedures.

Results

The unidimensionality of SSCE Mathematics

This first assumption was examined using a factor analysis, as this is a very important step prior to performing analysis. The Cronbach alpha was used to confirm the result giving by factor analysis, high internal consistency, 0.853, indicating that the SSCE Mathematics was unidimensional.

In the factor analysis, the initial communalities showed the variance in each variable are accounted for by all components. For principal components extraction, this was equal to 1.008 as the standard rule for correlation analyses. The Extraction communalities showed the estimates of the variance in each variable accounted for by the components. The principal component analysis revealed that the correlation matrix had its entire coefficients less than 0.3. That shows that the item loadings are considered relevant and contributed to the factor loadings.

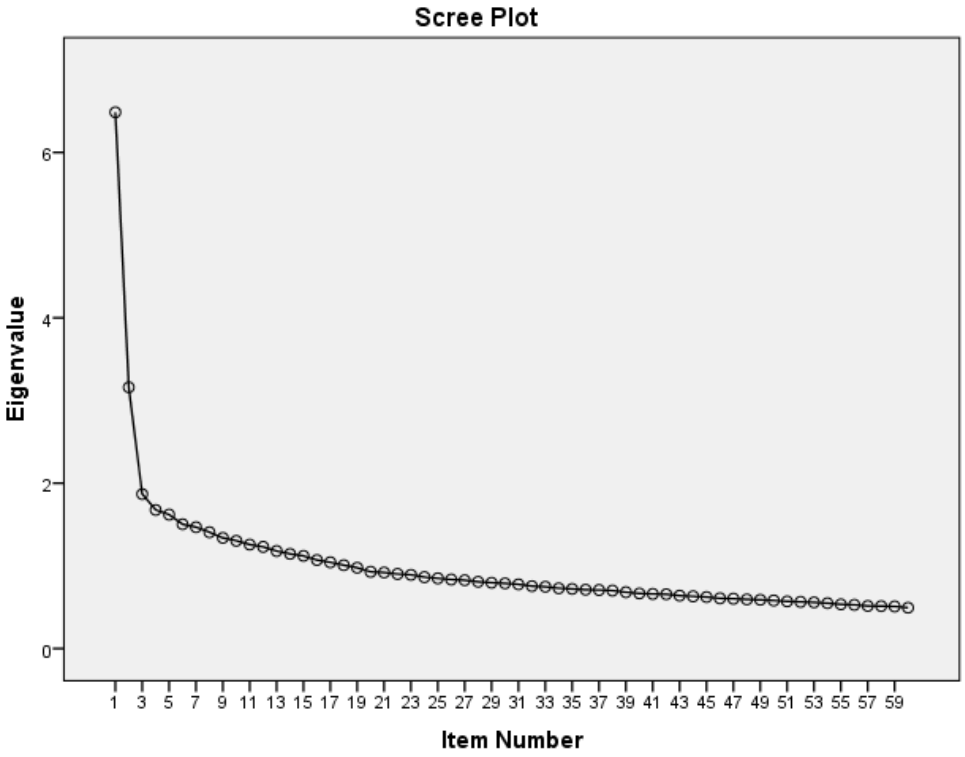
The extraction from the principal component analysis after interacting of communalities showed eighteen components with eigenvalues greater than 1 as revealed in the Scree plot (see Figure 1). This explained 10.810, 5.265, 3.113, 2.795, 2.700, 2.506, 2.447, 2.345, 2.231, 2.172, 2.096, 2.049, 1.966, 1.909, 1.868, 1.784, 1.738 and 1.680% of variance accounted for by each component to the total variance in all of the items. Furthermore, for the 60 multiple-choice Mathematics items, with respect to the eigenvalue greater than 1, the total percentage variance was 51.474.

Table 1: Factor Correlation Matrix of Mathematics 60 Multiple-Choice Items

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
F1	1.000																	
F2	.222	1.000																
F3	.197	.224	1.000															
F4	.112	.021	.134	1.000														
F5	.265	.228	.262	.167	1.000													
F6	.146	.137	.166	.282	.261	1.000												
F7	-.105	-.108	-.119	-.138	-.133	-.119	1.000											
F8	.158	.208	.170	.014	.199	.152	-.101	1.000										
F9	.129	.051	.028	.192	.125	.236	-.058	.060	1.000									
F10	.208	.174	.154	.079	.234	.135	-.101	.264	.092	1.000								
F11	.115	.096	.107	.203	.162	.237	-.086	.157	.189	.223	1.000							
F12	.158	.120	.167	.099	.155	.113	-.101	.165	.091	.131	.164	1.000						
F13	.170	.142	.101	.070	.156	.096	-.098	.197	.062	.181	.118	.179	1.000					
F14	.022	-.021	-.015	.074	-.027	.087	-.022	.078	.096	.103	.121	.052	-.003	1.000				
F15	.121	.113	.108	.109	.177	.125	-.079	.076	.118	.121	.094	.048	.083	.099	1.000			
F16	.173	.138	.109	.144	.205	.159	-.124	.209	.155	.259	.198	.104	.266	.117	.233	1.000		
F17	.094	.130	.154	.143	.244	.163	-.164	.126	.097	.164	.096	.104	.243	.048	.163	.324	1.000	
F18	.079	.094	.106	-.021	.106	-.019	-.042	.031	-.033	.047	-.082	.021	.101	-.090	.069	.098	.224	1.000

From Table 1, it could be seen that the correlation ranges from -0.003 to 0.324 which is less than correlation value of 0.35. This showed low correlation value and evidence that SSCE Mathematics is unidimensional.

Figure 1: Scree Plot for 60 dichotomous items



The Figure 1 is the scree plot for the 60 multiple-choice SSC Mathematics Examination items. The factor analysis that was performed on the items using extraction method of principal component analysis (see appendix iv) showed that the first factor having the initial eigenvalue (10.810) which clearly exceeded that of the second factor (5.265) as also revealed in Figure two. From Figure two, the Scree plot showed a visual of the total variance associated with each factor. The steep slope showed the large factors associated with the loading greater than the eigenvalue of 1. The gradual trailing off (scree) showed the rest of the factors lower than an eigenvalue of 1. There are thirteen factors whose values are greater than eigenvalue of l and one extracted communality factor distinctly highly than others, showing that the test is unidimensional in nature. Also, it can therefore be concluded that the 60 multiple-choice mathematics items is unidimensional.

Research Question 1: How comparable are the CTT-based and IRT-based item difficulty estimates?

Table 2 and 3 present the results addressing the first research question, by analyzing the comparability of correlations between the CTT- and IRT-based item difficulty estimates. Table 5 presents the n=1000 data while table 6 presents the n=100 data. To obtain the entries in tables 5 and 6, the following two steps were invoked: (a) for each of the 1000 and 100 samples, one-, two- and three-IRT based item difficulty parameter estimates and CTT based item difficulty parameter estimates were obtained using BILOG MG 3; (b) for each sample the CTT- and IRT – based item difficulty estimates were correlated for each of the sampling plan. Consequently, each of the table values is the correlations, except where the IRT model did not converged. The IRT-based item difficulty estimates were correlated with the CTT-based item difficulty estimate, p and the CTT-based normalized p values. The correlations between the IRT-based item difficulty estimates and the CTT-based item difficulty values are negative. However, these differences in scaling direction of the difficulty estimates are arbitrary.

Table 2: Comparability of Item Characteristics from the Two Measurement Frameworks: Correlations between CTT - and IRT - Based Item Difficulty Indexes (n = 1000)

	IRT Models					
	CTT p values			CTT Normalized p values		
Sampling Frame	1PL	2PL	3PL	1PL	2PL	3PL
Random samples	-0.988	-0.855	-0.836	-0.988	-0.855	-0.836
Gender group sampling						
Female	-0.800	-0.840	-0.753	-0.800	-0.840	-0.753
Male	-0.829	-0.799	-0.780	-0.829	-0.799	-0.780
Truncated Ability group sampling						
High-ability	-0.781	-0.713	NC	-0.781	-0.713	NC
Low-ability	-0.779	-0.702	NC	-0.779	-0.702	NC

Note: “NC” are models where all the items did not converge.

Table 2 shows correlations between the IRT-based item difficulty estimates and CTT-based item difficulty and normalized CTT-based item difficulty estimates for n=1000.

The IRT-based one-parameter item difficulty estimates had high correlations with the CTT-based item difficulty and normalized CTT-based item difficulty estimates. The correlations were, generally, in the -0.779 to -0.988 range.

The IRT-based item difficulty estimates for the two-parameter model had high correlations with the non-normalized and normalized CTT-based item difficulty estimates. The IRT-based two-parameter model correlations were, generally in the -0.702 to -0.855 range.

The IRT three-parameter item difficulty estimates was highly correlated with the CTT-based item difficulty estimates (non-normalized and normalized). The correlations were ranged from low-ability sample plan, -0.702 to random sample plan, -0.836.

Table 3: Comparability of Item Characteristics from the Two Measurement Frameworks: Correlations between CTT - and IRT - Based Item Difficulty Indexes (n = 100)

	IRT Models					
	CTT p VALUES			CTT NORMALISED p VALUES		
Sampling Frame	1PL	2PL	3PL	1PL	2PL	3PL
Random samples	-0.986	-0.834	-0.617	-0.986	-0.834	-0.617
Gender group sampling						
Female	-0.989	-0.874	-0.725	-0.989	-0.874	-0.725
Male	-0.950	-0.831	-0.782	-0.950	-0.831	-0.782
Truncated Ability group sampling						
High-ability	-0.699	-0.665	NC	-0.699	-0.665	NC
Low-ability	-0.985	-0.622	NC	-0.985	-0.622	NC

Note: NC are models where all the items did not converge.

Table 3 shows correlations between the IRT-based item difficulty estimates and CTT-based item difficulty estimates (normalized and non-normalized) when n=100. For the one-parameter model, the correlation between IRT-based item difficulty estimates had very high correlations with the CTT-based item difficulty estimates, the correlations were generally in the -0.950 to -0.989 range except in the high-ability sample plan that the correlation was moderate (r = -0.699).

For the two- parameter model, the correlation between the IRT-based item difficulty estimates with the normalized and non-normalized CTT-based item difficulty in the random sampling plan was high (r = -0.834); in the female sampling plan correlation was high (r = -0.874); in the male sample plan, the correlation was high (r = -0.831); the correlation in the high-ability sample plan was moderate (r = -0.665) and in the low-ability sample plan the correlation was moderate (r = -0.622).

For the three-parameter model, the IRT-based item difficulty estimate was correlated with normalized and non-normalized CTT-based item difficulty estimates, in the random sampling plan, the correlation was moderate (r = -0.617); in the female sample plan, the correlation was high (r = -0.725) and in the male sample plan, the correlation was moderate (r = -0.782). While in the truncated ability sampling plan there was no convergence in all the item parameters.

As it has been seen in the table, the two- and three-parameter IRT models produced lower correlations with the CTT-based item difficulty estimates than one-parameter IRT model.

Table 4: Comparability of Item Characteristics from the Two Measurement Frameworks: Correlations between CTT - and IRT - Based Item Difficulty Indexes (n = 1000)

	IRT Models					
	CTT P VALUES					
	Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	1PL	2PL	3PL	1PL	2PL	3PL
Random samples	-0.972	-0.855	-0.811	-0.972	-0.855	-0.811
Gender group sampling						
Female	-0.765	-0.803	-0.633	-0765	-0.803	-0.633
Male	-0.804	-0.756	-0.774	-0.804	-0.756	-0.774
Truncated Ability group sampling						
High-ability	-0.777	-0.784	NC	-0.777	-.784	NC
Low-ability	-0.707	-0.721	NC	-0.707	-0.721	NC

Table 4 shows the result of table 5 for n=1000 except that the sample correlations from table 7 have been corrected for bias using both the Fisher and Olkin and Pratt correction. All of the correlations generated using the Fisher correction matched those generated using the Olkin and Pratt correction.

For the one-parameter model, the correlation between IRT-based item difficulty estimates with the CTT-based item difficulty estimates was very high in the random sample plan (r = -0.972); in the female sample plan, the correlations was high (r = -0.765); the correlation was also high in the male sample plan (r = -0.884); in the high-ability plan, the correlation was moderate (r = -0.777) and was moderate in the low-ability plan (r = -0.707).

For the two- parameter model, the correlation between the IRT-based item difficulty estimates and the corrected CTT-based item difficulty estimates in the random sampling plan was moderate (r = -0.855); in the female sampling plan correlation was high (r = -0.803); in the male sample plan, the correlation was high (r = -0.756); the correlation in the high-ability sample plan was moderate (r = -0.784) and in the low-ability sample plan, the correlation was moderate (r = -0.721).

For the three-parameter model, the IRT-based item difficulty estimate was highly correlated with the corrected CTT-based item difficulty estimates in the random sampling plan (r = -0.811); in the female sample plan, the correlation was moderate (r = -0.633) and in the male sample plan, the correlation was high (r = -0.774). While in the truncated ability sampling plan there was no convergence in all the item parameters.

As it has been seen in the table, the two- and three-parameter IRT models also produced lower correlations with the CTT-based item difficulty estimates than one-parameter IRT model.

Table 5: Comparability of Item Characteristics from the Two Measurement Frameworks: Correlations between CTT - and IRT - Based Item Difficulty Indexes (n = 100)

IRT Models						
CTT P VALUES						
	Fisher Correction			Olkin and Pratt Correction		
Sampling Frame	1PL	2PL	3PL	1PL	2PL	3PL
Random samples	-0.967	-0.842	-0.615	-0.967	-0.842	-0.615
Gender group sampling						
Female	-0.970	-0.845	-0.756	-0.970	-0.845	-0.756
Male	-0.934	-0.801	-.687	-0.934	-0.801	-0.687
Truncated Ability group sampling						
High-ability	-0.788	-0.774	NC	-0.788	-0.774	NC
Low-ability	-0.980	-0.726	NC	-0.980	-0.726	NC

Table 5 shows the result of table 6 (n=100) except that the sample item estimates from table 8 have been corrected for bias using both the Fisher and Olkin and Pratt correction. All of the correlations generated using the Fisher correction matched those generated using the Olkin and Pratt correction.

For one parameter model, in the Random sample plan, the correlation was very high (r = -0.967); in the female sample it was very high (r = -0.970); in the male sample plan, the correlation was high (r = -0.934); in the high-ability sampling plan the correlation was -0.788 and in the low-ability sampling plan the correlation was very high (r = -0.980).

For the two-parameter model, the correlation between the IRT-based item difficulty estimates in the random sampling plan was high (r = -0.842); in the female sampling plan, correlation was high (r = -0.845); in the male sample plan, correlation was high (r = -0.801); while in the high-ability sample plan correlation was moderate (r = -0.774) and in the low-ability sample plan correlation was moderate (r = 0.726).

The IRT three-parameter item difficulty was moderately correlated with CTT-based item difficulty estimates in the random sampling plan (r = -0.615); in the female sample plan correlation was high (r = -0.756) and in the male sampling plans, correlation was moderate (r = -0.687). While in the truncated ability sampling plan there was no convergence in all the item parameters of the IRT model.

It was observed in the table, that the two- and three-parameter IRT models also produced lower correlations with the CTT-based item difficulty estimates than one-parameter IRT model.

Research Question 2: Are the CTT-based and IRT-based item discrimination estimates comparable?

Table 4 and 5 present the result addressing the second research question by analyzing the comparability of correlations between the CTT- and IRT-based item discrimination estimates. Table 3 presents the results for the n=1000 data. To obtain the entries in Table 3, the following two steps were taken: (a) for each of the 1000 samples the IRT one-, two- and three-parameter models estimates and CTT estimates were obtained; (b) for each sample the CTT- and IRT-based discrimination estimates were correlated for the same sampling plan. Consequently each of the tabled values is the correlation obtained, except where the IRT model did not converged. Note that the one-parameter IRT model does not estimate item discrimination, as so results for this model are indicated to be "not applicable" (N/A).

Table 6: Comparability of Item Characteristics from the Two Measurement Frameworks: Correlations between CTT - and IRT - Based Item Discrimination Indexes (n = 1000)

	IRT Models					
	Point-Biserial			Fisher Z Transformed Point-biserial		
Sampling Frame	1PL	2PL	3PL	1PL	2PL	3PL
Random samples	N/A	0.867	0.648	N/A	0.864	0.643
Gender group sampling						
Female	N/A	0.880	0.471	N/A	0.890	0.470
Male	N/A	0.858	0.931	N/A	0.885	0.925
Truncated Ability group sampling						
High-ability	N/A	0.452	NC	N/A	0.457	NC
Low-ability	N/A	0.430	NC	N/A	0.438	NC

Table 6 presented the results of the correlations between CTT-based item discrimination estimates (point-biserial and transformed point-biserial) and IRT-based item discrimination estimates for n=1000 data.

For the two-parameter model, correlation between the IRT-based item discrimination and the CTT-based item discrimination estimates (both point biserial and transformed point-biserial) was high (r = 0.867 and 0.864 respectively) in the random sample plan; in the female sample plan, the correlation coefficient was high (r = 0.880 and 0.890 respectively); correlation was low in the high-ability sample plan (r = 0.452 and 457 respectively) and the correlation coefficient in the low-ability sample plan was also low (r = 0.430 and 438 respectively).

For three-parameter model, the correlation between IRT-based item discrimination and CTT-based item discrimination estimates was moderate in the random sample plan (r = 0.648 and 0.643 respectively); in the female sample plan, correlation was moderately low (r = 0.471 and 0.470); in the male sample plan, correlation was high (r = 0.931 and 0.925 respectively).

Table 7: Comparability of Item Characteristics from the Two Measurement Frameworks: Correlations between CTT - and IRT - Based Item Discrimination Indexes (n = 100)

	IRT Models					
	Point-Biserial			Fisher Z Transformed Point-biserial		
Sampling Frame	1PL	2PL	3PL	1PL	2PL	3PL
Random samples	N/A	0.838	0.649	N/A	0.807	0.521
Gender group sampling						
Female	N/A	0.950	0.861	N/A	0.961	0.850
Male	N/A	0.930	0.406	N/A	0.939	0.401
Truncated Ability group sampling						
High-ability	N/A	0.544	NC	N/A	0.540	NC
Low-ability	N/A	0.531	NC	N/A	0.501	NC

Results in Table 7 show that for the n=100 data that, barring a some exception, demonstrated good relationships of item discrimination coefficients across measuring models, regarding of sampling plan.

For the two-parameter model, the IRT-based estimates of item discrimination and the CTT-based estimate of item discrimination (point-biserial and transformed point-biserial) were highly correlated in the random sample plan (r = 0.838 and 0.807 respectively), highly correlated in the in the female sample plan (r = 0.950 and 0.961 respectively) and in the male sample plan (r = 0.930 and 0.931 respectively). While moderately correlated in the high-ability sample plan (0.544 and 0.540 respectively) and moderately correlated in the low-ability sample plan (0.531 and 0.501 respectively).

However, the relationships weakened for the three-parameter IRT model; correlation coefficients in the random sample plan was moderate (r = 0.649 and 0.521 respectively); correlation was high in the female sample plan (r = 0.861 and 0.850 respectively) and in the male sample plan correlation was moderate (r = 0.406 and 0.401 respectively). All the items did not converge in the high-ability sample plan and low-ability sample plan. All of the correlations generated using original and fisher transformed point-biserial for different samples of examinees follow the same pattern except that correlation coefficients generated from Fisher transformed did not match the ones generated from the original point-biserial.

Discussion

Item response theory analysis can only be performed only when the test scores are unidimensional (Ojerinde, 2013). There are various ways for testing unidimensionality. However, unidimensionality can be established when one of two conditions is met from the results of an exploratory factor

analysis (Reckase, 1999): first, a factor analysis on the inter-item correlation matrix should show that the first factor accounts for at least 20% of the variance of the unrotated factor matrix or second the eigen value of the first factor should clearly exceed that of the second factor. Also, a high cronbach alpha indicated unidimensionlity.

The results of the factor analysis (the total percentage variance was 51.474) revealed that SSCE mathematics are evidences of unidimensionality. It can be concluded that the assumption of unidimensionality holds to a good extent in the test and in the mathematics items.

The results of correlations between CTT-based and IRT-based item difficulty estimates showed that the one- and two-parameter IRT item difficulty estimate provided results similar to the CTT-based item difficulty estimates. This result supports previous study that finding of Courville, (2004). This indicating that very similar mathematics achievement estimates would be obtained regardless of the measurement framework. Although, one-parameter IRT model provided the results that were more similar to CTT model counterparts compared to two-parameter IRT model. These results were found to be consistent with other finding of Marie (2004). The foregoing results resemble that of previous studies (Adedoyin, Nenty, & Chilisa, 2008; Nukhet, 2002; Fan, 1998). However, the difference lies in the choice of a two-parameter or three-parameter. Nukhet (2002) reported three-parameter as having the most comparable indices with CTT. Whereas Fan (1998) indicated that all three are comparable with CTT. Also, results from small samples, n=100, indicated that, small sample size used to compute the correlations are good estimates of what would be found in the large sample, n=1000. Therefore it can be said that CTT model was comparable to one-parameter and two-parameter IRT models.

Moreover, comparing the n = 1000 and n = 100 samples, both samples produced strong correlations between CTT-based and IRT-based two-parameter item discrimination estimates. But both produced lower, albeit strong correlations between the CTT-based and the IRT-based three-parameter item discrimination estimates. These results resemble that of previous work done that have found that both large and small samples produced very strong correlations between the CTT-based and IRT-based two-parameter item discrimination estimates but produced lower, albeit strong correlations between the three-parameter IRT-based and CTT-based item discrimination estimates.(Courville, 2004).

Findings

The major findings were:

1. For the small and large samples, the CTT-based and IRT-based item characteristics estimates were very comparable, indicating that an

analysis of the item statistic of examinees will lead to similar results across the different measurement theories.

2. The CTT-based item difficulty estimates and the one-parameter IRT item difficulty estimate provided very similar results. This showed that CTT and one-parameter IRT models could both be used independently to estimate the test item difficulty parameters.

3. The investigation of the item discrimination statistics in the comparability of item estimates produced strong correlations between the CTT-based and IRT-based two-parameter item discrimination estimates but produced lower, albeit strong correlations between the three-parameter IRT-based and CTT-based item discrimination estimates.

4. All the statistics indicated a progressive decay in the correlations as the sampling frameworks became more dissimilar.

5. Across all samples, the IRT-based item estimates in the one-parameter model were much more similar to the CTT-based item estimates.

6. Both CTT and IRT models can be used together in estimating item characteristics and in test development.

Conclusion

Based on the findings, the study concluded that CTT and IRT were comparable in estimating item characteristics of statistical and psychometric tests and thus could be used as complementary procedures in the development of national examinations.

Recommendations

The following are the recommendations;

1. The examination bodies using multiple-choice test instruments should employ the use of both IRT and CTT statistics in test development validation processes. This will ensure effective test development in that both statistics will complement one another.

2. For institutions and researchers that wish to use IRT in solving measurement problems should make efforts to conform to the assumptions before use especially property of unidimensionality.

3. Efforts should be made by examination bodies and educational institutions to train personnel in the applications of Item Response Theory.

4. More awareness and interest in Item Response Theory and its applications by making IRT core modules in both undergraduate and postgraduate programmes.

5. Department of Educational foundations and Councelling should make efforts in teaching her students on how to use statistical packages especially those that can handle Item Response Theory.

References:

1. Ashikhia, D. A. (2010). Students and teachers' perceptions of the causes of poor academic performance in Ogun State secondary schools (Nigeria): Implication for counseling for national development. Retrieved from http://www.eurojournal.com/ejss on august 28th, 2010.
2. Adedokun, J. A. (2002). Availability of learning environment resources as predictor of the level of teachers' accountability to science students' performance. *Nigerian Education Review, 3, 11-17.*
3. Adedoyin, O. O., Nenty, H. J., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review, 3*(2), 83-93.
4. Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
5. Aremu, O. A., & Sokan, B. O. (2003). *A multi-causal evaluation of academic performance of Nigerian learners:* issues and implications for national development. Department of Guidance and Counselling, University of Ibadan, Ibadan.
6. Chang, S., Hanson, B., & Harris, D. (2000, April). A Standardization Approach to Adjusting Pretest Item Statistics. Paper presented at the annual meeting of the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED 442 838).
7. Courville, T. G. (2005). *An empirical comparison of item response theory and classical test theory item/person statistics.* Unpublished doctoral dissertation, Texas A&M University. Retrieved February 5, 2015 from http://txspace.tamu.edu/bitstream/handle/1969.1/1064/etd-tamu-2004B-EPSY-Courville-2.pdf?sequence=1.
8. Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381.
9. Joseph, C. C., Jason, J. L., & Ron, D. H. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *HHS Public Access, 36(5), 648–662*.
10. Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Education Measurement: Issues and Practice, 12*(3): 38-47.
11. Hambleton R. K., & Swaminatthan, H. (1995). *Item response theory: Principles and application.* Boston: Kluwer.

12. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
13. Hinkle, D., Wiersma, W., & Jurs, S. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston: HoughtonMifflin.
14. Lawson, S. (1991). One Parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological development, 1*, 159-168.
15. MacDonald, P., & Paunonen, S. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921-943.
16. Marie, W. (2004). Classical Test Theory vs Item Response Theory: An Evaluation of the Theory test in swedish driving-license test. *Journal of Safety Research*, *37*(3), 285-291.
17. Nukhet, C. (2002). A Study of Raven Standard Progressive Matrices test's item measures under classic and item response models: An empirical comparison. *Ankara University, Journal of Faculty of Educational Science, 35*(2), 71-79.
18. Ojerinde (2013). *Classical Test Theory (CTT) VS Item Response Theory (IRT): An Evaluation of Comparability of Item Analysis Results*. Lecture Presentation at the Institute of Education, University of Ibadan.
19. Paul, M., & Sampo, V. (2002). A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and psychological measurement, 6*, 921-943.
20. Royce, H. (2009). Comparison of the item discrimination and item difficulty of the quick-mental aptitude test using CTT and IRT methods. *The International Journal of Educational and Psychological Assessment*. Vol. 1, Issue 1, pp. 12-18.
21. Sadler, D. Royce (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In Gordon Joughin (Ed.), *Assessment, Learning and Judgement in Higher Education* (pp. 45-63) Dordrech, Netherlands: Springer Science. doi:10.1007/978-1-4020-8905-3_4
22. Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*, 69-82.
23. Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

24. Zimmerman, D., Zumbo, B., & Williams, R. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica 24,* 133-158.