Herald Journal of Education and General Studies Vol. 2 (3), pp 107 - 114 August, 2013
Available online http://www.heraldjournals.org/hjegs/archive.htm

Full Length Research Paper

Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters

Omobola O. Adedoyin and J. A. Adedoyin

University of Botswana, Botswana

Accepted 13 July, 2013

This study assessed the comparability of test items parameter estimates between Classical test theory (CTT) and Item response theory (IRT) models using 2010 Botswana Junior Certificate (JC) mathematics paper 1 test items. A simple random sample of ten thousand (10,000) students was selected from the population of thirty six thousand, nine hundred and forty (36,940) students who sat for the JC mathematics examination paper 1 in 2010,using SPSS computer software. To estimate for the test items parameters in terms of item difficulty, item discrimination and the students' responses to the test items were analysed using both the CTT and IRT (3PLM) models. Pearson correlation coefficients were used to determine if the IRT and CTT test items parameter estimates were comparable and dependent t-test was used to find out whether the relationship between IRT and CTT test items parameter estimates were statistically significant. From the result of the analysis, it was found that the CTT and IRT item difficulty and item discrimination values were positively linearly correlated and there was no statistical significant difference between the item difficulty and item discrimination parameter estimates by CTT and IRT.

Keywords: IRT (Item response theory), CTT (Classical test theory), Item difficulty, Item discrimination.

INTRODUCTION

Despite its weaknesses, Classical test theory (CTT) has dominated the measurement of cognitive abilities in the educational system. In spite of this dominance, CTT has the limitation of sample dependency for estimating the test items parameters namely the item difficulty and item discrimination.This means, the various statistics used for describing or interpreting test scores such as the item difficulty (p-values), indexes of reliability/ validity and item discrimination values are sample dependent. CTT assumes that good test items discriminate through a wide range of abilities and it utilises the p-values (item difficulty) and point-biserial correlations to analyse and interpret the responses to test items. It is however, always very difficult for educators to assume that test items, constructed under CTT, measure what they are supposed to estimate because CTT is sample dependent.

*Corresponding Author E-mail: omobola_adedoyin@yahoo.com

In estimating CTT test items parameters, heterogeneous samples generally result in higher estimates of item discrimination indices as measured by point-biserial correlation coefficients, whereas item difficulty estimates rise and fall with high and low ability groups of examinees. According to Margno (2009),"CTT is based on the true score model, which depends on examinees aggregate score in a test and therefore does not permit a consideration of examinees responses to any specific item; providing no basis to predict how a given examinee will perform on a particular test item". At present, due to the limitations of CTT, item response theory (IRT) is receiving increasing attention from measurement experts for ability testing, test item selection, reporting of test scores and most especially in identifying the characteristics of test items. In IRT the item statistics depend to a great extent on the characteristics of the examinee sample used in the analysis. An important concern of test developers applying classical test theory is that the examinee sample should be representative of the overall population for whom the test is intended.

Some recent research studies compared the comparability of the item and person parameters using CTT and IRT. Fan (1998) examined the comparability of IRT and CTT statistics and test scores using the 1PL, 2PL and 3PL IRT models. Because Fan used real data sets, no comparison between true values and estimates could be made; rather he compared CTT and IRT results (e.g., the correlation between IRT difficulty parameters and CTT proportion correct values). Fan found a high degree of comparability between IRT theta-hat values and CTT values (the lowest correlation was 0.966) with 2PL estimates slightly less comparable than those of the 1PL and 3PL models. The item difficulty statistics were also highly correlated. The correlations were almost all 0.999 for the 1PL and were generally larger than 0.90. Again, the 2PL estimates were slightly less comparable than the 3PL estimates. Item discrimination parameters differed most from CTT item-total correlations with most correlations less than 0.90 and a few lower than 0.40.

In another empirical study by Courville (2005), the researcher found high correlations between the CTT and the IRT test item difficulties. The discrimination indices, however, correlated highly only when the spread of discriminations was large and the spread of difficulty values was small. MacDonald and Paunonen (2002), simulated data using 1PL and 2PL IRT models and then computed IRT and CTT statistics from these values. They performed three sets of correlations. First, they tested comparability of test scores, difficulty, and item discrimination by correlating estimated IRT and CTT statistics and they found very high comparability for test scores and difficulty and less comparability for item discrimination.

Item Response Theory (IRT)

IRT is the probability of answering an item correctly or of attaining a particular response level. It is modeled as a function of an individual's ability and the characteristics of the item. A paramount goal of IRT is predicting the probability of an examinee's level of correct response to an item of a particular difficulty. The latent traits can be measured on a transformable scale having a midpoint of zero, a unit measurement of one and arrange from negative infinity to positive infinity. While the theoretical range of ability is from negative infinity to positive infinity, practical considerations usually limit the range of values from –3 to +3 (Hambleton, Swaminathan and Rogers, 1991).

IRT begins with the proposition that an individual's response to a specific item or questions is determined by an unobserved mental attribute of the individual. Each of these underlying attributes, most often referred to as latent traits, is assumed to vary continuously along a single dimension usually designated by theta (θ) (Hambleton, Swaminathan and Rogers, 1991). There are traditionally three IRT mathematical equations termed, one, two, and three parameter models that are used to make predictions. The general IRT framework encompasses a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. These models relate the characteristics of individuals and the characteristics of the items to the probability of a person with given characteristics or level of an attribute choosing a correct response. For test items that are dichotomously scored, there are three IRT models, known as three-, two- and one-parameter IRT models. A primary distinction among the models is the number of parameter used to describe items. IRT models are mathematical functions that specify the probability of discrete outcome, such as a correct response to an item, in terms of persons and item parameters.

Item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptote). Three most commonly used IRT models are; one parameter logistic model (1PLM or Rasch model), two parameter logistic model (2PLM) and three parameter logistic model (3PLM). All three models have an item difficulty parameter (b), In addition, the 2PL and 3PLmodels possess a discrimination parameter (a), which allows the items to discriminate differently among the examinees. The 3PL model contains a third parameter, referred to as the pseudo-chance parameter (*c*). The pseudo-chance parameter (c) corresponds to the lower asymptote of the item characteristic curve (ICC) which represents the probability that low ability test takers will answer the item correctly and provide an estimate of the pseudo-chance parameter (Embretson and Reise, 2000). The values of item difficulty and item discrimination are often used to describe ICC. According to Adedoyin (2010), "The flatter the ICCs curve, the less the item is able to discriminate since the probability of correct response at the low ability levels is nearly the same as it is at high ability levels and the steeper the curve, the better the item can discriminate".

Purpose of the study

The main purpose of this paper is to assess the comparability of CTT and IRT in the estimation of test items parameters of 2010 Junior Certificate (JC) mathematics paper 1in Botswana, using students' responses. This study was guided by the following research questions:

1. What are the item parameter estimates of students' responses to JC mathematics paper 1 items based on CTT model?
2. What are the item parameter estimates of

students' responses to JC mathematics paper 1 items based on IRT (3PLM) model?
3. Is there any significant statistical comparability between the CTT and IRT (3PLM) item parameter estimates of students' responses to JC mathematics paper 1?

METHODOLOGY

Sample

The data for this study were the responses of 2010 Junior Secondary School (JSS) form three students in paper 1 mathematics multiple choice paper. These responses were obtained from Botswana Examinations Council. The examination mathematics paper 1 was administered at the end of form three to all students in JSS schools in Botswana. The population of this study was thirty six thousand, nine hundred and thirty- nine (36,939) JSS three students out of which,18271 were males and 18668 were females. Each of the JSS students who sat for the examination had an identification number comprising the school number or the center number. All the JSS students were included in the sampling frame, and they all had equal chances of been selected. A single stage sampling was done with each student selected independently by the use of SPSS software using select "Random sample of cases" menu. The selected sample of ten thousand (10,000) was analysed by descriptive statistics using the gender variable of the students, out of which 5436 were females and 4564 were males.

Instrument

The researchers collected the data for this study from Botswana Examination Council (BEC). The data were the responses from the 2010 mathematics paper one JC examination. The examination was a multiple choice paper which consisted of 40 items. The examination was administered for one and a half hours. Before the paper was administered an assessment syllabus was used to create a scheme of assessment (test blue print). In the scheme of assessment, the content area to be covered and the cognitive levels were shown to ensure a proper balance and emphasis of the syllabus.

Data analysis

In any analysis involving IRT, there are two basic assumptions that must be verified, the model fit and uni-dimensionality. Goodness of fit tests was used to examine how many items fitted the 1PLM, 2PLM and 3PLM and a confirmatory factor analysis was performed to test for uni-dimensionality of the test items. The forty (40) multiple choice items of 2010 mathematics paper 1 were assessed for model fit and analyses were performed on only the items that fitted the
3PLM model. They were then subjected to IRT model to find the item parameter estimates of the test items in terms of item difficulty and item discrimination. The responses of the ten thousand (10,000) students selected from the population of thirty six thousand, nine hundred and thirty- nine (36,939) JSS three students who sat for the mathematics paper one JC examination in government schools in Botswana were analysed to find CTT and IRT test items parameter estimates as follows:

CTT item parameter estimates

SPSS program version 16 was used to estimate the parameter estimates for CTT item difficulty, item discrimination, transforming the item parameters of both CTT and IRT into z-scores and point biserial correlation (r_{pb}) between CTT and IRT.

IRT item parameter estimates

MULTILOG-3.0 program was used to estimate the item difficulty (b-parameter) and item discrimination (a-parameter) using the IRT(3PLM).

Assumptions of IRT

IRT operates based on some assumptions and the first one is uni-dimensionality assumption. This assumption postulates that, only one ability is measured by the item that makes a test. Svend and Christensen (2002) pointed out that "what is required for the uni-dimensionality assumption to be met adequately is the presence of one dominant factor that influences test performance". Most IRT models assume that it is only a single latent trait that underlies performance on an item, and that responses to different items are independent given the latent trait. The second assumption is that of local independence. This assumption according to Ponocny (2002) states that, "when the abilities influencing test performance are held constant and the examinee's responses are equally held constant, then examinee's responses to any pair of items are statistically independent"

Another assumption of IRT is the correct utilization of models that fits the data. According to Chernyshenko, et al (2001)"researchers and practitioner need to pay attention to fundamental issue of model-data fit when using IRT models. For determining the model fit, Baker (2001) suggested that, "if the value of the obtained chi-square (or index) is greater than a criterion value, the item characteristic curve specified by the values of the item parameter estimates does not fit the data". Adedoyin

(2010) used chi-square test with probability greater than the alpha level of 0.05 significant level to selected items that fit 1-, 2- and 3PLM and Rasch model respectively. For this study statistical chi-square goodness of fit was used to evaluate the IRT model that fitted the data.

PRESENTATION OF RESULTS

Test for Uni-dimensionality

The method used to assess uni-dimensionality in this study was confirmatory factor analysis. It was performed to determine whether or not a dormant factor exists among all items as it is expected that the mathematics national examination would come up with one dominant factor. This factor would represent the construct underlining the mathematics skills measured by the examination. The confirmatory factor analysis performed on the 40 items of the 2010 JC mathematics paper one yielded nine eigen values greater than one. The first eigen value was 5.909 greater than the next eight eigen values (1.492, 1.096, 1.088, 1.060, 1.029, 1.022, 1.017 and 1.010). The first factor explained 14.772% of the variance in the data set. The second factor explained 3.73% of the remaining variance. The rest of the variance was explained by the other 38 factors with 24 factors each having an percentage of variance between 2 and 3 and 14 factors each having a percentage of variance of between 1 and 2. (Table 1)

A scree plot was produced to determine whether uni-dimensionality could be inferred. Scree plots provide a convenient way of visualising a dominant factor in principal component analysis. (Figure 1)

Test for model fit

The utility of the IRT model is dependent upon the extent to which the given responses reflect this model. To determine whether the test item fit the model, a Chi-square test was run on the data set using Bilog-M to establish whether the items fit the 1PL, 2PL and 3PL models. Table (2) showed the results of the chi-square statistics. The Chi-square goodness of fit analysis showed that only one item fitted the 1PL model, eleven items fitted the 2PL model and 23 items fitted the 3PL model. For the 2PL and 3PL model item 9 was omitted from the calibration as its initial slope was less than -0.15.

Extracting the number of items that fitted the 1PL, 2PL and 3PL models from the results of goodness of fit analysis resulted in table 3).

Since 23 items fitted the IRT (3PLM) model, the IRT (3PLM) model was used to estimate the item parameters. Table 4) answered the research questions 1 and 2, showing the CTT p-values, IRT b-values, CTT a-values and IRT a-values for the 23 test items.

Research question 1

1. What are the test items parameter estimates of 2. students' responses to JC mathematics paper 1 items using CTT model?

The result in Table (iv) for the CTT model, the item difficulty parameter estimates (p-values) ranged from 0.19 for item 26 to 0.71 for item 1.Most of the test items in table (iv) were of moderate difficulty, because their p-values were not too high. Only item 26 had a p-value of 0.19, which indicates this was a very difficult item. The discrimination parameter estimates (a-values) ranged from 0.14 for item 32 to 0.52 for item 40.The item discrimination can also be referred to as the point biserial correlation, which normally ranges from 0.00 to 1.00 and the higher the value, the more discriminating the item. A highly discriminating item should indicate that students with high test scores responded correctly whereas students with low test scores responded incorrectly. Brown (1996) suggested the following guidelines based on discrimination index (DI)in categorizing test items, ' the list of criteria below is based on a range of discrimination index (DI) that categorically defines the items in a test as follows: if DI ≥ 0.40, the item is functioning quite satisfactory; if 0.30≤ DI≤0.39, little or no revision of the item is required; if 0.20≤ DI≤0.29 the item is marginal and needs revision; if DI ≤ 0.19, the items should be eliminated or completely revised'. From table (iv), fourteen (14) items could be categorized as good items, because their discrimination index (DI) was from 0.30 ≤ DI ≥0.4, seven (7) items could be classified as items needing little or no revision. Items 30 and 32 had discrimination values less than 0.19 and they could be classified as poor items which should be eliminated or completely revised as suggested by Brown (1996).

Research question 2

What are the test items parameter estimates of students' responses to JC mathematics paper 1 items using IRT (3PLM) model?

In case of IRT model the item difficulty parameter estimates (p-values) ranged from 0.58 for item 23 to 1.82 for item 34. Test items with high b-values are normally hard items under IRT model; these are the items that low-ability examinees are unlikely to answer correctly. But items with low b-values are classified as easy items; these are items that most examinees including the low ability will have at least a moderate chance of answering correctly. From table (iv), all items with a b-value greater than 1.0 were classified as difficult items. Out of the twenty-three (23) items, eleven (11) items were difficult. The discrimination parameter estimates (a) values ranged from -0.20 for item 27 to 2.49 for item 39.The discrimination value expresses how well an item can differentiate among examinees with different ability

Table 1. Total Variance Explained by the result of factor analysis

	Initial Eigen values		
Component	**Total**	**% of Variance**	**Cumulative %**
1	5.909	14.772	14.772
2	1.492	3.730	18.502
3	1.096	2.740	21.242
4	1.088	2.720	23.963
5	1.060	2.649	26.612
6	1.029	2.573	29.185
7	1.022	2.554	31.739
8	1.017	2.543	34.282
9	1.010	2.524	36.806
10	.986	2.466	39.272
11	.964	2.410	41.682
12	.955	2.387	44.070
13	.952	2.381	46.450
14	.946	2.364	48.814
15	.934	2.334	51.149
16	.926	2.314	53.463
17	.912	2.281	55.744
18	.901	2.253	57.997
19	.890	2.224	60.222
20	.870	2.175	62.396
21	.859	2.147	64.543
22	.846	2.116	66.659
23	.833	2.082	68.742
24	.823	2.058	70.799
25	.818	2.045	72.845
26	.800	2.000	74.844
27	.791	1.978	76.822
28	.782	1.955	78.778
29	.777	1.943	80.721
30	.769	1.921	82.642
31	.754	1.885	84.527
32	.751	1.877	86.404
33	.724	1.811	88.215
34	.717	1.793	90.008
35	.710	1.775	91.783
36	.695	1.737	93.519
37	.679	1.697	95.216
38	.654	1.634	96.850
39	.639	1.598	98.448
40	.621	1.552	100.000

Extraction Method: Principal Component Analysis.

levels. Good items usually have discrimination values ranging between 0.5 to 2.0. In IRT, this is illustrated by an ICC, the higher an item's discrimination value. High discrimination indicates that higher scoring examinees tend to answer the item correctly, while lower scoring examinees tend to answer the item incorrectly. From table (iv), only five (5) items out of twenty-three (23) items were not able to discriminate among the examinees, since their discrimination values were lower than 0.5.

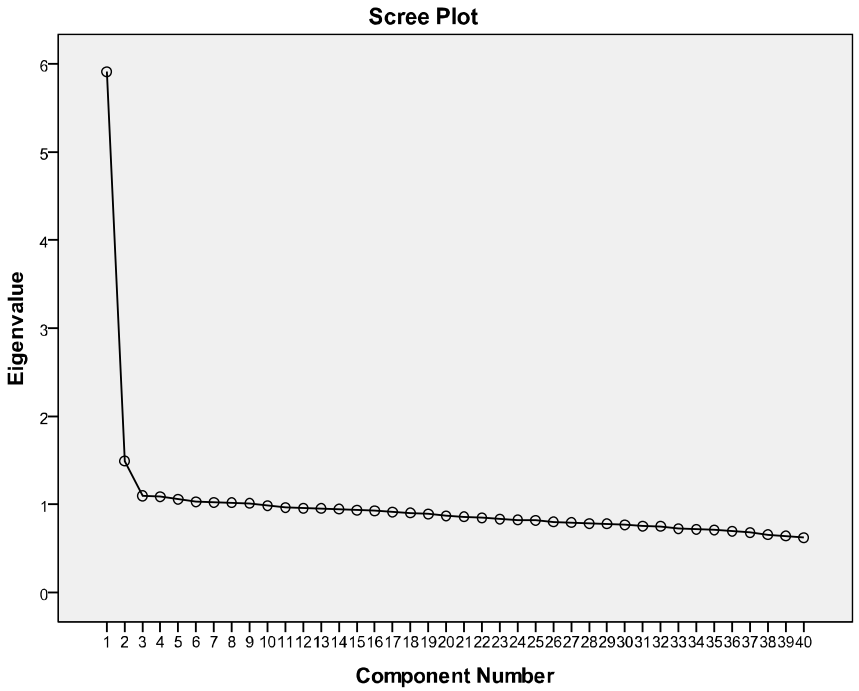

Figure 1. Scree Plot for the eigenvalues

Table 2. Results of the chi-square statistics for the 1PL, 2PL and 3PL IRT models

Items	1PL			2PL			3PL		
	Chi-square	**p**	**df**	**Chi-square**	**p**	**df**	**Chi-square**	**p**	**df**
1	60.3	0.0000	9.0	15.4	0.0812**	9.0	16.4	0.0585**	9.0
2	94.3	0.0000	9.0	34.2	0.0001	9.0	27.0	0.0014	9.0
3	361.1	0.0000	9.0	15.5	0.0000	9.0	13.7	0.0011	9.0
4	247.6	0.0000	9.0	25.0	0.0016	8.0	13.2	0.1556**	9.0
5	154.4	0.0000	9.0	78.8	0.0000	8.0	67.0	0.0000	8.0
6	82.3	0.0000	9.0	13.5	0.1409**	9.0	9.7	0.3741**	9.0
7	28.3	0.0008	9.0	15.5	0.0788**	9.0	5.7	0.773**	9.0
8	240.4	0.0000	9.0	61.8	0.0000	9.0	44.7	0.0000	9.0
9	428.4	0.0000	8.0						
10	16.4	0.0593**	9.0	12.3	0.1967**	9.0	7.9	0.5469**	9.0
11	63.3	0.0000	9.0	56.4	0.0000	9.0	41.6	0.0000	9.0
12	112.0	0.0000	9.0	21.8	0.0094	9.0	19.9	0.0182	9.0
13	329.6	0.0000	9.0	53.2	0.0000	9.0	67.4	0.0000	9.0
14	38.7	0.0000	9.0	8.6	0.4726**	9.0	6.7	0.6718**	9.0
15	45.0	0.0000	9.0	11.6	0.2356**	9.0	16.4	0.0596**	9.0
16	85.9	0.0000	9.0	58.1	0.0000	9.0	42.9	0.0000	9.0
17	429.8	0.0000	9.0	69.6	0.0000	7.0	35.0	0.0000	8.0
18	76.6	0.0000	9.0	24.8	0.0032	9.0	19.1	0.0242	9.0
19	222.5	0.0000	9.0	29.3	0.0003	8.0	28.7	0.0007	9.0
20	387.5	0.0000	8.0	20.6	0.0083	8.0	20.6	0.0146	9.0
21	21.7	0.0099	9.0	21.1	0.0122	9.0	6.6	0.6812**	9.0
22	405.9	0.0000	8.0	58.6	0.0000	8.0	39.2	0.0000	8.0
23	57.8	0.0000	9.0	15.3	0.0840**	9.0	12.8	0.1736**	9.0
24	93.6	0.0000	9.0	60.9	0.0000	9.0	12.7	0.1747**	9.0

25	140.3	0.0000	9.0	40.3	0.0000	9.0	14.1	0.1196**	9.0
26	57.7	0.0000	9.0	64.9	0.0000	9.0	5.8	0.7557**	9.0
27	46.8	0.0000	9.0	14.4	0.1080**	9.0	15.6	0.0749**	9.0
28	252.5	0.0000	9.0	39.3	0.0000	9.0	27.7	0.0011	9.0
29	47.6	0.0000	9.0	18.2	0.0330	9.0	6.0	0.7404**	9.0
30	143.8	0.0000	9.0	31.5	0.0002	9.0	8.3	0.4999**	9.0
31	180.5	0.0000	9.0	20.9	0.0075	8.0	10.3	0.3234**	9.0
32	194.3	0.0000	9.0	28.8	0.0007	9.0	9.0	0.4404**	9.0
33	47.9	0.0000	9.0	10.3	0.3294**	9.0	4.6	0.8708**	9.0
34	128.6	0.0000	9.0	98.4	0.0000	9.0	12.8	0.1723**	9.0
35	171.7	0.0000	9.0	23.5	0.0028	8.0	14.3	0.1126**	9.0
36	103.9	0.0000	9.0	41.7	0.0000	9.0	40.8	0.0000	9.0
37	146.5	0.0000	9.0	29.6	0.0003	8.0	24.8	0.0032	9.0
38	68.7	0.0000	9.0	19.4	0.0222	9.0	14.1	0.1204**	9.0
39	97.8	0.0000	9.0	4.0	0.9111**	9.0	8.4	0.4980**	9.0
40	136.1	0.0000	9.0	17.0	0.0298	8.0	15.1	0.0893**	9.0

**The items with probability greater than the alpha level of 0.05 significant level.

Table 3. The number of items fitting each model

IRT model	1PL	2PL	3PL
Items fitting the model	10	1, 6, 7, 10, 14, 15, 23, 27, 33, 39	1, 4, 6, 7, 10, 14, 15, 21, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40
Number of items fitting the model	1	10	23

Table 4. Item Parameters estimates using CTT and IRT (3PLM) models

ITEMS	CTT p-values	IRT p- values	CTT a-values	IRT a-values
1	0.71	0.69	0.38	0.26
4	0.38	1.33	0.49	0.65
6	0.30	0.90	0.38	0.99
7	0.38	0.97	0.38	1.03
10	0.44	0.65	0.28	1.28
14	0.36	0.60	0.21	1.71
15	0.60	0.86	0.24	1.18
21	0.45	1.15	0.32	1.05
23	0.51	0.55	0.37	-0.14
24	0.39	1.08	0.25	1.61
25	0.32	1.39	0.41	0.94
26	0.19	1.38	0.26	1.81
27	0.58	0.59	0.38	-0.15
29	0.46	0.91	0.23	1.42
30	0.37	1.25	0.16	2.07
31	0.32	1.20	0.42	0.96
32	0.28	1.32	0.14	2.24
33	0.56	0.84	0.38	0.24
34	0.21	1.99	0.21	1.78
35	0.42	1.04	0.40	0.75
38	0.32	1.07	0.32	1.03
39	0.32	0.48	0.32	2.58
40	0.52	0.82	0.52	0.24

Table 5. Result of the dependent t-test between $CTTz_p$ and $IRTz_b$;$CTTz_a$ and $IRTz_a$,

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
$CTTz_p$-	$IRTz_b$	.00174	.88060	.18362	-.37906	.38254	.009	22	.993
$CTTz_a$-	$IRTz_a$	.10348	.97735	.20379	-.31916	.52612	.508	22	.617

Research question 3

Is there any significant statistical comparability between the CTT and IRT (3PLM) item parameter estimates of students' responses to JC mathematics paper 1?

The CTT p-values and IRT a-values test items parameter estimates were changed into standard z-values, after which $CTTz_p$ and $IRTz_b$ were correlated, and also the standardized z-values of CTT item discrimination a-value ($CTTz_a$) and IRT item discrimination a-value ($IRTz_a$) were also correlated to find the extent to which CTT test item parameter estimates was comparable to IRT test item parameter estimates (Table 5)

It was found that $CTTz_p$ and $IRTz_b$ were positively correlated with correlation coefficient value of $r(22)= .612$, $p <.01$ which indicated that the correlation was statistically significant. This also showed that the p-values of CTT and b-values of IRT were comparable and they could both be used independently to estimate the test items parameters. The values of $CTTz_a$ and $IRTz_a$ were also positively correlated with correlation coefficient value of $r(22)= .521$, $p< .05$, which indicated that the correlation was statistically significant. Apart from correlating the values of $CTTz_p$ and $IRTz_b$;$CTTz_a$ and $IRTz_a$, dependent t-test was used to find out if the item difficulty and item discrimination parameter estimates by CTT and IRT were statistically significant. It was found from the analysis that there was no statistical significant difference between the item difficulty parameter estimates by CTT and IRT ($t_{22} = .009$, $p > .05$), and it was also revealed that there was no statistical significant difference between the item discrimination parameter estimates by CTT and IRT($t_{22} = .508$, $p > .05$) from table below

DISCUSSION AND CONCLUSION

This study assessed the comparability between the estimation of test items parameters using Classical test theory (CTT) and Item response theory IRT (3PLM) models, using 2010 Botswana JC mathematics paper 1 test items. From the results of the test items parameter estimates of CTT and IRT, the item difficulty and item discrimination values of the two measurement theories correlated positively. It was evident that there was no significant difference between CTT and IRT item parameter estimates. This also showed that the p-values of CTT and b-values of IRT were comparable and also the a-values of CTT were comparable with the a-values of IRT. It can then be concluded that both the item parameter estimates of CTT and IRT could be used independently to estimate the test items parameters, which revealed that the parameter estimates of the two measurement frameworks were comparable. These findings seemed to be consistent with the previous studies (e.g Fan (1998)) who indicated that CTT test items parameter estimates were comparable to IRT (3PLM) test items parameter estimates. Other studies, like Stage (1999) and MacDonald and Paunonen (2002), also indicated that CTT and IRT measurement theories often produce quite similar results in computing for test items parameter estimates. It can be concluded from this study that CTT test items parameter estimates are quite similar and comparable to IRT (3PLM) test items parameter estimates. There is a further need to investigate or assess the comparability of CTT with IRT (1PLM) and IRT (2PLM) in estimating test items parameters.

Results of the study could be affected by research instruments not been standardised for different ethnic or cultural examinees. The issue of standardization of the research instruments in terms of ethnic or cultural differences of the examinees could also affect the assessment of the comparability between CTT and IRT test item parameters estimates is the validity and reliability. According to Tilov, et al (2012), 'standardized research instruments facilitate data collection and monitoring, yet standardization requires that research instruments be adapted for use in different cultures because of the increase in diversity in research that examines practices across different cultures'. These researchers specified that there could be challenges in terms of specific properties when selecting and adapting research instruments using psychometric properties like validity, appropriateness, reliability and responsiveness as well as feasibility and acceptability of the research. The overall results on assessing the comparability of test items parameter estimates between the two measurement theories CTT and IRT could also be affected due to ethnic and cultural background of the examinees, because ethnic, language and cultural background of the examinees are also an important

factor of psychological and educational testing.

REFERENCE

Adedoyin OO (2010). Using IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics *Educational Research and Reviews*. 5 (7): 385-399.

Baker FB (2001). *The Basic of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, USA.

Brown JD (1996). *Testing in Language Programs*. New Jersey: Prentice Hall Regents.

Chernyshenko OO, Stark S, Chan K, Drasgow F, William B (2001). Fitting Item Response Theory Models to Two Personality Inventories: Issues and Insight. *Multivariate Behavioral Research*. *36* (4): 523-562

classical test theory: A study of the SweSAT test READ. (Educational Measurement No 31).Umea University, Department of Educational Measurement.Conditions, consequences and Goodness of fit test. *Methods of psychological Research online* 7: 1-12.

Embretson SE, Reise SP (2000). *Item Response Theory for Psychologists.* New Jersey: Lawrence Erlbaum Associates.

Fan X (1998). Item response theory and classical test theory: A comparison of their item/person

Hambleton RK, Swaminathan H, Roger HJ (1991). *Fundamentals of item response theory*. Newbury Park. C.A: Sage

MacDonald P, Paunonen S (2002). A Monte Carlo comparison of item and person

Magno C (2009). Demonstrating the difference between Classical Test Theory and Item

Ponocny I (2002). The Applicability of some IRT Models for Repeated Measurement Designs; Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment*.**1** (1): 1-11

Stage C (1999). A comparison between item analysis based on item response theory and statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*. *62*(6): 921-943.

statistics. *Education and Psychological Measurement*. 58: 357-381.

Svend K, Christensen KB (2002). Analysis of local Independence multidimensional in graphical loglinearRasch Models. *Education and Psychological Measurement*. **4**5 (6): 856-865.

Tilov B, Dimitrova D, Stoykova M, Totnjova B, Foreva G, Stoyanov D (2012). Cross-cultural validation of the revised temperament and character inventory in the Bulgarian language, *JECP* 18,: 1180-1185.