THE INVARIANCE OF ITEM RESPONSE THEORY (IRT) PARAMETER ESTIMATES AND CLASSICAL TEST THEORY (CTT) STATISTICS

Kesamang Monamodi E. E. Ph.D.*

Abstract

The paper intends to investigate the invariance of the parameter estimates generated from Item Response Theory (IRT) models and those generated from the Classical Test Theory (CTT) model. There have been problems associated the Classical Test Theory or number-right scoring system used by many assessment bodies in Africa. These problems include the need for parallel test forms each year which themselves are not easy to construct. The statistics from CTT are examinee dependent and the performance on a test depends on the ability of examinees. This study investigated the invariance of CTT item/person statistics and IRT item/person parameter estimates.

It has been assumed rather casually that IRT parameter estimates are more reliable to CTT item/person statistics. Recent research (for example, Fan, 1998) paint a different picture that the two assessment framework actually yield estimates that are comparable than previously thought. The item and person statistics and parameter estimates for the tests are compared to establish whether there exists a significant difference amongst them. The sample of the study consist of about 2000 Form Three (Grade 10) examinees that were about to sit for their final JCE examinations in 2008. The other data set was the national data from the 2005 and 2006 cohort. The instruments were the 2005 and the 2006 Botswana Junior Certificate objective science papers.

* **Ph.D.Researcher**

The study intended to test the Invariance of Item/Person Parameter Estimates for CTT and IRT. The CTT p-values (difficulty) and their corresponding IRT b-parameter estimates are highly correlated for all the IRT models. For the CTT item discrimination index a, (item-test, point-biserial correlation-r_{pbs}) and the IRT item discrimination a-parameter (item slope parameter), the two statistics are highly correlated for the 1-PL and 2-PLbut not for the IRT 3-PL model. Notwithstanding this, the CTT item/person statistics and IRT item/person parameter estimates are invariant and similar conclusions could be drawn irrespective of which method was used to estimate the examinee's ability.

Key words: Item Response Theory, Classical Test Theory, Invariance, Comparability

Background

This empirical study compares Classical Test Theory (CTT) item/person statistics and Item Response Theory(IRT)item/persons parameter estimates. The IRT scoring procedures are gaining recognition because of the problems associated with the (Number-right) type of scoring. CTT as a testing framework makes implausible assumptions which many test-developers find difficulty in accepting. Standard setting procedures have been using the CTT methods in setting performance standards. The number-right type score is a Classical Test Theory framework and gives equal weightings to all items irrespective of their difficulty. CTT based scoring assumes that two examinees who scores a 1 on an item or who scores a 0 on an item are of the same ability. This assumption ignores the fact that the two examinees did not have the same probability of either getting the item correct (1) or the item incorrect (0). IRT as an item-based framework and probabilistic in nature, assumes that the two examinees may not necessarily have the same probability of either responding correctly or incorrectly to an item even though their responses were similar.Warm (1978) argues and rightly so, that, a score of 1 or 0 does not reflect either 100% or 0% ability on that item. He opines based on IRT that a true measure of ability could be measured by the examinee's degree of certainty in attempting an item. If the degree of certainty is 50%, that examinee should be awarded a partial credit of .5 on that item as a measure of his/her ability on that item. He argues that, an examinee's degree of certainty or the probability of getting an item correct, '…[P(θ)] might be interpreted as a measure of his knowledge, and is called his true score on the item. The sum of his/her true scores is his/her true

test score. His/her true test score is the raw score he/she would get, if there were no measurement error in the test' (p.59).

Item Response Theorylogistic models estimates the person's ability (θ) by approximating the trait level using the Likelihood Estimation (LE). This LE uses the response function or vector of an examinee to calculate the approximate or Likelihood ability level by the use of Estimation-Maximization (EM) algorithm. The response function of a dichotomously scored test is a series of 1's and 0's indicating the items scored correctly (1) and incorrectly (0) by an examinee in a test.Classical Test Theory item statistics are examinee's dependent. A group of high ability examinees will tend to score high on a test than a group of low ability examinees on the same test. CTT scoring is based on the total number-correct scores. This is because the individual examinee is scored on the item he/she answers correctly. In IRT the examinee's score is not only dependent on the total examinee's score, but also on the statistical characteristics of items scored correctly or incorrectly (Weiss and Yoes, 1991). CTT also relies on reliability based on the true score. The true score cannot be directly measured, but can only be estimated from the observed score. The coefficient of reliability is also examinee dependent. Standard error of measurement (SEM) is calculated from this reliability coefficient.

Classical Test Theory,also known as the true score theory does not detect item bias, this is because the validity of scores from such measurements are influenced by many factors aside from ability. The scores are examinee and item dependent. The interpretation of such scores may be misleading. This could lead to some subgroups e.g. rural, minorities being disadvantaged by being denied access, and not treated equally because they scored lower even when they are of the same ability as the other sub-groups. According to Hambleton and Jones (1995) therefore, it

is obviously desirable to have (i) item statistics that are not group dependent, (ii) scores describing examinees proficiency that are not dependent on test difficulty, (iii) test models that provide the basis for matching test items to ability level, (iv) test models that are not based on implausible assumptions, and (v) test models that do not require strictly parallel tests for assessing reliability (p.418).

Classical Test Theory measurements of examinees are test dependent and that of the items or tests are examinee dependent. Examinees will tend to score low in a very difficult test, for example and on the other hand the nature of the items or test determines the performance of the examinees. The difficulty of an item in CTT is not inherent in the item or test itself, but is related to the ability of examinee taking it (Nenty, 2004). On the other hand, ability that is an inherent attribute of an individual is dependent on the type of test administered to him/her. Therefore a score of 60 may be equivalent to another score of 30 in another test measuring the same latent trait. A human being has only one trait level on any attribute, but CTT would tend to designate two or more trait levels (60 and 30) to on an individual. According to Nenty (2004), CTT estimates of this trait level are fundamentally flawed.

Item Response Models try to estimate the likelihood of some reaction by an examinee with a cognitive latent trait (ability) level encountering an item with some cognitive resistance overcoming such an item. This likelihood attempts to predict the outcome of such an encounter (Nenty, 2004). The probability of success is dependent on the ability level and item resistance. If the response pattern of an examinee to some set of items is known, the ability parameter could be estimated by IRT models. The IRT models provide the quantitative basis for computing the probability that an examinee will answer a specific item correctly based on the characteristics of the item as a function of the examinee's ability.

The invariance nature of IRT models

Aside from the assumptions of unidimensionality and local independence, IRT property of invariance of item and ability parameters is the basis for the strength of the item response models. This property stands on the premise that item parameters a, b, and c do not depend on the ability (θ) distribution of examinees and that the parameter that characterises an examinee does not depend on any particular set of items.

This is a very important property in that according to its premise, it does not matter whether examinees sit two different sets of items as long as the IRT model used fits the data, the same Item Characteristic Curves (ICC) will be obtained over the distribution of ability in the group of examinees used to estimate item parameters. The item parameters will remain invariant across

the two groups of examinees. The property of invariance also states that the θ level of an examinee is not dependent on any set of test items an examinee takes. This property makes it possible to make an estimation of item parameters and hence an estimation of ability of an examinee. The maximum likelihood estimate (MLE) gives an estimation of θ for an examinee given some response pattern (U), Warm (1978). On the other hand the classical item difficulty p-value is dependent on the ability of an examinee.

Selected Literature on IRT and CTT Invariance Property

Comparison of CTT and IRT item/person statistics

There has been some interest in investigating the item/person statistics for the CTT and IRT as measurement frameworks. Work done by researchers such as Hambleton and Jones (1993); Fan (1998); Stage 2003, 1998a, 1998b, 1999; Bechger, Gunter, Huub, andBeguin, 2003; Wiberg 2004; Mellenbergh1996; Adedoyin, Nenty and Chilisa 2008 and others suggest that this field of research has gained some interest in recent times. Recent work has been centred on the comparability of CTT item/personstatistics and IRT item/person parameter estimates. Fan 1998, investigated empirically the comparability of CTT and IRT statistics and their invariance. Fan 1998, acknowledges that despite the theoretical weaknesses of CTT as compared to IRT, there are few studies which have empirically examined the similarities and differences between the two frameworks. Hambleton and Jones 1993 pointed out that the attraction for CTT is its relatively weak theoretical assumptions which make it easy to apply in many testing situations. These weak assumptions make it easier for one to work with and interpret test scores. Lawson 1991, empirically compared item/person CTT statistics and IRT parameter estimates on three data sets using the Rasch model and found a strong correlation between the two methods, similar results across some independent samples of candidates were obtained by Adedoyin2004, working on the 2004 Botswana Junior Certificate Examinations (JCE) mathematics. Even though Lawson 1991, used small data set, Fan 1998, using many large sample sizes (each sample consisting of 1000 examinees) arrived at the same conclusions that the statistics from the two frameworks are comparable. Fan 1998, used the data from Texas Assessment of Academic Skills (TAAS) data base which was administered to 11 year old students. TAAS is a criterion referenced test battery for maths, reading and writing skills.

The work of Fan 1998, indicated that (i) the person and the item statistics for the two frameworks were highly comparable, (ii) the difficulty parameters are more correlated than the discrimination parameter for the two methods, (iii) the two methods produce statistics that are invariant across samples, but the discrimination parameters are less invariant than the difficulty parameter. This could mean that IRT models are more invariant and robust in estimating discrimination parameter than CTT. The weakness of this study was that this comparability on the statistics was not subjected to some form of significance testing, which could have given it more authority.

Stage 1998a, 1998b, 1999, 2003, compared CTT and IRT statistics on subtests and the total test on the Swedish national achievement tests known as *SweSAT* (Swedish Scholastic Aptitude Test) which tests proficiencies in maths, interpretation of diagrams, tables, maps, English reading comprehension, Swedish reading comprehension, and a vocabulary test. She found that the two frameworks are comparable and has this to say, "In the studies reported in this paper, the CTT statistics were not only comparable to the IRT parameters, they were generally more invariant between different samples of test takers. One possible explanation of these results is that the IRT model did not fit the test data. But even if the results are due to poor model fit, the only reasonable conclusion is that for SweSAT data, CTT seems to work better than IRT" (p.25). She reached a conclusion that, '…since the model data fit was somewhat dubious, especially for the total test, there was nothing to be gained by switching from CTT to IRT'.

The question of goodness of fit arises when the IRT model does not fit the test data. Hambleton,Swaminathan and Rogers 1991, states that a poorly fitting IRT model will not yield invariant parameters, and that "In many IRT applications reported in the literature, model-data fit and the consequences of misfit have not been investigated adequately. As a result, less is known about appropriateness of particular IRT models for various applications than might be assumed from the voluminous IRT literature" (p.53). The authors goes on further and warn against placing too much confidence in statistical tests, which also have a serious flaw, sensitivity to examinee sample size. The authors simulated some data set on five different sample sizes, and found that small samples do not lead to item misfit, but large sample lead to more items being rejected for model-data misfit.

Monte Carlo simulation by MacDonald and Paunonen2002, on the item/person CTT and IRT statistics showed that they are highly comparable, invariant and accurate in the test conditions simulated. Item discrimination was only accurate with CTT in some conditions. This conclusion is similar to the one reached by several authors such as Lawson 1991, and Fan 1998. Ndalichako and Rogers 1997, found that item and person parameter estimates of these two frameworks are not only comparable, but the two parameters correlate almost perfectly. Encouraged by such results, they favoured the continued use of CTT for test scoring and item analysis.

Studies on IRT have been more focussed on the applications of IRT on stability of item/person parameter estimates, test equating and Computer Adaptive Test (CAT) but not on the comparability of CTT and IRT analysis and item/person estimates. Fan 1998, found this disturbing by saying "It is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scare. It appears that the superiority of IRT over CTT in this regard has been taken for granted in the measurement community, and an empirical scrutiny has been deemed unnecessary. The empirical silence on this issue seems to be an anomaly, p361.

Research hypothesis

The purpose of this comparative study is to investigate the CTT item/person statistics and IRT item/person parameter estimates for invariance. Can the same conclusions be drawn or reached in the estimation of examinee ability irrespective of which measurement framework is employed?

The following research hypothesis were formulated and tested at an alpha level of 0.05, and these are,

1. Ho: There is no significant relationship in the CTT item statistics and IRT item parameter estimates for a) the Botswana JCE 2005 science paper and b) the Botswana JCE 2006 science paper.

2. Ho: There is no significant relationship in the CTT-based person statistics (total score) and IRT person parameter estimates (theta) for a) the Botswana JCE 2005 science paper and b) the Botswana JCE 2006 science paper.

Limitation of the study

The study envisages applying IRT analysis model on test items, which were constructed using CTT framework or guidelines. It is quite possible that most items/persons may not fit the chosen IRT model used. If this proves to be the case, then the interpretation of the findings may not be very valid or may be open to some criticism.

Research methodology

The study employs a quantitative analysis in the estimation of CTT and IRT item/person parameter estimates tests. For IRT,1-PL, 2-PL and 3-PL IRT models wereemployed to calibrate item parameter estimates. The programs that were used are Parscale and Xcalibre.Iteman program was used to generate the CTT item/person statistics. The population of the study was the Form Three examinees of 2008. The sample in this study refers to about 2500 students from this group who took the tests. The data was collected by the use of two instruments. The JCE data refers to the data from the Junior Certificate Examination cohort for 2005 and 2006 examination years.

Hypotheses- 1

Correlation of CTT item statistics and IRT item parameter estimates for a) the Botswana 2005 and b) 2006 national science examinations

a) H_0: There is no significant relationship in the CTT item statistics and IRT item parameter estimates for the Botswana 2005 National Science Examination.

b) H_0: There is no significant relationship in the CTT item statistics and IRT item parameter estimates for the Botswana 2006 National Science Examination.

The CTT item statistics and IRT item parameter estimates were correlated. These statistics are the (i) the difficulty p-value from CTT and the item difficulty b-parameter (item location parameter) from IRT 1-PL, 2-PL and 3-PL models, and (ii) the CTT item discrimination index a, (item-test, point-biserial correlation) and the IRT item discrimination a-parameter (item slope parameter). The r_{pbs} (*r-point biserial*) was bias corrected by removing the contribution of an item to the total score before calculating the r_{pbs} for the item.

Hypotheses-2

Correlation of CTT person statistics and IRT person parameter estimates for a) the Botswana 2005 and b) 2006 national science examinations

a) H_0: There is no significant relationship in the CTT-based person statistics (total score) and IRT person parameter (theta) estimates for the Botswana 2005 National Science Examination.

b) H_0: There is no significant relationship in the CTT-based person statistics (total score) and IRT person parameter (theta) estimates for the Botswana 2006 National Science Examination.

The CTT person statistics and IRT person parameter estimateswere compared. These statistics are the total score, T for an examinee from CTT and the ability or the theta (θ)parameter from IRT. The comparability was done for the1-PL, 2-PL and the 3-PL models. The total score T, was the total raw examinee score on the test. The theta of examinees was correlated to the total score T.

Results

Item response theory imposes some restrictions on the type of data used for analysis. The data used for IRT models should (i) fit the type of IRT model used, (ii) adhere to the assumption of unidimensionality, that is, the data set must be a measure of only a single trait and (iii) satisfy sampling adequacy and sphericity. The instruments for the study have been summarised in Table 1. The two papers are the 2005 and the 2006 Botswana national science examination objective papers. The papers were also sat for by a sample in 2008 and the JCE data comes from the national data base for the two cohorts that sat for the 2005 and 2006 papers respectively.

Table 1: The Summary Results of the Two Instruments

	NO# of Items	Paper	Sample/cohort	Year administered
Sample	40	2005	**1386**	2008
	40	2006	**1287**	2008
JCE	40	2005	**36380**	2005
	40	2006	**36383**	2006

Principal Component Analysis (PCA)was run on the data set to determine whether the data set measures one or more constructs. It is expected that the two Botswana National Science Examination papers would come up with one dominant factor. This factor would represent the construct underlying the science skills measured by the examinations. The two papers produced one dominant factor, the scree plots showed this characteristic steep slope for this factor.

A Chi-square test was run on the data sets to establish whether the test data/items fit the 1 PL, 2-PL and the 3-PL IRT Xcalibre and Parscale IRT Models. Chi-square test for Model-data Fit for the two instruments (2005 and 2006) indicated that the instruments fit the 1-PL, 2-PL and the 3-PL Xcalibre andParscale IRT models. For example, for the 3-PL Parscale Model, only six items out of the forty (40) items did not fit this IRT model for the 2005 data set and five items for the 2006 data set did not fit this model.

1. Testing hypotheses on the invariance of item parameter estimates for CTT and IRT for the 2005 and 2006 Botswana JCE science examination papers

(a) XCalibre output- 1-PL, 2-PL &3-PL

H_0:1 There is no significant relationship in the CTT item statistics and IRT item parameter estimates for the Botswana 2005 National Science Examination.

H_0:2 There is no significant relationship in the CTT item statistics and IRT item parameter stimates for the Botswana 2006 National Science Examination.

CTT and IRT item difficulty parameter estimates

The CTT item statistics were correlated to the IRT item parameter estimates (from Xcalibre program).Table 2 shows the results for the Pearson correlation coefficients$r_{XY}$for CTT p-values vs IRT b-parameter and the CTT item-total correlation (the point-biserial correlation), and the IRT slope parameter (the a-parameter) for both years of the examinations for the sample group and the national data from the 2005 and 2006 cohorts. The item-total correlation was corrected for item-bias by removing the item when this correlation was run. The CTT item/person statistics were compared with the IRT item/person parameter estimates for all the three models.

Table 2: *The Pearson CorrelationAnalysis for CTT statistics vs IRT parameters for both 2005 and 2006Examinations from* ***Xcalibre***

			Sample	JCE Examination
CTT	Vs	IRT	r_{xy}	r_{xy}
2005				
p-value		b-parameter (1-PL)	.950*	.819*
p-value		b-parameter (2-PL)	.967*	.906*
r_{pbs}		a-parameter (2-PL)	.906*	.954*
p-value		b-parameter (3-PL)	.834*	.856*
r_{pbs}		a-parameter (3-PL)	.032	.035
2006				
p-value		b-parameter (1-PL)	.949*	.915*
p-value		b-parameter (2-PL)	.985*	.953*
r_{pbs}		a-parameter (2-PL)	.897*	.851*
p-value		b-parameter (3-PL)	.912*	.954*
r_{pbs}		a-parameter (3-PL)	.180	.310

**Correlation significant at the 0.05 level (2-tailed).*

The p-values from CTT and their corresponding IRT b-parameter for the sample have *r* = .950 and .949 (1-PL) for the 2005 and 2006 respectively. These are relatively high correlation and this would mean that it would not matter which framework one uses for estimating the item difficulty, the same conclusions will still be made. The CTT p-values and IRT item difficulty parameters are strongly correlated for the two examinations. The correlations for the CTT and the IRT 2-PL and 3-PL model are also very high. The 3-PL statistics are slightly lower when compared to those obtained from the 1-PL and 2-PL models, but are still relatively high.Similar results are obtained for the JCE national examination. The CTT framework gives similar estimates about item difficulty as IRT models. It should be noted though that the 2-PL model appears more robust in estimating item difficulty than the 3-PL model. The 1-PL gives relatively high correlations for both data sets. All the coefficients are statistically significant.

CTT and IRT item discrimination parameter estimates

The CTT item discrimination statistics (item–total correlation or the r-point-biserial) were correlated to the IRT item slope parameter, the a-parameter for the two examinations for the 2-PL and 3-PL models. For the sample the *r*= .906 and .897 for the 2005 and 2006 were obtained

for the CTT values and the IRT 2-PL model. These statistics are strongly correlated. It therefore would not matter which framework is used to estimate the item discrimination among examinees, as the same conclusions will still be made. All the coefficients are statistically significant. For the 3-PL model the *r*= .032 and .180 for the 2005 and 2006 were obtained for the CTT values and the IRT model. Similar results were obtained for the JCE national examination. These statistics (r_{XY}) are far too low and not significant, and these could indicate that the 3-PL model produces seriously unstable item-slope parameters estimates. Very unreliable conclusions could be drawn if both the CTT item discrimination statistics and IRT item slope parameter estimates from the 3-PL model are used.

(b) Parscale Output- 1-PL, 2-PL & 3-PL

A similar correlation was carried out using the Parscale analysis. Table 3 shows the Pearson correlation coefficients for CTT p-values vs IRT b-parameter and the CTT item-total correlation, and the IRT the a-parameter for both 2005 and 2006 examinations from Parscale. The analysis shows the results from the sample and the JCE examination data.

CTT and IRT item difficulty parameter estimates

The p-values from CTT and their corresponding IRT b-parameter for the sample are*r* = .919 and .870 (2-PL) for the 2005 and 2006 respectively. The 1-PL model shows very high correlation coefficients all very close to near perfect. These are relatively high correlation and this would mean that it would not matter which framework one uses for estimating the item difficulty, the same conclusions will still be made. Correlations for the CTT and the IRT 3-PL model are *r* = .850 and .937 for the 2005 and 2006 respectively. These statistics are slightly lower when compared to those obtained from the 1-PL and 2-PL models, but are still relatively high. The CTT framework gives similar estimates about item difficulty as all IRT models. Relatively similar results are obtained for the national data.

Table 3: *The Pearson Correlation Analysis for CTT statistics vs IRT parameters for both 2005 and 2006 Examinations from* ***Parscale***

	Sample	**JCE Examination**

CTT	Vs	IRT	r_{XY}	r_{XY}
2005				
p-value		b-parameter (1-PL)	.998	.998
p-value		b-parameter (2-PL)	.919*	.906
r_{pbs}		a-parameter (2-PL)	.884*	.953
p-value		b-parameter (3-PL)	.850*	.860
r_{pbs}		a-parameter (3-PL)	.391	.232
2006				
p-value		b-parameter (1-PL)	.996	.997
p-value		b-parameter (2-PL)	.870*	.963
r_{pbs}		a-parameter (2-PL)	.878*	.816
p-value		b-parameter (3-PL)	.937*	.955
r_{pbs}		a-parameter (3-PL)	.244	.312

**Correlation significant at the 0.05 level (2-tailed)*

CTT and IRT item discrimination parameter estimates

The CTT item discrimination statisticsfor the sample were correlated to the IRT item a-parameter for the two examinations for both the 2-PL and 3-PL models. The *r*= .884 and .878 for the 2005 and 2006 were obtained for the CTT values and the IRT 2-PL model respectively. These statistics are strongly correlated and statistically significant. It therefore would not matter which framework is used to estimate the item discrimination among examinees, as the same conclusions will still be arrived at. The correlation coefficients though are slightly lower than those obtained from the Xcalibreanalysis. For the 3-PL model the *r*= .391 and .244 for the 2005 and 2006 were obtained for the CTT values and the IRT model. These correlations are quite high if compared to similar correlation coefficients obtained from the Xcalibre program. These low correlation coefficients could indicate that the 3-PL model for both Xcalibre and Parscale produce seriously unstable item-slope parameter estimates. Very unreliable conclusions could therefore be drawn if the IRT item slope parameter estimates from the 3-PL model and their CTT counterparts (*r*- point-biserial) are used.

IRT item parameter estimates from Xcalibre correlated better with CTT statistics than the Parscale statistics. But Parscale gives better correlation coefficients for the 3-PL model a-parameter and CTT item-total correlation coefficient than Xcalibre. The same applies for the JCE examinations.

(c) The 3-PL Poor Estimation of the a-parameter

The 3-PL model both in the Xcalibre and Parscale programs seems to overestimating or underestimates the a-parameter. Figure 1 shows the scatter plot for 2005 JCE examination r_{pbs} and the a-parameter from the Parscale 3-PL model. The two variables are poorly correlated r = .035. A similar plot is observed with 3-PL Xcalibre output.

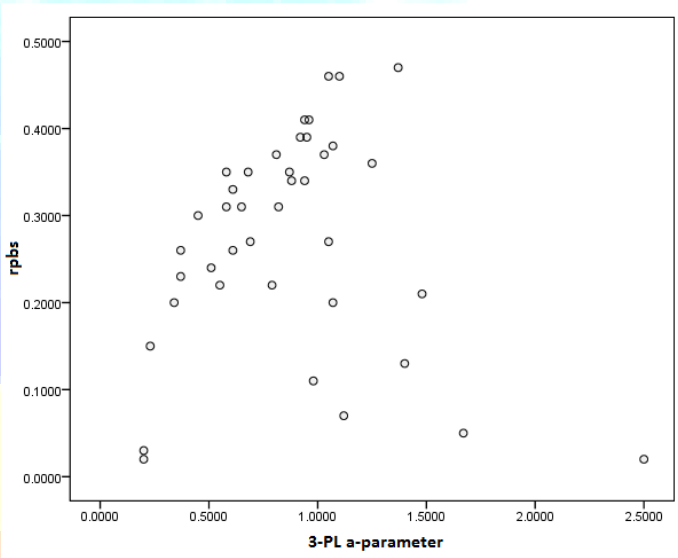

Figure 1:The scatter plot for 2005 JCE examination r_{pbs} and the a-parameter from Parscale 3-PL.

Table 4 shows the correlation of the Parscale 2-PL and 3-PL a-parameter outputs. The a-parameters are from Parscale 2005 JCE examinations. The two data set are poorly correlated as Pearson correlation coefficient r = .029is too low and not significant. The 3-PL tend to give higher a-parameter values than the 2-PL for an item. This may explain why the 3-PL correlates poorly with r_{pbs} from CTT.

Table 4: *The Pearson Correlation Coefficient Statistics forthe 2-PL and 3-PL a-parameter from Parscale Model-2005 JCE Examination*

Variable		3-PL
2-PL	Pearson Correlation	0.029
	Sig. (2-tailed)	.860
	N	40

*: *correlation is significant at 0.01 level (2-tailed).*

When the outputs of the 2-PL and 3-PL from Xcalibre and Parscale are correlated, it becomes clear that the 2-PL outputs correlate better than the 3-PL outputs from the two programs. This is shown in Table 5.

Table 5: *The Pearson Correlation Coefficient Statistics for the 2-PL and 3-PL Modelsfrom Xcalibre and Parscale for 2005 JCE Examination*

Variable		2-PLPars	3-PLX-Cal	3-PLPars
2-PLX-Cal	Pearson Correlation	**0.991***	0.029	0.235
	Sig. (2-tailed)	**.000**	.860	.144
	N	**40**	40	40
2-PLPars	Pearson Correlation		0.035	0.268
	Sig. (2-tailed)		.832	.094
	N		40	40
3-PLX-Cal	Pearson Correlation			**0.585***
	Sig. (2-tailed)			**.000**
	N			**40**

*: *correlation is significant at 0.01 level (2-tailed).*

Table 6 shows the CTT p-value, r_{pbs} and the 2-PL and the 3-PL a-parameter from both Xcalibre and Parscale for the first 20 items. On a closer look, it seems that for more difficult items (green) the 3-PL overestimate the a-parameter and for easier items (yellow) it underestimates the a-parameter.This may explain the low relationship between the 3-PL a-estimates from the two IRT programs and the CTT r_{pbs}. It could also be that the 3-PL takes care of guessing by

examinees. It may also be possible that the 3-PL model approximate the a-parameter better than the 2-PL model, this is more so because in the 2-PL model, the r_{pbs} forces a linear relationship on something known to be non-linear, the Item Response Function (IRF). Both the easier items and the most difficult items are highlighted. Items with a p-value < .30 were regarded as difficult and those with p-value > .60 were less demanding.

Table 6: *The p-value, the 2-PL and the 3-PLa-Parameter from both* Xcalibre *and Parscale*

			Xcalibre		**Parscale**	
	CTT		**2-PL**	**3-PL**	**2-PL**	**3-PL**
Item	p-values	r_{pbs}	a	a	A	a
1	.34	.36	0.61	1.25	0.555	1.32
2	.69	.39	0.77	0.95	0.793	0.894
3	.19	.13	0.27	1.40	0.206	1.465
4	.32	.02	0.20	2.50	0.041	0.043
5	.54	.24	0.42	0.51	0.345	0.468
6	.21	.21	0.32	1.48	0.319	1.542
7	.42	.38	0.62	1.07	0.585	1.102
8	.56	.22	0.41	0.55	0.314	0.517
9	.42	.41	0.69	0.94	0.655	0.895
10	.61	.37	0.65	1.03	0.634	1.006
11	.22	.07	0.22	1.12	0.101	1.392
12	.60	.22	0.40	0.79	0.303	0.753
13	.46	.31	0.49	0.58	0.439	0.547
14	.45	.23	0.41	0.37	0.314	0.341
15	.40	.27	0.44	1.05	0.375	1.027
16	.59	.46	0.87	1.10	0.907	1.12
17	.53	.41	0.71	0.96	0.708	0.906
18	.46	.15	0.31	0.23	0.207	0.183
19	.52	.30	0.47	0.45	0.443	0.37
20	.46	.34	0.54	0.88	0.503	0.869

Green= Difficult items

Yellow= Easy items

2. Testing hypotheses on the invariance of person parameter estimates for CTT and IRT for the 2005 and 2006 Botswana junior certificate science examination papers

H_0:1 There is no significant relationship in the CTT-based person statistics (total score) and IRT person parameter (theta,θ) estimates for the Botswana 2005 National Science Examination.

H_0:2 There is no significant relationship in the CTT-based person statistics (total score) and IRT person parameter (theta,θ) estimates for the Botswana 2006 National Science Examination.

Sample and JCE National Data (Xcalibre and Parscale outputs)

The correlation of examinee's ability was also investigated. The Pearson correlation coefficient for CTT and the IRT 1-PL, 2-PL and the 3-PL models were compared for the sample and the JCE national data. If the coefficient are high then the CTT and IRT ability statistics would mean that the two frameworks give similar statistics and therefore similar conclusions could be drawn irrespective of which framework is used.Table 7 shows the CTT total scores and the IRT theta values (ability estimates) for the first twenty (20) examinees for the two instruments for the sample from both Xcalibre and Parscale respectively. The IRT ability estimates were run for the 1-PL, 2-PL and the 3-PL models.

Table 7: *The CTT Total Scores and the IRT Theta Values for the First Twenty (10) Examinees for the 2005 and 2006 Botswana National Science Examinations for the Sample from* ***Xcalibre***

Sample		**2005**				**2006**		
Examinee No#	Total score	θ 1-PL	θ 2-PL	θ 3-PL	Total score	θ 1-PL	θ 2-PL	θ 3-PL
1	13	-0.85	-0.80	-0.69	19	-0.11	-0.11	0.13
2	13	-0.85	-0.90	-0.93	29	1.08	1.25	1.24
3	11	-1.13	-1.10	-6.55	23	0.39	0.43	0.67
4	10	-1.28	-1.11	-1.15	18	-0.23	-0.44	-0.47
5	15	-0.59	-0.46	-0.14	11	-1.28	-1.18	-1.63

6	21	0.02	0.24	0.47	17	-0.47	-0.61	-0.98
7	25	0.53	0.70	0.83	27	0.93	0.95	1.04
8	26	0.8	0.96	1.08	18	-0.35	-0.31	-0.05
9	14	-0.72	-0.69	-0.75	21	0.14	-0.08	0.00
10	18	-0.22	-0.32	-0.22	23	0.26	0.28	0.42

The CTT total score was correlated with the person parameter (Theta-θ) from the 1-PL, 2-PL and 3-PL IRT Models.

Correlation of CTT Total examinee scores and IRT person parameter estimates (Theta or ability)

Table 8 shows the summary of the results for all the examinees. The Pearson correlation coefficients for CTT and the 1-PL, 2-PL and the 3-PL IRT models yielded higher correlation coefficients, except r = .629 for the correlation coefficient for the 2005 3-PL model. The 1-PL model yields the highest r_{xy}-values perfect to near perfect correlations. For all IRT models, similar conclusions could be made irrespective of which framework was used to estimate the ability levels of a group of examinees. The 1-PL and 2-PL models seem to provide more robust estimates compared to the 3-PL models in ability estimation.

Table8: *The Pearson Correlation Analysis for CTT Total Score vs IRT Ability Score for both 2005 and 2006 Examinations from Xcalibre and Parscale for both the Sample and JCE Cohorts*

Sample				**JCE Examination**	
2005					
CTT	**Vs**	**IRT**	**Xcalibre** r_{XY}	**Xcalibre**	r_{XY}
Total score		θ (1-PL)	.991*	.997*	
Total score		θ (2-PL)	.985*	.985*	
Total score		θ (3-PL)	.795*	.972*	
2006			**Xcalibre**	**Xcalibre**	
Total score		θ (1-PL)	.990*	.991*	
Total score		θ (2-PL)	.977*	.990*	
Total score		θ (3-PL)	.837*	.975*	
2005			**Parscale**	**Parscale**	

CTT	Vs	IRT		
Total score		θ (1-PL)	.999*	1.000*
Total score		θ (2-PL)	.988*	.977*
Total score		θ (3-PL)	.629*	.968*
2006		**Parscale**		**Parscale**
Total score		θ (1-PL)	.999*	.999*
Total score		θ (2-PL)	.985*	.989*
Total score		θ (3-PL)	.973*	.968

**Correlation significant at the 0.05 level (2-tailed).*

The correlation coefficient from the national cohort is in agreement with the results from the sample of examinees from the same population. The strong correlation for the sample indicate that the sample of over 1200 examinees was large enough to yield reliable IRT parameter estimates similar to those from the whole population. This support the findings of Lord (1969), Warm (1978) and Fan (1998) that a sample of over 1000 examinees is sufficient enough for most IRT models to yield reliable estimates.

Discussions

The purpose of the study was to investigate the invariance of the CTT statistics and the IRT item/person parameter estimates. The item parameter estimates from IRT were the b-parameter (item location parameter) and the a-parameter (item slope parameter). The CTT statistics were the p-value (item difficulty) and the discrimination index (item-total correlation). The item/person parameter estimates from the IRT 1-PL, 2-PL and 3-PL models were used in these comparisons.

The invariance of item/person parameter estimates for CTT and IRT for the 2005 and 2006 Botswana junior certificate science examination papers

This study intended to establish whether it would make any difference if either CTT statistics or IRT item/person parameter estimates are used for national examination analysis and scoring to estimates examinee's ability. The results indicate that it would not matter whether CTT or IRT

frameworks are used in the estimation of examinee's ability as similar conclusions will be made. The results of this study suggest that it would not matter which framework is used in estimating examinee's ability, test difficulty and its discrimination power, one would still get the same results. This then would mean that the item/person parameter estimates or statistics are invariant for both CTT and IRT frameworks.

The p-values from CTT and their corresponding IRT b-parameters are highly correlated and the CTT item discrimination statistics (item–total correlation or the r-point-biserial) correlated highly with the IRT item a-parameter for all IRT models except the 3-PL models. The CTT item discrimination statistics correlated highly with the IRT item slope parameter for the 2-PL model. The b-parameter and the p-values correlated highly when compared to the correlation coefficient from the discrimination indices and the a-parameter. The 1-PL model had very high correlation coefficient compared to the other two models. The CTT item discrimination statistics from the 3-PL model yields very low correlation indices, and therefore very unreliable conclusions will be made. The examinees ability yields high Pearson correlation coefficient for CTT and the IRT 1-PL, 2-PL and the 3-PL models.

This findings are consistent with the work of Lawson 1991, whose study empirically compared item/person CTT statistics and IRT parameter estimates on three data sets using the Rasch model and found a strong correlation between the two methods, similar results across some independent samples of candidates were obtained by Adedoyin2004, working on the 2004 Botswana Junior Certificate Examinations (JCE) mathematics. Fan 1998, using many large sample sizes (each sample consisting of 1000 examinees) arrived at the same conclusions that the statistics from the two frameworks are comparable.

Stage,2003, 1998a, 1998b, 1999, compared CTT and IRT statistics on subtests and the total test on the Swedish national achievement tests known as *SweSAT* (Swedish Scholastic Aptitude Test) found that the two frameworks are.

Monte Carlo simulation by MacDonald and Paunonen,2002 on the item/person CTT and IRT statistics showed that they are highly comparable, invariant and accurate in the test conditions

simulated. Item discrimination were only accurate with CTT in some conditions. This conclusion is similar to the one reached by several authors such as Lawson,1991 and Fan, 1998. Ndalichako and Rogers,1997 found that item and person parameter estimates of these two frameworks are not only comparable, but the two parameters correlate almost perfectly.

Conclusions

As a way of concluding on the study findings, a note should be made on the main focus of the study, the invariance of the CTT statistics and IRT parameter estimates. The following conclusions were deduced from the study findings:

(a) The p-values from CTT and their corresponding IRT b-parameters are highly correlated. It therefore would not matter which framework one uses for estimating the item difficulty, the same conclusions will still be made. This applies for both for the 1-PL, 2-PL and the 3-PL models. The 1-PL models produced very high correlation coefficients compared to the other two models.

(b) The CTT item discrimination statistics (item–total correlation or the r-point-biserial) correlated highly with the IRT item a-parameter for the 1-PL and the 2-PL models. It therefore would not matter which framework one uses for estimating the item discrimination, the same conclusions will still be made. The 3-PL a-parameter and the CTT item discrimination statistics yields very low correlation indices, and therefore very unreliable conclusions could be drawn if both the CTT statistics and IRT item a-parameter estimates from the 3-PL model are used.

(c) The examinees ability yields high Pearson correlation coefficient for CTT and the IRT 1-PL, 2-PL and the 3-PL models. For all the IRT models, similar conclusions could be arrived at irrespective of which framework was used to estimate the ability levels of a group of examinees.

(d) A sample of just over 1000 examinees gives similar IRT item/person parameter estimates as the whole population of examinees.

The traditional or the number right scoring framework compares favourably with IRT models in estimating the test difficulty and examinee's ability. The traditional method of scoring is still useful in scoring and grading or selecting examinees for progression. One would still make the

same decision in placing examinees at different performance standards levels even when IRT models for scoring and grading have been employed.

References

- Adedoyin, O. O., (2006). Investigating the invariance of the item difficulty parameter estimates based on CTT and IRT. Unpublished Doctoral Thesis.
- Adedoyin, O. O., Nenty, H. J., &Chilisa, B., (2008). Investigating the invariance of the item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review*. 3(2), 83-93.
- Bechger, T. M., Gunter, M.., Huub, H., &Beguin, A., (2003). Using classical theory in combination with item response t.heory.*Applied Psychological Measurement*, 27(5), 319-334.
- Fan**,** X., (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58 (3), 357-381.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J., (1991): *Fundamentals of item response theory*. London: SAGE Publications.
- Hambleton, R.K., &Jones, R.W., (1993). Comparison of classical test theory and item response theory and their applications to test development.*Educational Measurement: Issues and Practice,* 12(3), 38-47.
- Hambleton, R. K., & Jones, R.., (1995): Item bias review. Practical assessment and evaluation, 4(6).Retrieved on 08/03/2006 from file http//PAREonline.net/getvn.asp?v=4&n=6.
- 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 developments* 1, 159-168. Greenwich, CT: JAI.
- Lord, F. M.,(1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem).*Psychometrika*, 34(3).259-299.

- MacDonald, P., &Paunonen, S.V., (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory.*Educational and Psychological Measurement*, 62(6), 921-943.
- Mellenberg, G. J., (1996). Measurement precision in test score and item response models.*Psychological Methods*, 1(3), 293-299.
- Ndalichako, J. L., & Rogers, W. T., (1997).Comparison of finite state score theory, classical test theory and item response theory in scoring multiple choice questions. *Educational and Psychological Measurement*, 57, 580-589.
- Nenty, H. J., (2004). *The application of IRT in strengthening assessment's role on the implementation of national education policy*. Paper presented at the 22nd AEAA Conference, Gaborone, Botswana.
- Stage, C., (2003). Classical test theory or item response theory: The Swedish experience. *Educational Measurement* 42.Umeå, Sweden: University of Umeå, Department of Educational Measurement. Retrieved on 06/09/2008: accessed from www.umu.se/edmeas/publikationer/pdf/em%20no%2042.pdf.
- Stage, C., (1998a).*A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT sub-test WORD. Educational Measurement* No 30.Umeå: Umeå University, Department of Educational Measurement.
- Stage, C., (1998b).*A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT sub-test ERC. Educational Measurement* No 30.Umeå: Umeå University, department of Educational Measurement.
- Stage, C., (1999).*A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT sub-test READ. Educational Measurement* No 30.Umeå: Umeå University, Department of Educational Measurement.
- Warm, T. A., (1978). *A primer of Item response theory*. Oklahoma City: U.S. Coast Guard Institute.
- Weiss, D. J., &Yoes, M. E., (1991). In R. K. Hambleton and J. Zaal (Eds).Advances in educational and psychological testing. Boston: Kluwer Academic Publishers.