International Scholars Journals

International Journal of Education Research and Reviews ISSN 2329-9843 Vol. 7 (11), pp. 001-011, November, 2020. Available online at www.internationalscholarsjournals.org © International Scholars Journals

Full Length Research Paper

Investigating the invariance of item difficulty parameter estimates based on CTT and IRT

O.O. Adedoyin[1]*, H. J.Nenty[2] and B Chilisa[2]

[1]Molepolole College of Education, Botswana.
[2]University of Botswana, Gaborone, Botswana.

Accepted 11 September, 2019

This is a quantitative empirical research study validating the invariance of item difficulty parameters estimates based on the two competing measurement frameworks, the classical test theory (CTT) and the item response theory (IRT) . In order to achieve the set goal, one fifty five (155) different independent samples were drawn from the population of students (35,262) who sat for the 2004 Paper 1 Botswana Junior Secondary School Certificate in Mathematics. These samples were selected based on gender, gender by educational regions, ability groups, and educational regions). The item difficulty parameter estimates from CTT and IRT were tested for invariance using repeated measure ANOVA at 0.05 significant levels. The study focussed on two research questions which were: (i) which of the two test theories CTT or IRT item difficulty parameter estimates vary across different samples of persons? And (ii) which of the two test theories CTT or IRT item difficulty parameter estimates vary across sample sizes? These research questions were answered through testing of hypothesis derived from each research questions. The research findings were that the item difficulty parameter estimates based on CTT theoretical framework were variant across the different independent samples. The item difficulty parameter estimates based on IRT theoretical framework were invariant across the different independent groups and also the item difficulty parameter estimates for IRT were invariant across groups with varying sample sizes. Overall, the findings from this study discredited the CTT theoretical framework for its inability to produce item difficulty invariant parameter estimates.

Key words: CTT, IRT Invariance Item and person parameters

INTRODUCTION

Classical test theory (CTT) and item response theory (IRT) are widely perceived as representing two very different measurement frameworks. However, few studies have empirically examined the similarities and differences in the parameter estimates using the two frameworks. CTT is based on the true score theory, which views the observed score as a combination of the true score and error. The true score reflects what the examinee actually knows, but it is always contaminated by different sources of errors. The test reliability is expressed as a ratio between the true score variance and observed score variance. CTT utilizes measures of item difficulty and item discrimination, the values of which are dependent upon the distribution of examinee proficiency within a sample.

*Corresponding e-mail: omobola_adedoyin@yahoo.com

Although the assumptions upon which classical test theory is based allowed it to be applied to an assortment of test construction situations, these same assumptions create weaknesses in the classical test theory model. The CTT based statistical indices are easy to compute, manipulated and understood by lay persons, but they vary from sample to sample.

According to Hambleton and Jones (1993), “the major advantages of CTT are its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations. And while classical models have proven very useful in test development they have several important limitations. The two statistics that form the cornerstones of most classical test theory, item difficulty and item discrimination are both sample dependent.” In particular, the classical test theory model, because it lacks information regarding how examinee is predicted to perform on a particular item, cannot accommodate tests

that target an examinee's proficiency level and, because item parameter indices are sample dependent, it lacks invariance of item parameters across groups of examinees (Hambleton et al., 1991).

Although CTT has served the measurement community for most of the century, CTT is not without critics. A primary criticism of CTT is related to the instability of item and person statistics produced within its theoretical framework. For years it was believed that the item statistics derived in CTT, such as item difficulty and item discrimination were depended on the sample of respondents selected to answer the items.

During the last decades a new measurement theory, the item response theory (IRT) was developed and has become an important complement to CTT in design, interpretation and evaluation of tests or examinations. The interest in IRT grew out of a combination of the concern relative to the weaknesses inherent in classical test theory by measurement professionals and the increased availability of computing power. IRT has strong mathematical basis and depends on complex algorithms that are more efficiently solved via computer. IRT (Hambleton and Swaminathan, 1985; Harris, 1989) is a group of measurement that describes the relationship between an examinee's test performance (observable) and the traits assumed to underlie performance on an achievement tests (unobservable) as a mathematical function called an item characteristics curve (ICC). IRT rests on two basic postulates:

a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits or abilities.

b) The relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC) (Hambleton et al., 1991).

IRT is more theory grounded than CTT and it models the probabilistic distributions of examinee's success at the item level. As its name indicates, 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.

REVIEW OF LITERATURE

In theory, IRT provides a way to overcome the sample dependence found in classical test theory if a certain set of strong assumptions of the IRT measurement model are met. These assumptions are unidimensionality and local independence. Unidimensionality assumes that a single latent trait underlies test performance an local independence assumes that an examinee's performance on one test item has no effect on performance on other test items. In practice, these assumptions are rarely met, and the advantages of using an IRT model are realized only when the model provides a reasonable fit to the test data (Hambleton et al, 1991). The IRT framework compasses of 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.

Theoretical, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person statistics. As a result, in theory, IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This invariance property of item and person statistics of IRT has been illustrated theoretically (Hambleton and Swaminathan, 1985; Hambleton et al., 1991) and has been widely accepted within the measurement community.

One great advantage of IRT is the item parameter invariance. The property of invariance of ability and item parameters is the cornerstone of IRT, and it is the major distinction between IRT and CTT (Hambleton, 1994). This means that the properties of tests and items derived from IRT e.g. item and test statistics are not theoretically sensitive to examinee characteristics, unrelated to ability such as gender or average group performance.

The practical advantage of invariance includes the possibility to generate optimal individual scores, to tailor tests and to examine test validity via a wider range of item and score statistics. IRT scoring reduces score bias related to group composition and allows comparison of individual scores across different tests. Individual scores are based on predictions from IRT item parameter estimates and the pattern of responses given to test items. When IRT models estimates of item statistics are used in test development, item selection is unbiased by the composition of the pilot sample who provide data for calibrations. Due to the invariance property of IRT, item difficulty based on a separate sub-population will be equivalent up to a linear transformation of scale (Rudner, 1983). In addition, it is possible to assign more precise scores to individuals (with smaller errors of measurement). IRT analyses allow for testing empirically some aspects of score validity that cannot be made explicit using CTT model (Hambleton, 1984). The invariance property of IRT model parameters makes it theoretically possible to solve important measurement problems that have been difficult to handle within the CTT framework such as those encountered in test equating and compu-terized adaptive testing (Hambleton et al., 1991). Despite the theoretical advantages attributed to item response theory (IRT), over classical test theory(CTT), little has been done to demonstrate empirically on the superiority of IRT to CTT in the measurement community.

STATEMENT OF THE PROBLEM

Invariance is the bedrock of objectivity in physical measurement, and the lack of it tends to raise a lot of question about the scientific nature of educational measurement. Measurement that changes in results or findings when used across different objects cannot contribute to the growth of science or to the growth of objective knowledge in any area. In measurement theory, analysis based on CTT has been used over the years and is still useful nowadays in test construction, although the trend is definitely towards item response theory (IRT) that provides for sample free and item free measurement. It is presently common to refer to IRT as the "modern" method of item analysis, with the obvious implication being that CTT is not modern. Not modern does not mean that CTT is no more useful in measurement theory. A primary criticism of CTT is the instability of its item and person statistics, that is, item statistics derived with CTT such as item difficulty and discrimination, are dependent on the sample of respondents.

Due to the instability of CTT item and test statistics, many researchers assumed that invariance characteristics of IRT parameter estimates makes it superior to CTT in educational measurements. However, the empirical studies especially in Africa on the superiority of IRT to CTT in measurement theory are very scarce to support this assumption. The empirical studies available, however, have primarily focused on the application in test equating and very few studies have compared CTT and IRT for item analysis and test design. According to Fan (1998), "It is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scarce. It appears that the superiority of IRT over CTT in this regard has been taken for granted in the measurement community, and no empirical scrutiny has been deemed necessary. The empirical silence on this issue seems to be anomaly."

There is a limited number of empirical studies directly or indirectly addressing the invariance issue, There is an obvious lack of systematic investigation of item and person statistics obtained from either CTT or IRT frameworks There is also lack of studies that empirically examine the relative invariance of item and person statistics obtained from CTT and those from IRT. The major criticism for CTT is its inability to produce item/person statistics that are invariant across examinee/ item samples. This criticism has been the major impetus for the development of IRT models and for the exponential growth of IRT research and applications in recent decades.

Despite theoretical differences between IRT and CTT, there is a lack of empirical knowledge about how, and to what extent, the IRT and CTT based item and person statistics behave differently. The degree of invariance of item parameter estimates across samples, usually considered as theoretically superiority of IRT models in measurement theory should be investigated, using empirical studies.

The purpose of this study is to find out whether the item difficulty parameter estimates are invariant across samples of persons on samples drawn from a large sample of real data like the Botswana Junior Secondary School Certification examination using CTT and IRT theoretical framework.

Item parameter estimates based on each model will be cross-validated using randomly replicated samples. In addition, the models will be tested across groups, where possible group related item and total score bias is an important issue (gender, location). Finally, the practical impact of each measurement model will be assessed by identifying differences in their parameter estimates (item difficulty index, b-values). This study is intended to investigate empirically the invariance of the parameter estimates of IRT and CTT using the data from 2004 Mathematics Paper 1 Botswana Junior Secondary School Certificate Examination in an attempt to verify the invariance parameter estimates of IRT models to CTT.

RESEARCH QUESTIONS

The questions which the study is concerned with are stated. Answers to each of these questions will be sought through testing a research hypothesis derived from each of these questions.

(i)Which of the two test theories CTT or IRT item difficulty parameter estimates vary across different samples of persons?
(ii) Which of the two test theories CTT or IRT item difficulty parameter estimates vary across sample sizes?

RESEARCH HYPOTHESES

To determine whether or not, the item difficulty or person parameter estimates based on CTT and IRT theories are significantly invariant across different samples of items or persons, six research hypotheses were tested using repeated measure analysis of variance (RMANOVA) at an alpha level of 0.05. These hypotheses compared the item difficulty parameter estimates among the selected independent groups within the same theoretical framework (CTT and IRT). In the null form, the hypotheses were;

H_{11}: Differences in Groups have no significant influence on the item difficulty parameter estimates based on CTT across different samples of the examinees.

H_{12}: Differences in Groups with varying sample sizes have no significant influence on the item difficulty parameter estimates based on CTT across different samples

of the examinees.

H_{13}: Differences in Groups have no significant influence on the item difficulty parameter estimates based on IRT across different samples of the examinees.

H_{14}: Differences in Groups with varying sample sizes have no significant influence on the item difficulty parameter estimates based on IRT across different samples of the examinees.

SIGNIFICANCE OF THE STUDY

This is a quantitative empirical research study that will be determining the invariance properties based on CTT and IRT theoretical measurement frameworks, using the JC 2004 Mathematics Paper 1 examination responses. Theoretically, the property of item or person parameter invariance is the most valuable in test construction and test analysis. The invariance property, which is the bedrock of objective measurement will be empirically determined and established based on CTT and IRT theoretical frameworks. Given the limited number of empirical studies directly or indirectly addressing the invariance issue, there is an obvious lack of systematic investigation about the invariance of the item and person statistics obtained from either CTT or IRT frameworks, and a lack of studies that empirically compares the relative invariance of item and person statistics obtained from CTT versus those from IRT. It is envisaged that this study will determine and establish the invariance properties of CTT and IRT. The property of invariance parameter estimates across different samples will be determined based on CTT and IRT, using one of the high stake examinations in Botswana, the 2004 Junior Secondary School Mathematics assessment responses with a population of thirty six thousand (36,000). Since the population of the subjects to be used in this study is very large, it is envisaged that the findings of this research study will be reliable, objective and valid. This will be of great importance to researchers in educational measurement community who have been seeking for objective, reliable and valid measurement approach in analyzing, interpreting test/examination scores. Since empirical studies examining the invariance characteristics of item and person statistics from the two measurement frameworks CTT and IRT are very scarce.

It is also envisaged that the findings of this research study will increase the empirical knowledge based on CTT and IRT theoretical frameworks. Classical test theory (CTT) and item response theory (IRT) are widely perceived as representing two very different measurement frameworks. However, few studies have empirically examined the similarities and differences in the parameter estimates using the two frameworks. Empirical research for the truth on whether IRT or CTT item/person parameter estimates are comparably the same or different will be established in this study.

METHODOLOGY

The general sampling procedure

The invariance parameter estimates by CTT and IRT were estimated using a total number of one hundred and fifty five (155) different samples, taken from the population of the students who sat for Paper1, 2004 Botswana Junior Secondary School Certificate Examinations in Mathematics as shown in Table 1 using nine different sampling plans. Some of these different independent samples were of the same sample sizes, and some were of varying sample sizes. These different independent samples were based on gender, educational regions, gender by educational regions and ability levels of the students in mathematics.

Data analysis

For the CTT item difficulty estimates, it was calculated as the proportion of correct response by the examinees. The software MULTILOG VERSION 7.0 was used to estimate the IRT item difficulty parameter estimates. The repeated measure ANOVA was used for testing hypothesis on the item difficulty parameter estimates. The reason been that the correlation statistics used by other researchers in testing for invariance have been criticised by Rupp and Zumbo (2004) as not good enough to test for invariance. According to this source, “Pearson Product-Moment Correlation Coefficient (PPMCC) is insufficient for the purpose of testing for invariance.”

PRESENTATION AND DISCUSSION OF RESULTS

H_{11}: Differences in groups of examinees have no significant influence on the item parameter estimates based on CTT

Table 2 presents the results of the analysis for testing Hypothesis One on the item difficulty parameter estimates based on CTT theoretical framework across different independent groups. The different independent groups were gender, population, educational regions, gender by educational regions and the ability groups. As presented on Table 2, the item parameter estimated based on the following independent groups: gender, population samples with varying sizes, central educational region, low ability samples and high ability samples with varying sizes are not significantly different. But for the remaining samples, population, educational regions (Southern, Northern, North Western, South Central), gender by educational regions with varying sample sizes (500, 1000 and 1500), high ability group such difference are significant. That is, out of the seventeen (17) different independent samples, for eight samples, the differences are not significant at 0.05 alpha level, whereas for the remaining nine samples such differences are significant. Using the nine samples for which the differences are significant, the trend in the lack of invariance tends to

Table 1. Sampling plan for the different independent samples for this study.

A	Gender sampling with the same sample size of 1000	Number of samples	
	Male [M]	10	M1,M2,M3M4,M5, M6,M7,M8,M9,M10
	Female [F]	10	F1,F2,F3,F4,F5, F6,F7,F8,F9,F10
B	Sampling from the population [P] with same sample size of 1000	20	P1,P2,P3,P4,P5,P6,P7, P8,P9,P10,P11,P12,P13, P14,P15,P16,P17,P18,P19,P20
C	Population sampling with varying sizes [PS] from sample size of 1000 to 1900	10	PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS8, PS9, PS10
D	Educational regions with the same sample sizes 1000 each Central[C], North[N] South central[SC], South[S]North western [NW]	15	C1,C2,C3.N1,N2,N3, SC1,SC2,SC3,S1,S2, S3,NW1,NW2,NW3
E	Education regions with varying sample sizes of 1000, 1500, 2000, 2500.	20	C1S,C2S,C3S,C4S, N1S,N2S,N3S,N4S SC1S,SC2S,SC3S,SC4S S1S,S2S,S3S,S4S NW1S,NW2S,NW3S,NW4S
F	Ability groups with same sample size of 1000 High [HA]		
		10	HA1,HA2,HA3,HA4,HA5, HA6,HA7,HA8,HA9,HA10
	Low [LA]	10	LA1,LA2,LA3,LA4,LA5 LA6,LA7,LA8,LA9,LA10
G	Ability groups with different sample sizes from 1000 to 1900 High [HAN]	10	HAN1,HAN2,HAN3,HAN4,HAN5, HAN6,HAN7,HAN8,HAN9,HAN10
	Low [LAN	10	LAN1,LAN2,LAN3,LAN4,LAN5 LAN6,LAN7,LAN8,LAN9,LAN10
I	Gender by educational regions of varying sample sizes from 500 to 1500 Sample size of 500	30	MC1, MN1, MNW1,MS1,MSC1 FC1,FN1,FNW1,FS1,FSC1
	Sample size of 1000		MC2,MN2,MNW2,MS2,MSC2 FC2,FN2,FNW2,FS2,FSC2
	Sample size of 1500		MC3,MN3,MNW3,MS3,MSC3 FC3,FN3.FNW3.FS3.FSC3

show especially in the samples based on gender and educational regions. It may be possible that there is interaction effect between gender and the educational regions. Another reason for this maybe that the test items constructed for JC examinations were gender and educational regions biased, which suggests that there may be group differences effect on the test items.

H_{12}: Differences in groups of examinees with varying sample sizes have no significant influence on the item parameter estimates based on CTT

Table 3 shows that the CTT item difficulty parameter estimates for the varying sample sizes are invariant across the different groups (population samples) (PS1- PS10)

Table 2. Repeated measure ANOVA of the group influence of CTT item difficulty parameter estimates (p-values)

Source of variables	SS	df	MS	F	p
Female groups	.003	9	.000	1.494	.162*
Error	.018	90	.0002		
Total	.021	99			
Male groups	.003	9	.000	1.870	.068*
Error	.013	81	.0002		
Total	.016	90			
Gender groups	.008	19	.000	.624	.885*
Error	.112	171	001		
Total	.120	190			
Population samples	.031	19	.002	4.999	.000
Error	.062	190	.000		
Total	.093	209			
Population samples with varying sizes (PS)	.001	9	.000	.550	.834*
Error	.011	90	.000		
Total	.012	99			
Education region samples (C1,C2,C3,C1S,C2S,C3S,C4S)	.002	6	.000	1.618	.158*
Error	.011	60	.000		
Total	.013	66			
Education region samples (N1,N2,N3,N1S,N2S,N3S,N4S)	.002	6	.000	2.401	.038
Error	.009	60	.000		
Total	.011	66			
Education region samples (S1,S2,S3,S1S,S2S,S3S,S4S)	.004	6	.001	2.893	.015
Error	.013	60	.000		
Total	.017	66			
Education region samples (SC1,SC2,SC3,SC1S ,SC2S,SC3S,SC4S)	.008	6	.001	7.899	.000
Error	.011	60	.000		
Total	.019	66			
Education region samples (NW1,NW2,NW3,NW1S, NW2S,NW3S,NW4S)	.004	6	.001	3.031	.012
Error	.012	60	.000		
Total	.016	66			
Education region samples Gender(M/F) by Educational regions of sample size 500	.083	9	.009	11.144	.000
Error	.075	90	.001		
Total	.158	99			
Education region samples Gender(M/F) by Educational regions of sample size 1000	.048	9	.005	5.398	.000
Error	.088	90	.001		
Total	.136	99			

Table 2 continue

Education region samples Gender(M/F) by Educational regions of sample size 1500	.043	9	.005	6.600	.000
Error	.066	90	.001		
Total	.109	99			
High ability (HA) samples	.006	9	.001	2.933	.004
Error	.019	90	.000		
Total	.025	99			
Low ability (LA) samples	.007	9	.001	1.869	.067*
Error	.037	90	.000		
Total	.044	99			
High ability samples with varying sizes (HAN)	.002	9	.000	1.869	.066*
Error	.011	90	.000		
Total	.013	99			
Low ability with varying sizes (LAN)	.002	9	.000	.922	.510*
Error	.018	90	.000		
Total	.020	99			

*(samples for which the differences in p-values are not significant at 0.05 alpha level)

Table 3. Repeated measure ANOVA of the group influence of CTT item difficulty parameter estimates (p-values) for varying sample sizes

Source of variables	SS	df	MS	F	p<
Population samples with varying sizes (PS)	.001	9	6.960E-05	.550	.834*
Error	.011	90	.000		
Total	.012	99			
Education region samples (C1,C2,C3,C1S,C2S,C3S,C4S)	.002	6	.000	1.618	.158*
Error	.011	60	.000		
Total	.013	66			
Education region samples (N1,N2,N3,N1S,N2S,N3S,N4S)	.002	6	.000	2.401	.038
Error	.009	60	.000		
Total	.011	66			
Education region samples (S1,S2,S3,S1S,S2S,S3S,S4S)	.004	6	.001	2.893	.015
Error	.013	60	.000		
Total	.017	66			
Education region samples (SC1,SC2,SC3,SC1S,SC2S,SC3S,SC4S)	.008	6	.001	7.899	.000
Error	.011	60	.000		
Total	.019	66			
Education region samples (NW1,NW2,NW3,NW1S,NW2S,NW3S,NW4S)	.004	6	.001	3.031	.012
Error	.012	60	.000		
Total	.016	66			
High ability samples with varying sizes (HAN)	.002	9	.000	1.869	.066*
Error	.011	90	.000		
Total	.013	99			
Low ability with varying sizes (LAN)	.002	9	.000	.922	.510*
Error	.018	90	.000		
Total	.020	99			

*(samples for which the differences in p-values are not significant at 0.05 alpha level)

with F value of .550, p-value of .834, Central region samples (C1-C4S) with F value of 1.618, p-value of .158, high ability samples with varying sample sizes (HAN1-HAN10), F value of 1.869, p-value of .066 and the low

Table 4. Repeated measure ANOVA of the group influence of IRT item difficulty parameter estimates (p-values)

Source of variables	SS	df	MS	F	p<
Female groups	.040	9	.004	.939	.495*
Error	.429	90	.005		
Total	.469	99			
Male groups	.037	9	.004	1.027	.425*
Error	.361	90	.004		
Total	.398	99			
Gender groups	.127	19	.007	.619	.889*
Error	2.054	190	011		
Total	2.181	209			
Population samples	.136	19	.007	.986	.479*
Error	1.381	190	.007		
Total	1.517	209			
Population samples with varying sizes (PS)	.082	9	.009	1.494	.162*
Error	.555	90	.006		
Total	.637	99			
Education region samples (C1,C2,C3,C1S,C2S,C3S,C4S)	.061	6	.010	1.678	.142*
Error	.365	60	.006		
Total	.426	66			
Education region samples (N1,N2,N3,N1S,N2S,N3S,N4S)	.035	6	.006	2.258	.055*
Error	.154	60	.003		
Total	.169	66			
Education region samples (S1,S2,S3,S1S,S2S,S3S,S4S)	.084	6	.014	1.481	.200*
Error	.565	60	.009		
Total	.649	66			
Education region samples (SC1,SC2,SC3,SC1S,SC2S,SC3S,SC4S)	.070	6	.012	1.433	.217*
Error	.485	60	.008		
Total	.555	66			
Education region samples (NW1,NW2,NW3,NW1S,NW2S,NW3S,NW4S)	.055	6	.009	1.284	..278*
Error	.431	60	.007		
Total	.486	66			
Education region samples Gender (M/F) by Educational regions of sample size 500	.136	9	.015	.943	.493*
Error	1.436	90	.016		
Total	1.572	99			
Education region samples Gender(M/F) by Educational regions of sample size 1000	.174	9	.019	1.478	.168*
Error	1.178	90	.013		
Total	1.352	99			

ability group with varying sample sizes (LAN1-LAN10) with F value of .922, p-value of .510. For all the other samples the difference are significant at 0.05 alpha level (the Northern, Southern, South central, North Western educational regions). For all the educational regions except the Central region these are significant, which implies that there is a trend in the lack of invariance within the educational regions.

Table 4. continue

Education region samples Gender(M/F) by Educational regions of sample size 1500	.129	9	.014	1.03	.391*
Error	1.205	90	.013		
Total	1.334	99			
High ability (HA) samples	1.257	9	.140	1.048	.409*
Error	11.993	90	.133		
Total	13.250	99			
Low ability (LA) samples	14.453	9	.1606	1.019	.431*
Error	141.844	90	.1576		
Total	156.297	99			
High ability with varying sample sizes (HAN)	.563	9	.063	1.571	.136*
Error	3.586	90	.040		
Total	4.179	99			
Low ability samples with varying sizes	11.951	9	1.328	1.471	.171*
Error	81.261	90	.903		
Total	93.212	99			

*(samples for which the differences in p-values are not significant at 0.05 alpha level)

H_{13}: Differences in groups of examinees have no significant influence on the item parameter estimates based on IRT

Table 4 presents the results of the analysis for testing Hypothesis Five on the item difficulty parameter estimates based on IRT theoretical framework across different independent groups. The different independent groups are: gender, population, educational regions, gender by educational regions and the ability groups. Table 4 reveals that for all the different independent groups the differences are not significant, which means that the item difficulty parameter estimates based on IRT theoretical framework are invariant across the different independent groups. This implies the IRT item difficulty estimate do not depend on the sample or group selected and used to estimate the parameter. That is, regardless of the groups groups, the estimation of IRT item difficulty will always be the same value, and this is the concept of invariance.

H14: Differences in groups of examinees with varying sample sizes have no significant influence on the item parameter estimates based on IRT

The above hypothesis was tested using repeated measure ANOVA on different independent groups with varying sample sizes based on the IRT item parameter estimates. The results of this analysis which is presented on Table 5 shows that the IRT item difficulty parameter estimates for the varying sample sizes are invariant across the different groups.

SUMMARY OF RESEARCH FINDINGS

Based on testing the four hypotheses posited for this study on the invariance of parameter estimates based on CTT and IRT theoretical frameworks, the following were the research findings (6) below.

The main purpose of this study was to determine the invariance of each item parameter across different samples of examinees. It was intended that the findings emanating from this study would contribute to the attempt by measurement scientists to validate the claims by the relatively few IRT in comparison to the traditional CTT as to the invariance of item and person parameter estimates. Such invariance property is seen to be the most desirable scientific property of any measurement, and for educational measurement to claim scientific status, its parameter estimates must be seen to attain this all important invariance status.

In testing, the concept of invariance is that the difference between the parameter of any two items does not depend upon the ability parameter of a particular set of persons whose responses to the items are used to estimate the item/person statistics, and the difference between the ability parameters of any two persons does not depend on the difficulty parameter of the particular item or items the persons selected. Hence, with invariance there is “sample free item calibrations” and “item or test-free person measurement”

From this study it can be concluded that:

(i) The CTT person parameter estimates failed to exhibit the invariance property across all the different subsets of items.

(ii) The CTT item difficulty parameter estimates were variant across different independent samples of persons.

(iii) The IRT item difficulty parameter estimates were in-

Table 5. Repeated measure ANOVA of the group influence of IRT item difficulty parameter estimates (p-values) with varying sample sizes

Source of variables	SS	df	MS	F	p
Population samples with varying sizes (PS)	.082	9	.009	1.494	.162*
Error	.555	90	.006		
Total	.637	99			
Education region samples (C1,C2,C3,C1S,C2S,C3S,C4S)	.061	6	.010	1.678	.142*
Error	.365	60	.006		
Total	.426	66			
Education region samples (N1,N2,N3,N1S,N2S,N3S,N4S)	.035	6	.006	2.258	.055*
Error	.154	60	.003		
Total	.169	66			
Education region samples (S1,S2,S3,S1S,S2S,S3S,S4S)	.084	6	.014	1.481	.200*
Error	.565	60	.009		
Total	.649	66			
Education region samples (SC1,SC2,SC3,SC1S,SC2S,SC3S,SC4S)	.070	6	.012	1.433	.217*
Error	.485	60	.008		
Total	.555	66			
Education region samples (NW1,NW2,NW3,NW1S,NW2S,NW3S,NW4S)	.055	6	.009	1.284	.278*
Error	.431	60	.007		
Total	.486	66			
High ability with varying sample sizes (HAN)	.563	9	.063	1.571	.136*
Error	3.586	90	.040		
Total	4.149	99			
Low ability samples with varying sizes	11.951	9	1.328	1.471	.171*
Error	81.261	90	.903		
Total	93.212	99			

*(samples for which the differences in p-values are not significant at 0.05 alpha level).

invariant across different independent samples of persons.

(iv) The IRT item difficulty estimates were invariant across varying sample sizes of persons.

Overall, the findings from this study discredited the CTT theoretical framework for its inability to produce item difficulty invariant parameter estimates in all the selected independent samples. In case of IRT theoretical framework, it was able to show the invariance parameter estimates for the item difficulty estimates and samples with varying sizes.

It is also envisaged that the findings of this research study will increase the empirical knowledge based on CTT and IRT theoretical frameworks for educational measurement analysis. It can then be recommended that:

(i) For more objective educational measurement IRT theoretical framework should be incorporated by Examination Boards into educational measurement practices, tests or examinations in Africa.

(ii) The issue of IRT parameter estimates is still new in Africa, therefore, workshops, seminars and conferences should be organised for researchers in educational testing.

(iii) It is high time for experts in educational measurement in Africa to rise to the challenges posed by the measurement community and be fully aware of the usefulness of IRT in constructing and scoring of tests or examinations.

(iv) Invariance, always claimed for IRT parameter estimates, if validated, would be of tremendous importance for testing in Africa.

(v) Encourage the development of items based on IRT

test develop guidelines as such test is more likely to meet the invariance property claimed by this theory.

REFERENCES

Fan X (1998). Item response theory and classical test theory: A comparison of their item/person statistics. Education and Psychological Measurement, 58,357-381

Harris D (1989). Comparison of 1-, 2- and 3-parameter IRT models. Educational Measurement Issues and Practice, 8: 35-41.

Hambleton RK (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema*, *6* (3), 535-556.

Hambleton RK, Jones RW (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.

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

Hambleton RK, Swaminathan H (1995). Item response theory: Principles and application. Boston: Kluwer.

Lord FM (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord FM, Novick MR (1968). Statistical theories of mental test scores. Reading, MA: Addison- Wesley.

Lord FM (1983). Small justifies Rasch model. In D.J. Weiss (Ed.), New horizons in testing*,* New York: Academic Press. pp. 51-62.

Masters N (1982). A Rasch model for partial credit scoring. Psychometrika*,* 47(2): 149-174.

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*, (6), 921-943.

Rupp AA, Zumbo BD (2004). A note on how to quantify and report whether IRT parameter Invariance holds: When Pearson correlations are not enough. Educational and Psychological Measurement, 64 (4), 588-599.