© Kamla-Raj 2010
PRINT: ISSN 0975-1122 ONLINE: 2456-6322

Int J Edu Sci, 2(2): 107-113 (2010)
DOI: 10.31901/24566322.2010/02.02.07

Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories

O.O. Adedoyin

Department of Educational Foundations, University of Botswana, Gaborone, Botswana
E-mail: omobola_adedoyin @ yahoo.com

KEYWORDS Invariance. Classical Test Theory (CTT). Item Response Theory (IRT)

ABSTRACT This is a quantitative empirical research study validating the invariance of person parameter estimates based on the two competing measurement frameworks, the classical test theory (CTT) and the item response theory (IRT). In order to achieve this goal, 11 items that fitted the 2PL model from the 40 items of Paper 1 Botswana junior secondary mathematics examinations, were used to estimate the person ability for a sample of five thousand examinees (5000) out of a total of thirty- five thousand, five hundred and sixty- two (35562) who sat for the examination. The person parameter estimates from CTT and IRT were tested for invariance using repeated measure ANOVA at 0.05 significant level.The IRT person parameter estimates based on IRT were invariant across subsets of items, for the examinees randomly selected. In conclusion, the findings from this study show that, there is gross lack of invariance when classical test theory (CTT) is used to estimate person parameter or ability. IRT person parameter estimates exhibited the invariance property across subset of item. This study advocates for the use of IRT in test construction and measurements of achievement abilities.

INTRODUCTION

The general concept of invariance is that the change following or preceding a distinguished phenomenon must not be unique, but share some properties with changes associated with similar phenomena. Ordinarily, invariance is some kind of correspondence between two types of mathematical objects, so that two “similar” things correspond to one and the same object. That is, the notion of invariance is the quality of being resistant to variations under a set of transformations. Scientifically, the notions of invariance have played a central role in the investigation of statements considered suitable to scientific laws. For instance, the classical concept of “dimensional invariance” has been widely used, via the method of dimensional analysis, in the search for lawful numerical relations among physical variables. For instance, in physics, invariants are usually quantities conserved (unchanged) and the physicist wants to be able to track what does and what does not change under a set of transformation. Without invariance principles, there would be no laws of physics because physicists rely on the results of experiments remaining the same from day to day and place to place. The philosophical idea of invariance is that some points remain fixed under all transformations, and that such points would contribute absolutely to knowledge and values.

In testing, the concept of invariance is that the estimate of the parameter of an item across two or more groups of population of interest disparate in their abilities must be the same. And similarly, the estimate of the ability/ person parameter of the same testees based on items which are disparate in their difficulties must also be the same. Hence, with invariance there is “sample free item calibrations” and “item or test-free person measurement.”

As succinctly stated by Horn and McArdle (1992), “The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying a phenomenon, measurement yields measures of the same attributes. If there is no evidence indicating presence or absence of measurement invariance - the usual case - or there is evidence that such invariance does not exist, then the basis for drawing scientific inference is severely lacking. Findings of differences between individuals and groups cannot be unambiguously interpreted” (p.117).

According to Nenty (2004), ‘Invariance is the bedrock of objectivity in physical measurement, and the lack of it raises a lot of questions about the scientific nature of psychological measurement. Measurement which when repeated assigns significantly different values to the same characteristic of an individual, and for which such assignments depend on the particular set of items designed to measure the trait, cannot contribute to the growth of science or to the growth of objective knowledge in any area’.

Over the past 30 years, the field of educational measurement has undergone changes and new innovations have been created to meet the increasing demand for valid interpretation of individual score from educational tests or examinations.

There are two measurement theories, which have been theoretically and technologically developed in analyzing or standardizing tests, examinations within measurement frameworks. These are the Classical Test Theory (CTT) and Item Response Theory (IRT). CTT and IRT are widely perceived as representing two very different measurement frameworks. At the item level, the CTT model is relatively simple; CTT does not invoke a complex theoretical model to relate an examinee's ability to success on a particular item. Instead, CTT collectively considers a pool of examinees on an item.

During the last couple of 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. IRT has strong mathematical basis and depends on complex algorithms that are more efficiently solved via the computer. IRT (Hambleton and Swaminathan 1985: 11; Harries 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; and

(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: 7).

PROBLEM OF THE STUDY

Theoretically, the invariance property of person parameter estimates is most valuable in testing, and for any objective measurements of abilities or test item parameters. It is necessary for such measurements to be consistent and should not vary. Measurement that changes in results when used across different objects cannot contribute to the growth of science or to the growth of objective knowledge in any area. Item calibration that is independent of the calibrating sample, and person measurement that is independent of a particular set of items, are the main advantages of objective measurement. In other words, the concept of invariance in measurement is that the estimates of all item / person parameters should not vary significantly over different samples of people, and from any calibrated subset of items.

The purpose of this study is to find out whether the person parameter estimates based CTT and IRT theoretical framework are invariant across samples of items. This is a quantitative empirical study concerned with estimating the person parameters based on CTT and IRT, and testing these estimated parameters for invariance. To determine whether or not the person parameter estimates based on CTT and IRT theories are significantly invariant across different subsets of items, this study addressed the research question:

Which of the Two Test Theories CTT or IRT Person Parameter Estimates Vary Across Different Subsets of Items?

This research question was answered by testing two research hypotheses derived from it using an alpha level of 0.05 within each measurement theoretical frameworks. These hypotheses compared estimates of person parameter estimated based on different subsets of validated items. The research hypotheses were:

(i) H_1: Different subsets of items from the Botswana JC Examination in mathematics have no significant influence on the person parameter estimates based on CTT theoretical framework.

(ii) H_2: Different subsets of items from the Botswana JC Examination in mathematics have no significant influence on the person parameter estimates based on IRT theoretical framework.

Review of Literature

Classical Test Theory (CTT)

CTT as a simple useful model in measurement theory that describes how errors in measurement can influence observed scores or measurements. CTT is based on the true score theory which views the observed score (X) as a combination of the true scores (T) and an error component (E). This is often expressed mathematically as:-

$$X = T \pm E \qquad (1)$$

The true score, T, reflects whether the examinee's amount of knowledge or ability is the true measurement of the examinee, but it is always contaminated by random errors, which can result from numerous factors, such as guessing, fatigue or stress (Lord 1980, pp. 4-5). The observed score X is often called a fallible score, fallible because there is a degree of error involved. The true score is often defined as the score that would be obtained if there were no error of measurement. CTT utilizes traditional item and sample dependent statistics, which include item difficulty, item discrimination, item test inter-correlations and a variety of related statistics.

Assumptions of Classical Test Theory

There are three main assumptions made in classical test theory. The first assumption is that the error is uncorrelated with the true score. That is the error and the true score obtained in measurements are uncorrelated (Lord 1980)

$$\sigma_{ET} = 0 \qquad (2)$$

The second assumption is that the error terms have an expected mean of zero; that is, these random errors over many repeated measurements are expected to cancel out such that, in the long run, the expected mean of measurement errors would be equal to zero. If the error is zero, the observed value becomes the true value {X=T}. Third assumption is that the errors from parallel measurements are uncorrelated. The definition of true value according to CTT model is defined as the expected value of the observed value {T = E{X}} (Lord 1980: 6).

A key element of classical test theory is the definition of parallel measurements. Two measurements X and X' are said to be parallel measurements if the expected values of the two observed scores X and X' are equal (E[X] = E[X']), which means that the two observed scores X and X' will have the same true scores {T=T'} and equal observed variances

$$\sigma^2[X] = \sigma^2[X'] \qquad (3)$$

The error variance for the two parallel scores are usually equal {$\sigma_E^2 = \sigma_{E'}^2$} for every population of examinees (Lord 1980: 6).

In classical test theory, reliability is determined by the correlation coefficient between the observed scores on two parallel measurements. As the reliability of a measurement increases, the error variance becomes relatively smaller. When error variance is relatively small, an examinee's observed score is very close to the true score. However, when error variance is relatively large, observed score gives a poor estimate of true scores (Lord 1980, pp. 6-7).

Limitations of Classical Test Theory

Hambleton et al. (1991) identified four limitations of classic test theory. The first limitation is that the item statistics such as item difficulty and item discrimination depend on the particular examinee samples based on which they were obtained, i.e. they are group dependent and test dependent. The average level of ability and the range of ability scores among a sample of examinees influence, often substantially, the values of the item statistics. For example, item difficulty levels will be higher when examinee samples used to obtain the statistics have higher ability than the average ability level of the examinees in that population. Also, item discrimination indices tend to be higher when estimated from an examinee sample heterogeneous in ability than from an examinee sample homogeneous in ability.

The second limitation is that the definition of reliability is established through the concept of parallel tests, and this is difficult to achieve in practice. This is because individuals are never exactly the same on a second administration of a test because people tend to forget, they develop new skills or their motivational or anxiety level might change.The third limitation is that the standard error of measurement is assumed to be the same for all subjects and does not take into account the variability in error at the different trait levels (Hambleton et al. 1991) while the fourth limitation reflects the focus on test level information to the exclusion of item level information. Test level information is an additive process i.e. it is the sum of the information across items, and item level information is only the information for a particular item. With these limitations, it is obvious that classical test theory deals with the individual's total score, and not their ability at the individual item level (Hambleton et al. 1991). An alternative to CTT is IRT.

Item Response Theory (IRT)

IRT provides an alternative to classical test theory as a basis for examining the relationship between item responses and the ability of the examinee being measured by the test or scale (Hambleton and Swaminathan 1985) That is, the essence of IRT is that the probability of answering an item correctly or of attaining a particular response level is modeled as a function of an individual's ability and the characteristics of the item. And a paramount goal of IRT is predicting the probability of an examinee of a given ability level responding correctly 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. (Hambleton et al.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 (q) (Hambleton et al. 1991). There are traditionally three IRT mathematical equations termed, one, two, and three parameter models that are used to make predictions.

Assumptions of IRT

Like any other measurement theory, IRT has its own assumptions. The first is the assumption of uni-dimensionality. This assumes that a test measures one and only one latent trait or ability at a time. This means that the test or group of items should measure one and only one latent trait. According to Hambleton et al. (1991:9), a latent trait is a hypothetical and unobservable characteristic like mathematical ability. The second assumption is the concept of local independence, which means that statistically, the only factor influencing the individual's response to each item is independent of response to another item in the test (Hambleton et al. 1991). That is, answering one question correctly or incorrectly will not influence the probability of answering any other question correctly or incorrectly in a test. For example, calculating a correct answer to a particular question of the set of items cannot rely on information obtained from answering the previous question from the same set of items correctly. Local independence also applies to the examinees, which means that all examinee's scores are independent of all other examinees' scores.

The third assumption is that the item characteristics curve (ICC) is a mathematical monotonically increasing function that describes the relationship between an examinees item performance and the trait underlying the item performance. ICCs allow the researcher to provide estimates of ability that are independent of the particular set of items which have been selected from a larger set of items, all of which measure the same trait or ability (Hambleton et al. 1991).

General Item Response Theory Models

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 parameters used to describe items.

The equation of the ICC for one parameter logistic model is given as:

The two-parameter model equation is:

$$P_i(\theta) = \frac{1}{1+e^{-D_{a_i}(\theta-b_i)}}$$

The three parameter IRT model equation is:

$$P_i(\theta) = c_i + (1-c_i)\,\frac{1}{1+e^{-D_{a_i}(\theta-b_i)}}$$

where,

$P_i(\theta)$ - the probability of a correct response for the ith item.

b_i - is the difficulty parameter for the ith item

a_i - is the discrimination parameter for the ith item

c_i - is the guessing parameter for the ith item

θ - is the ability level

D - represents a scaling factor, which is set to 1.7

METHODOLOGY OF THE STUDY

Data Source

The data used in this study were the students' responses to the items in Paper 1 of 2004 Botswana junior secondary school certificate examinations in mathematics. The 2004 JC Mathematics Paper 1 which was a multiple choice examination paper, consisted of forty items (40 items) based on the three year JC mathematics curriculum. The current structure of education in Botswana is seven years of primary education, three years of junior secondary education, two years of senior education, and four years of university education (7+3+2+4). The Botswana Junior Certificate (JC) examination is administered at the end of the third year of the Junior Certificate (JC) course to measure the achievement level of candidates at that point. The examination is used for two purposes: as a tool to select students who are to proceed to the next level of education, which is the senior secondary, and also as an assessment mechanism that measures the extent to which basic competencies and skills have been acquired.

Sampling Procedure

Selection of Students' Responses

To determine the extent to which CTT- and IRT-based person parameter estimates are invariant across different samples of test items (i.e. test forms) a computer-based simple random sample of responses of five thousand students (5000 students) from a total population of 35,262 students who took the examination were selected.

Selection of Items

Out of a total of 40 items in the examination, a subtest of 11 items that were found to fit the 2-parameter IRT model were used in this study. Out of these 11 items, five subtests of five items each (SS1 to SS5) were generated electronically through the simple random sampling procedure (Table 1).

Data Analysis

The person ability estimates based on the two test theories were estimated as follows: for the CTT, the person ability was estimated as the sum of each examinees' responses to all the test items in each subtest. These were transformed to percentiles for comparability of the different samples. The IRT person ability (the theta values) of examinees based on each subtest were estimated using the IRT computer Multilog (version 7.0) software (2PL) by (Thissen 1991), and these were transformed to z-score values for comparability among the different samples. Hypotheses were tested using repeated measure ANOVA to investigate the invariance of person parameter estimates across the five subtests for each of the two theories.

Table 1: Sampling plan for five different subtests

Samples for the subtests	*Sample size*
SS1	5 (items 1,2,3,4,5)
SS2	5 (items 2,4,5,12,15)
SS3	5(items12,15,21,22,37)
SS4	5 (items 1,3,21,22,27)
SS5	5 (items 2,3,4,15,37)

SS: Subtest samples

RESULTS

H_{11}: Differences across subsets of items from the Botswana JC Examination in Mathematics have no significant influence on the person parameter estimates based on CTT theoretical framework.

To find out whether or not the person parameter estimates based on CTT are invariant across subsets of items, five subsets consisting of five items each from the eleven items that fitted the IRT model were used. The five subset scores of the 5000 students transformed to percentiles were compared using the repeated measure ANOVA across a certain raw score. Table 2 presents the results of this analysis for CTT theoretical framework. The table shows that the p-values for all the raw score values were less than alpha level 0.05, which is very significant, that is, the person parameter estimates based on the five subsets of the 11 items were significantly different. This shows that differences in ability estimates based on each of the five subtests of items for the 5000 examinees from items of the Botswana JC Examination in Mathematics were significantly different. Hence, subsets of test items developed to measure the same ability have significant influence on the estimate of such ability for the examinees. This implies that the examinees' score or ability is dependent on the particular set of items administered, that is, it is test-dependent.

In that case, the CTT person parameter estimates were found to vary cross the item groups from the same test.

H_{12}: Different subsets of items from the Botswana JC Examination in mathematics have no significant influence on the person parameter estimates based on IRT theoretical framework.

To test the second hypothesis, the same analysis was repeated for IRT-based estimates of person ability based on the same five subsets of five items (see Table 3). This revealed that differences across subset of items from the Botswana JC Examination in Mathematics have no significant influence on the person parameter estimates based on IRT theoretical framework. In other words, the person parameter estimates based on IRT were found to be invariant across subsets of items**,** since all the p-values (as shown in Table 3) were greater than 0.05 which shows that they were not significant. This implies that an examinee's ability estimate is the same across the different subsets of test items.

Table 2: Repeated measure ANOVA of the different subsets of items on the person parameter estimates based on CTT theoretical framework

Raw score	*SS*	*df*	*MS*	*F*	*p<*
1					
Factor	22.923	4	5.731	247.009	0
Error	463.925	19996	0.023		
Total	486.848	20000			
2					
Factor	0.781	4	0.195	8.474	0
Error	14.931	648	0.023		
Total	15.712	752			
3				10.538	0
Factor	1.313	4	0.328		
Error	39.007	1252	0.031		
Total	40.32	1256			
4					
Factor	2.806	4	0.702	20.515	0
Error	55.258	1616	0.034		
Total	58.064	1620			
5					
Factor	5.989	4	1.497		
Error	72.699	2412	0.03	49.673	0
Total	78.688	2416			
6					
Factor	7.117	4	1.779		
Error	83.795	2888	0.029	61.32	0
Total	90.912	2892			
7					
Factor	6.774	4	1.693		
Error	78.106	3072	0.025	66.605	0
Total	84.88	3076			
8					
Factor	4.029	4	1.007		
Error	66.227	2956	0.022	44.964	0
Total	70.236	2960			
9					
Factor	1.22	4	0.305		
Error	33.484	2320	0.014	21.134	0
Total	34.704	2324			
10					
Factor	0.294	4	0.074		
Error	8.842	1636	0.005	13.607	0
Total	9.136	1640			
11					
Factor	0	4	0		
Error	0	812	0	0	0
Total	0	816			

Table 3: Repeated measure ANOVA of the different subsets of items on the person parameter estimates based on IRT theoretical framework

Raw score	*SS*	*df*	*MS*	*F*	*p*
1					
Factor	0.98	4	0.245	1.79	.131*
Error	33.94	248	0.137		
Total	34.92	252			
2					
Factor	0.351	4	0.088	0.753.556*	
Error	75.47	648	0.116		
Total	75.821	652			
3					
Factor	0.569	4	0.142	1.24	.292*
Error	143.486	1252	0.115		
Total	144.055	1256			
4					
Factor	0.091	4	0.023	0.169.954*	
Error	216.606	1616	0.134		
Total	216.697	1620			
5					
Factor	0.771	4	0.193	1.547.186*	
Error	300.559	2412	0.125		
Total	301.33	2416			
6					
Factor	0.243	4	0.061		
Error	388.124	2888	0.134	0.452.771*	
Total	388.367	2892			
7					
Factor	0.496	4	0.124		
Error	398.946	3072	0.13	0.955.431*	
Total	399.442	3076			
8					
Factor	0.496	4	0.124		
Error	398.946	3072	0.13	0.955.431*	
Total	399.442	3076			
9					
Factor	0.535	4	0.134	1.18	.317*
Error	263.092	2320	0.113		
Total	263.627	2324			
10					
Factor	0.449	4	0.112	0.891.468*	
Error	205.993	1636	0.126		
Total	206.442	1640			
11					
Factor	1.326	4	0.331	2.739	208*
Error	98.269	812	0.121		
Total	99.595	816			

DISCUSSION

The main purpose of this study was to determine the invariance of person parameter estimates across different subsets of items based on the two measurement theories CTT and IRT. 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 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.Based on testing the two hypotheses posited for this study on the invariance of parameter estimates based on CTT and IRT theoretical frameworks, the following were the research findings:

(i) The CTT person parameter estimates did not exhibit the invariance property, since all the raw score groups were variant. That is, differences across subsets of items have significant influence on the person parameter estimates based on CTT.

(ii) The IRT person parameter estimates were all invariant, that is, differences across subsets of items have no significant influence on the person parameter estimates based on IRT.

The findings from this research study indicated that the invariance of person parameter estimates by CTT varied across the different subsets of items. That is the various subsets of items have significant influence on the person parameter estimates using the CTT theoretical framework. Whereas the person parameter estimates by IRT were invariant, that is according to the findings of this study, the various subsets of items have no significant influence on the person parameter estimates based on IRT theoretical framework.

The findings of this study are similar to other studies on measurement invariance (Stark et al. 2006) which emphasised that "Unless measurement invariance is established, conducting cross-group comparisons of mean differences or other structural parameters is meaningless. The degree to which instruments are invariant across use in different situations and with different groups of people has been greatly facilitated by the development of several analytic techniques including item response theory and confirmatory factor analysis (CFA)". According to Millsap (2010, pp. 5-9), "If the item response function varies across time, the probabilities of various item responses for a given latent variable score have changed over time. Measurement invariance is then violated, and observed changes in item scores over time are ambiguous and difficult to interpret. Change could stem from either real change among the individuals involved or the changing item parameters. However, if the item response function is invariant over time, then the observed changes in item scores cannot result from changes in item parameters and must have other explanations (such as real growth)".

From the result of this current study, it is only the IRT person parameter estimates that exhibited the invariance property, whereas the CTT person parameter estimates did not exhibit the invariance property. And for objective measurements, the invariance property of person statistics is imperative, and there is the need for further research work for more empirical investigation on the invariance of person parameter estimates based on CTT and IRT.

REFERENCES

Hambleton RK, Swaminathan H 1985. *Item Response Theory*: *Principles and Application*. Boston: Kluwer.

Hambleton RK, Swaminathan H, Rogers HJ 1991. *Fundamentals of Item Response Theory*. Newbury Park, California: Sage Publications.

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

Horn JL, McArdle JJ 1992. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18: 117-144.

Lord FM 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Millsap R 2010. Testing Measurement Invariance using IRT in Longitudinal Data: An Introduction. *Child Development Perspectives* 4(1): 5-9.

Nenty H J 2004. From classical test theory (CTT) to item response theory (IRT): An introduction to a desirable transition. In: OA Afemikhe, JG Adewale (Eds.): *Issues in Educational Measurement and Evaluation in Nigeria* (In honour of Wole Falayajo). Institute of Education, University of Ibadan, Ibadan, Nigeria, pp.372-384.

Stark S, Chernyshenko OS, Drasgow F 2006. Detecting DIF with CFA and IRT: Towards a unified strategy. *Journal of Applied Psychology*, 91: 1292–1306.

Thissen D 1991. *Multilog Version 7.0. Multiple Categorical Item Analysis and Test Scoring Using Item Response Theory.* Chicago: Scientific Software International.