ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Comparison of Classical Test Theory and Item Response Theory:A Review of Empirical Studies

[1]Ado Abdu Bichi, [2]Rahimah Embong, [3]Mustafa Mamat, [4]Danjuma A. Maiwada

[1]Faculty of Islamic Contemporary Studies,Universiti Sultan ZainalAbidin, 21300 Kuala Terengganu, Malaysia
[2]Research Institute of Islamic Product & Civilization,Universiti Sultan ZainalAbidin, 21300 Kuala Terengganu, Malaysia
[2]Faculty of Informatics and Computing, Universiti Sultan ZainalAbidin, 21300 Kuala Terengganu, Malaysia
[3]Faculty of Education, Northwest University, PMB 3220, Kano-Nigeria

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form 26 December 2014
Accepted 28 January 2015
Available online 1 April 2015

Keywords:
Classical Test Theory, Item Response Theory, Psychometrics, Educational Measurement

ABSTRACT

Background: The practice of testing has become increasingly common and the reliance on information gained from test scores to make decision has made an indelible mark on our culture. The entire educational system is today highly concerned with the design and development of the tests, the procedures of testing, instruments for measuring data, and the methodology to understand and evaluate the results. In theory of measurement in education and psychology there are two competing measurement frameworks, namely Classical Test Theory (CTT) and Item Response Theory (IRT). The techniques of the two frameworks are applied in assessment situations to improve test analysis and test refinement procedures. **Objective:** The main purpose of this paper is to provide a critical review of relevant empirical studies conducted to compare the two theories in test development. **Results:** Findings reveals that CTT and IRT are highly comparable; however, no study provides enough empirical evidence on the extent of disparity between the two frameworks and the superiority of IRT over CTT despite the theoretical differences. **Conclusion:** the inability of these empirical studies to provide enough evidence of superiority of the IRT over CTT may result from the instruments they used in conducting the studies. It is recommended that further studies be conducted with different tools to further explore the true picture of the framework and provide enough evidences to justify or prove the theoretical stands of the two frameworks in the field of educational and psychological measurement.

To Cite This Article: Ado Abdu Bichi, RahimahEmbong, Mustafa Mamat, Danjuma A. Maiwada, Comparison of Classical Test Theory and Item Response Theory: A Review of Empirical Studies. ***Aust. J. Basic & Appl. Sci.***, *9(7): 549-556, 2015*

INTRODUCTION

Assessment of students learning is very important in education. The assessment of students' cognitive abilities, academic skills and intellectual development involves certain techniques employed to sample students' performance on a particular learning outcome targeted by the instructional objectives one of that techniques is test, the test is expected to sample students' behaviours. Thus creating quality tests is very important in assessing the students' performance; many indices have been developed in order to construct valid and reliable items during test development. These indices developed mostly rely on the two popular statistical frameworks Classical Test Theory and Item Response Theory. The two frameworks are associated with the item development process in the field of educational and psychological test. These frameworks are widely been used in test development to ensure quality of measuring instruments and discuss in various literatures in the field of psychological and educational measurements on their suitability and effectiveness in test development process. In the theories the models associated with each have been described and compared, and the ways in which test development generally proceeds within each frameworks have demonstrated (Hambleton and Swaminathan, 1993) the existence of the theoretical as well as empirical differences and similarities of the two frameworks were extensively described in many studies. This paper provides a critical review of the existing empirical studies conducted to describe and compare the two popular frameworks.

Theoretically, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person statistics (Fan 2008).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

Corresponding Author: RahimahEmbong, Research Institute of Islamic Product & Civilization, Universiti Sultan ZainalAbidin, 21300 Kuala Terengganu, Malaysia.
Tel: +60199109727; E-mail: rahimahembong@unisza.edu.my

person statistics of IRT has been illustrated theoretically (Hambleton and Swaminathan, 1985; Hambleton, Swaminathan, and Rogers, 1991). A search of literatures revealed that several studies have been conducted to empirically examine the comparability of IRT-based and CTT-based item and person statistics i.e. (Lawson 1991; Fan 1998; MacDonald and Paunonen 2001; Guler*et al*., 2014; Courville 2004; Ojerinde*et al*., 2012; Ojerinde, 2013; and Magno 2009) all the studies compared IRT-based and CTT-based item and person statistics using different data sets and settings, and their findings revealed a strong relationships between the IRT and CTT, it Suggest that information from the two approaches about items and examinees might be very much the same. However, Cook, Eignor, and Taft(1988)reported Lack of invariance for both CTT-based and IRT-based item difficulty estimates.

Despite the number of empirical studies conducted to directly or indirectly provide empirical evidences of the relationship between CTT and IRT, there are not enough studies that provide evidence theoretical superiority of IRT over CTT. This inability of the studies to provide a clear distinction between the two measurement frameworks as theoretically been established leaves much to be ask as relate to the suitability of the tools used in the studies.

Objective:

The main purpose of this paper is to critically review the previous empirical studies that were conducted to clarify some aspects of classical test theory (CTT) and item response theory (IRT) modeling especially with regards to item development.

Significant of the study:

The results from this review of empirical comparison of CTT and IRT will provide measurement specialists, test developers and the researchers with information regarding the suitability and comparability of the two frameworks from the practical point of view and provide basis for further research to improve measurement practice.

The following questions guide the study:

(1) How comparable are the CTT and IRT frameworks in terms of item and person parameters?

(2) Is there enough empirical evidence on the extent to which CTT and IRT behave differently?

2.0 Concept of CTT and IRT:

Classical test theory (CTT) and item response theory (IRT) are generally perceived as the two popular statistical frameworks for addressing Measurement Problems, both the two approaches describe characteristics of an individual and analyze abilities and latent attributes and enable to predict outcomes of psychological and educational tests by identifying item parameters which are item difficulty, discrimination and the ability of the examinees. Although CTT has been used for most of the time in educational and psychological measurement, in recent decades IRT has been gaining ground, there by becoming a favourite measurement framework. The weak theoretical assumptions are the major arguments against CTT which make it according to Hambleton and Jones (1993), easy to apply in many testing situations.In their views, the person statistic is item dependent and the item statistics such as item difficulty and item discrimination are sample dependent. On the other hand, IRT is more theory grounded and models the distribution of examinees' success at the item level. As its name implies, IRT mainly focuses on the item-level information in contrast to CTT's principal focus on test-level information.Notwithstanding the recent growth and theoretical superiority of the item response theory (IRT), classical test theory(CTT) continues to be an important framework for test construction (Bechger*et al*.,2003).It is therefore pertinent to give a brief explanation of the two frameworks in order to give a clear notion of the relationship between CTT and IRT for researchers and item writers who are frequently familiar with the frameworks to appreciate the two.

2.1 What is Classical Test Theory?

According to Hambleton and Jones (1993) Classical test theory is a theory about test scores that introduces three concepts- (1) test score (often called the observed score), (2) true score, and (3) error score. Within that theoretical framework, models of various forms have been formulated. For example, in what is often referred to as the "classical test model," a simple linear model is postulated linking the *observable test score(X)* to the sum of two unobservable (or often called *latent)* variables, *true score (T)* and *error score (E),* that is:

$X = T + E.$

Because the true score is not easily observable, instead, the true score must be estimated from the individual's responses on a set of test items. Therefore the equation is not solvable unless some simplifying assumptions are made. The major assumptions underlines the CTT are (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on the parallel tests are uncorrelated.

The major advantages of CTT as highlighted by Hambleton and Jones (1993) are its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations.The benefit of using CTT in test development as given by Schumacker (2010), are (1) when compared to item response theory models, analyses can be performed with smaller representative samples of examinees. This is particularly important when field-testing a measuring instrument. (2) Classical test analysis employs relative simple mathematical procedures and model parameter estimations are conceptually straightforward.(3) Classical test analysis is often referred to as “weak models” because the assumptions are easily met by traditional testing procedures.

Fan (1998) summarizes the major limitation of CTT as circular dependency: (a) The person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent. This circular dependency poses some theoretical difficulties in CTT’s application in some measurement situations (e.g., test equating, computerized adaptive testing).

The major focus of CTT is on test-level information, however item statistics (i.e., item difficulty and item discrimination) are also an important part of the CTT model (Fan, 1998)

2.2 What is Item Response Theory?

Hambleton and Jones (1993) describe Item response theory as a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test. Item responses can be discrete or continuous and can be dichotomously or polychotomously scored; item score categories can be ordered or unordered; there can be one ability or many abilities underlying test performance; and there are many ways (i.e., models) in which the relationship between item responses and the underlying ability or abilities can be specified. Within the general IRT framework, many model shave been formulated and applied to real test data.

The characteristics of Item Response Models as summarised by Hambleton and Swaminathan (1985) are, first, an IRT model must specify the relationship between the observed response and underlying unobservable construct. Secondly, the model must provide a way to estimate scores on the ability. Thirdly, the examinee’s scores will be the basis for estimation of the underlying construct. Finally, an IRT model assumes that the performance of an examinee can be completely predicted or explained from one or more abilities. In item response theory, it is often assumed that an examinee has some latent, unobservable trait (also called ability), which cannot be studied directly. The purpose of IRT is to propose models that permit to link this latent trait to some observable characteristics of the examinee, especially his/her faculties to correctly answering to a set of questions that form a test (Magis 2007).

Item Response Theory, Item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptote). Three most commonly used IRT models are; one parameter logistic model (1PLM or Rasch model), two parameter logistic model (2PLM) and three parameter logistics model (3PLM).

All three models have an item difficulty parameter (b), In addition, the 2PL and 3PLmodels possess a discrimination parameter (a), which allows the items to discriminate differently among the examinees. The 3PL model contains a third parameter, referred to as the pseudo-chance parameter (*c*). The pseudo-chance parameter (c) corresponds to the lower asymptote of the item characteristic curve (ICC) which represents the probability that low ability test takers will answer the item correctly and provide an estimate of the pseudo-chance parameter (Embretson and Reise, 2000)

3.0 Comparison of CTT and IRT:

According to Sohn (2009) one of distinguishing characteristics of item indices under CTT and IRT frameworks is whether they are sample dependent or invariant. The item parameters under CTT are regarded as sample dependent, because they are based on the total score of the test which is the person parameter in CTT and has a variant attribute. Another way of saying this is that the values of the item parameters are different across the samples collected for the test. This characteristic may be a threat to the reliability of the test. So, in order to generalize the results of the test, random sampling is assumed for CTT. The item parameters under IRT, however, are not considered to be dependent upon the ability level of the examinees responding the item (Baker, 2001). In other words, the item parameters are regarded as sample invariant. If an item measures the same latent trait for groups, the estimated item parameters are assumed to be the same. Because the item difficulty parameter under IRT is independent of the samples, it is considered easier to interpret than that under CTT. Baker and Kim (2004) argue that the concepts of item difficulty in CTT and the location parameter *i b* in IRT are not completely interchangeable. Under CTT, an easy item is defined by a low ratio of the correct response to an item in the total population. On the other hand, under IRT, an item is defined as easy when the magnitude of the item difficulty parameter is less than the average level of ability. Considering item discrimination parameter, however, it is regarded as the parameter which makes it possible to establish a distinction among examinees’ different ability under both CTT and IRT. Thus, IRT has been considered to hold advantages over CTT at least in terms of theoretical point of view. Lord (1980) argued that IRT provides the methods of optimally discriminating items in the scope of a passing score. However, practical researches comparing CTT and IRT have not shown there is consistent superiority of IRT measurement statistics.

Table 1: Main Difference between CTT and IRT Models, source :(Hambletonand Jones 1993).

Area	CTT	IRT
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (i.e easy to meet with test data)	Strong (i.e more difficult to meet with data)
Item-ability relationship	Not specified	Item characteristics functions
Ability	Test scores or estimated true scores are reported on the test-score scale (or a transformation test score scale)	Ability scores are reported on scale – ∞ to + ∞ (or transformed scale)
Invariance of item and person statistics	No-item and person parameters are sample dependent	Yes-item and person parameters are sample independent, if model fits test data
Item statistics	*p*, *r*	*b*, *a* and *c* (for the three-parameters model) plus corresponding item information functions
Sample size (for item parameters estimation)	200 to 500 (in general)	Depends on the IRT model but larger samples i.e over 500, in general are needed

4.0 Review of Empirical Studies:

Many studies have been conducted to investigate the comparability of item and person parameter estimates by CTT and IRT approaches. The studies conducted by different scholars are critically reviewed and discuss in this study.

Lawson (1991) comparesRasch model item- and person-parameters to CTT difficulty and number-right in three sets of examination data. His CTT and IRT results showed that the correlation coefficient (r= -.9949) of the level of item difficulty through CTT and IRT (Rasch model) were very high. This result indicated that item difficulty estimates behaved very similarly in two different approaches, CTT and IRT. CTT and IRT have also been compared under simulated conditions.

Fan (1998) uses a large-scale test database from a statewide assessment program to examined and also compared CTT and IRT estimates from different subsamples of N=1000 to investigate the invariance properties of CTT and IRT parameter *estimates*. He created samples varying in their representativeness by sampling a larger dataset (e.g., random selections vs. men and women vs. high and low scores). Of course, IRT parameters are known to be invariant but Fan was testing the empirical invariance of the IRT item parameter *estimates* and comparing them to estimates of CTT statistics, which are thought not to be invariant. Fan found that *both* CTT and IRT difficulty, and to a lesser degree, discrimination statistics displayed invariance. For item difficulty, CTT estimates were *closer* to perfect invariance. Overall the Fan's Finding shows that both the CTT and IRT produced similar results in terms of comparability of item and person statistics and also on the degree of invariance of the item statistics from the two approaches. In his conclusion, Fan questioned whether IRT had the "advertised" advantages over CTT.

Idowu*et al*. (2001) apply the Classical Test Theory and Item Response Theory to evaluate the quality of an assessment constructed by the researchers to measure National Certificate of Education (NCE) students' achievement in Mathematics. A sample of 80 students was drawn for the study from the Abia State College of Education, Arochukwu. The instrument used was the Mathematics Achievement Test (MAT) for College students developed by the researchers. Data was analysed in two dimensions. First, the psychometric properties of the instrument were analyzed using CTT and IRT and the detection of item bias was performed using the method for Differential Item Functioning (DIF). The results showed that although Classical Test Theory (CTT) and Item Response Theory (IRT) methods are different in so many ways; outcome of data analysis using the two methods in this study did not say so. Items which were found to be "bad items" in CTT came out not fitting also in Rasch Model. Overall results of analysis showed that the achievement test in its generality was a good test. Although there are items removed, revised, and rephrased, most of the items came out to be "good items". These were also the items that turned out to have extreme logit measures qualifying it to be unfitting in the latent trait model. Surprisingly, some of the items came out to be biased as detected in the DIF analysis.

MacDonald and Paunonen (2002) felt that prior research might be influenced by the fact that they were real-data studies. In particular, these researchers were interested in the effects of the specific items used in the study. They also wished to examine accuracy, which was not possible in the previous, real-data studies. Therefore, they simulated data using 1PL and 2PL IRT models and then computed IRT and CTT statistics from these values. They performed three sets of correlations. First, they tested comparability of test scores, difficulty, and item discrimination by correlating estimated IRT and CTT statistics; they found very high comparability for test scores and difficulty and less comparability for item discrimination. Next, they correlated values obtained from different samples to test invariance; they found exceptional invariance with CTT exhibiting slightly closer to perfect invariance as compared to IRT.

Courville (2004) in the examination of empirical comparison of item response theory and classical test theory item/person statistics, the study focused on two central themes: (1) how comparable are the item and person statistics derived from the item response and classical test framework? and (2) How invariant are the item statistic from each measurement framework across examinee samples? The ACT Assessment test composed of four tests: English, Mathematics, Reading, and Science were used for the study. Random samples of 80,000 examinees composed of 40,000 males and 40,000 females were drawn from the population of 322,460. The results of this study indicate high correlations between CTT-based and IRT-based estimates, at least for the one-parameter and two-parameter models. This result holds for either small sample clinical trials or large sample assessment situations. Similarly the CTT item difficulty estimates, for the random sampling plan, had a higher degree of invariance than the IRT-based item difficulty estimates, especially for the two- and three-parameter models. The discrimination indices, however, correlated highly only when the spread of discriminations was large and the spread of difficulty values was small

Progar*et al. (2008)* in their study titled An empirical comparison of Item Response Theory and Classical Test Theory ,the researchers used a real data set from the Third International Mathematics and Science Study (TIMSS 1995) to address the following questions: (1) How comparable are CTT and IRT based item and person parameters? (2) How invariant are CTT and IRT based item parameters across different participant groups? (3) How invariant are CTT and IRT based item and person parameters across different item sets? The findings indicate that the CTT and the IRT item/person parameters are very comparable, that the CTT and the IRT item parameters show similar invariance property when estimated across different groups of participants, that the IRT person parameters are more invariant across different item sets, and that the CTT item parameters are at least as much invariant in different item sets as the IRT item parameters. The results furthermore demonstrate that, with regards to the invariance property, IRT item/person parameters are in general empirically superior to CTT parameters, but only if the appropriate IRT model is used for modelling the data

Zaman*et al*. (2008) in their study to compare the CTT and IRT for students ranking on the basis of their abilities on objective type test in Physics at secondary level taking a random sample of 400, 9th grade students from variety of population in Pakistan using a content valid test of 80 multiple choice item, found out that, CTT-Based and IRT-based examinee ability estimates were very comparable and highly correlated (0.95), indicating that the ability level of individual examinees will lead to similar results across the different measurement theories.

Magno (2009) has conducted a study to demonstrate the difference between classical test theory (CTT) and item response theory (IRT) approach using a random sample of 219 junior higher school students in Philippines, and actual test data for chemistry. The CTT and IRT were compared across two samples and two forms of test on their item difficulty, internal consistency, and measurement errors. The results demonstrate certain limitations of the classical test theory and advantages of using the IRT. It was found in the study that, IRT estimates of item difficulty do not change across samples as compared with CTT with inconsistencies; difficulty indices were also more stable across forms of tests than the CTT approach; IRT internal consistencies are very stable across samples while CTT internal consistencies failed to be stable across samples; IRT had significantly less measurement errors compared to CTT.

Adedoyin (2010) investigates the invariance of person parameter estimates based on Classical Test and Item Response Theories, 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, a random sample of five thousand examinees (5000) were drawn from the population 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. The findings of the 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.

Ojerinde*et al*. (2012) evaluate the use of English pre-test data so as to compare indices obtained using the 3-parameter model of the Item Response Theory (IRT) with those from the Classical Test Theory (CTT) approach and verify how well the two can predict actual test results and the degree of their comparability, using a sample of 1075 test takers that took one version of the pre-test in use of English of the UTME. The findings in this study have indicated that the person and item statistics derived from the two measurement frameworks are quite comparable. The degree of invariance of item statistics across samples, usually considered as the theoretical superiority of IRT models, also appeared to be similar for the two measurement frameworks but the IRT model provided a better idea about the internal consistency of the test than CTT. However, the 3PL model was found to be more suitable in multiple-choice questions in ability test but involved more complex mathematical estimation procedure than the CTT. In the overall, indices obtained from both approaches gave valuable information with comparable and almost interchangeable results.

Pido (2012) conducted a study to determine and compare the item parameters of the MCQ in the 2011 Uganda Certificate of Education (UCE) examinations using the CTT and IRT approaches. Four subjects, Geography, Chemistry, Physics and Biology Paper were used. 480 scripts of the examinees in each subject were selected as sample for the study. The examinees' responses were analysed using the Xcalibre 4.1.7.1 software to determine item parameters based on the CTT and IRT approaches. The correlation coefficient and the inspection methods were used to compare the difficulty and discrimination indices obtained from both the CTT and IRT approaches. The results revealed a very high correlation between the difficulty indices obtained using CTT and IRT approaches. Similar result was also found between discrimination indices. The overall result revealed a strong relationship between the values of item parameter estimated using CTT approach and those estimated using IRT approach.

Abedalazizand Leng (2013), in the study to examine the Relationship between CTT and IRT Approaches in Analysing Item Characteristics, the aim was to compare the item difficulty and item discrimination of the Mathematical ability scale using the two methods across 1, 2, and 3 parameters. The instrument was administered to tenth grade sample of N=602. The data gathered was analysed for possible relationship of the item characteristics using CTT and IRT methods. Results indicate that the 3-parameter logistic model has the most comparable indices with CTT, furthermore, CTT and IRT models (1-parameter logistic model and 3-parameter logistic model) can be used independently or altogether to describe the nature of the items characteristics.

Nenty and Adedoyin (2013) in their study to compare and testfor significance the invariance between the item parameter estimates from CTT and those from each of 2-,3- IRT across models/theories for inter and intra model validation of the two test theories. 10,000 junior secondary school pupils were randomly selected from the population of 36,940 pupils who sat for 2010 mathematics paper 1 examination in Botswana. The estimated CTT and IRT item parameters were tested for significance with respect to invariance concept using dependent t-test with respect to the two theory models, the result It showed that, the inter validation of item parameter estimates between CTT and 2-3-IRT models were not significant. This showed that the inter invariance concept of item parameter estimates for CTT and 2-, 3-IRT models was established. The item difficulty parameter estimates between 2-3-IRT models were not statistically significant, but there was a statistical significant difference in the item discrimination parameter estimates. This showed that the intra invariance concept of item discrimination parameter estimates for 2-, 3-IRT models could not be established.

In a study conducted to Assess the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters Adedoyinand Adedoyin(2013) found out that, the CTT and IRT item difficulty and item discrimination values were positively linearly correlated and there was no statistical significant difference between the item difficulty and item discrimination parameter estimates by CTT and IRT.

Ojerinde (2013) conducted a study to evaluate the psychometric utility of data obtained using the two models in the analysis of UTME Physics Pre-test so as to examine the results obtained and determine how well the two can predict actual test results and the degree of their comparability. The researcher also verified the conditions to be fulfilled for IRT to be usefully applied with real test data. Findings showed that the result obtained using the IRT model was found to be more suitable in multiple-choice questions (MCQs) in ability test but involved more complex mathematical estimation procedure than the classical approach. In the overall, indices obtained from both approaches gave valuable information with comparable and almost interchangeable results in some cases.

In another study, Guler*et al*. (2014) compare classical test theory and item response theory in terms of item parameters, the aim of their study was to empirically examine the similarities and differences in the parameters estimated using the two approaches, a random sample of 1250 students from the group of 5989 students who had taken 25-item Turkish high schools entrance exam (HSEE) in 2003 were used In the study, the findings reveals that, the highest correlations between CTT and 1-parameter IRT model (0.99) in terms of item difficulty parameters, and between CTT and 2-parameter IRT model (0.96) in terms of item discrimination parameters. The result also shows the lowest level of correlation between the 3-parameter model and CTT although the 3-parameter model was identified as the most congruous one in terms of model-data fit. In the light of their findings, it may be said that there is not much difference between using 1 or 2-parameter IRT model and CTT. However, in cases where the probability of guessing is high, there is a significant difference between 3-parameter model and CTT.

The researchers correlated the estimated and true statistics to examine accuracy, the studies demonstrated that the item and person parameters generated by CTT and IRT were very accurate and highly comparable across all conditions. However in the case of the item discrimination indices, the statistic under IRT accurately estimated the discrimination value in all conditions. On the other hand, the item discrimination value under the CTT framework obtained an accurate estimate only when the potential item pool had a narrow range of item difficulty levels. Generally the studies show a high degree of measurement accuracy of the CTT and IRT framework, and provide no substantial evidence of superiority of IRT over CTT.

5.0 Discussion:

Fan (1998) claims that, the CTT has served the measurement community for most of this century and IRT has witnessed an exponential growth in recent decades. In comparison of theories it is determined that IRT models are more informative than CTT models if samples are big enough to allow their application, if the items obey the laws defining the models, and if detailed information about the itemsis sought (Steyer, 1999).

Thus, to summarise the prior literatures, all the findings shows very high correlations between the CTT-Based person/item parameters, and IRT-Based Person/item parameters, the studies further found no evidence of a higher invariance of the IRT item parameters in comparison to the CTT item parameters. The discrimination indices, however, correlated highly only when the spread of discriminations was large and the spread of difficulty values was small. Moreover, the CTT discrimination estimates were in some conditions (i.e., at a large spread of difficulties) less accurate than the IRT estimates. Similarly, Fan concludes that IRT may not have many advantages over CTT for traditional tasks, like test construction, while MacDonald emphasized the advantages of using IRT to construct tests.

However, few studies found that IRT provided more dependable statistical information than the CTTespecially where the probability of guessing is high, (MacDonald and Paunonen 2001; Guler*et al*., 2014; Ojerinde*et al*., 2012; Ojerinde, 2013; and Magno2009). With regards to the invariance property, IRT item/person parameters are in general empirically superior to CTT parameters, but only if the appropriate IRT model is used for modeling the data (Progarand Soča, *2008;* andAdedoyin, 2010).

Lawson's (1991) and Fan's (1998) results therefore seem damning for IRTat least in terms of the inherent superiority generally afforded to IRT. However, Lawson's study used only the Rasch model and Fan's comparison of the Rasch model with the 2PL and 3PL suggests that the Rasch model may be least different from CTT. Neither Lawson's nor Fan's study simulated the response data, so they were unable to compare the estimates to population parameters.

However the critical analysis of all these studies conducted to compare the two frameworks, had their own share of limitations that may potentially undermine the validity of their findings.

First of all, the characteristics of the test items used in many studies are out of date. Most of the studies used out of date data because their aim was to compare the theories, therefore they used the data only as a research instrument.

Secondly, most of the studies used limited item pool. Ideally, the test item pool should be larger and more diverse in terms of item characteristics so that items can be sampled from the pool to study the behaviours of CTT and IRT item statistics under different conditions of item characteristics (Fan, 1998).

Finally, all the studies utilises data from standardised test items for their analyses. Fourthly, some of the studies used limited or small examinee sample their analysis (i.e N<500 examinees).

6.0 Recommendations:

Despite the theoretical differences between IRT and CTT and the superiority of IRT over CTT, from the above review of the existing literature it appears that there is not enough empirical knowledge about how, and to what extent, the IRT- and CTT-based item and person statistics behave differently.

It is recommended that further studies be encourage in the same field to utilises a recent or updated test data, large item pool, large examinee sample (i.e N>500 examinees) and a calibration of a newly teacher made achievement test (Non-standardised test items) to compare the CTT and IRT item parameters and to investigate the invariance of the parameters across such frameworks. This would greatly enhance the interpretability and objectivity of tests developed by teachers at all levels of education from primary, secondary and tertiary educationto improves test analysis and test refinement procedures.

Conclusion:

From the review of these empirical studies a common thread emerges. The CTT-based and IRT-based item and person parameters were very accurate and highly comparable across all conditions. Thus, showing a high degree of measurement accuracy of the CTT and IRT, however the studies show that, in the case of item discrimination indices, the statistic under IRT accurately estimated the discrimination value in all conditions, as against CTT framework which obtained an accurate estimate only when the potential item pool had a narrow range of item difficulty levels. Similarly, no study provides enough empirical evidence on the extent of disparity between the two frameworks and the superiority of IRT over CTT despite the theoretical differences. This inability to provide the evidence may result from the instruments and methods used in their studies.

REFERENCES

Abedalaziz,N.,C.H.Leng, 2013.The Relationship between CTT and IRT Approaches in Analysing Item Characteristics.The Malaysian Online Journal of Educational Sciences, 1(1).

Adedoyin, O.O., 2010. Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories. International Journal of Educational Sciences, 2(2): 107-113.

Adedoyin, O.O., J.Adedoyin, 2013.Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters.Herald Journal of Education and General Studies, 2(3): 107-114.

Baker, F.B., 2001.The Basic of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, USA.

Baker, F.B., S.H. Kim, 2004.Item response theory: Parameter estimation techniques (2nded.). New York: Marcel Dekker.

Bechger, T.M., G.Maris, H.H.Verstralen, A.A.Béguin, 2003.Using classical test theory in combination with item response theory.Applied Psychological Measurement, 27(5): 319-334.

Cook, L.L., DR.Eignor, HL.Taft, 1988. A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. Journal of Educational Measurement, 25: 31-45.

Courville, T.G., 2004. An empirical comparison of item response theory and classical test theory item/person statistics.Ph.D Dissertation, Texas A & M University.

Crocker, L., J.Algina, 1986.Introduction to classical and modern test theory.New York: Holt, Rinehart & Winston.

Erguven, M., 2014. Two approaches to psychometric process: Classical test theory and item response theory. Journal of Education, 2(2): 23-30.

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.

Güler, N., G.K.Uyanık, G.T.Teker, 2014.Comparison of classical test theory and item response theory in terms of item parameters.European Journal of Research on Education, 2(1): 1-6.

Hambleton, R.K., H.Swaminathan, 1985.Item response theory: Principles and applications (Vol. 7): Springer.

Hambleton, R.K., R.W. Jones, 1993.Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12(3):3847.

Idowu, E.O., A.N.Eluwa, B.K.Abang, 2011. Evaluation of Mathematics Achievement Test: A Comparison Between Classical Test Theory (CTT) and Item Response Theory (IRT) Journal of Educational and Social Research,1(4):99-106.

Kinsey, T.L., 2003.A comparison of IRT and RASCH procedures in a mixed-item format test: Unpublished Doctoral Thesis, University of North Texas.

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 Press.

Lord, F.M., 1980.Applications of item response theory to practical testing problems.Hillsdale, NJ: Lawrence Erlbaum.

MacDonald, P., S.Paunonen, 2002. A Monte Carlo Comparison of item and person statistics based on item response theory versus classical test theory, Educational and Psychological Measurement, 62: 921-943.

Magis, D., 2007. Influence, Information and Item Response Theory in Discrete Data Analysis.Retrieved on 3 June, 2014 fromhttp://bictel.ulg.ac.be/ETD-db/collection/available/ULgetd-06122007-100147/.

Magno, C., 2009. Demonstrating the Difference between Classical Test Theory and Item Response Theory using Derived Test Data.The International Journal of Educational and Psychological Assessment, 1(1): 1-11.

Mead, A.D., AW.Meade, 2010.Item selection using CTT and IRT with unrepresentative samples.Paper presented at the twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA.

Nenty, H., O.O.Adedoyin, 2013.Test for invariance: inter and intra model validation of classical test and item response theories. Asia PacificJournal of Research,I(IX).

Ojerinde, D., 2013. Classical test theory (CTT) VS item response theory (IRT): An evaluation of the comparability of item analysis results. A guest lecture presented at the Institute of Education, University of Ibadanon 23rd May.

Ojerinde, D., K.Popoola, P.Onyeneho, 2012. A comparison between classical test theory and item response theory: experience from 2011 pre-test in the use of English language paper of the unified tertiary matriculation examination (UTME).Journal of educational assessment in Africa, 7: 173-191.

Pido, S., 2012.Comparison of item analysis results obtained using item response theory and classical test theory approaches. Journal of educational assessment in Africa, 7: 192-209.

Progar, S., G.Socan, M.Slovejija, 2008.An empirical comparison of item response theory and classical test theory.Horizons of Psychology**,** 17(3): 5-24.

Sohn Y., 2009.A Comparison of Methods for Item Analysis and DIF using Classical Test Theory, Item Response Theory and Generalized Linear Model. M.Ed Thesis, University of Georgia.

Zaman, A., A.U.R. Kashmiri, M.Mubarak, A.Ali, 2008. Students Ranking, Based on their Abilities on Objective Type Test: Comparison of CTT and IRT. Retrieved on 23 April, 2014 from http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1051&context=ceducom