GROUP AND ITEM INVARIANCE OF ITEM DIFFICULTY PARAMETER BASED ON ITEM RESPONSE AND CLASSICAL TEST THEORIES

[1]Ijeoma Joy Chikezie (Ph.D) & [2]Eme U. Joseph (Ph.D)

[1]School of Education and Ancillary Studies
National Institute for Nigerian Languages
P.M.B. 7078, Aba, Abia State.
joyeddy@yahoo.com, 08030975128
&
[2]Department of Educational Foundations, Guidance & Counseling
University of Uyo,
P.M.B. 1017, Uyo, Akwa Ibom State.
emeujoseph@yahoo.com, 08023263002

ABSTRACT

The main concern of any assessment procedure is to adopt a measurement approach that will yield valid and reliable test items and test scores on which decisions about the examinees are based. Classical Test Theory (CTT) and Item Response Theory (IRT) are two major measurement frameworks employed in psychometrics. CTT, used over the years, has been theoretically criticized for its inability to solve measurement problems such as test equating, differential item functioning, item banking, and invariance among other. The emergence of IRT as a preferred framework sparked off a debate in the psychometric community on the superiority of IRT over CTT, particularly in the provision of ability estimates and item parameters that are independent of test and sample respectively. This study, therefore, compares the CTT and IRT in terms of item/person parameters. Descriptive survey research design was adopted. The sample comprised 1150 senior secondary three students drawn from 50 schools in Abia State using multistage sampling technique. Two parallel Chemistry Achievement Tests (CAT A & B) developed by the researchers were used for data collection. Each instrument consisted of 60 items of a 4-option multiple choice test. Kuder-Richardson formula 20 reliability estimates of CAT A and CAT B yielded coefficients of 0.88 and 0.90 respectively. Four hypotheses guided the study. The item difficulty and person parameters from IRT and CTT were tested for invariance using independent and dependent t-tests at 0.05 alpha levels.

Findings indicated a significant difference between item/person parameters of CTT and IRT. Furthermore, the item difficulty parameter and ability estimates of IRT were invariant as against those of CTT. Based on the findings, it was therefore, concluded that, given that the data fits the IRT model used, IRT is empirically superior to CTT. Logical recommendations were highlighted which include that IRT should be used in solving measurement problems.

Keywords: Classical test theory, Item response theory, item parameter invariance, group invariance.

INTRODUCTION

Much research and debate have been on-going in the psychometric community motivated by the question of the superiority of Item Response Theory (IRT) over Classical Test Theory (CTT). Classical test theory and item response theory are two major statistical frameworks employed in addressing measurement challenges. Although CTT has been the basis for most measurement procedures, it has been criticized because its item and person statistics are sample dependent. Classical test theory, also known as true score theory relates the observed test score (X) of an examinee for a certain test to the true score (T) and the error score (E) (Duong, 2004). The relationship is expressed in a simple equation as: $X = T \pm E$. The true score are unobservable variable, thus they are unknown variables and make it difficult to solve the equation. However, there are three fundamental assumptions of CTT models that would make the equation solvable: (1) true score and error score are uncorrelated (2) the average error score over population of examinees is zero and (3) error scores on parallel tests are not correlated (Hambelton & Jone cited in Metibemu & Oluwatayo, 2016).

According to Tractenberg (2010), CTT focuses on total test score. Classical theoretic constructs operate on the summary of items, individual scores are not considered. Moreover, the test-score emphasized implies that when an outcome measure is established, characterized, or selected on the basis of its reliability, tailoring the assessment is not possible, and in fact, the items in the assessment must be considered exchangeable. Another feature of CTT is that it utilizes measures of item difficulty and item discrimination, the values of which are dependent upon the distribution of examinee proficiency within a sample (Adedoyin, Nenty & Chilisa, 2008). Based on the two item statistics, items are chosen as the test generates desired test score distribution and has high item-total score correlation. Put differently, items with high discrimination indices are selected while level of difficulty is controlled by the purpose of the test and the predicted ability distribution of the target population of examinees on the abilities being tested.

The application of CTT models in measurement issues suggests that they have some advantages that can be attributed to them. The models are relatively easy to use,

understand and apply in testing practice (Brennan, 2010). Computationally, they are simple and do not require strict goodness-of-fit study to ensure the good fit of a model to actual test data (Duong, 2004). Besides, CTT models require relatively small sample size. While CTT models have proven useful in these areas, they have serious shortcomings. The item and person characteristics (item difficulty parameter and examinee sores) are not discernible (An & Yung, 2014). Put differently, item and person statistics are largely dependent on the sub-population in question. If high ability sample is used, all the items would appear easy. On the other hand, if a low ability sample is used, the set of items will appear difficult. This limitation of CTT makes it difficult to estimate examinees' abilities using different test forms. In addition, after calibrating items from a population, the scores of subjects from that population can be compared directly even if they respond to different subsets of the test (An & Yung, 2014).

Awareness of the limitations of CTT models and attempts to overcome them ushered in IRT as an emerging trend in assessment and in psychometrics in general. In recent years, IRT has gained popularity and increasing attention as it is presented as modern and superior alternative to CTT (Embretson & Reise, 2000). Harris (1989) in Adedoyin, Nenty & Chilisa, (2008) sees IRT as a group of measurement that describes the relationship between an examinee's test performance and trait assumed to underlie the performance. In other words, it models the relationship between the latent variable (trait) being measured and the item response. Magno (2009) expressed the relationship as

$$P_i(?) = b_i + a_i(?);\ i = 1 \ldots n$$

Where

$P_i(?)$ = the probability that an examinee with a given latent trait, ability (?), will answer

item i correctly;

b_i = the item difficulty index;

a_i = the item discriminating index

n = the number of items.

The primary interest of IRT is not test level information rather item-level information. It considers the pattern of response to individual items (Tomkomiak & Wright, 2007cited in Metibemu & Oluwatayo, 2016). In IRT approach, each item on a test has its own Item Characteristic Curve (ICC) that describes the probability of an examinee responding correctly to a randomly selected item from a population of items supposed to measure the same ability. Basically, IRT focuses on the three- two- and one-parameter models in test calibration (estimating item and ability parameter). The main distinguishing factor of IRT models from those of CTT is the mathematical form of ICCs. The number of item parameters- item difficulty (b), item discrimination (a), and guessing (c), required to describe an ICC depends on the chosen IRT model. The three-parameter model describes a-, b-, and c- item parameters, two-parameter model describes a- and b- item parameters

while one-parameter model describes only b-item parameter. Due to the robust nature of the IRT models, test calibration is more effective and accurate if assumptions underlying the framework are met.

Having explored the characteristics, advantages and limitations of each of the frameworks, they can be effectively compared on some vital points to showcase the superiority of IRT over CTT. Item response theory, as the name implies, focuses mainly on the item level information in contrast to the CTT's principal focus on test level information (Morales, 2009, Nenty, 2004). Put in another way, CTT links test scores to true scores rather than items scores to true scores as is the case of IRT which is item centered in estimation of person's ability. The implication is that two examinees who have the same total number of items correct by CTT in the same test may not be assigned the same ability estimate with IRT. Item response theory assumes that there is a correlation between the score gained by a candidate for an item/ test and their overall ability on the latent trait which underlies test performance.

Item response theory procedure for item analysis consists of determining sample-invariant item parameters using relatively complex mathematical techniques and large sample sizes (which is evidence of robustness), and utilizing goodness-of-fit criteria to detect items that do not fit the specified response model (Hambleton & Jones, 1993 cited in Metibemu & Oluwatayo,2016). Critically, the characteristics of an item are said to be independent of the ability of the examinees that were sampled. Put in another way, sample in the contest of IRT is invariant, large and need heterogeneous sample (Baker, 2001; Partchev, 2004). Item analysis, in the framework of CTT consists of determining sample-specific item parameters by employing simple mathematical technique and moderate sample sizes, and deleting items based on statistical criteria. Classical test theory obtains four main indices from students' responses to test items; these are an index of item difficulty (or facility), an index of item discrimination, item validity and effectiveness of distraction (Izard, 2005). The major limitation is that item statistics depend to a great extent on the characteristics of the examinee sample used in the analysis.

Whereas reliability, in the context of CTT, refers to the precision of measurement, Item response theory makes it clear that precision is not uniform across the entire range of test score. Item response theory advances the concept of item and test information to replace reliability. Put in another way, item and test information show how precise the measurement of an individual's ability is. The standard error estimation is the reciprocal of the test information of a given trait level. More information implies less error of measurement. (de Ayala, 2009). The discrimination parameter plays a vital role in the function for two-and three-parameter models. In general, while highly discriminating items contribute greatly but over a narrow range, less discriminating items provide less information but over a wider range of ability (Warm, 1978 in Amajuoyi, 2015). According to Suen (1990), the informativeness of the item at a particular ? value is influenced by the amount of error associated with the measurement of that ? value. The

lower the measurement error of an item at a certain ? value, the more informative is that item at the ? value.

Classical test theory framework item and person parameters are sample dependent whereas in IRT item and person parameters are invariant (independent) if model fits the test data (University Testing Service, 2000). A practical implication of invariance principle is that a test located anywhere along the ability scale can be used to estimate an examinee's ability (Baker 2001). The approach of CTT holds that an examinee should obtain high score and low score for an easy and difficult test respectively. By this the underlying ability of the examinee cannot be ascertained. In contrast to IRT, an examinee could take an easy or difficult test and obtain approximately the same ability estimate. This indicates that the examinee's ability is fixed and invariant with respect to the items used to measure it (Baker, 2001). Further improvement of IRT as it concerns invariance of item and person statistics is that it provides significantly greater flexibility in situations where different samples or test forms are used.

Earlier studies on the comparison of CTT and IRT were based on item and person parameters (which is the important scientific property of any measurement). Progar and Socan (2008), using two-parameter model found that IRT item and person parameters were invariance as against those of CTT. They concluded that IRT is empirically superior to CTT. Magno (2009) demonstrated the difference between CTT and IRT compared the two theoretical framework across independent samples and two forms of test on item difficulty, internal consistency and measurement errors using three-parameter model. He found that IRT estimates of item difficulty do not change across samples as compared with CTT which were inconsistent. Moreover, difficulty indices were also more stable across forms of test than in CTT. Similarity, Adeboyin, Nenty & Chilisa (2008) in their study, investigated the invariance if item difficulty parameters based on CTT and IRT. They found that item difficulty based on IRT frame work, were invariant across the different independent samples while item difficulty estimates based on CTT were variant. They concluded that the finding discredited CTT frame work for its inability to produce p item difficulty invariant parameter estimates. As a departure, studies of Stage (2003); Wiberg (2004); Courville (2004) revealed high correlations between CTT and IRT item and ability parameter estimates indicating comparability.

The summary of the review shows that there are different outcomes and inconclusive results on the comparability of CTT and IRT frameworks probably because of the model used and the statistics used in analysis the data. The present study therefore, investigated the superiority of IRT over CTT by comparing the invariance of the item/person statistics using independent test forms independent samples of examinees.

Statement of the Problem

Classical Test Theory (CTT) and Item Response Theory (IRT) are two major measurement frameworks employed in psychometrics. CTT, used over the years, has been theoretically criticized for its inability to solve measurement problems such as test equating, differential item functioning, item banking, and invariance among other. The emergence of IRT as a preferred framework sparked off a debate in the psychometric community on the superiority of IRT over CTT, particularly in the provision of ability estimates and item parameters that are independent of test and sample respectively. Evidences from review of literature revealed inadequate empirical studies directly or indirectly assessing item and group invariance based on how the item and person statistics behave differently. The purpose of the study therefore, was to the compare IRT and CTT in terms of item and person parameters. Specifically, the study was to determine the group invariance across independent samples of gender and ability groups of examinees and item invariance across different samples of items of two forms of test based on the two measurement frameworks.

Hypotheses

The following hypotheses were tested at 0.05 alpha levels

1. The item difficulty estimates based on IRT do not significantly vary across independent samples of examinees in terms of gender and ability groups.
2. The item difficulty estimates based on CTT do not significantly vary across independent samples of examinees in terms of gender and ability groups.
3. The ability estimates of examinees based on IRT do not significantly vary across different samples of items.
4. The ability estimates of examinees based of CTT do not significantly vary across different samples of items.

Methods

Design of the Study: This study made use of descriptive survey research design in which information was captured from a representative sample of a population and inferences so generated were generalized over the entire population.

Participants: The population of the study consisted of all the students in SS3 in 2013/2014 session in public Senior Secondary Schools in Abia State. There are 216 senior secondary schools with approximate population of 11,666 students. Multi-stage sampling technique was adopted to draw the sample for the study. Firstly, simple random was used to select two education zones, Umuahia and Aba, from the three existing education zones of Abia State. These zones had four and nine Local Government Areas (LGAs) respectively. Secondly, two and four LGAs were randomly sampled from

Umuahia and Aba education zones respectively. The LGAs were Umuahia South and Ikwuano from Umuahia zone and Aba North, Aba South, Osisioma and Obingwa from Aba zone. Thirdly, The LGAs were further stratified by schools and 17 schools were randomly drawn from Umuahia zone while 33 were selected from Aba zone, giving a total of 50 schools. Finally, purposive sampling technique was used to draw SS3 students offering chemistry from the fifty schools. These made up a sample size of 1,150 students. Out of this sample, 635 were males while 515 were females.

Instrument: The instrument for data collection consisted of two parallel Chemistry Achievement Tests (CAT A & B) developed by the researchers following due process of test construction. These tests are 4-option multiple choice tests with 60 items each. Copies of CAT A and CAT B and marking guides were validated by three subject specialists and three experts in Measurement and Evaluation of the Department of Educational Foundations, Guidance and Counselling, University of Uyo, Akwa Ibom State. The corrections and suggestions were taken into consideration and integrated into the final drafts of the test. Kuder-Richardson formula 20 reliability estimates of CAT A and CAT B yielded coefficients of 0.88 and 0.90 respectively. The reliability coefficients were high enough to consider the instruments as reliable.

Procedure for Data Collection: The researchers sought the permission of the principals and the assistance of the chemistry teachers of the sampled schools. The dates for administering the two CAT test forms were announced seven days to first testing. This was to enable the students prepare for the test. The aim of the test was explained to the students as well as the description of the CAT test forms. The students were given identification numbers which were also used to number the test booklets. This was necessary for easy matching of the test for each examinee after testing. The teachers at the schools assisted in administering the test at the time that was convenient to the school. West African Senior School Certificate Examination (WASSCE) allocates one hour for fifty multiple choice chemistry items, therefore, an hour thirty minutes was given to the students to ensure that they attempted the items to the best of their ability. CAT B was administered two weeks after; almost under the same testing conditions. The responses to the test were retrieved and organized for scoring. CAT A and B were dichotomously scored. Items correctly responded to were scored 1 while 0 was given to wrong responses. The score per item per respondent was obtained.

Data Analyses: Maximum Likelihood Estimation Technique of BILOG MG V 3.0 procedures were used for 3-parameter model to estimate item difficulty parameter and the ability estimate for IRT framework; CTT item difficulty estimates was calculated as the proportion of correct response by the examinees and the ability estimate calculated as the total number of item scored correctly. The research questions were answered, the independent and paired t-tests scores were used for testing hypotheses on the item

difficulty parameter and ability estimates respectively. The reason for employing t-test statistics in the present study was that correlation statistics used by earlier researchers in testing for invariance have been criticized as not good enough and insufficient for the purpose of testing for invariance (Rupp & Zumbo, 2004 cited in Adedoyin, Nenty and Chilisa, 2008).

Results

The descriptive information of participants is presented in Table 1.

Table 1: *Demographic Information of participants*

Variables	Categories	Sample size (N)	Percentage
Gender	Male	635	55.2
	Female	515	44.3
CAT A			
IRT	High ability	574	49.9
	Low ability	576	50.1
CTT	High ability	898	78.1
	Low ability	252	21.9
CAT B			
IRT	High ability	548	47.7
	Low ability	602	52.3
CTT	High ability	618	53.7
	Low ability	532	46.3

Table 1 showed the different independent samples with their sample sizes and percentages. These independent samples for the frameworks, IRT and CTT, were based on gender and ability levels of students in CAT A and CAT B respectively. For IRT, participants with positive values on the ability scale were regarded as having high ability whereas those with negative value on the ability scale were regarded as having low ability. Conversely, for CTT percentage of the sum of scores obtained on the items in the test was used to determine the ability groups. Participants with 50% above were regarded as having high ability whereas those with scores eequal to or less than 40% were regarded as having low ability.

Hypothesis 1: The item difficulty estimates based on IRT do not significantly vary across independent samples of examinees in terms of gender and ability groups.

Table 2: *Independent t -test Analysis of Group Invariance based on IRT across Independent Samples*

Variables	N	Mean	Std. Deviation	Df	t-cal.	t-crit.	Decision
CAT A							
Male	635	- 3.97	101.24				Not
				1148	0.16*	1.96	Significant
Female	515	- 4.97	112.41				
High ability	574	0.76	0.56				Not
				1148	1.33*	1.96	Significant
Low ability	576	5.14	106.20				
CAT B							
Male	635	1.13	20.90				Not
				1148	1.18*	1.96	Significant
Female	515	.049	0.93				
High ability	548	0.86	0.56				Not
				1148	1.31*	1.96	Significant
Low ability	602	- 4.96	- 103.96				

*Not Significant at $p \geq 0.05$

Table 2 presented the results of the analysis for testing hypothesis one on the item difficulty parameter estimates of two test forms (CAT A and CAT B) based on IRT framework across different independent groups. The independent groups are gender and the ability groups. The result revealed that for all the independent groups the differences are not significant, which means that the item difficulty parameter estimates based on IRT framework are invariant across gender and ability groups. That is, regardless of the groups or sample of examinees used, the estimation of IRT item difficulty will always be the same value.

Hypothesis 2: The item difficulty estimates based on CTT do not significantly vary across independent samples of examinees in terms of gender and ability groups.

Table 3: *Independent t -test Analysis of Group Invariance based on CTT across Independent Samples*

Variables	N	Mean	Std. Deviation	Df	t-cal.	t-crit.	Decision
CAT A							
Male	635	29.50	5.85				
				1148	31.98*	1.96	Significant
Female	515	39.86	4.93				
High ability	898	37.43	5.13				
				1148	36.60*	1.96	Significant
Low ability	252	24.29	4.70				
CAT B							
Male	635	34.63	7.91				
				1148	9.23*	1.96	Significant
Female	515	30.18	8.86				
High ability	618	36.41	4.08				
				1148	54.30*	1.96	Significant
Low ability	532	23.01	4.28				

*Significant at $p \leq 0.05$

Table 3 presented the results of the analysis for testing hypothesis two on the item difficulty parameter estimates on two parallel tests (CAT A and CAT B) based on CTT framework across gender and the ability groups. The results indicated that for all the independent groups and test forms the differences are significant, which means that the item difficulty parameter estimates based on CTT framework are variant across gender and ability groups, that means that the estimates are sample dependent. This implied that the item difficulty parameter values vary with varying samples.

Hypothesis 3: The ability estimates of examinees based of IRT do not significantly vary across different samples of items.

Table 4: *Paired t-test Analysis of Item Invariance based on IRT across Different Samples of Items*

Variables	N	Mean	Std. Deviation	Df	t-cal.	t-crit.	Decision
CAT A	60	0.23	0.48				Not
				59	1.20	1.99	Significant
CAT B	60	0.84	7.02				

*Not Significant at p $\geq$ 0.05

Table 4 presented the results of the analysis for testing hypothesis three on the ability estimates of examinees based on IRT framework across two samples of tests: CAT A and CAT B. The result showed that the difference between ability estimation on the two forms of test was not statistically significant. This implied that the value of ability estimation does not vary irrespective of the sample of items used.

Hypothesis 4: The ability estimates of examinees based of CTT do not significantly vary across different samples of items.

Table 5: *Paired t-test Analysis of Item Invariance based on CTT across Different Samples of Items*

Variables	N	Mean	Std. Deviation	Df	t-cal.	t-crit.	Decision
CAT A	60	0.58	0.16				
				59	3.53	1.99	Significant
CAT B	60	0.51	0.20				

*Significant at p = 0.05

Table 5 presented the results of the analysis for testing hypothesis four on the ability estimates of examinees based on CTT framework across two samples of tests: CAT A and CAT B. The result indicated a statistically significant difference between the ability estimates on the two forms. This implied that the value of ability estimates vary with respect to the sample of items used. It further means that the abilities of the examinees are dependent on the sample of items and difficulty level of the items used for estimation.

Discussion

The main purpose of this study was to determine the group invariance and item invariance of independent groups of examinees and item difficulty parameter across different samples of items respectively.

The results in Table 2 showed that there was no significant difference between the independent groups: male and female, high ability and low ability, of the examinees based on IRT framework. The finding implied that difficulty parameter estimates based on IRT framework are invariant across the different independent groups of gender and ability. This implies the IRT item difficulty estimate do not depend on the sample or group used to estimate the parameter. The study also revealed in Table 4, based on IRT framework, a no significant difference between ability estimation on the two forms of test. These results indicated that IRT framework presents ability and item parameter estimates that are independent of the item difficulty across different sub-sets of items.

Table 3 and Table 5, based on CTT framework, and indicated that there was a significant difference between the independent groups: male and female, high ability and low ability, of the examinees; and ability estimation on the two forms of test respectively. The implication of these results was that ability estimate of examinees dependent on the item difficulty across different sub-sets of items while item difficulty parameter estimates were dependent on the different sub-groups of examinees. The significant differences observed suggested lack of invariance. This may be that that test items were gender biased.

Whereas these findings for IRT and CTT were consistent with those of Progan and Socan (2008); Adeboyin, Nenty and Chilisa (2008); Magno (2009) who in their separate studies upheld the principle of invariance, they negated those of Stage (2003); Courville (2004); Wiberg (2004) who established comparability for the two frameworks. The negation or difference could be as a result of the IRT parameter adopted or statistical tool used as the use of correlation has been criticized as insufficient for testing invariance.

Conclusion

In the light of the findings, it could be seen that IRT framework showed the invariance property of both item difficulty and person parameter estimates while CTT could not produce estimates that are sample and person independent. Therefore, it is concluded that IRT models is superior to CTT models.

Recommendations

Based on the findings it was recommended that:

1. IRT models, having proved superior to CTT models, should be adopted by psychometric community in test development and measurement practices for more objective measurement.

2. Test developers should aim at constructing items based on IRT guidelines so that the test will meet the acclaimed invariance property of IRT.
3. Researchers and stakeholders in testing should organize workshops, seminars and conferences in order to help measurement community catch-up with this emerging measurement framework, IRT.

REFERENCES

Adedoyin, O. O., Nenty, H. J., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on classical test and item response theories. *Educational Research and Review*, 3(2), 83 – 93. Retrieved on September 17, 2016 from http://www.academicjournals.org/ERR.

Amajuoyi, I. J. (2015). *Construction and validation of a diagnostic chemistry achievement test for senior secondary three students using classical test and item response theories*. An Unpublished Doctoral Thesis. University of Uyo, Akwa Ibom State.

An, X. & Yung, Y. (2014). Item response theory: What it is and how you can use the IRT procedures to apply it. SAS Institute Inc. Retrieved on September 17, 2016 from https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Baker, F. B. (2001). *The basics of item response theory.* (2nd ed.), USA: ERIC Clearinghouse on Assessment and Evaluation.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24 (1), 1 – 21. Available at http://dx.doi.org/10.1080/08957347.2011.532417.20. Retrieved December, 2012.

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics.* Unpublished Doctoral Thesis. Texas: A & M University. Retrieved on July 19, 2016 http://oaktrust.library.tamu.edu/bitstream/handle/1969.1/1064/etd-tamu-2004B-EPSY-Courville-2.pdf?sequence=1

deAyala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford press.

Duong, M. (2004). Introduction to item response theory and its applications. *CEP900: Proseminar in Learning, Technology and Culture*. Retrieved on September 12, 2016 *https://msu.edu/~dwong/StudentWorkArchive/CEP900F04-RDP/Mihn-ItemResponseTheory.htm*

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Izard, J. (2005). Trial testing and item analysis on test construction. In: Ross, K. N. (Ed) *Quantitative research methods in educational planning.* UNESCO International

Institute for Educational Planning. Retrieved on December 20, 2012 from http://www.unesco.org/riep..

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 Association,* 1(1), 1 – 11.

Metibemu, M. A. & Oluwatayo, G. K. (2016). Comparison of classical test and item response theories in the scoring of physics achievement test. A Paper Presented at *the 18th Annual Conference of Association of Educational Researchers and Evaluators of Nigeria (ASSEREN)* held at Imo State University, Owerri.

Morales, M. A. (2009). Evaluation of Mathematics Achievement Test: A Comparison Between CTT and IRT. *The International Journal of Education and Psychological Assessment,* 1 (1): 19 – 26.

Nenty, H. J. (2004). From classical test theory (CTT) to item response theory (IRT): An introduction to a desirable transition. In: Afemikhe, O.. and Adewale, J. g. (Eds.), *Issues in educational measurement and evaluation in Nigeria (In Honor of Wole Falayejo),* pp. 37 – 384, Ibadan: Institute of Education, University of Ibadan, Nigeria.

Partchev I. (2004). *A visual guide to item response theory*. Jena: Frederick – Schiller – Universitat. Retrieved on December 20, 2012 from https://www.metheval.uni-jena.de/irt/VisualIRT.pdf.

Progar, S. & Socan, G. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology,* 17(3), 05 – 24. Retrieved on July 17, 2016 from http://psiholoska-obzorja.si/arhiv_clanki/2008_3/progar_socan.pdf.

Stage, C. (2003).*Classical test theory or item response theory: The Swedish experience*. Santiago Chile: Centro de Estudies Publicos. Retrieved on July 17, 2016 from http://www.sprak.umu.se/digitalAssets/59/59524_em-no-42.pdf.

Suen, H. K. (1990). *Principles of Test Theories.* New Jersey: Lawrence Erlbane Association Publishers.

Tractenberg, R. E. (2010). Classical and modern measurement theories: Patient reports and clinical outcomes. *Contemp Clin Trails,* 31(1), 1 – 3. Retrieved on September 12, 2016 from https://www.ncbi.nlm.nih.gov/pubmed/20129315

University Testing Services; (2000). Academic Testing: Item Response Theory Approach. PennState: Schreyer Institute for Teaching Excellence. Retrieved on December 20, 2012 from http://www.schreyerinstitute.psu.edu.

Wiberg, M. (2004). Classical Test Theory vs. Item Response Theory: An evaluation of the theory test in the Swedish driving-license test. Retrieved on July 19, 2016 from http://hbanaszak.mjr.uw.edu.pl/TempTxt/Wiberg_2004_CTTvsIRTSwedishDrivingLicenceTest.pdf