A Comparison of Classical and Item Response Theory Person/Item Parameters of Physics Achievement Test for Technical Schools

Mfonobong Umobong (PhD) ***and*** **Sunday Sunday Jacob**
Department of Educational Foundations,
University of Uyo, Uyo
Nigeria.
mfonobongumobong@uniuyo.edu.ng and sundayjacobs@yahoo.com

+2348035079349 and +2348065169183

ABSTRACT

The study was carried out to compare classical test and item response theory based person/item parameters of technical schools' Physics Achievement Test(PAT). The survey research design was used for the study. The population of this study comprised 1553 senior technical two (ST2) students from six technical schools in Akwa Ibom State from where a stratified random technique was used to select a sample of 1000 students. The instrument 'Physics Achievement Test (PAT)' was used for data collection with a reliability coefficient of 0.74 obtained using Kuder-Richardson 20 formula. BILOG MG was used to estimate the item and person parameters of classical test and item response theory, after which the item and person parameters generated from the two measurement frameworks were compared. The findings of the study showed that the two measurement frameworks produced very similar item and person statistics for the PAT. Based on these findings, it was concluded that the classical test and item response theory person and item parameters of technical school physics achievement test are comparable and can be used interchangeably.

KEY WORDS: Classical Test Theory, Item Response Theory, Physics Achievement Test, Item parameters, Person Parameters

Introduction

The fact that no country can advance in technology without a sound science background has led to the increased teaching and learning of science globally. This probably made the National Policy on Education (2004) to advocate for improvements in the teaching and learning of Science and technology in order to create the foundation for technologically oriented workforce in line with the needs of national development. Within the context of science education, Physics has been identified as a very important science subject and its importance in scientific and technological development of any nation has been widely reported (Adesoji and Olatunbosun, 2008). Physics makes significant contributions through advances in new technologies that arise from theoretical breakthroughs. For example, advances in the understanding of electromagnetism or nuclear physics led directly to the development of new products which have dramatically transformed modern-day society, such as television, computers, domestic appliances, and nuclear weapons; advances in thermodynamics led to the developments in industrialization; and advances in mechanics inspired the development of calculus.

In spite of the importance of Physics, students' performance in Physics in both internal and external examinations has not been impressive over the years (Effiong, 2013). A close observation of students' performances in Physics in NABTEB examinations and Akwa Ibom State central promotion examination revealed that majority of the students fails to record a credit pass in Physics. This trend has always generated concern among scholars, parents, educators, scientists and the government and could be blamed on instructional method, instructional materials or assessment techniques (Jegede, 2012). This unfortunate development could be attributed to the measurement theories used for generating item and person parameters.

The word 'measurement' is most often associated with the application of one kind of instrument or the other to gauge the quantity but rarely the quality of something possessed by the body being measured. If what is to be measured is visible, touchable and could be measured by having the measuring instrument to make some form of contact with, then the measurement is said to be physical; and being physical, it is objective, in order words, it is error free (Nenty, 1998). On the other hand, measurement of behavioural characteristics, because of their indirect nature, is inferential and as such error prone. This is because what is observed during the measurement process is used to predict, infer or estimate what is being looked for. Hence, in educational measurement, unlike in physical measurement, there is need for a theory of measurement to provide some guide and direction for measurement to estimate a given trait level possessed by the object based on such measurement.

Test theories and related models are important to the practice of educational and psychological measurement because they provide a framework for considering issues and addressing technical problems. One of the most important issues is the handling of measurement errors. A good theory or model can help in understanding the role that measurement errors play in estimating examinees ability and how the contributions of error might be minimized (Hambleton and Jones ,1993).Lord and Novicks (1968) asserted that since the scores that result from measurement efforts, to some varying degree, are not errorless in representing the trait levels of individuals being measured, there is need to be interested in something other than such scores in order to predict these trait levels more validly. According to Nenty (2004) this cannot be done without the guidance of an operationalizable theory or model and two of such theories are widely used with their accompanying models: the classical test theory (CTT) and the item response theory (IRT).

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 scores from educational tests or examinations. The classical test theory and the item response theory have been theoretically and technologically developed in analyzing or standardizing tests, examinations within measurement frameworks (Adedoyin, 2010). Sharkness and DeAngelo (2011) asserted that classical test theory and item response theory are the two primary measurement theories that researchers employ to construct measures of latent traits. Due to the fact that latent traits are by their very nature unobservable, researchers must measure them indirectly through a test, task, or survey. The reason unobservable traits can be assessed in such a way is because the traits are assumed to influence the way that people respond to test or survey questions. While no perfect measure of a latent variable can ever exist, by examining how a person responds to a set of items relating to a single underlying dimension, researchers can create scores that approximate a person’s ‘‘level’’ of the latent trait. Classical test theory and item response theory are both tools that can be used to do this, but beyond their common purpose the two measurement systems are quite dissimilar.

Eluwa, Akubuike and Bekom (2011) observed that classical test theory (CTT) and item response theory (IRT) are commonly perceived as representing two very different measurement frameworks. Although CTT has been used for most of the time by the measurement community, in recent years IRT has been gaining ground, thereby becoming a favorite measurement framework. Warm (1978) stated that most testing practitioners use classical test theory, whether they know it or not as the basic tools in use are the p and d-values, mean, standard deviation of examinees scores, skewness, kurtosis and reliability of the test. Warm (1978) further stated that the problem with these statistics is that they are relative to the characteristics of the test and of the examinees.

Similarly, Hambleton and Jones (1993) remarked that most of the work in classical test theory has focused on models at the test-score level (in contrast to item response theory). That is, the models have linked test scores to true scores rather than item scores to true scores. Classical test theory does not invoke a complex theoretical model to relate an examinee's ability to succeed on a particular item, instead, it collectively considers a pool of responses of examinees on an item. Item response theory (IRT) has become an important complement to classical test theory in design, interpretation and evaluation of tests or examinations. According to Ojerinde, Popoola, Ojo, and Onyeneho (2012), IRT is generally regarded as an improvement over CTT. There is nothing CTT can accomplish that IRT cannot. The converse is however not true. For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information. Some applications such as computer adaptive testing are enabled more easily by IRT and may not be performed with CTT. Anastasi and Urbina (2002) noted that another fundamental feature of item response theory can help to produce a test that has the desired precision of measurement at the defined ability level. With respect to test scoring too, the item response theory-based methods are more useful compared to classical test theory, that is, they offer considerable advantages over the "number right" scoring methods typically used in classical test theory -based tests. The major argument against classical test theory is its rather weak theoretical assumptions which makes it easy to apply in many testing situations (Umobong, 2004, Adedoyin, 2010).

In comparing classical test theory (CTT) to item response theory (IRT) empirically, Fan (1998) sought for "appreciable differences" between the item and person statistics produced with each model along with evidence of the invariance that is claimed by the item response theory. The study empirically examined the behaviors of the item and person statistics derived from these two measurement frameworks. Using two multiple choice subtests from the 1992 Texas assessment of academic skills test data (with 193000 participants), he devised an extensive sampling plan with 40 random samples, 80 different gender samples and 80 high-low ability level samples – each including 1000 response vectors. Ability estimates from each of the IRT models were reported as highly correlated (all > 0.96) with the CTT ability estimates. CTT item difficulty estimates were reported as correlating extremely well with the 1PL/Rasch estimates (all >0.998) and moderately well with the 2PL and 3PL estimates (0.830 < r < 0.957). Correlations between CTT point biserials and 2PL and 3PL discrimination parameters were not as strong or consistent.

Adedoyin and Adedoyin (2013) conducted a study to assessed the comparability of test items parameter estimates between Classical test theory (CTT) and Item response theory (IRT) models using ten thousand (10,000) students who sat for the 2010 Botswana Junior Certificate (JC) mathematics paper 1 test. Pearson correlation coefficients were used to determine if the IRT and CTT test items parameter estimates were comparable and dependent t-test was used to find out whether the relationship between IRT and CTT test items parameter estimates were statistically significant. From the result of the analysis, it was found 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.

Eluwa, Akubuike and Bekom (2011) applied the classical test theory and item response theory to evaluate the quality of items on the National Certificate of Education (NCE) students' achievement in Mathematics for 80 students. Data was analyzed in two dimensions. First, the psychometric properties of the instrument were analyzed using classical test theory and item response theory 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 where found to be "bad items" in CTT came out not fitting also in the Rasch Model.

Although the assumptions and mathematics of item response theories are more complex, costly and time consuming, several authors such as Morales (2009) and Kaplan and Saccuzzo (2005) have argued that their empirical benefits are sufficient to warrant their usage. Moreover, a closer examination of literature reveals that there is no empirical study on the comparative analysis of classical test and item response theories in National Business and Technical Examination Board (NABTEB) examinations and with the persistent poor performances obtained by candidates in Physics in both central promotion examinations and NABTEB examination, the researchers were therefore prompted to empirically compare classical test and item response theory person/item parameters for technical schools physics achievement test.

Purpose of the Study

The purpose of this study was to carry out a comparative analysis of classical test and item response theory based person/item parameters of technical schools Physics Achievement Test in Akwa Ibom State. Specifically, the study sought to achieve the following objectives:

1. To compare classical test and item response theory based item difficulty parameters
2. To compare classical test and item response theory based item discrimination parameters
3. To compare classical test and item response theory based person parameters

Research Questions

The following research questions were formulated to direct the study:

1. How comparable are classical test and item response theory based item difficulty parameters?
2. How comparable are classical test and item response theory based item discrimination parameters?
3. How comparable are classical test and item response theory based person parameters?

METHODS

Design of the Study

The research design adopted for the study was the descriptive survey research design. This design was considered appropriate because it allowed the researchers to collect data from few individuals known as sample and generalize same to the entire population. This design is suitable for situations that involve objective description of existing phenomena. Hence, the reason it was being adopted for the study.

Population and Sampling

The population of the study comprised 1553 senior technical two (ST2) Students from six technical schools in Akwa Ibom State. The stratified random sampling technique was used to draw a sample size of 1000 students representing 64.5 % of the population.

Instrument for Data Collection

The researcher-developed instrument known as Physics Achievement Test (PAT) was used for the study. The instrument which initially consisted of 100 items in a multiple choice format covering the various aspects of Physics was developed by the researchers. After trial testing, the items with undesirable qualities were deleted, and some were reframed. The items in the instruments covered mechanics, waves, optics, heat and sound wave. Finally, a total of fifty (50) items were used for the study.

Reliability and validity

Content validity was used for validating this instrument. The procedure involved using test blue print as a guide in the development of the items. From each of the contents, 10 items were developed while 15, 10, 10, 5, 5, and 5 items were developed from the comprehension, knowledge, application, analysis, synthesis and evaluation respectively. A reliability coefficient of 0.74 was obtained for the instrument using the Kuder-Richardson 20 formula.

Administration of the Instrument

The instrument was administered by the researchers with the help of Physics teachers in all the schools visited. At the end, a total of one thousand (1000) copies of the instrument were administered and retrieved.

Method of Data Analysis

BILOG MG which is a statistical package for analyzing item and person parameters was used to estimate the item and person parameters of both the Item Response Theory item and the classical test theory. The person/item parameters of the two measurement framework were thereafter compared using Pearson Product Moment Correlation.

RESULTS

Research Question One

How comparable are the item difficulty parameters in physics generated by classical test and item response theory?

In order to answer this question, item difficulty parameters of classical test and item response theories were compared. Using the decision rule of Williams (2013),who stated that under classical test theory, items whose difficulty index falls between (0-3.99), (0.40-0.599) and (0.60-0.99) are regarded as difficult, moderate and easy items respectively while under item response theory, items with indices ranging from (1.0 and above), (0.0-0.99) and (less than 0.0) are regarded as difficult, moderate and easy respectively. The items and their respective difficulty parameters as well as the result of the analysis are as presented in Table 1:

Table 1: Comparison of Item Difficulty Indices of Classical and Item Response Theories

ITEMS	P-value	CTT	b-parameter	IRT
1	0.719	Easy	-0.635	Easy
2	0.672	Easy	-0.84	Easy
3	0.391	Difficult	4.608	Difficult
4	0.688	Easy	-1.154	Easy
5	0.764	Easy	-1.971	Easy
6	0.385	Difficult	3.732	Difficult
7	0.451	Moderate	2.554	Difficult
8	0.654	Easy	-0.774	Easy
9	0.804	Easy	0.737	Moderate
10	0.545	Moderate	-0.144	Easy
11	0.591	Moderate	-0.412	Easy
12	0.36	Difficult	3.353	Difficult
13	0.61	Easy	-0.641	Easy
14	0.553	Moderate	0.398	Moderate
15	0.492	Moderate	0.401	Moderate
16	0.521	Moderate	0.302	moderate

17	0.502	Moderate	0.355	Moderate
18	0.549	Moderate	0.075	Moderate
19	0.496	Moderate	0.601	Moderate
20	0.38	Difficult	2.077	Difficult
21	0.506	Moderate	0.375	Moderate
22	0.55	Moderate	0.393	Moderate
23	0.476	Moderate	0.439	Moderate
24	0.20	Difficult	2.438	Difficult
25	0.283	Difficult	1.004	Difficult
26	0.27	Difficult	1.663	Difficult
27	0.41	Moderate	-2.024	Easy
28	0.367	Difficult	1.852	Difficult
29	0.289	Difficult	1.898	Difficult
30	0.518	Easy	-1.732	Easy
31	0.498	Moderate	0.022	Moderate
32	0.523	Moderate	0.27	Moderate
33	0.588	Moderate	-1.5	Easy
34	0.313	Difficult	0.59	Moderate
35	0.298	Difficult	1.999	Moderate
36	0.376	Difficult	1.093	Difficult
37	0.314	Difficult	1.68	Difficult
38	0.322	Difficult	1.658	Difficult
39	0.39	Difficult	0.528	Moderate
40	0.538	Moderate	0.193	Moderate

41	0.447	Moderate	0.949	Moderate
42	0.293	Difficult	2.305	Difficult
43	0.497	Moderate	0.032	Moderate
44	0.324	Difficult	3.531	Difficult
45	0.567	Moderate	-0.315	Moderate
46	0.48	Moderate	0.377	Moderate
47	0.668	Easy	-0.799	Easy
48	0.505	Moderate	-0.01	Easy
49	0.577	Moderate	0.264	Moderate
50	0.477	Moderate	0.111	Moderate

Result in Table 1 shows that under classical test theory, nine out of 50 items were regarded as easy, 24 items moderate and 17 items difficult since their difficulty indices were within (0.60 – 0.99), (0.40 – 0.599) and (0 – 3.99) respectively. The result further shows that under item response theory, 13 items were regarded as easy, 22 moderate and 15 were regarded as difficult since their difficulty indices were within (less than 0.0), (0.0 – 0.99) and (1.0 and above) respectively. The table also revealed that 79.5% of the items that were easy, moderate and difficult under classical test theory were also easy, moderate and difficult respectively under item response theory. This finding implies that item difficulty parameters under the two measurement frameworks are very comparable.

To further compare item difficulty parameters generated from the two measurement frameworks, Pearson Product Moment Correlation was used for correlating the two difficulty parameters of classical and item response theories where X and Y represented P-value and b-parameters respectively. The result of the analysis is as presented in Table 2:

Table 2: Pearson Product Moment Correlation analysis of item difficulty parameters of classical test and item response theory

(n = 50)

Variables	$\sum X$	$\sum X^2$	$\sum XY$	r-value
	$\sum Y$	$\sum Y^2$		
Classical Test Theory	23.89	12.23		
			8.17	-0.77*
Item Response Theory	31.91	123.94		

**Significant at .05 level; df = 49; Critical r value is .288.*

The result of the analysis of correlating item difficulty indices (p-value) generated from classical test theory and item difficulty parameters (b-parameter) generated from the item response theory shows a high coefficient of -0.77 which indicated that the correlation was statistically significant. This result implies that the item difficulty estimates from the two measurement frameworks are comparable. This means that the two theoretical frameworks could be used interchangeably to generate item difficulty parameters.

Research Question Two

How comparable are the item discrimination parameters in physics generated by classical test and item response theory?

In order to answer this question, item discrimination parameters of classical test and item response theories were compared using the decision rule of Williams (2013) that under classical test theory, items with discrimination indices that falls between (0-0.29), (0.3-0.59) and (0.60-0.99) are regarded as low, moderate and high respectively while under item response theory, items whose discrimination parameters falls between(.01-.64), (.65-1.34) and (1.35 and above) are regarded as low, moderate and high respectively. The items and their respective discrimination parameters as well as the result of the analysis are as presented in Table 3:

Table 3: Comparison of the item discrimination indices of the classical and Item response theories

ITEMS	d-value	CTT	a-parameter	IRT
1	0.312	Moderate	0.512	Moderate
2	0.461	Moderate	0.601	Moderate
3	0.109	Low	0.059	Low
4	0.339	Moderate	0.45	Low
5	0.281	Low	0.38	Low
6	0.035	Low	0.076	Low
7	0.043	Low	0.077	Low
8	0.344	Moderate	0.571	Low
9	0.386	Moderate	1.195	Moderate
10	0.614	High	0.974	Moderate
11	0.333	Moderate	0.084	Low
12	0.138	Low	0.013	Low
13	0.254	Low	0.045	Low
14	0.533	Moderate	1.37	High
15	0.62	High	1.35	High
16	0.67	High	1.03	High
17	0.633	High	1.367	High
18	0.692	High	0.626	Moderate
19	0.454	Moderate	0.66	Moderate
20	0.182	Low	0.096	Low
21	0.609	High	1.369	High
22	0.43	Moderate	1.39	High
23	0.557	Moderate	0.626	Moderate
24	0.499	Moderate	0.257	Low
25	0.242	Low	0.114	Low

26	0.341	Moderate	0.064	Low
27	0.525	Moderate	0.635	Moderate
28	0.047	Low	0.067	Low
29	0.256	Low	0.014	Low
30	0.21	Low	0.425	Low
31	0.542	Moderate	0.892	Moderate
32	0.453	Moderate	1.584	Moderate
33	0.307	Moderate	0.465	Moderate
34	0.289	Low	0.684	Low
35	0.170	Low	0.408	Low
36	0.044	Low	0.075	Low
37	0.145	Low	0.104	Low
38	0.114	Low	0.122	Low
39	0.231	Low	0.133	Low
40	0.31	Moderate	0.806	Moderate
41	0.103	Low	0.138	Low
42	0.363	Moderate	0.651	Moderate
43	0.35	Moderate	0.679	Moderate
44	0.145	Low	0.126	Low
45	0.427	Moderate	0.596	Low
46	0.154	Low	0.134	Low
47	0.161	Low	0.31	Low
48	0.371	Moderate	0.684	Moderate
49	0.401	Moderate	0.685	Moderate
50	0.495	Moderate	0.732	Moderate

Results in Table 3 show that under classical test theory, 21 out of 50 items had low discrimination indices, 23 items had moderate indices and 6 items had high discrimination indices because their discrimination indices were within (0 – 2.99), (0.3 – 0.59) and (0.60 – 0.99) respectively. The result further shows that under item response theory, 27 items had low discrimination indices, 17 items had moderate indices and 6 items had high discrimination indices because their discrimination indices were within (.01 - .64), (.65 – 1.34) and (1.35 and above) respectively. The table also revealed that 76.4% of the items that were low, moderate and high under classical test theory were also low, moderate and high respectively under item response theory. This finding implies that item discrimination parameters under the two measurement frameworks are very comparable.

To further compare the item discrimination parameters in Physics generated by classical test and item response theory, Pearson Product Moment Correlation was used for correlating the item discrimination parameters generated from the two measurement frameworks where X and Y represented d-value and a-parameters respectively. The result of the analysis is as presented in Table 4:

Table 4: Pearson Product Moment Correlation analysis of item discrimination parameters of classical test and item response theory

Variables	**$\sum X$**	**$\sum X^2$**	**$\sum XY$**	**r-value**
	$\sum Y$	$\sum Y^2$		
Classical Test Theory	16.72	7.22		
			11.91	0.77*
Item Response Theory	26.51	23.75		

*Significant at .05 level; df = 49; Critical r value is.288

The result of the analysis of correlating item difficulty indices (d-value) generated from classical test theory and item difficulty (a-parameter) generated from item response theory shows a high coefficient of 0.77 which indicated that the correlation was statistically significant. This high coefficient shows that a correspondence exist between CTT–based discrimination indices and IRT–based discrimination parameters. This result implies that the item discrimination estimates from the two measurement frameworks are comparable. This means that the two measurement frameworks can be used interchangeably to generate item discrimination parameters.

Research Question Three

How comparable are the person parameters in physics generated by classical test and item response theory?

In order to answer this question, person parameters of classical test and item response theory were compared using the decision rule that under CTT, those who scored less than 20 were regarded as low ability students and those who scored 31 and above high ability students while under item response theory, those with theta values less than 0 were regarded as low ability students while those with theta values greater than 0.6 were high ability students. Out of the 1000 test takers, 325 test takers were regarded as low ability students while 192 were high ability students under CTT whereas under IRT, 395 were regarded as low ability students while 201 were regarded as high ability students.

To further compare person parameters from the two measurement frameworks, Pearson Product Moment Correlation Analysis was adopted where X and Y represented person parameters generated from CTT and IRT respectively. The result of the analysis is as presented in Table 5:

Table 5: Pearson Product Moment Correlation analysis of Person Parameters of Classical Test and Item Response Theory

Variables	$\sum \mathbf{X}$	$\sum \mathbf{X}^2$	$\sum \mathbf{XY}$	**r-value**
	$\sum \mathbf{Y}$	$\sum \mathbf{Y}^2$		
Classical Test Theory	23883	628225		
			6238.77	0.87*
Item Response Theory	0.20	875.42		

**Significant at .05 level; df = 999; Critical r value .195*

The result of the analysis of correlating person parameters generated from classical test theory and that generated from item response theory shows a high coefficient of 0.87 which indicated that the correlation was statistically significant. This high coefficient implies that a high correspondence exist between CTT–based person parameter and IRT–based person parameter. This result implies that the person parameters from the two measurement frameworks are comparable. Hence, the two measurement frameworks can be used interchangeably.

Discussion of Findings

The result of the analysis of the first research question revealed that the item difficulty estimates from the two measurement frameworks are comparable. This means that there is no much difference in the item difficulty parameter generated from the two measurement framework. This is because over 75% of the items that were easy, moderate and difficult in one measurement framework were also found to be the same in the other measurement framework. The result of this findings is in line with that of Eluwa, Akubuike and Bekom (2011) ,whose 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 where found to be “bad items” in CTT came out not fitting also in the Rasch Model. The findings of this study also agrees with that of Adedoyin and Adedoyin (2013), who found out that there was no statistical significant difference between the item difficulty parameter estimates of classical test theory and item response theory which means that the p-values of classical test theory and b-values of item response theory were comparable and they could both be used independently to estimate the test items parameters.

Result also revealed that the item discrimination estimates from the two measurement frameworks are comparable. This is because over 75% of the items that fell under low, moderate and high discrimination index under the classical test theory also fell under low, moderate and high discrimination index under the item response theory respectively. The result of this study agrees with that of Adedoyin and Adedoyin (2013), who found no statistical significant difference between the item discrimination parameter estimates by CTT and IRT which means that the d-values of CTT and the a-values of IRT were comparable and they could both be used independently to estimate the test item parameters. The result of this study is however in contrast with that of Fan (1998), who found that the relationship between CTT and IRT item discrimination indices is weaker. Furthermore, the relationship between CTT and IRT item discrimination indices shows considerable variation across tests, across sampling conditions, and across IRT models (two- vs. three-parameter IRT models). Fan (1998), further explained that the lower comparability between the discrimination indices derived from CTT and IRT implies that, in some cases, CTT and IRT may yield noticeable discrepancies with regard to which items have more discrimination power, which, in turn, may lead to the selection of different items for a test, depending on which framework is used in the estimation of item discrimination.

Result further revealed that the person parameter estimates from the two measurement frameworks are comparable. This therefore implies that majority of the people that the classical test theory framework classified as having low ability were also found to have low ability under the IRT framework. The findings of this study supports Fan (1998), who found out that the CTT- and IRT-based examinee ability estimates correlate extremely highly with each other, for different samples, and for all three (one-, two-, and three-parameter) IRT models, with average correlations between CTT- and IRT-based ability estimates greater than .96 for all conditions. These very high correlations indicated that CTT- and IRT-based person ability estimates are very comparable with each other. In other words, regardless of which measurement framework we rely on, the same or very similar conclusions will be drawn regarding the ability levels of individual examinees.

CONCLUSION

From the findings of the study that the item difficulty estimates, item discrimination parameters and the person parameters from the two measurement frameworks are comparable. It can be concluded that the classical test and item response theory person and item parameters of technical school physics achievement test are comparable and can be used interchangeably.

RECOMMENDATIONS

Based on the conclusion of the study, the following recommendations were made:

1. Government and school proprietors should organize workshops and seminars regularly for teachers on principles of test construction using the two measurement frameworks; this will equip teachers with the knowledge of how to use the two measurement frameworks to construct test items that can measure absolute ability.

2. The State Ministry of Education and other public examination bodies that are still using only one test theory for test construction and development, should use the two test theories simultaneously to ensure the development of objective instrument.
3. The two measurement frameworks should be taught simultaneously at undergraduate and post graduate levels for all the students in education and psychology and not only to measurement and evaluation students. This will help to equip the students with adequate knowledge of test development using the two measurement frameworks.

REFERENCES

Adedoyin, O. (2010). Investigating the Invariance of Person Parameter Estimates based on Classical Test and Item Response Theories. *International Journal of Education Science,* 2(2), 107-113.

Adedoyin, O.O and Adedoyin, J.A. (2013). Assessing the comparability between classical testtheory (CTT) and item response theory (IRT) models in estimating test item parameters. Herald Journal of Education and General Studies. 2 (3),107 – 114. Available at http://www.heraldjournals.org/hjegs/archive.htm. Retrieved August, 2013.

Adesoji, F. A. and Olatunbosun, S. (2008). Student, Teacher and School Environmental Factors as Determinants of Achievement in Senior Secondary School Physics in Oyo State, Nigeria. *The Journal of International Social Research.* 1 (2), 13-29.

Anastasi, A. and Urbina, S. (2002). *Psychological Testing*. New York: Prentice Hall.

Bertrand, R. and Blais, J. G. (2004). Modèles de mesure. L'apport de la théorie des Reponses aux items. Québec: Presses de l'Université du Québec. Bock, R. D., Thissen, D., and Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34, 197-211.

Effiong, U.G. (2013). Correlates of Students Achievement in Physics in Technical Schools in Akwa Ibom State. Unpublished Seminar Submitted to Faculty of Education, University of Port Harcourt, Rivers State Nigeria.

Eluwa, O. I., Akubuike, N. E. and Bekom, K. A. (2011). Evaluation of Mathematics Achievement Test: A Comparison Between Classical Test Theory (CTT) and Item Response Theory (IRT). Proceedings of the 2011 International Conference on Teaching, Learning and Change (c) International Association for Teaching and Learning (IATEL). *Journal of Education and Social Science Research,* 1 (4), 99-106.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Parameters. *Journal of Educational and Psychological Measurement*, 58, 357–381.

Federal Republic of Nigeria. (2004). National Policy on Education. Abuja: Government Printers.

Hambelton, R. and Jones, R. (1993). Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development. *Educational Measurement: Issues and Practice,* 12, 38-47

Hambleton, R. K (1989). Principles and Selected Applications of Item Response Theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., 147-200). New York: Macmillan

Jegede, S. A. (2012). Remediation of Students' Weaknesses for Enhanced Achievement in Physics. *Greener Journal of Educational Research,* 2 (4), 95-99.

Kaplan, R. M. and Sacuzzo, D. P. (2005). *Psychological testing; principles, applications and issues. 6th ed.* Belmont: Thomas Wadsworth.

Kpolovie, P. J. (2010). *Advanced Research Methods*. Owerri: Springfield Publishers.

Laveault, D. and Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (2nd edition). Brussels: De Boeck.

Lord, F. M., and Novicks, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

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 O. A. Afemikhe and J. G. Adewale (Ed.), *Issues in educational measurement and evaluation in Nigeria.* (in honour of Professor Wole Falayajo) (Chapter 33, pp. 371 -384). Institute of Education, University of Ibadan: Ibadan, Nigeria.

Nenty, H.J. (1998). Attributional Analysis of Mathematics Achievement Related Behaviour Among Secondary School Students in Lesotho. *BOLESWA Educational Research Journal*, 15, 1-12.

Ojerinde, D., Popoola, K., Ojo, F., Onyeneho, P. (2012). *Introduction to item response theory: Parameter models, estimation and application*. Abuja: Marvelouse Mike Press Ltd.

Umobong, M. E. (2004). Item Response Theory: Introducing Objectivity into Educational Measurement. In O. A. Afemikhe and J. C. Adewale (Eds.), *Issues in Educational Measurement and Evaluation in Nigeria* (in Honour of Professor Wole Falayajo) (384–398). Ibadan: Institute Education, University of Ibadan, Nigeria.

Warm, T. A. (1978). *A Primer of Item Response Theory*. Oklahoma City: United States Government Printing Office.

Williams, O.D. (2013). *Fundamentals of Classical Test and Item Response Theory*. New York: The Frankford Press.