EMPIRICAL COMPARISON OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY USING GEOGRAPHY ACHIEVEMENT TEST IN BENUE STATE

Yeke, C. M.
College of Agricultural and Science Education,
Joseph Sarwuan Tarka University Makurdi,
Benue State-Nigeria

Onah, D. O.
College of Agricultural and Science Education,
Joseph Sarwuan Tarka University Makurdi,
Benue State-Nigeria

Emaikwu, S. O.
College of Agricultural and Science Education,
Joseph Sarwuan Tarka University Makurdi,
Benue State-Nigeria

Obinne, A. D. E.
College of Agricultural and Science Education,
Joseph Sarwuan Tarka University Makurdi,
Benue State-Nigeria

Abstract

This study focused on Empirical Comparison of Classical Test Theory and Item Response Theory using Geography Achievement Test in Benue State. In carrying out the study, four research questions were posed and two hypotheses formulated. The study adopted Ex-post-facto research design. The sample size for the study was 581 secondary school Geography students representing 60% of the population for the study. This was drawn from 986 geography students randomly selected from146 secondary schools in Benue North-East Education Zone. A - 50 item Geography achievement test was adopted and used for data collection. The reliability coefficient of 0.93 was obtained using Kuder Richardson 20 formula (K-R_{20}). The instrument was administered to respondents with the aid of research assistants. The research questions posed were answered using percentages, difficulty and discrimination indices with the aid of Bilog-MG statistical package while the hypotheses formulated were tested using independent t-test statistic at 0.05 level of significance. The result revealed that, statistical significance exists between the CTT and IRT in item discrimination and item difficulty indices in favour of IRT. All the hypotheses tested were rejected as they were all statistically Significant in favour of IRT. Based on the findings it was concluded that the two-parameter logistic model can successfully be applied in the determination of item statistics and the items were good for measuring achievement of Geography students with little modification. It is therefore recommended among others that: The item difficulty and item discrimination estimates of students' achievement test in Geography based on CTT revealed that few of the items had poor difficulty and discrimination indices, the Item difficulty and item discrimination estimates of students' Geography achievement test (GAT) based on IRT revealed that, majority of the items have better difficulty and discrimination indices to measure students' ability than CTT.

Keywords: Empirical Comparison, Classical test theory, Item Response Theory, achievement test.

Introduction

In school system, instruments are developed in line with standard procedures and used to elicit information about the students. This instruments are further subjected to thorough analysis in order to establish the best indices in line with the relevant framework for standard achievement test development. According to Amadi (2012), standardised achievement test development involves putting together a number of questions in line with the syllabus and specific objectives anchored on a related theory. Also, it has to strictly follow standard procedures for developing an acceptable measurement instrument. These procedures according to Anikweze (2012) ought not to be violated otherwise, the objective of such instrument would be defeated. However, where several frameworks are involved and there are doubts on the validity of one, it becomes obvious to make a comparison in order to determine the most suitable framework that fits the purpose. So, the work "comparative analysis of geography achievement test based on Classical test theory and Item response theory" is one of the attempts to determine a more valid framework to be used in developing achievement test. This is achieved by using parameters of the two measurement frameworks, where item statistics are compared and for this study, the comparison was in terms of item difficulty and discrimination. Also because of the uncommon denominator inherent in the models, the students' achievement test analysis is compared.

Therefore, in this study attempt was made to compare classical test theory (CTT) and item response theory (IRT) with the aim of identifying the better framework to use in achievement test production, administration and scoring. The main goal of psychometricians and psychologists is to provide specific objectivity in measurement. This problem of objectivity has taken its root from measurement theories in which the examinees' characteristics and test characteristics are seen to be inseparable as one can not estimate the item parameter without using the number of examinee sample (Amadi,2012).

Theories are principles used in the development and application of standard rules that can be widely accepted (Ekwonye & Eguzo, 2011). However, they do not operate in isolation but on models. For example, Classical Test Theory operates on true score model while item response Theory operates on a number of parameters such as 1-PL, 2-PL and 3-PL which are central in the application of IRT. In this study, comparison of the two frameworks is made based on the 2-PL parameters using students' achievement derived from the geography achievement test in order to identify a more suitable framework that should be used by examination bodies. The revelation emerged from application of Classical Test Theory (CTT) and Item Response Theory (IRT) is that, it is possible with the two frameworks to produce achievement test by test developing institutions. In Classical test Theory (CTT), Statistical procedures are involved with detailed description of theoretical and mathematical characteristics of models and item indices are used to check quality of items put in the item-bank. Investigation revealed that in NECO, the most adopted practical procedure of test construction is based upon Classical Test Theory (CTT) and its concept of reliability (Onah & Amadi, 2017). Classical Test Theory dwells more on the reliability of psychological test and it is (CTT) defined as the body of related psychometric theories that predict outcomes of psychological testing such as enhancing essay comprehension and improvement of psychological test (Ekwonye & Eguzo, 2011). However,

the theory gives information only on test level which has the tendency to reduce its predictability, but findings revealed that, it is widely used by almost all the examination bodies in Nigeria in the development of item banks. According to Gregory (2011), Classical Test Theory models assume that, each person has a true score that would be obtained if there were no errors in measurement. A person's true score (T) is defined as the expected number and correct score over an infinite number of independent administrations of test. Unfortunately, test users never use a person's true score but only an observed score, defined as X=T± E.

In the other hand, Item Response Theory (IRT) according to Dibu et al (2012) is a family of statistical procedures for analyzing and describing test performance. It has three major characteristics that distinguish it from CTT. IRT refers to the family of latent trait models used to establish psychometric properties of items and scales. It is sometimes referred to as modern psychometrics because of its usage in large – scale assessment, testing programmes and professional testing firms. This is why IRT has almost replaced CTT (Kpolovie, 2010). IRT focuses on performance of individual items, rather than only on whole tests. It describes item performance at each level of student's ability; and it is model – based. The most common IRT model, called the one-parameter logistic model (1-PL) or the Rasch model assumes that, the probability of responding correctly or wrongly is a function of a person's ability and the difficulty of the item. The two parameter model (2-PL) and the three parameter model (3PL) are all models used in IRT for empirical item analyses. In a related submission, Kpolovie (2010) said IRT is a modern theory on development of test items that is anchored on the relationship between the individual examinee's latent psychological trait and his/her response to an item on a test which measures that specific attribute. This theory postulates that: (a) examinees test performance can be predicted and explained by a set of factors called trait, latent traits or abilities and (b) the relationship between the examinee item performances and these traits can be described by a monotonically increasing function called item characteristics function.

According to Dibu, et al, (2012), the task of assembling all components of the formulae, understanding and general application in item development are tedious and therefore constitute some of the drawbacks in the use of Item Response Theory (IRT). The authors added that, the drawbacks could be traced to the detailed nature of the theory which makes it difficult to be easily understood by many test developers unlike the Classical Test Theory (CTT). This might be responsible for the wide use of CTT in achievement test development among the examination bodies. An investigation in the examination bodies, namely: NECO, WAEC and NABTEB has shown that, almost all the test instrument used in Nigeria by these bodies during national examinations are developed using the CTT models even with the advent of IRT. Thus, this study compares basically the two frameworks using item parameters from Geography Achievement Test (GAT) adopted from NECO, which is the actual index of achievement in this case. The difficulty index according to Akinyele (2019) is the range of item parameters that determines the frequency count of the numbers of individuals choosing each option together with the number not answering the item at all. The difficulty index within the range of 0.30 – 0.70 which is widely considered moderate was adapted for the CTT in the study with $-0.2 \leq b \leq 0.2$ for IRT. While item discrimination index according to the author is the difference between the proportions of the candidates scoring an item correct in the upper group and those scoring the item correct in the lower group. Thus, for CTT, discrimination indices ranges from 0.25

– 1.00 and for IRT, any discrimination index that is greater or equal to 0.20 (i.e $a \geq 0.20$) is widely considered adequate. In a study on investigating the invariance of item difficulty parameter estimates based on CTT and IRT; Nenty (2008) opined that, item parameter estimates vary across the measurement frameworks but can be compared using both item indices and person achievement. So, there is the possibility of comparing parameters from the two frameworks using item parameters based on the students' achievement. Thus, the Geography Achievement Test (GAT) developed by NECO is believed to have the right item statistics to be the desired achievement test. However, because of the criticism and students' achievement in recent past, it has become obvious to compare the two measurement frameworks in order to identify which one is more suitable.

Thus, the two measurement frameworks might appear to be incomparable owing to lack of common denominator but at person parameter level, comparability might be possible because of the common interpretation associated with the tests' scores. Also, with the achievement test scores, items analysis is carried out using measurement framework and the indices of difficulty and discrimination are established from each of the frameworks which present the tendency for easy comparison. Therefore, indices from the two frameworks were compared to determine the good statistical parameters that could guide the test developers on the use of suitable framework. So, the scores are the bases for comparison in this study since the models of the frameworks are not the same.

Purpose of the study

The objective of this study is to compare Classical Test Theory (CTT) and Item Response Theory (IRT) in estimating test item parameters in Geography in North- East Education Zone of Benue State. Specifically, the study sought to ascertain;

i. The item difficulty and discrimination estimates of Geography Achievement Test based on CTT model
ii. The item difficulty and item discrimination estimate of Geography Achievement Test based on IRT model
iii. The statistical mean difference between the CTT and IRT item discrimination estimates of Achievement Test in Geography
iv. The statistical mean difference between the CTT and IRT item difficulty estimates of Achievement Test in Geography

Research Questions

The following research questions were formulated to guide the study.

i. What are the item difficulty and discrimination estimates of Achievement Test in Geography based on CTT model?
ii. What are the item difficulty and discrimination estimates of Achievement Test in Geography based on IRT model?
iii. What is the mean difference between CTT-based and IRT-based item discrimination estimates in Geography Achievement Test?
iv. What is the mean difference between CTT-based and IRT-based item difficulty estimates in Geography Achievement Test?

Hypotheses

The following research hypotheses guided the conduct of the study and were tested at .05 level of significance.

i. There is no statistical significant mean difference between the CTT and IRT models based on item discrimination estimates of Achievement Test in Geography.

ii. There is no statistical significant mean difference between the CTT and IRT models based on item difficulty estimates of Achievement Test in Geography.

Literature Reveiw

Application of CTT to Test Development: There are various stages in the construction and development of tests using CTT approach. These among others include: Preparation of Table of specification in conjunction with the subject syllabus and weighting, item writing by subject experts, test validation/ moderation of test items and test administration among others. Within a CTT framework, item statistics are examinee sample-dependent for CTT models. This means that test item statistics are very dependent on the sample of examinees used in item calibration. But it would facilitate test development if the item statistics were not directly tied to the choice of examinee sample (Embretson, 2006).

Application of Item Response Theory (IRT) in Test Development: Item response theory (IRT) is arguably one of the most influential developments in the field of educational and psychological measurement. IRT provides a foundation for statistical methods that are utilized in contexts such as test development, item analysis, equating, item banking, and computerized adaptive testing. Its applications also extend to the measurement of a variety of latent constructs in a variety of disciplines (Bradburn, 2009). The pendulum swing in test development techniques is from Classical Test Theory (CTT) to Item Response Theory (IRT).

According to Nworgu (2010), application of IRT in test development process is a recent trend which marks a departure from the traditional practice of basing test development on CTT. With IRT, items are calibrated without reference to the sample but in terms of the trait level or ability level of an individual referred to as theta (e) and item parameter estimates.

Research Methods

Research Design: The ex-post facto design was used for the study in collecting the data.

This designs were considered suitable for the study since event in the research had already taken place. Ex-post facto research according to Emaikwu (2015) is a systematic empirical inquiry in which the researcher does not have direct control of independent variables because their manifestations have already occurred or because they are inherently not manipulated. In the context of educational research, ex-post facto also known as 'after the fact' or 'retrospective' investigate possible cause-and-effect relationships by observing an existing condition or state of affairs and searching back in time for plausible causal factors.

The study was carried out in Benue North-East Education Zone, Nigeria.The zone is divided into three traditional districts, Kwande, Jeechira and Sankera. The population of the study is 968 Senior Secondary Three (SS 3) students who registered for Geography in the 2019/2020 academic session from the seven Local Government Areas in the zone. They

consisted of 503 males and 465 females Senior Secondary (SS 3) Geography students in Benue State

Instruments of Data Collection was one standardized geography achievement adapted from NECO past question papers. The instrument was administered on the SS III students. A total of 581 copies of the Geography Achievement Tests (GAT) were administered on students in the selected schools in Benue North – East education zone.

Data Analysis Techniques: BILOG-MG software was first used to compute the item parameters (item difficulty and discrimination indices) which were used to answer the research questions. Then results from the analysis were subjected to independent sample t-test. The use of independent t-test was based on the two independent groups involved in the comparison.

Results

Research Question 1: What are the item difficulty and discrimination estimates of Achievement Test in Geography based on CTT model? Answer to this question is presented in table 1.

Table 1: Summary of CTT Item Parameters for Geography Achievement Test (GAT)

Item parameters	N	Good items	%	Poor items	%	Total
Difficulty index	581	43	86.0	7	14.0	50
Discrimination index	581	45	90.0	5	10.0	50

Table 1 present the summary of item parameters for Geography achievement test (GAT) taken by 581 students. The Table revealed that, out of 50 items for CTT based model, 43 (86.0%) items have good difficulty index and 7(14%) items were poor, which means there were either too easy or too difficult for the test takers. Also, the discrimination index revealed that 45(90.0%) items discriminated well and 5(10.0%) items discriminated poorly. This shows that, the Geography achievement test (GAT) items did meet the standard of a good test as the bad items were not much and the percentage of poor discrimination index (10.0%) was also on the low side.

Research Question 2: What are the item difficulty and discrimination estimates of Achievement Test in Geography based on IRT model? Answer to this question is presented in table 2

Table 2: Summary of IRT Item Parameters for Geography Achievement Test (GAT)

Item parameters	N	Good items	%	Poor items	%	Total
Difficulty index	581	42	84.0	8	16.0	50
Discrimination index	581	47	94.0	3	6.0	50

Table 2 presents the summary of item parameters for Geography achievement test (GAT) taken by 581 students. The Table revealed that, out of 50 items for IRT based model, 42 (84.0%) items have good difficulty index and 8(16%) items were poor, which means there were either too easy or too difficult for respondents. Also, the discrimination index revealed that 47(94.0%) items discriminated well and 3(6.0%) items discriminated poorly. This shows that, the Geography achievement test (GAT) items did meet the standard of a good

test as the bad items were not much and the percentage of poor discrimination index (6.0%) was also low.

Research Question 3: What is the mean difference between CTT-based and IRT-based item discrimination estimates in Geography Achievement Test? Answer to this question is presented in table 5.

Table 3: Summary of CTT and IRT Discrimination index for Geography Achievement Test (GAT)

Item parameters	N	Good items	%	Poor items	%	Total
CTT	581	45	90.0	5	10.0	50
IRT	581	47	94.0	3	6.0	50

Table 3 present the summary of differences in CTT and IRT discrimination index for Geography Achievement Test (GAT) taken by 581 students. The Table revealed that, out of 50 items for CTT model, there were 45 (90.0%) good items compared to IRT model which has 47 (94.0%) good items because of their acceptable discrimination indecis ($r \geq 0.20$ and $a \geq 0.20$). CTT has the highest number of 5 bad items (5 items) or 10% compared to IRT 3 items or 6% with the difference of 4 poor items.

Research Question 4: What is the mean difference between CTT-based and IRT-based item difficulty estimates in Geography Achievement Test? Answer to this question is presented in table 4.

Table 4: Summary of CTT and IRT Difficulty index for Geography Achievement Test (GAT)

Item parameters	N	Good items	%	Poor items	%	Total
CTT	581	43	86.0	7	14	50
IRT	581	42	84.0	8	16	50

Table 4 presents the summary of item parameters for Geography achievement test (GAT) taken by 581 students. The Table revealed that for CTT, out of 50 items, 43 (86.0%) items have good difficulty index and 7(14%) items were poor, which means there were either too easy or too difficult for the test takers. The Table also revealed that for IRT, out of 50 items, 42 (84.0%) items have good difficulty index and 8(16%) items were poor, which means there were either too easy or too difficult for the examinees.

Hypothesis 0ne: There is no statistical significant mean difference between the CTT and IRT models based on item discrimination estimates of Achievement Test in Geography. The independent t- test of significance is presented in Table 5.

Table 5: Independent t-test of Significant Mean Differences between CTT and IRT Based Item Discrimination estimates

Parameters	N	Mean	Std	Df	t	P-value	A	Remark
CTT	50	.2005	.29004					
IRT	50	.7503	1.09204	98	7.235	.000	0.05	**Significant**
Total	100							

P<0.05

The result in Table 5 revealed independent t-test results of the mean difference between CTT and IRT Based on item discrimination estimates of students' responses to achievement test in Geography. The finding indicates a statistical significant mean difference between CTT and IRT Based on item discrimination estimates of students' responses to achievement test in Geography ($t = 7.235$, $df = 98$, $p = .000<0.05$). Thus, the hypothesis which states that, there is no statistically significant mean difference between the CTT and IRT based on item discrimination estimates of students' responses to Achievement Test in Geography is rejected. This implies that, there is statistically significant mean difference between CTT and IRT based on item discrimination estimates of students' responses to achievement test in geography in favour of IRT.

Hypothesis two: There is no statistical significant mean difference between the CTT and IRT models based on item difficulty estimates of Achievement Test in Geography. The independent t- test of significance is presented in Table 6.

Table 6: Independent t-test of Significance Mean Differences between CTT and IRT Based on Item Difficulty estimates

Parameters	N	Mean	Std	Df	T	P-value	A	Remark
CTT	50	.1224	.24806					
IRT	50	.3397	1.71013	98	3.002	.000	0.05	**Significant**
Total	100							

P<0.05

Table 6 revealed the independent t-test results of the mean difference between CTT and IRT based on item difficulty estimates of students' responses to geography achievement test (GAT). The result indicates a statistical significant mean difference between CTT and IRT Based on item difficulty estimates of students' responses to achievement test in geography ($t = 3.002$, $df = 98$, $p = .000<0.05$). So, the hypothesis which states that, there is no statistically significant mean difference between the CTT and IRT based on item difficulty estimates of students' responses to Achievement Test in Geography is rejected. This implies that there is statistically significant mean difference between CTT and IRT based on item difficulty estimates of students' responses to geography achievement test (GAT) in favour IRT.

Discussion of Findings

Discussion of findings was based on the research questions raised and the formulated research hypotheses. Findings from research question one as presented revealed that, few items have poor difficulty indices and a few items discriminated very poorly. The number of

bad items based on CTT are higher than IRT. Based on the result, it is possible that, the resultant effect of the higher number of poor items is because, CTT model only measures students' ability based on their grades or total achievement in a particular subject not minding the item quality whether there are questions that are above their standard. The finding agrees with the work of Eleje, Onah and Abanobi (2018), the authors conducted a comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis result and found that, the two frameworks are only comparable in terms of scores of the respondents. The result supports the work of Felix (2018) who worked on statistical results estimated using CTT approach. The study checked statistical characteristics of achievement test using the traditional measurement framework. The author submitted that, CTT which is traditional framework and IRT are comparable with IRT having advantage only with the scores obtained from the two frameworks. Thus, the use of only CTT by the NECO in item development could also be responsible for the poor performance recorded by Geography students in year 2013, 2014, 2015, 2016, 2017, 2018, 2019 and 2020 examinations. IRT framework should as well be employed in test development to check students' poor achievement.

Research question two revealed that, GAT based on IRT model as presented has a high number of items having good difficulty and discrimination indices with a few poor items that should have been modified or deleted. Findings from the comparative analysis showed statistically significant result in favour of IRT so that, IRT item calibration produced more good items and fewer items for modification or deletion than CTT item calibration. The finding presupposes that, IRT models measure students' ability based on individual items unlike CTT that only looks at the total achievement of the students in a particular subject. Therefore, IRT could easily be embraced based on the detailed nature of the framework. This finding is in tandem with the findings of Ogomaka, Onah and Amadi (2017) who worked on the comparison of the development of chemistry achievement test using item response theory and classical test theory. The authors found that, the item characteristic curve and information function in IRT enhance the reliability and validity of the achievement test. Thus, since students' achievement is to a large extent, a function of good measurement instrument, items developed using only CTT framework are capable of exerting negative impact on the students' achievement.

The finding from research question three as presented revealed that CTT-based and IRT- based item discrimination estimates are not comparable as IRT-based item discriminate better than CTT. The study agreed with the work of Fan (2001) who noted that "because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between the IRT and CTT-based item person statistics". Nevertheless, the finding is contrary to the submission of Troy (2004) who upheld that, an empirical comparison of item response theory and classical test theory item/person statistics can yield uniform result. Similarly, Dibu (2013) argued in a study on Classical Test Theory Versus Item Response Theory: An Evaluation of the Comparability of Item Analysis Results that, CTT and IRT are comparable and almost interchangeable in some cases. Thus, it could be seen from the result that, most of the items from both frameworks are good to be used in measuring students' ability. Therefore, the use CTT alone in item development would not be responsible for the students' poor performance. Other attribute of test such like standard error of measurement should also be considered as opined by Obinne (2008) who

worked on the psychometric analysis of two major examinations conducted in Nigeria by NECO and WAEC.

Result from the comparative analysis revealed difference in the difficulty indices of the items based on the two frameworks which suggests lack of comparability. However, the finding disagrees with Awopeju and Afolabi (2016) who compared Classical Test Theory (CTT) and Item Response Theory (IRT)-estimated item difficulty and item discrimination indices in relation to the ability of examinees in Senior School Certificate Examination (SSCE) in Mathematics with a view to providing empirical basis for informed decisions on the appropriateness of statistical and psychometric tests. The finding by Afolabi (2016) revealed that, CTT-based item difficulty estimates and IRT based item difficulty estimates were comparable. Results indicated that CTT-based and two-parameter IRT-based item discrimination estimates were comparable. The study concluded that CTT and IRT were comparable in estimating item characteristics of statistical and psychometric tests and thus could be used as complementary procedures in the development of national examinations. This result was disagreed by Mirnah (2018) in a similar study. The author pointed out variations that exist in the formulae and models used in the two frameworks.

The findings from research hypothesis one as presented revealed that there is statistically significant mean difference between the CTT and IRT based item discrimination estimates of students' responses to Geography achievement test (GAT) and as such the hypothesis is rejected. This implies that there is statistically significant mean difference between CTT and IRT Based item discrimination estimates of students' responses to GAT. The finding disagrees with the finding of Guler, Uyanık and Teker (2014), the authors carried out a study on comparison of classical test theory and item response theory in terms of item parameters. It was found that there was not much difference between using 1 or 2-parameter IRT model and CTT which is the opposite of the present study that saw significant difference in the item parameter estimates. Nevertheless, the result of the study by Guler, Uyanık and Teker (2014) was disagreed in a recent work by Mirnah (2018) who worked on an empirical comparison of item Response Theory and Classical Test Theory in Geography. The author found that, properties of items from the two frameworks vary. Thus, since the variation is significant, one can safely believe that, through IRT model, test constructors would be able to generate more reliable items than in the CTT model that is being currently used and ultimately the test scores of examinees would be more reliably estimated in IRT.

Result from the test of hypothesis two as presented showed that, there is statistical significant mean difference between the CTT and IRT based item difficulty estimates of students' responses to geography achievement test (GAT). The study agreed with the work of Adegoke (2013), who conducted a study on Comparison of Item Statistics of Mathematics Achievement Test using Classical Test and Item Response Theory Frameworks. This implies that there is statistically significant mean difference between CTT and IRT based item difficulty estimates of students' responses to achievement test in Geography. The finding disagrees with Osarumwense and Oyedeji (2015), who worked on empirical comparison of methods of establishing item difficulty index of test items using classical test theory (CTT) empirically compare two methods of computing the item difficulty index of test items based on Classical test Theory (CTT). The authors revealed that there was significant difference between the means of the item difficulty indices obtained by using the two methods.

Conclusion

Based on the results, the following conclusions were drawn: The two-parameter logistic model was used in the calibration of students' responses to GAT based on CTT and IRT model. The students' responses to GAT based on CTT revealed that majority of the test items have good item difficulty and item discrimination indices to measure student ability. This is because CTT does not examined item characteristics in details like the IRT does, the validity and reliability of the test is based upon the total test scores regardless of students' ability. The IRT produced test items with better item difficulty and items discrimination indices than CTT. The high result of good test items in IRT as compared to CTT is due to the fact that IRT focuses on item by item analysis and the validity of the test items are assessed for each item with the reliability calculated for each person's ability.

Recommendations

It is therefore recommended that;

1. The item difficulty and item discrimination estimates of students' achievement test in Geography based on CTT revealed that few of the items have poor difficulty and discrimination indices to measure students' ability. This implies that based on the comparative analysis, the instrument was reliable. Therefore, can be used in measuring students' ability.
2. Item difficulty and item discrimination estimates of students' responses to Geography achievement test (GAT) based on IRT revealed that, majority of the items have better difficulty and discrimination indices to measure than CTT. So, IRT framework is most suitable and should be used by test developers.
3. There is statistically significant mean difference between the CTT and IRT based on item discrimination estimates of students' responses to Geography achievement test (GAT) in favour of IRT. Thus, IRT is recommended to instrument developers.
4. There is statistically significant mean difference between the CTT and IRT based on item difficulty estimates of students' responses to geography achievement test (GAT) in favour of IRT. This shows suitability of IRT framework and should therefore be preferred by the examination bodies.

References

Adegoke, C. (2013). Investigating the invariance of persons parameter estimates based on classical test and item response theories 2010. *An International Journal on Education Science 2 (2),107-113*

Akinyele, G. (2019). Using reliability, validity and item analysis to evaluate a teacher-Made test in Geogreaphy. Retrieved from *http://stu.westga.edu/ahinson/reseacharticle* on 20/10/2019

Amadi, V. C. (2012). Evaluation of the implementation of continuous assessment in secondary schools in Owerri Education 1&2. *Unpublished M.Ed Thesis;Imo State University: Owerri*

Anikweze, C. M. (2012). *Measurement and evaluation: for teacher education.* Ibadan: Malijoe Soft Print.

Bradburn, N. M. (2009). *The structure of psychological wellbeing*. Chicago: Aldine Publishing.

Dibu, O., Kunmi, P., Francis, O., & Patrick, O. (2012). *Introduction to item response theory: Parameter models, estimation and application*. Goshen Print media Ltd

Ekwonye, E. C. & Eguzo, G.C. (2011). *Basic test, the ones in measurement and evaluation*. Oweri: Joe Mankpa Publisher.

Eleje, I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item. Analysis result European *Journal of Educational & Social Sciences 3(1),* 34-43

Emaikwu, S.O. (2015). *Fundamentals of test, measurement and evaluation with psychometric theories*, Makurdi: SAP Ltd.

Embretson, S. E. (2006). *Item response theory for psychologists.* Mahwah: Lawrence Erlbaum Associates.

Fan, X. (2001). Item response theory and classical test theory: An empirical comparison of their item/person parameters. *Educational and Psychological Measurement,* 58, 357-381.

Felix, F. (2018). Statistical results estimated using CTT approach. Retrieved from www.standfonline.com on 25/07/2019

Kpolovie, P. J. (2010). *Advance research method.* Oweri: Spring Field Publishers.

Awoeju, P., & Afolabi, S, (2016). 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*.*

Mirnah, O. K. (208). An empirical comparison of item response theory and classical test theory in geography. Retrieved from www.*znanstveniempiričnoraziskovalni prispevek..com* on 11/04/2019

Afolabi, D. C., (2016). Effect of incorporating practical into Mathematics evaluation on senior

National Examination Council (2013, 2014, 2015,2016 &2017). *Result Analysis Report.* Makurdi Zonal Office: Unpublished.

Nenty, J. H. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT.http://www.academicjournals.org/ERR Retrieved on the 12th august,2021

Nworgu. B. G. (2010). *Educational measurement and evaluation: theory and practice.* Nsukka: Hallman Publishers.

Obinne, A.D.E. (2008). A psychometric analysis of two major examinations in Nigeria: Standard error of measurement. Retrieved on 3rd September, 2019 *from krespublishers.com/IJES-03-2-137-11-044-Obinne-A-D-E-Tt.pmd*

Ogomaka, P. M., Onah, F.E. & Amadi, V. C. (2017). The comparison of the development of Chemistry achievement test using item response theory and classical test theory. *Nigerian Journal of Educational Research and Evaluation*. 16(2), 52-60

Guler, C. S., Uyanik, B. & Teker (2014). A comparison between item analysis based on item response theory and classical test theory. *A study of the Swe SAT Teast READ.* Retrieved from *www.edusci.umu.se on The 8/2/ 2019*.

Osurumwase, F. E. & Oyedeji, V.C. (2017). Item response theory and classical test theory: An emprical comparison of their item/person statistics. Educational and psychological measurement Retrieved from www.Researchgate.Net on 8/2/2019

Onunkwo, G. N. (2002). *Fundamentals of educational measurement and evaluation*. Onitsha: Cape Publisher International.

Troy, G.C. (2004). An empirical comparison of item response theory and classical test theory item/person statistics. *Journal of educational and psychological measuremen. 58(3)*54-62

Onah, F. E. & Amadi, V. C. (2017). Item response theory and classical test theory: An emprical comparison of their item/person statistics. Educational and psychological measurement Retrieved from www.Researchgate.Net on 8/2/2019