THE EFFECTIVENESS OF CLASSICAL TEST AND ITEM RESPONSE THEORIES IN SECONDARY SCHOOL GEOGRAPHY ACHIEVEMENT TEST CONSTRUCTION IN OYO STATE

Sunday O. Adedigba, Ayoka M. Olusakin, & Ngozi A. Osarenren
Department of Educational Foundations
University of Lagos Akoka, Lagos
08033864458/adedigbasunday842@yahoo.com

Abstract

Measurement theories are important to practice in educational measurement because they supply a background for addressing measurement problems. One of the most essential challenges is handling the Measurement Errors. A reliable theory can help in understanding the role of errors they play in measurement. Therefore, the purpose of this study was to investigate the effectiveness of Classical Test and Item Response Theories in secondary school Geography Achievement Test construction, in Oyo State. The descriptive survey research design was used for this study. The population for the study comprised of all SS2 students offering geography in the three educational zones of the State. The sample for the study was 1200 students. Multi stage stratified sampling technique was used to get this sample. A 100 item GAT was the instrument for the study. The hypotheses were tested using paired sampled t- test. Results revealed that item statistics obtained from the two theoretical frameworks were comparable. However, item statistic obtained from IRT model showed more stability than those from CTT. Moreover, for item selection process, IRT model led to deletion of fewer items than CTT model. This result indicates that test developers and public examining bodies should incorporate IRT model into their test development processes because through IRT model test constructors would be able to generate more reliable items than in the CTT model which is being employed currently in schools. It was further recommended that CTT framework could be used as a complement to IRT.

Keywords: Item analysis, Classical test theory, Item response theory, Item statistics, Item parameters.

Introduction

Evaluation is observed as a qualitative description of pupils' behaviour. Mehrens and Lehman (2009) asserted that irrespective of how efficient the teacher is, how intelligent the learners are, and how adequate the audio-visual equipment, if no provision is formed for evaluation of the students' progress, all teaching efforts could be completely invalidated. Evaluation concerns determining the standard of the curriculum, facilities, and performance of pupils using various tools which include test. Test is an indispensable tool for evaluating the learning outcomes and the change in behaviour of learners. Anastasi cited in Okoli (2005) defines a test as a set of standardized items or inventories administered on an individual for the purpose of measuring or obtaining quantitative information about several aspects of the individual's behaviour.

Testing is important in education and other science fields because many selections and policies are made in line with the results of testing. It's a fundamental part of the teaching-learning process not as a basis for ranking students at the end of the teaching – learning process but to guide teaching and aid in the development of curriculum, as well as appraisal of needs, learning difficulties, level of mastery and differences among students. Different kinds of tests are used for assessment and consequently evaluation.

A variety of tests are employed in education but the use of multiple choice tests is in vogue all over the world. According to experts (Steven, Richard, Paul & Bud, 1991; Akinyele, 2015), tests play important role in giving feed backs to stakeholders in education on various aspects of educational objectives including the cognitive, the affective, and the psychomotor domains. It is a good instrument for measuring the students' Intelligent Quotient. Tests are either criterion-reference based or norm reference based. A criterion based test is a test whose purpose is to determine the numbers of students that have mastered certain contents while norm based test differentiates among students.

In educational measurements, there are two main frameworks through which tests can be developed, validated and ultimately used for assessing examinees' performance. These are Classical Test Theory (CTT) and also the Item Response Theory (IRT). The Classical Test Theory involves three concepts. These are: test (observed) score, true score and error score. Hambleton and Jones (1993) opined that within these three concepts, several models are formulated, of which the central model is the "classical test model". This model connects the observed test score (x) to the sum of the two unobserved (or often called latent) variables, true score (T) and error score (E). Classical test theory is roughly synonymous with true score theory. This theory assumes that every individual features a true score which might be obtained if there have been no errors in measurement. However, because measuring instruments are imperfect, the score observed for every person may differ from an individual's true ability.

The Classical Test Theory (CTT) analyses are the best and most generally used variety of analysis. The statistics could be computed by readily available statistical packages or perhaps by hand. Classical analyses are performed on the test as a whole rather than on the item (High Achievers or Low Achievers from the score within the test). Although item statistics could be generated, they apply only to group of learners on the collection of items. CTT is based on the true score model and utilizes some statistics such as Difficulty index, Discrimination index and Reliability. Most importantly in CTT, we assume that the error is: (a) Normally distributed (b) uncorrelated with true score and (c) has a mean of zero. In the usage of CTT, Ojerinde (2013) stated that the ability of the students is dependent on the type of test items employed and the parameters of the items are dependent on the samples of test used by the students.

Item Response Theory (IRT) on the other hand refers to a family of latent trait models employed to establish psychometric properties of items and scales. It is sometimes regarded as the strong true score theory or modern mental test theory because IRT is a newer theory and makes stronger assumptions when compared to classical test theory. IRT is a general statistical theory about examinee, item, test performance and the way

performance relates to the talentss that are being measured by the items within the test (Hambleton & Jones, 1993). Item Response Theory relies on the concept that the probability of a correct/keyed response to an item will be a mathematical function of person and item parameters. The person parameter is named latent trait or ability, intelligence or the strength of an attitude. Item parameters include difficulty (location), discrimination (slope or correlation) and pseudo guessing (lower asymptote).

Statement of the Problem

Construction and validation of multiple choice tests using CTT and IRT have been in existence for many years in the developed countries. In Nigeria, measurements of students' achievement by teachers and public examining bodies, until recently, have always been focused on classical test theory framework in validating their test items. Construction of valid school- based test in Secondary Schools has not been given much attention and one of the greatest problems is that teachers make use of unvalidated teacher-made test for internal examinations and students performance in their examination do not often appear to correctly predict their performance in their Senior Secondary School Certificate Examinations that public examining bodies in Nigeria conduct.

Examinees' scores have always been based on number- correct scoring method of CTT (Adegoke, 2013). Although, during test development, item statistics such as difficulty and discrimination parameters are also assessed in CTT framework, these parameters have not often been used in the estimation of examinees' scores. The use of IRT and some studies have shown that estimation of examinees' scores using CTT is error prone and therefore the use of IRT has been recommended.

An aspect which research in Geography Education has not focused much on is how assessment practices in terms of test construction and validation affect students responses to test items and their ultimate scores. This shows that the procedures and frameworks for test development and how test items are constructed can have impact on the effectiveness of the test (Adedoyin & Adedoyin, 2013; Idowu, Eluwa & Abang,

2011). It is as a result of this that research much shift focus towards the assessment procedures being adopted by classroom teachers and public examining bodies. This is because the assessment practices could be one of the reasons why students are performing poorly in Geography. No doubt, poorly worded test items with ambiguous answers may be confusing to test takers and if tests are not properly scored, examinees final scores in a test may not reflect their actual ability. In view of the forestated, this study investigated the effectiveness of Classical Test and Item Response Theories in secondary school Geography Achievement Test construction, in Oyo State.

Purpose of the Study

The purpose of this study was to investigate the effectiveness of Classical Test and Item Response Theories in secondary school Geography Achievement Test construction, in Oyo State. Specifically, this study is designed to:

i. determine the difference between CTT-based item discrimination index and IRT –based item discrimination estimates.
ii. examine the difference between CTT-based item difficulty estimates and IRT-based item difficulty estimates.

Research Hypotheses

The following hypotheses formed the basic assumptions for this study:

i. There is no significant difference between CTT- item discrimination index and IRT- item discrimination estimates.
ii. CTT – based item difficulty estimates will not significantly differ from IRT based item difficulty estimates.

Methodology

The entire procedure used to carry out this study was discussed here.

Research Design

The research design for the study was the descriptive survey design. This design is suitable because it is capable of studying large and small populations (or universe) by selecting and studying samples chosen from the population to discover the relative incidence, distribution and interrelations of sociological and psychological variables (Ilogu, 2005).

Population of the Study

The population for the study comprised all Senior Secondary School Two Geography students in public schools in Oyo State, Nigeria. The Senior Secondary two students were chosen because as at the time of the study they were expected to have covered most of the Geography topics in the syllabus on which questions were based and were available for the period of the assessment.

Sample and Sampling Technique

The sample for this study comprised of 1,200 SSII students both (male and female) from 60 secondary schools in three educational zones and Multistage sampling technique was used. The first stage of the multistage process was the selection of three educational zones out of the six educational zones through simple random sampling method of hat and draw process. The selected zones were educational zones 3,4 and 5.The second stage involved the selection of twenty (20) senior secondary schools from each of the selected educational zones using simple random sampling technique of lucky dip process with replacement.

A total of 20 schools were selected in zone 3 out of 46 schools, 20 senior secondary schools were also selected in zone 4 out of 91 schools and 20 schools in zone 5 out of 120 schools. In all, a total of sixty schools were involved in the study. Thirdly, stratified random sampling technique process was used to select students from each of the schools randomly selected and twenty SSII students offering Geography in all the 60 schools were selected to ensure equal representativeness of the various sub-groups making up the population.

Instrumentation

The instrument adopted was 100-Item Draft Geography Achievement Test (DRA – GAT). The following steps were taken in the draft of 100 – items GAT development.

Step One: Preparation of the 100 – test item pool. Two Secondary School Geography graduate teachers who are also examiners with WAEC read the draft copy of the test items and their corrections were noted. The information provided on each of the test items was used to re-write some of the items. The test blue print was presented in Table below.

Step Two: The test items were refined by subjecting them to item analysis. Item analysis involves calculating index of difficulty and index of discrimination of each test item. An index of difficulty ranging from 0.40 to 0.60 is considered good, while discrimination index of +0.30 to 1 is good (Okoli, 2005). The good items from the item analysis were then validated by administering it in the selected schools.

Table 1:

Test Blue Print for 100 – item DRA – GAT.

		Behavioural Objectives				
Content	**Weight %**	**Knowledge 36%**	**Comprehension 30%**	**Application 23%**	**Analysis 11%**	**Total 100**
Solar System	8%	46, 49, 59 (3)	47, 56 (2)	60, 94 (2)	79 (1)	8

Latitude and Longitude	17%	9, 25, 37, 65, 91 (5)	11, 15, 21, 26, 64 (5)	10, 16, 34, 53 (4)	54, 74, 75 (3)	17
The earth	8%	20, 51 (2)	50, 52, 58 (3)	95, 99, 100 (3)	-	8
The rocks	13%	2, 4, 32,70 (4)	7, 24, 55, 73 (4)	22, 23, 81 (3)	48, 98 (2)	13
Weather and climate	26%	12, 13, 14, 39, 62, 63, 72, 89, 93 (9)	17, 19, 61, 71, 77, 82, 86,90 (8)	23, 33, 68, 69, 92, 97 (6)	38, 40, 67 (3)	26
Weathering	6%	30, 35, 36 (3)	18 (1)	78, 87 (2)	-	6
Internal processes of land forms	22%	3, 8,28, 29,31,57,66,88,96 (9)	1,5,6,45,76,80,83 (7)	27,42,44,84 (4)	43,85 (2)	**22**
TOTAL	**100%**	**36**	**30**	**23**	**11**	**100**

The instrument was administered on the selected participants by the researcher and the trained research assistants. The administration of the instruments was done during the normal time scheduled for Geography on the school official Time Table. This was to avoid disruptions to the school activities. During the administration, the students were assigned one and half hour (90 minutes) to complete the test. The data collected from the study was analyzed using both descriptive and inferential statistics. The hypotheses were tesed using paired sampled t-test and Psychometric package of R language and environment for statistical computing.

Result and Discussion

Hypothesis One: There is no significant difference between CTT- item discrimination index and IRT- item discrimination estimates

To test the hypothesis, the discrimination indices of the test items under CTT and IRT were estimated. The Discrimination parameters of the test items were estimated using

Psychometric package of R Language and Environment for Statistical Computing. And the IRT based item discrimination parameters were estimated with 3-parameter logistic model of IRTPRO using Maximum Marginal Likelihood estimation (MML). Considering the fact that the discrimination index of CTT and IRT are not on the same metric (CTT has values from -1 to 1 and IRT has values from –infinity to + infinity), comparing CTT and IRT-based item discrimination required converting CTT-based discrimination indices to the metric of IRT. To achieve this,

$$a_{ctt} = \frac{item\ biserial\ correlation}{\sqrt{1 - item\ biserial\ correlation^2}}.$$

Thereafter, the converted CTT discrimination indices were compared with the IRT discrimination indices of the 100-item GAT. The results are presented as follow:

Table 2:

Mean and standard deviation of GAT under CTT and IRT estimated item discrimination

Item (D)	N	$\bar{x}$	SD
CTT	100	0.10	0.07
IRT	100	0.53	0.42

Table 2 presents the item discrimination parameters of the 100-item GAT under CTT and IRT. The Table showed that the item discrimination indices of the 100-item GAT was higher when estimated with IRT framework (Mean = 0.53; SD = 0.42) than when the discrimination indices of the items were estimated with CTT (Mean = 0.10; SD = 0.07). In order to assess whether the observed difference in the estimates obtained under IRT and CTT method of estimating item discrimination, paired samples t-test statistic was conducted. The result is presented in Table 3 below.

Table 3:

Paired samples t-test of CTT and IRT estimated item discrimination

	Mean Diff	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	Df	Sig. (2-tailed)
				Lower	Upper			
CTT Pair 1 - IRT	-.43440	.35761	.03576	-.50536	-.36344	-12.147	99	.000

The result presented in Table 3 showed that the difference observed in the discrimination indices of the 100-items GAT estimated with CTT and IRT was statistically significant (t = -12.147, df = 99, p = 0.000). Thus, the hypothesis which states that "There is no significant difference between CTT- item discrimination index and IRT- item discrimination estimates" was rejected.

Hypothesis Two: CTT – based item difficulty estimates will not significantly differ from IRT based item difficulty estimates

To test this hypothesis, the difficulty indices of the test items under CTT and IRT were estimated. The Difficulty parameters of the test items were estimated using Psychometric package of R Language and Environment for Statistical Computing and the IRT based item difficulty parameters were estimated with 3-parameter logistic model of IRTPRO using Maximum Marginal Likelihood estimation (MML). Considering the fact that the difficulty index of CTT and IRT are not on the same metric (CTT has values from 0 to 1 and IRT has values from –infinity to + infinity).Therefore, comparing CTT and IRT-based item difficulty requires converting CTT-based difficulty

indices to the metric of IRT. To achieve this, $b = \frac{\frac{\ln\left(\frac{p}{1-p}\right)}{1.7}}{item\ biserial\ correlation}$.

Thereafter, the converted CTT difficulty indices were compared with the IRT-based difficulty indices of the 100-item GAT and the results are presented as follow:

Table 4:

Mean and standard deviation of GAT under CTT and IRT estimated item difficulty

Item (P)	**N**	**$\bar{x}$**	**SD**
CTT	100	-12.74	127.50
IRT	100	- 0.18	15.85

Table 4 presents the item difficulty parameters of the 100-items GAT under CTT and IRT. The table showed that the items were more difficult when estimated with the IRT framework (Mean = -0.18; SD = 15.85) than when they were estimated using the classical test theory approach for item analysis (Mean = -12.74; SD = 127.50). In order to assess whether the observed difference in the difficulty estimates obtained under IRT and CTT method of estimating item difficulty, paired samples t-test statistic was conducted. The result is presented in Table 8 below.

Table 5:

Paired samples t-test of CTT and IRT estimated item difficulty

	Mean Diff	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference	T	Df	Sig. (2-tailed)

				Lower	Upper			
CTT Pair 1 - IRT	-12.56013	127.07827	12.70783	-37.77521	12.65496	-.988	99	325

The result presented in Table 5 showed that the difference observed in the difficulty indices of the 100- GAT items estimated with CTT and IRT was not statistically significant (t = -0.988, df = 99, p = 0.325). Thus, the hypothesis which states that "There is no significant difference between CTT- item difficulty index and IRT- item difficulty estimates" was not rejected.

Discussions

Hypothesis one stated that there is no significant difference between CTT item discrimination index and IRT item discrimination estimates. The result of the finding indicated that there was a significant difference between CTT item discrimination index and IRT item discrimination estimates. Hence, the hypothesis was rejected. This finding was supported by Hambleton and Jones (1993), and Wilberg (2004) who stated that CTT based discrimination index is comparable with the IRT based discrimination parameter. They were of the opinion that the correlation coefficient of the relationship between a-values and point biserial correlation should be high and positive. However, using CTT-based item statistics estimates more items were deleted from the 100 items GAT than when IRT-based item statistics estimates were used. This finding lay credence on the observation of test experts such as Hambleton and Jones (1993) and Ojerinde (2013) that despite the popularity of classical item statistics as an integral part of standardized test and measurement technology, it is fraught with so many limitations.

Hypothesis two stated that CTT – based item difficulty estimates will not significantly differ from IRT based item difficulty estimates. The result of the analysis showed that the items were more difficult when estimated with the IRT framework than when they

were estimated using the CTT approach. Furthermore, the result indicated that there was no significant difference between the CTT item difficulty model and IRT item difficulty estimates. Therefore the hypothesis was not rejected. The finding was contrary to Hambleton and Jones (1993) who was of the view that the correlation should be high and negative and also the results of past studies such as Wilberg (2004) and Stages (2003) laid credence to this that as the value of P increases, b, decreases.

Conclusion

In line with the findings of this study, the study concluded that:

i. There was a significant difference between CTT item discrimination index and IRT item discrimination estimates.
ii. CTT – based item difficulty estimates would not significantly differ from IRT based item difficulty estimates.

Recommendations

In line with the findings of the study, the following recommendations were made:

i. Examination bodies using multiple choice test instruments should adopt the use of both IRT and CTT statistics in test development processes.
ii. Geography achievement tests constructed by teachers that are used to examine students' performance compared to educational standards should be made to pass through all the processes of standardization and validation.
iii. Item analysis should be maintained in test development and evaluation because of its relevance in the investigation of reliability and in minimizing measurement errors.
iv. Training on test construction and development should be regularly organized for teachers to be more proficient in test construction, marking and grading of students scripts.

References

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

Adegoke, B.A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice*, 4 (22), 87 – 96.

Akinyele, O.A. (2015). An overview of classical test theory and item response theory in test development. *Nigeria Journal of Educational Research and Evaluation, 14* (3), 147-157.

Anastasi, A. (1982). *Psychological testing*. New York. Macmillian

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Education measurement, Issues and Practice*, 12 (3), 253 – 262.

Idowu, E.O., Eluwa, A.N., & Abang, B.K. (2011). Evaluation of mathematics achievement test: A comparison between classical test theory and item response theory. *Proceedings of International Association for Teaching and Learning*, 134 – 141.

Ilogu, G.C. (2005). *Educational research and evaluation: a comparison for students*. Yaba: Mandate communication Ltd.

Mehrens, W.A., & Lehmann, I.J. (2009). *Measurement and evaluation in educational psychology*. New York: Reinehart, Holt & Winston.

Ojerinde, D. (2013). Classical test theory (CTT) vs item response theory (IRT): An evaluation of the comparability of item analysis results. *A paper presented at the Institute of Education*, University of Ibadan.

Okoli, C.E. (2005). *Introduction to educational and psychological measurement*. Lagos: Behenu Press and Publishers.

Stage, C. (2003). Classical test theory or item response theory: the Swedish experience. Umea: Kluwer Academic Publisher.

Steven, J.B., Richard, R.S., Paul, F.M., & Bud, W. (1991). How to prepare better multiple-choice test items: guidelines for university faculty, *European Journal of Social Sciences, 15*(2), 61-72.

Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedishdriving-license test*. Umea: Kluwer Academic Publications.