MODE OF TEST ADMINISTRATION AS A FACTOR IN PRE-DEGREE STUDENTS MATHEMATICS TEST SCORE VALIDITY AND RELIABILITY IN SOUTH WEST UNIVERSITIES, NIGERIA

ALADENUSI, O.

Department of Educational Psychology
Federal College of Education (Technical) Akoka, Lagos-State.
kemab2004@yahoo.com

Abstract
Scores generated from test instruments should measure what it is purported to measure and consistently too. When scores failed to achieve these two, they are meaningless and cannot be used for accurate decision making. Investigating and providing evidences of score validity and reliability are the main objective of this study. Thus, this study examined the mode of test administration on score validity and reliability using tests among pre-degree students in southwest, Nigeria. A survey research design was adopted. The population of the study comprised 14,532 students across the nine public universities in south-west where pre-degree programme is mounted. Stratified, systematic and simple random sampling techniques were used to select 400 participants. Mathematics Multiple Choice Test was used to collect data. Participants were grouped into two; Paper Pencil Test (PPT) and Computer Based Test (CBT). Same instruments were administered to the two groups with different instructions and procedures of administration. Two hypotheses were tested in the study. Scores generated were factor analysed and interpreted. Results indicated that Mathematics test administered via PPT was higher in validity and reliability than its CBT counterpart. It was recommended that PPT should still be encouraged as a mode of test administration for Mathematics achievement test.

Keywords: Validity, Reliability, Computer Based Test, Paper Pencil Test

Introduction

Scores form a fundamental premise upon which decisions about examinee are made in various life situations. The scores from test will be meaningless if they are not accurate, truthful, dependable and consistent. In testing, scores are widely used for educational accountability and decision making. Scores generated from academic achievement test are used for several purposes by the teachers, students, curriculum planners, educational administrators among others. Scores obtained from tests results in consequential decisions about individual students for promotion, selection, placement, awarding high school diplomas and degree certification. The quality of the scores yielded by test (instrument) is always of primary importance and should be verified. Investigating and improving the quality of that scores should be the interest of measurement specialists, an essential concern for test users and consumers of test scores. The scores from a class of Mathematics for example, would indicate whether learning has taken place after instruction has been given and whether or not they are adequately prepared for the next stage of instruction.

Validity and reliability are psychometric properties inherent inscores generated from aninstrument, example of such instruments are; tests, questionnaire, observer rating among

others. Examinees are subjected to school entrance examination of which Mathematics is a core subject at all levels of educational system. Validity and reliability are psychometric characteristics to be evaluated, revered when it comes to judging the quality of test information. According to Anastasi & Urbina, (2012),the validity of a test concerns what the test measures, how well it does so and what can be inferred from the test scores. Whenever a test user wishes to make an inference from test scores, the validity of those inferences must be verified for the scores to be meaningful. All evidences provided strengthens the argument that the construct of interest is actually the construct the scores represent. In addition, and in order that scores be valid, it must be interpretable and useful. Hubley & Zumbo, (2011) posited that validity is about the inferences, interpretations, actions, or decisions that are based on a test score and not the test itself. They further posited that violations of score validity severely impact the function and functioning of score interpretations. According to them, score validity inadequacies impact even more serious consequences on score interpretation than its score reliability counterpart.

Score reliabilityis of utmost importance in measurement because it is a necessary but not sufficient condition for scorevalidity, any weakness in score reliability will impactthe validity of scores yielded from an instrument (Russell, 2008). These distinctions illuminate the inextricable link between validity and reliability. Perfectly unreliable scores even with some degrees of validity measure nothing (Thompson, 2006). In other words, poor score reliability oftencompromise the effectiveness of the scores obtained to measure the intended constructs, hence, poor score reliability estimate may compromise score validity.The validity ofany score is influenced directly by the reliability of the data and none of these things can be correctly interpreted without examining the reliability of one's data(Nilsson, Schmidt &Meek, 2002). It is imperative that those who use tests are able to evaluate whether the data they obtain so cleverly are any good in the first place (Cone & Foster, 1991).If the test scores are not valid, they misrepresent students' true knowledge which might result in inappropriate decisions that could have negative, temporary or lasting effect on the students. This is why Jimoh&Omoregie (2012), posited that any action that undermines examinations poses a great threat to the validity and reliability of the examination results and its certification.

Also, lack of score reliability has a direct consequence on the uses of test scores. Ghiselli (1964) stated that unreliable scores are of little value when we wish to compare two or more individuals on the same test, to assign individuals to groups or classes, to predict other types of behaviour, to compare different traits of an individual, or to assess the effects of various systematic factors upon an individual's performance.

There are two types of mode of item delivery in most levels of education in Nigeria. They are the traditional paper and pencil type (PPT) which is also known to be the conventional mode of item delivery and the computer base testing (CBT), which is the electronically mode of item delivery.The name PPT was derived from its mode of delivery where items (questions) to be responded to by the examineesare presented to them in a paper and its instructions are all in the question paper. The examinees will provide answers to the items by writing in the answer sheet(s) or answer booklets using a pencil or biro depending on the instruction given to the examinees by the examiner. Its administration is carried out by tutored invigilators, who are not only going to the examination hall to distribute the question papers to the examinees butto also invigilate properly to avoid irregularities that can lead to abuse of the test scores. However, the PPT is faced with an increasing number of limits among which includes the drastically increasing number of students and the conventional

examination method became time consuming in terms of the examination time for evaluation and assessment (Oduntan, Ojuawo&Oduntan, 2015).

The work environment of the 21st century has changed due to the progress made in the field of information technology and in order to satisfy the needs in measurement and evaluation of the 21st century, a basic and qualitative change is required (Beller, 2013). As a result, psychometric testing has since adopt Computer Base Testing(CBT) as its mode of item delivery, instructions and time allowed is computerised, itappears on the computer screen and the computer prompts the client to answer series of questions and this can only be done by pressing allocated keys on the keyboard or by using the mouse or touchpad to select the answer. CBT allows for immediate scoring and reporting of results (Foster, 2010; Wang, 2010).

The use of CBT for Mathematics mode of item administration by examination bodies in Nigeria (such as theUnified Tertiary Matriculation Examination (UTME) by the Joint Admission Matriculation Board (JAMB))could be observed with a number of shortcomings that may have undermined its success.As it is, the advantages of CBT may still not confirm the sanctity of the yielded scores validity and reliability.Furthermore, researchers have studied the effect and score equivalence on the mode of item administration on achievement tests over the years. However, findings related to the score equivalence between CBT and PPT is inconclusive. Some of the studies indicated that the CBT scores were equivalent to the PPT scores (Bergstrom, 1992; Boo &Vispoel, 1998; Choi &Tinkler, 2002; Johnson & Green, 2004), whereas, other studies indicated that the results from CBT and PPT could not be used interchangeably (Pommerish& Burden, 2000).

There are inconclusive findings on these positions which have not been able to separately provide comprehensive explanation to the issue of score inconsistency among learners. Indeed, examination bodies like Joint Admission Matriculation Board (JAMB) have fullyengagedthe use of computer administration. This study is designed to examine the mode of test item administration in determining the level of score validity and reliability. Specifically, the study investigated the extent to which mode of test item delivery interplay in determining score validity and reliability for Mathematics multiple choice tests among the pre-degree students in South-west, Nigeria.

Research Hypotheses

The following hypotheses were formulated for the study.

1. There is no significant difference in the validity of Mathematics Multiple choice tests scores administered via PPT and those administered via CBT.
2. There is no significant difference in the reliability of Mathematics Multiple choice test scores administered via PPT and those administered via CBT.

Methodology

The research design adopted for this study was a descriptive research design. It involved the selection of a sample of pre-degree students and the result of the study applies on the entire pre-degree students in South-West,Nigeria. The population of the study consists of students in public universities in South-West, Nigeria where pre-degree programme is offered. The public universities include all the Federal and State Universities. The lists of the universities are presented in Table 1.

Table 1: List of Universities in South-West, Nigeria

SN	Universities	Location	Public Type	No of Students on roll
1	Olabisi Onabanjo University (OOU), Ago-Iwoye.	Ogun	State	638
2	Tai-Solarin University of Education (TASUED), Ijebu Ode.	Ogun	State	146
3	Federal University of Agriculture, Abeokuta (FUNAAB).	Ogun	Federal	1560
4	Ladoke-Akintola University of Technology (LAUTEC), Ogbomoso.	Oyo	State	4600
5	Obafemi-Awolowo University (OAU), Ile-Ife.	Osun	Federal	2600
6	Osun-State University (UNIOSUN), Osogbo.	Osun	State	1030
7	Federal University of Technology Akure (FUTA)	Ondo	Federal	2500
8	Adekunle Ajasin University Akungba (AAUA),	Ondo	State	980
9	Ekiti-State University (EKSTU), Ado-Ekiti	Ekiti	State	478
	Total			**14,532**

There are nine public universities as at 2016/2017 academic session. The sample for this study comprised of 400 pre-degree students selected through systematic sampling in the public Universities where there are pre-degree centres and computer centers for the CBT mode of testing in South-west, Nigeria. These Universities were selected through stratified and simple random sampling techniques. The division of state and federal universities in the south-west were adopted as strata.

Thereafter, simple random sampling technique, hat and draw method was used to select two federal universities and two state universities. Thus, four universities were selected. After the selection of the universities, systematic sampling was used to select participants for the study. This led to the selection of 100 students in each institution and a total of 400 pre-degree students participated in the study. Besides, two groups were created in each institution and assigned Paper and Pencil Test (PPT) and Computer Based Test (CBT).

Table 2: Distribution of Participants Based on Institutions and Groups

University	Groups				Total
	PPT		CBT		
	Code	Participants	Code	Participants	
1	A1	50	A2	50	100
2	B1	50	B2	50	100
3	C1	50	C2	50	100
4	D1	50	D2	50	100
Total					400

Table 2 shows 50 participants were chosen in each institution for PPT and CBT respectively. The researcher developed instrument titled Mathematics Multiple Choice Test (MMCT) that was used to collect relevant data. The MMCT had two sections, namely Section A and B. Section A was used to collect personal information about the respondents such as; gender, age, name of institution, etc. Section B contained 50 Mathematics multiple choice questions. The items in the MMCT were developed and refined by the researcher with reference to selected topics in UTME Mathematics Syllabus. UTME syllabus was used since all pre-degree students were expected to sit for Unified Tertiary Matriculation Examination before they can gain their admission into the university. During the pilot study, an item analysis was conducted. The items with the indices of difficulty ranging from 0.2 to 0.8 and positive values were used for discrimination index for the items MMCT. The MMCT was content validated using the Table of Specification in table 3 and Cronbach alpha was used to determine the internal consistency of the MMCT. The process yielded a correlation coefficient of 0.78.

Table 3: Table of Specification for a 50-Item MMC test

CONTENTS	**BEHAVIOURS**				**Total Items**
	Knowledge	**Comprehension**	**Application**	**Analysis**	
1).Number and numeration.	2	3	3	2	10
2). Algebra	3	2	4	1	10
3).Geometry/Trigonometry	2	2	3	3	10
4).Calculus	2	2	3	3	10
5).Statistics	2	2	3	3	10
Total Items	11	11	16	12	**50**

Mathematics Multiple Choice Test was administered in the four universities selected for the study by the researcher with the assistance of four proctors and four computer technologists. MMCT was administered on Groups A1, B1, C1 and D1 in the four selected universities using the PPT mode first and CBT mode later. In addition, GroupsA2, B2, C2 and D2 had theirs administered to them using CBT mode and later the PPT mode. Instruction that was given to participants in PPT groups differs from that of CBT groups based on the nature of the test mode. Duration of 60 minutes was given to all the participants based on the test modes. The participants writing the PPT modes were to read the instruction on the question paper given to them. They do not need any password to log on the question paper given to them.

However, the CBT Groups need to log on before they can gain access to the questions. Aside that, they need to log on with their different password and click on the test they are to attempt at that particular time. Timing for the participants was automatic from the point of logging on. Besides, there was "submit" option for those who finished before the expiration of time. The scripts of the participants in PPT groups were marked and scored with the help of the research assistants, while the CBT was scored electronically and immediately generated. Scores from PPT and CBT on Mathematics multiple choice were factor analysed and co-related.

Descriptive and inferential statistics were used to find the mode effect on the score validity and reliability. In examining the mode effect of PPT and CBT, data were analysed based on item response theory approach to establish its validity and reliability. Thus for validity, construct evidence in form of factor analysis was established for each mode of administration. Also, the equivalence of the factors generated by the two modes PPT and CBT were examined using Turker's Exact Test. The difference in performances between the modes was established through analysis of variance. Reliability was determined through the internal consistency using Cronbach Alpha. Differences in mode effect was also determined by comparing the Alpha value through t-test of correlated values.

Results

Hypothesis One: *There is no significant difference in the validity of Mathematics Multiple choice test scores administered via PPT and those administered via CBT.*

Factor analysis was conducted on the scores of Mathematics for the PPT and the CBT administered test. The PPT administered Mathematics multiple choice test measures more than one latent trait of the examinees. To identify the number of specific factors that underlie the test, Parallel Analysis (PA) was conducted. The result is presented in Figure 1.

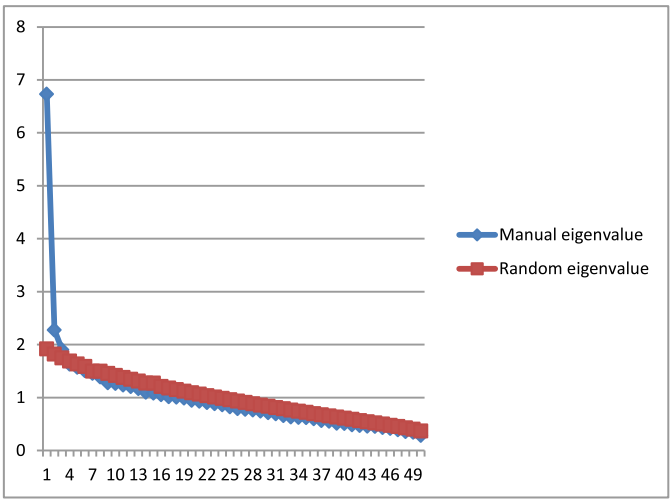

Figure 1: Number of factors of PPT administered using MMCT

Figure 1 show that there are three Eigenvalues that were observed to be above the point where the random eigenvalue intercepted the Eigenvalues of the real data set (PPT, MMCT). Thus, there were three factors that were extracted to underlie the PPT MMCT.
Similarly, the result showed that the CBT version of MMCT measures more than one latent trait of the examinees. To identify the number of specific factors that underlie the test, PA was conducted. The result is presented in Figure 2

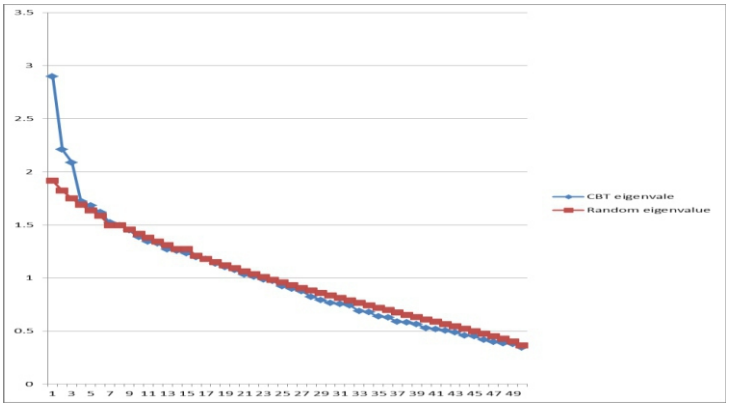

Figure 2: Number of factors of CBT using MMCT

Figure 2 show that there are six eigenvalues that are observed to be above the point where the random eigenvalue intercepted the eigenvalues of the real data set (CBT MMCT). Thus, there are six factors that underlie the CBT Mathematics multiple choice tests.

Table 3: Comparison of Factor loadings of PPT and CBT using MMCT

Factor									
PPT				CBT					
	1	2	3	1	2	3	4	5	6
M1	0.05	-0.12	**0.60**	0.25	0.30	-0.13	0.21	0.20	-0.01
M2	-0.15	0.22	0.27	0.21	-0.13	-0.09	0.11	-0.01	0.19
M3	-0.08	0.10	**0.45**	**0.38**	0.06	0.00	0.16	-0.09	-0.12
M4	**0.34**	0.01	0.04	0.09	-0.03	-0.08	0.24	-0.01	0.08
M5	0.06	**0.35**	-0.05	-0.03	-0.08	0.01	0.00	**0.34**	0.07
M6	0.07	-0.17	**0.50**	-0.10	-0.04	0.03	-0.10	0.18	0.08
M7	0.04	0.26	0.04	-0.07	**0.41**	0.09	-0.15	-0.02	-0.05
M8	0.06	0.13	**0.42**	0.09	-0.11	0.06	**0.37**	0.13	0.22
M9	-0.01	**0.43**	0.16	**0.52**	-0.13	-0.04	-0.05	0.04	-0.01
M10	0.20	0.24	-0.03	-0.05	-0.01	-0.03	0.08	0.06	0.07
M11	0.05	**0.38**	0.03	0.15	0.05	-0.12	**-0.44**	0.11	0.01
M12	0.31	0.09	0.03	**0.43**	-0.17	0.08	-0.07	0.02	-0.04
M13	0.23	0.19	0.13	0.14	0.04	-0.09	0.02	-0.13	-0.03
M14	-0.13	**0.45**	-0.02	0.30	0.18	0.00	-0.22	-0.05	0.16
M15	0.08	0.22	0.29	0.11	0.03	0.04	**0.38**	-0.14	-0.02
M16	-0.13	**0.55**	0.05	0.04	0.05	0.04	0.00	-0.21	0.31
M17	0.12	0.10	0.31	0.23	0.10	0.09	0.08	0.08	-0.09
M18	0.21	0.11	0.07	0.21	-0.01	0.14	0.05	0.08	0.03
M19	0.11	0.18	0.15	0.02	0.13	0.03	0.02	-0.05	0.07
M20	**0.43**	-0.03	0.00	0.29	0.07	-0.16	0.07	-0.13	0.19
M21	-0.05	0.28	0.04	0.20	-0.14	-0.03	0.02	0.14	0.13
M22	-0.04	**0.45**	0.14	-0.07	0.15	0.17	0.13	0.03	0.14
M23	0.10	0.31	0.04	0.24	0.03	0.14	0.03	-0.26	-0.06
M24	-0.10	0.29	0.19	0.30	0.15	0.03	-0.04	-0.05	-0.05
M25	0.15	0.24	0.12	-0.14	0.21	0.15	0.13	0.01	-0.01
M26	**0.51**	-0.24	0.09	0.01	-0.04	-0.05	0.09	0.04	0.25
M27	0.25	0.07	-0.01	-0.12	0.14	0.24	0.29	0.03	0.08
M28	**0.44**	-0.03	0.20	-0.03	**0.34**	0.03	0.06	0.05	-0.10
M29	**0.56**	-0.22	0.15	**0.38**	-0.07	**0.35**	0.05	-0.05	-0.08
M30	**0.40**	0.02	0.06	0.17	-0.02	**0.34**	-0.06	-0.01	0.03
M31	**0.36**	0.09	0.00	-0.07	0.03	0.17	0.00	0.12	-0.05
M32	**0.43**	-0.07	0.06	0.05	-0.05	0.02	-0.07	**0.38**	-0.04
M33	0.12	0.24	-0.11	0.00	0.01	-0.02	0.16	0.29	-0.02
M34	0.24	0.17	-0.01	0.08	0.15	0.01	-0.04	**0.34**	-0.03
M35	**0.40**	0.06	-0.05	0.05	-0.03	**0.36**	0.07	0.00	0.19
M36	0.12	0.27	0.04	-0.04	0.22	0.22	-0.08	-0.04	0.04
M37	0.24	0.07	0.13	-0.05	**0.37**	-0.12	-0.02	0.08	0.09
M38	**0.53**	0.09	-0.07	0.04	-0.14	0.20	-0.20	-0.06	0.02
M39	0.30	**0.36**	-0.09	-0.02	-0.03	0.11	0.05	0.05	**0.37**
M40	-0.04	**0.36**	-0.18	0.11	0.06	**0.33**	-0.03	0.24	-0.06
M41	**0.35**	0.11	-0.09	-0.02	-0.06	**0.40**	0.06	-0.07	0.07
M42	**0.47**	0.04	-0.08	-0.13	0.21	-0.08	0.01	-0.07	0.27
M43	**0.53**	0.01	-0.18	0.08	**0.41**	-0.06	0.12	-0.10	-0.08
M44	0.22	0.17	0.02	0.21	0.31	0.15	-0.06	0.01	0.06
M45	**0.35**	-0.09	0.06	-0.03	0.12	0.11	-0.11	0.01	0.23
M46	0.25	-0.02	-0.01	-0.06	-0.01	0.11	0.11	-0.23	0.28
M47	**0.41**	0.15	-0.08	0.07	0.06	0.08	-0.17	0.23	0.17
M48	0.24	0.14	-0.25	-0.08	-0.02	0.10	0.02	0.13	**0.34**
M49	0.06	0.17	0.09	0.03	**0.32**	-0.16	0.07	-0.01	0.04
M50	0.14	0.20	-0.08	-0.08	**0.39**	-0.01	-0.16	-0.10	0.04

Table 3 showed that, under the PPT version of the test, all the three factors have three or more loadings greater than or equal to 0.32. The results further showed that under the CBT version of MMC test, five factors (factor 1, factor 2, factor 3, factor 4, and factor 5) out of the six extracted factors have three or more loadings greater than or equal to 0.32 while (factor 6) have two loadings greater than or equal to 0.32. These results showed that the PPT and CBT MMCT have three and five well defined factors that underlie them respectively. These results show that the PPT and CBT MMCT measure three and five dominant traits of the examinees respectively. This showed that the test scores validity of MMC test under the PPT mode of administration was different from the test scores validity obtained from the same test under the CBT mode of administrations. The more valid of the PPT and CBT versions of the test was determined. The results are presented in Figure 3

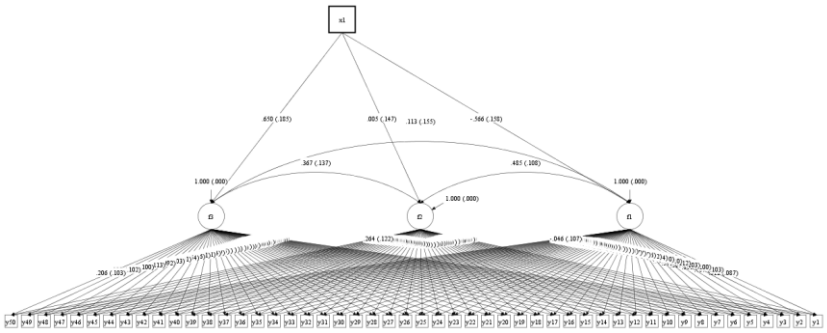

Figure 3: Exploratory Factor Analysis with covariate (university type: federal and state) of PPT version of MMCT

χ^2 (1182) = 1172.316, P = 0.1592); RMSEA= 0.012 (90% CI = 0.000 - 0.020, probability of RMSEA$\leq$0.05 = 1.000), CFI= 0.98, TLI= 0.97

Figure 3 shows a MIMIC model with covariate (type of university, federal was coded 1 and state coded 2) where the covariate was hypothesized not to affect the factors (Factor 1, Factor 2, and Factor 3). The figure shows that adding respondents' type of university to the 3-factor model do not distort the model (χ^2 (1182) = 1172.316, P = 0.1592); RMSEA= 0.012 (90% CI = 0.000 - 0.020, probability of RMSEA $\leq$ 0.05 = 1.000), CFI= 0.98, TLI=0.97). Therefore, the consistency of factors underlying the MMCT among students of state and public universities was assessed. The results are presented in Table 4.

Table 4: Model result of Exploratory Factor Analysis with covariate (university type: federal and state) of PPT version of MMCT

Factor			Estimate	S.E.	Est./S.E.	Two - Tailed p-value
F1	X1	ON	-0.194	0.158	-1.228	0.219
F2	X1	ON	0.005	0.147	0.037	0.97
F3	X1	ON	0.189	0.185	1.024	0.306

Table 4 shows that all the factors extracted to underlie the PPT MMCT among the students of state university were consistent with the factors found to underlie the test among students of federal university (F1 = -0.194, p> 0.05, F2 = 0.005, p>0.05 and F3 = 0.189, p > 0.05).

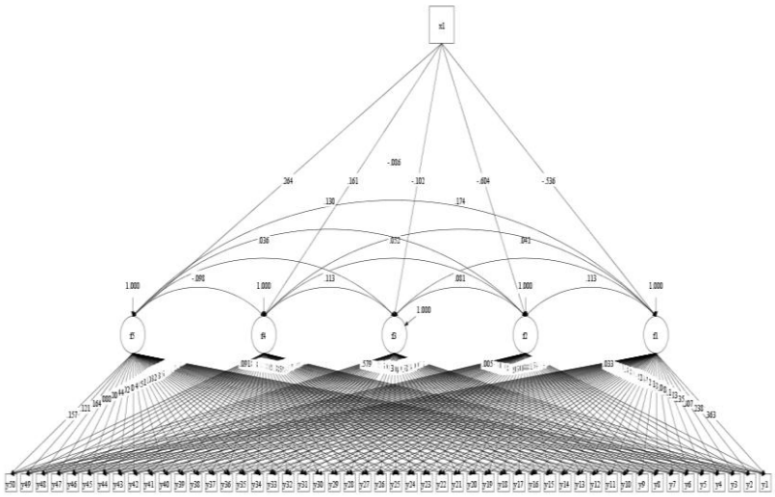

Figure 4: Exploratory Factor Analysis with covariate (university type: federal and state) of CBT version of MMCT

χ^2 (1030) = 1060.644, P = 0.2472); RMSEA= 0.010 (90% CI = 0.000 - 0.019, probability of RMSEA≤ 0.05 = 1.000), CFI= 0.999, TLI=0.976

Table 5: Model result of Exploratory Factor Analysis with covariate (university type: federal and state) of CBT version of MMCT

Item	Covariate		Estimate	S.E.	Est./S.E.	Two-Tailed p-value
F1	X1	ON	-0.536	0.199	-2.69	0.007
F2	X1	ON	-0.604	0.2	-3.018	0.003
F3	X1	ON	-0.102	0.184	-0.558	0.577
F4	X1	ON	0.161	0.186	0.865	0.387
F5	X1	ON	0.264	0.201	1.317	0.188

Table 5 shows that three out of the factors underlying the MMC test, were consistent among students of state and federal universities (F3 = -0.102, p> 0.05 F4 = 0.161, p> 0.05 and F4 = 0.264, p> 0.05). The results further showed that two of the factors extracted to underlie the PPT MMC test among students of state university were inconsistent with the factors found to underlie the test among students of federal university (F1 = -0.536, p <0.05, F2 = -0.604, p < 0.05).

These results showed that the PPT version test consistently measure the same traits among identifiable subgroups of students. The CBT version of the Mathematics measured different traits in the two identified subgroups. This implies that the same Mathematics test had varying validity under PPT and CBT modes of administration. However, the test under the PPT mode of administration was more valid. Therefore, the hypothesis that there is no significant difference in the validity of Mathematics multiple choice test scores based on mode of delivery was rejected. Consequently, there was significant difference in the validity of Mathematics multiple choice test scores based on mode of delivery, with the PPT version of the Mathematics test having higher validity.

Hypothesis Two: *There is no significant difference in the reliability of Mathematics Multiple choice test scores administered via PPT and those administered via CBT.*

Table 6: Comparison of reliability coefficients estimate of MMCT for PPT and CBT mode of administrations

					95% confidence interval	
	Alpha	x^2	df	p-value	lower bound	upper bound
PPT	0.85	65.2729	1	0.0000	0.83	0.88
CBT	0.62				0.55	0.68

Table 6 shows that the Reliability estimate of test scores of the MMCT administered via PPT (α_{PPT} = 0.85) was greater than the reliability estimate of test scores of the same MMCT administered via CBT (α_{CBT} = 0.62). Dependent alpha formula showed that the difference observed in the reliability estimates of the tests' scores in the two modes of administration was significant (χ^2 (1)=65.2729,p<0.05).26

Therefore, the hypothesis which states that there is no significant difference in the reliability of Mathematics MCQ test scores based on mode of delivery was rejected. Hence there was significant difference in the reliability of Mathematics test scores based on mode of delivery. Consequently, MMCT administered via the PPT mode of administration was found to be more reliable than its CBT counterpart.

Discussion of Findings

This finding suggests that MMC test administered via PPT mode had its test scores validity differed significantly from the test scores validity of the same MMCT administered via CBT mode. Where the test scores of Mathematics administered via PPT was more valid than the test scores of MMCT administered via CBT. These findings is supported by Magure, Smith and Brailer (2010), whose findings revealed that there was a significant difference in the performances of students who completed their assessment via CBT and PPT. In contrast to these findings was Adeniji and Ubulom (2016), during their study of senior secondary school students' academic performance in general mathematics using different testing modes

observed that a change in mode of testing does not alter the mean performance of SSCE candidates. The differences observed in the assessment of which mode of administration produced the better test scores validity of Mathematics achievement tests between PPT and CBT could be attributed to many factors prominent, among these factors could be the level of interaction required by the question papers with the PPT mode of administration for Mathematics which may not occur in CBT. In the CBT mode, the level of interaction is limited but in the PPT mode, examinees have the liberty to interact with the papers the way they like within the confines of examination ethics. Hence, Mathematics test being what it is, examinees have to work out the answers using calculators, mathematical sets, four figure tables among others. In the PPT mode, these Mathematics instruments can be used without additional skills. But in the CBT mode, additional skills other than the ones required to provide answers to the questions might be needed. The implications of these findings are that test scores validity of MMCT administered via PPT mode differs significantly from the test scores validity of MMCT administered via CBT.

Furthermore, the study revealed significant difference in the reliability of Mathematics Multiple choice test scores administered via PPT and those administered via CBT. The comparison of the test scores reliability based on mode of delivery PPT and CBT was compared using Mathematics MCQ tests presented to the examinees. This finding suggests that the observed reliability difference in the test scores of Mathematics administered via PPT and CBT was statistically significant. The results revealed that the Mathematics MCQ test administered via PPT was more reliable than its CBT counterpart. The scores of Mathematics test administered via PPT werebetter than the test scores of Mathematics test administered via CBT. The implication of this finding is that the reliability of cognitive test scores administered via PPT and CBT differs significantly from one another. This result is opposed to the results of Öz and Özturan, (2018) in their comparability studies English Foreign Language (EFL) using Spearman Rank order and Mann-Whitney U tests which indicated that test-delivery mode did not have any impact on the reliability and validity of the tests administered in either way that indicated that there was not any significant difference in test scores between participants who took the computer-based test and those who took the paper-based test.

Conclusion

The general goal of all test users is to ensure a better test score validity and reliability. Scores should be able to measure objectively the intended construct it was meant for and consistently too. Therefore, based on the findings of this study, it was concluded that the mode of item administration used for entrance examination cannot be limited to CBT mode of item delivery. Results of this study clearly revealed that mode equivalence with respect to tests score validity and reliability is statistically significant. The factor structure of each of the scores indicated that candidates did not have the same ability/traits to attempt tests using the same mode of item administration. Therefore, the PPT mode of item administration cannot be totally abandoned as a mode of item administration because it is also significantly higher in validity and reliability with MMCT.

Recommendations

Based on the findings of this study, the following recommendations were made.

1. CBT mode of test administration does not totally improve test score validity of Mathematics multiple choice tests for entrance examination into any tertiary institution. Therefore, based on the result it was recommended that as CBT is being used as a mode of Mathematics multiple choice test administration, PPT should still be in use as a mode of test administration especially for Mathematics multiple choice test.
2. Strongly recommend that scores yielded for admission and placement be factor analysed to bring out enough arguments that can support the use of its inferences for appropriate decision making.

References

Adeniji, J. K., &Ubulom, W. J. (2016). Comparative analysis of senior secondary school students' academic performance in general mathematics using different testing modes. *International Journal of Innovative Information Systems& Technology Research, 4*(2), 24-28.

Anastasi, A., & Urbina, S. (2012). *Psychological Testing* (7th Ed.). New Delhi: PHI Learning Private Limited.

Beller, M. (2013).*Technologies in large-scale assessments: New directions, challenges, and opportunities. In: The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*. Netherlands: Springer, 25–45.

Bergstrom, B. (1992). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Boo, J., &Vispoel, W. P. (1998). *Computer versus paper-pencil assessment of educational development: Score comparability and examinee preference*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Choi, S. W., &Tinkler, T. (2002).*Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA

Cone, J.D., & Foster, S.L. (1991). Training in measurement: Always the bridesmaid. *American Psychologist, 46*, 653-659.

Foster, D. F., (2010). Worldwide testing and test security issues: Ethical challenges and solutions. *Ethics & Behaviour*, *20*(3/4), 207-228.

Ghiselli, E.E. (1964). *Theory of psychological measurement*. New York: McGraw-Hill.

Hubley, A. M., & Zumbo, B. D. (2011). *Validity and the Consequences of Test Interpretation and Use*. Springer Science+Business Media B.V. 2011. Published online: 103, 219-230.

Jimoh, B. O., &Omoregie, O. E. (2012).Factors that predispose secondary school teachers to examination malpractice in Edo State, Nigeria.*Review of European Studies,4* (1), 245-254

Johnson, M.,& Green, S. (2004). *On-line assessment: The impact of mode on students' strategies, perceptions, and behaviours*. Paper presented at the annual meeting of the British Educational Research Association, Manchester, Great Britain.

Maguire, K. A., Smith, D. A., &Brailler, S. A. (2010). Computer-based testing: A comparison of computer-based and paper-and pencil assessment. *Academy of Educational Leadership Journal, 14*(4), 117-125.

Mazzeo, J.,& Harvey, A. L. (1988).*The equivalence of scores fromconventional and automated educational and psychological tests: A review of literature (College Board Report No. 88-8)*. Princeton, NJ:Educational Testing Service.

Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement, 62*(4), 647-658.

Oduntan, O. E., Ojuawo, O. O., &Oduntan, E. A. (2015). A Comparative Analysis of Student Performance in Paper Pencil Test (PPT) and Computer Based Test (CBT) Examination System. *Research Journal of Educational Studies and Review, 1*(1), 24-29.

Öz, H., &Özturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies, 14*(1), 67-85.

Pommerich, M. & Burden, T. (2000).*From Simulation to Application: Examinees React to Computerized Testing*. Presented at Annual Meeting of the National Council on Measurement in Education (NCME) 2000.Retrieved from https://www.learntechlib.org/p/88572/.

Russell, W. (2008). *Reliability Generalization(RG) Analysis: The test is not Reliable.* A paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans.

Thompson, B.(2006). *Foundations of behavioral statistics: An insight-based approach*. New York: The Guilford Press.

Wang, H. (2010).Comparability of computerized adaptive and paper-pencil tests. *Test Measurements and Research Services Bulletin, 13*(1), 17.