OPEN ACCESS
Journal of Advances in Education and Philosophy
Abbreviated Key Title: J Adv Educ Philos
ISSN 2523-2665 (Print) |ISSN 2523-2223 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

Measurement Models and the Psychometric Properties of 2016 And 2017 Neco Mathematics Instrument in Calabar Education Zone of Cross River State

Asuquo, Umo Bassey[1*], Emmanuel Joseph (Ph.D)[2], Henry Emmanuel Bassey[3]

[1]Measurement and Evaluation Unit Department of Science Education Faculty of Education Ebonyi State University, Abakaliki, Nigeria
[2]Department of Continuing Education and Development Studies Faculty of Education, University of Calabar, Calabar
[3]Department of Educational Foundations and Childhood Education Cross River State University of Technology Cross River State

DOI: 10.36348/jaep.2022.v06i06.001 | **Received:** 27.04.2022 | **Accepted:** 04.06.2022 | **Published:** 18.06.2022

***Corresponding author:** Asuquo, Umo Bassey
Measurement and Evaluation Unit Department of Science Education Faculty of Education Ebonyi State University, Abakaliki, Nigeria

Abstract

This study examined the basic measurement theories (classical test and item response theories) and psychometric properties of the adopted 2016 - 2017 National Examination Council (NECO) mathematics examination instruments in Calabar Education Zone of Cross River State, Nigeria. The research design adopted for the study was the instrumentation design. The stratified random sampling technique was used to select a sample of one thousand three hundred and fifty one (1,351) students across the local government area in the zone. Two instruments, namely the 2016 and 2017 NECO mathematics multiple choice tests were used for data collection. A trial test of the instruments on 60 comparable non sample group showed reliability (KR_{20}) indices of .91 and .94 for the 2016 and 2017 versions respectively. Difficulty indices ranged between .20 and .85 as well as .10 and .91 respectively for the two years. Discrimination indices were .42<d< 1.00 and .31< d< .89 respectively for the two versions. The final administration yielded data that were collated and analyzed using 2-way repeated measure ANOVA and chi-square statistics. Statistical analysis, showed that the item difficulty and item location, item discrimination indices and 'a' parameter and the option distraction powers of the 2016 and 2017 NECO mathematics multiple choice tests were significantly different. It was concluded that significant difference exist between psychometric properties using the different measurement theories. Based on these findings, it was recommended that NECO should use the item response theory (IRT) in assessment for its relative merit over the CTT among others.

Keywords: Item Response Theory (IRT), Classical Test Theory (CTT), measurement Models, Psychometric Properties, Mathematics Instrument, Calabar Education Zone.

INTRODUCTION

Mathematics is an indispensable subject of study; it plays an important role in forming the basis of all other sciences which deals with the quantification of the material substance of space and time. The knowledge of mathematics is necessary for the study of physical sciences. Mathematical computation and calculations play an important role in architectural activities. With basic mathematical skills we can keep record of our day- to- day expenses and experiences on. Although mathematics is so important, it is one of the least understood academic subject. It is both detested and feared by many students and the yearly failure of students in Senior Secondary Certificate Examination (SSCE) is a cause of serious concern for parents, guidance, counselors, teachers, administrators and government. It is sad to note that for some years now the performance of students writing SSCE mathematics is poor. The scenario is becoming worrisome to education stakeholders as nobody knows where to place the blames.

Going down memory lane, in 2002 at NECO's maiden edition, most candidate passed their registered subject including English language and mathematics which led to the public accusing the examination body of awarding cheap results to students, but after that year the performances of candidates in mathematics and English have subsequently deteriorated every year, down to 11.3 percent in 2011\2012 result, a situation which calls for concern among stakeholders. A statistical breakdown of the awful trends revealed that

in the 2009 May/June NECO senior secondary school certificate examinations, only 12% candidates recorded the mandatory five credits including English and mathematics. Details of the result further showed that only 4,223 candidates out of the 234,683 candidates had five credits that included mathematics and English language representing 1.8%. In the June/July 2010 secondary school certificate examination conducted by NECO, 79% failed to get credit pass in Mathematics. During NECO June/July 2011 SSCE school based examination, less than 25% of the 1,160,561 candidates had passed at credit level including the two core subject of Mathematics and English language across the country. This is quite unacceptable. The government of Cross River State increased budgetary allocation from N16.5m to N60.0m in 2007 as a mark of concern for education coupled with incentives accruable to science teachers. Apart from N40 million spent by government for payment of WEAC examination fees of all students of Cross River State origin, N80 million was also spent for supply of science and technical equipment to secondary schools in the state.

Some people attribute student's low performance in Mathematics to low accountability, poor parental monitoring, low reading skills, poor learners attitude towards Mathematics teachers attitude and instructional method what of the assessment instruments used? Does it fit any measurement model? There is also lack of confidence in public examinations because of continuous failure. Where lies the problem of poor Mathematics achievement? To what extent does the Mathematics multiple choice test items constructed by NECO satisfy the criterion of quality as epitomized by the psychometric properties of test instrument, using CTT and IRT models. This study therefore tries to find out the psychometric properties of the 2016 and 2017 NECO Mathematics examination items/tests in Calabar Education Zone of Cross River State using measurement theories.

Purpose of the study

The purpose of the study was to find the psychometric properties of the 2016 - 2017 NECO Mathematics tests using the measurement theories. The study:

i. Compared the item difficulty in the measurement theories for the 2016 and 2017 NECO Mathematics multiple-choice test.
ii. Compared the measurement theories and year of examination on item discrimination indices of 2016 and 2017 NECO Mathematics multiple-choice test items.
iii. Determined the association between the option distraction power of the 2016 and 2017 NECO Mathematics multiple-choice test items across content areas.

Research questions

For the study to have a focus, the following research questions were formed.

i. How do measurement theories and year of test influence item difficulty of 2016 and 2017 NECO Mathematics tests differ?
ii. How do measurement theories and year of test influence item discrimination of 2016 and 2017 NECO Mathematics tests differ?
iii. How do the option distraction powers of the 2016 and 2017 NECO Mathematics multiple-choice test items across content areas relate?

Statement of hypotheses

The following null hypotheses were formulated to guide the study:

i. There is no significant influence of measurement theories and year of tests on item difficulty of 2016 and 2017 NECO Mathematics tests.
ii. There is no significant influence of measurement theories and year of tests on item discrimination of 2016 and 2017 NECO Mathematics tests.
iii. There is no significant association between the option distraction powers of 2016 and 2017 NECO Mathematics multiple choice test items across content areas.

METHODOLOGY

Research design: The researcher adopted the instrumentation design for the study. This is because the research seeks to investigate measuring instruments to ascertain some level of certainty concerning their properties. Instrumentation design according to Mehrens and Lehman (1996) in Okeme (2009) is a type of design which aims at developing and certifying the efficacy of an instrument for the measurement of a given behavior or construct.

Area of the study: The research area for this study was Calabar Education Zone of Cross River State, Nigeria. The Calabar Education Zone is located in the Southern part of Cross River State, Nigeria and lies between latitude $4^0 28'$ and $6^0 35'$ North of the equator and longitude $7^0 50'$ and $9^0 28'$ East of the Greenwich meridian, with an area of 20,156km^2 (National Board of Technical Education, 2010). Calabar Education Zone is politically known as Southern Senatorial district. The zone is bounded in the North by Yakurr and Abi Local Government Areas, in the South by the Bonny and Atlantic Ocean, in the East by the republic of Cameroon and in the west by Ebonyi, Abia and Akwa Ibom States respectively. The Zone is made up of seven (7) local government areas namely; Akamkpa, Akapbuyo, Bakassi, Biase, Calabar Municipality, Calabar South and Odukpani. The inhabitants of this study area especially those in the rural areas like Akamkpa, Akpabuyo, Bakassi, Biase and Odukpani are predominantly farmers, fishermen and traders. Those in the urban areas are mostly civil and public servants with few in private outfits.

The education zone has educational institutions like the University of Calabar, Cross River University of Technology, Calabar (with 3 campuses outside the zone), Cross River State College of Education, Akamkpa, State College of Health Technology, Calabar and the study center of National Open University of Nigeria in Calabar. Secondary schools of note include Hope Waddle Training Institute, Calabar, Government Secondary Schools at Akamkpa and Greek Town and St. Patrick's College Ikot Ansa, Calabar.

Population of the study: The population of the study consisted of 7,591 senior secondary three (SSIII) students in Calabar Education Zone of Cross River State Nigeria in the 2018/2019 academic year (Department of Planning, Research and Statistics, Secondary Education Board Calabar, 2018). This population spread across eighty-nine (89) public secondary schools. This population comprises of both male and female students. The population distribution of subjects is as shown in Table 1.

Sample and sampling procedure: The sample comprised 1,351 senior secondary three (SSIII) students drawn by stratified random sampling technique from 20 out of the 89 secondary schools in the education zone. The student sample size was 17.78% of the total population 7,591. According to Sandelowski (1995) a good maximum sample size is usually around 10% of the population. This is as shown in population and sample distribution of student subjects in Table 1.

The sampling technique adopted in selecting the sample for this study was stratified random sampling technique. The stratification was done base on local government areas, within the zone. Next, the number of schools was selected in proportion to population of schools and of students. Then simple random sampling technique was used to select the student in each local government area. To do this, the register of students in each of the sampled schools was used and all the student in the class register form the sample of the study. This was done in each of the sampled schools.

Table-1: Population and sample distribution of student (subjects)

L.G.A	No. of schools	No. of schools selected	Population of students	No. of students sampled		Total No. of students sampled
				Male	Female	
Akamkpa	20	3	634	53	60	113
Akpabuyo	7	4	364	25	39	64
Bakassi	3	1	81	5	9	14
Biase	17	1	512	41	51	92
Calabar Municipality	17	5	3,284	235	350	585
Calabar South	8	4	2,095	180	193	373
Odukpani	17	2	621	40	70	110
Total	89	20	7,591	579	772	1,351

Instrumentation

The instruments for this study was Mathematics multiple-choice test items of both 2016 and 2017NECO Examinations. The instruments were fully adopted for the study by the researcher. The instrument was constructed by the examination body NECO (National Examination Council). There were sixty items each of the 2016 and 2017 NECO examination being multiple choice items of five point response options. Candidates were required to encircle the letter bearing the answer in each item.

Validity and reliability of the instrument

The instrument is a standardized test, thus its validity was assumed. However, face validity was further explored by giving the instrument with the title, purpose and research questions to two independent experts in Cross River University of Technology, Calabar for vetting. One of the experts was a senior lecturer in mathematics education and the other, a professor of educational measurement and evaluation. Their independent comments were used to correct typographical and other errors to ensure test quality.

Since the instrument is a standardized test, the reliability was already established by the examination board. However, the researcher carried out a confirmatory reliability test using Kuder Richardson 20 (KR_{20}) reliability method to determine the reliability of both instruments. This was done by administering both instruments to sixty (60) SSIII students selected from another education zone. The Students were given the NECO 2016 and 2017multiple choice test items simultaneously. The reliability coefficients for both instruments were computed to be .91 and .94 respectively. The difficulty indices range from 0.20 to 0.85 and 0.00 to 0.91 respectively for 2016 and 2017. The discrimination indices range from 0.42 to 1.00 and 0.31 to 0.89 for 2016 and 2017 instruments respectively.

Procedure for data collection

The researcher visited each of the selected schools for self and mission introduction and requested for permission to administer the instruments. In each school, the researcher first reported to the principal,

introducing herself with student identity card as evidence. The purpose of the visitation was also made known, with a verbal appeal for assistance. Where the principal was not available, any of the vice principals was consulted for this purpose. In each school visited, the principal or vice showed readiness to help. Assisted by the vice principal and in some cases the subject teacher, the researcher administered the instruments to the students of each sampled school in their respective classrooms. The instruments were administered in one piece. The assistant in each class addressed the students and asked them to co-operate by completing the instrument diligently as it will help them to improve their capacity in mathematics and also help the school to be listed as a cooperating institution. The subjects were advised to start with instrument one (2016 NECO Mathematics multiple choice test) to be completed in one hour, before proceeding to instrument two (2017 NECO Mathematics multiple choice test) also scheduled to take one hour.

Procedure for data analysis

The data was keyed into the winsteps software (version 4.0.0), and the SPSS software (version 22.0). These software were used to analyze the data. Specifically, ANOVA was used to analyze hypotheses one and two, while chi-square was used to analyze hypotheses three.

RESULT AND DISCUSSION

This section shows hypothesis-by-hypothesis presentation of the results of data analysis. Each of the hypotheses were restated and the results presented and interpreted. All hypotheses were tested at .05 level of significance with applicable degrees of freedom.

Hypothesis one

There is no significant influence of measurement theory and year of test on the item difficulty in the NECO mathematics test.

The independent variables in this hypothesis are the measurement theory which is categorical (CTT and IRT) and year of test which is categorical (2016 and 2017) while the dependent variable is difficulty index which is continuous. To test the significant influence of measurement theory and year of test on item difficulty for the NECO mathematics test, 2-way repeated measure ANOVA was applied with results as shown in Table 2.

Table-2: 2-way repeated measure ANOVA of significant influence of measurement theory and year of test on item difficulty in NECO mathematics test.

Source factor		SS	df	MS	F	Prob.
Year	Sphericity assumed	2.299	1	2.299	17.792	.000
	Greenhouse-Geisser	2.299	1.000	2.299	17.792	.000
	Huynn feldt	2.299	1.000	2.299	17.792	.000
	Lower bound	2.299	1.000	2.299	17.792	.000
Error (year)	Sphericity assumed	7.624	59	.129		
	Greenhouse-Geisser	7.624	59.000	.129		
	Huynn feldt	7.624	59.000	.129		
	Lower bound	7.624	59.000	.129		
Model theory	Sphericity assumed	.010	1	0.10	.080	.048
	Greenhouse-Geisser	.010	1.000	0.10	.080	.048
	Huynn feldt	.010	1.000	0.10	.080	.048
	Lower bound	.010	1.000	0.10	.080	.048
Error (theory)	Sphericity assumed	7.531	59	1.28		
	Greenhouse-Geisser	7.531	59.000	1.28		
	Huynn feldt	7.531	59.000	1.28		
	Lower bound	7.531	59.000	1.28		
Year *Model theory	Sphericity assumed	.037	1	.037	.318	.035
	Greenhouse-Geisser	.037	1.000	.037	.318	.035
	Huynn feldt	.037	1.000	.037	.318	.035
	Lower bound	.037	1.000	.037	.318	.035
Error (year *model theory)	Sphericity assumed	6.903	59	.117		
	Greenhouse-Geisser	6.903	59.000	.117		
	Huynn feldt	6.903	59.000	.117		
	Lower bound	6.903	59.000	.117		

Table 2 shows the summary of repeated measures ANOVA effect with corrected F-values. The table is split into sections for each of the effect in measurement theory, year of test and their associated error terms. From the table, it was observed that there was a significance in year of examination (prob. = .000), measurement theory (prob. = 0.048) and the interaction yr. *model (prob. = 0.035) this shows that there is a significant influence of measurement theory, year of examination, as well as interaction (year *theory) and the item difficulty of the NECO Mathematics tests. Therefore, the null hypothesis of no significant influence is rejected and the alternative hypothesis upheld. This means that difficulty is not the same for the different measurement theories and year of examination that is different theories may produce different item difficulties just as different years may. The main effect for both the year and the model theory as well as the interaction (year *model) were significant indicating that for each examination year the item difficulty differ from the other year. Item difficulty also differs with measurement theory and the combined effect is also significant. This may be so owing to both differences in theoretical positions, perspectives and approaches. Whereas CTT sees item difficulty as a measure of the proportion of testees that found the item easy, IRT sees it as measure of item location in the item response function (IRF). It is a point in the ability scale where testees have 50 percent probability of answering the item correctly. The analysis also revealed that the item difficulty indices of 2016 and 2017 range from 0.18 to 0.38 and 0.08 to 0.47 respectively and the item location of 2016 and 2017 range from 0.4 to 0,8 and 0.5 to 0.9 However it was discovered that item location has the higher mean b-value of 0.14 in the 2017 test while the classical item difficulty had a higher mean value of 0.32 in the 2016 test. This shows that viewed from IRT the 2017 mathematics test item had higher difficulty values than the 2016 test items but viewing from CTT 2016 test items had higher difficulty (easiness) values.

Results of the study corroborate the work of Michael (2005) who conducted a study on the influence of attitude on item difficulty and other item parameters using 150 third grade high school mathematics students. Results showed that response to the test and item difficulty differed significantly between high and low level attitude subjects (students) though the study relate item difficulty with attitude of testee, it differs from the current one in the sense that it focuses on one measurement theory (CTT) only whereas the present study is a comparative study of item difficulty viewed from two perspectives (CTT and IRT).

These findings agree with Sotaridona, Pornel and Valleyo (2003) who in their study observed that two examinees may have the same person score but the difficulty of the items may be different and hence the ability of the examinees in item response theory may be different and item response pattern also different. Also, the works of Beck (1978), Ekpenyong (1991) and Standley and Hopkins (1972) are concurrent pointing out that the difficulty level of an item has a significant influence on the performance of the testees, thereby influencing the other test item parameters.

On the other hand, the finding disagrees with the study of Bandele and Adewale (2013) that compared the item level of WAEC, NECO and NABTEB mathematics achievement examination in Nigeria. Using analysis of variance (ANOVA), it was revealed that there was no significant difference in the item difficulty level of WAEC, NECO and NABTEB Mathematics achievement examination. This may have been so because the researcher used CTT model to assess the psychometric properties of the items, whereas, in the present study item difficulties are compared from the perspectives of two measurement models.

The implication of this finding is that since item difficulty is relevant in determining whether students have learned the concept being tested, a teacher needs to ensure that when constructing a test, the item must neither be too easy nor too difficult. If items are too difficult almost everyone gets the item wrong, suggesting guessing and if they are too easy, almost everyone gets it right and this makes it difficult for the item to discriminate students who know the tested concept from those who do not know. Test items with poor difficulty index should be reviewed as difficulty is the characteristic of both the item and the sample taking the test.

Hypothesis two: There is no significant influence of measurement theory and year of test on the item discrimination in the NECO mathematics test.

The independent variables in this hypothesis are the measurement theory which is categorical (CTT and IRT) and the year of test which is also categorical (2016 & 2017), while the dependent variable is discrimination index which is a continuous variable. To test the null hypothesis of significant influence of measurement theory and year of examination on item discrimination, the 2-way repeated measure ANOVA was applied with results as shown in Table 3.

Table-3: 2-way repeated measure ANOVA of significant influence of measurement theory and year of test on item discrimination in NECO mathematics test.

Source factor		SS	Df	MS	F	Prob.
Year	Sphericity assumed	31.081	1	31.081	790.422	.000
	Greenhouse-Geisser	31.081	1.000	31.081	790.422	.000
	Huynn feldt	31.081	1.000	31.081	790.422	.000
	Lower bound	31.081	1.000	31.081	790.422	.000
Error (year)	Sphericity assumed	2.320	59	.039		
	Greenhouse-Geisser	2.320	59.000	.039		
	Huynn feldt	2.320	59.000	.039		
	Lower bound	2.320	59.000	.039		
Model (theory)	Sphericity assumed	.029	1	.029	.819	.019
	Greenhouse-Geisser	.029	1.000	.029	.819	.019
	Huynn feldt	.029	1.000	.029	.819	.019
	Lower bound	.029	1.000	.029	.819	.019
Error (theory)	Sphericity assumed	2.100	59	.036		
	Greenhouse-Geisser	2.100	59.000	.036		
	Huynn feldt	2.100	59.000	.036		
	Lower bound	2.100	59.000	.036		
Year *Model (theory)	Sphericity assumed	.227	1	.227	7.109	.002
	Greenhouse-Geisser	.227	1.000	.227	7.109	.002
	Huynn feldt	.227	1.000	.227	7.109	.002
	Lower bound	.227	1.000	.227	7.109	.002
Error year *model (theory)	Sphericity assumed	1.887	59	0.32		
	Greenhouse-Geisser	1.887	59.000	0.32		
	Huynn feldt	1.887	59.000	0.32		
	Lower bound	1.887	59.000	0.32		

Table 3 shows the summary of repeated measures ANOVA effect with corrected F-values. The table is split into sections for each of the effect in measurement theory, year of test and their associated error terms. From the table, it was observed that there was a significance in year of examination (prob. = .000), measurement theory (prob. = .01) and the interaction yr. *model (prob. = .00) this shows that there is a significant influence of measurement theory, year of examination, as well as interaction (year *theory) and the item discrimination in the NECO mathematics tests. Therefore, the null hypothesis of no significant influence is rejected and the alternative hypothesis upheld.

The results of this study showed that there was a significant influence of the measurement theory and year of test on the item discrimination in NECO Mathematics test. Main effect of measurement theory and year of test was each significant as well as interaction. This means that item discrimination is not the same in CTT and IRT, neither is it the same for 2016 and 2017. The combined effect of theory and year of test is also significant. Item discrimination in the lenses of CTT and IRT produce different values. This is so because the sensitivity of the two measurement frameworks in pulling apart the bright and the dull students is not the same. It was also revealed that from the CTT viewpoint the item discrimination indices of 2016 and 2017 NECO Mathematics tests ranges from 0.12 to 0.58 and 0.10 to 0.39 respectively while the 'a' parameters of 2016 and 2017 test items range from 2.1 to 1.4 and 0.66 to 1.15 respectively. Using IRT test items of 2016 discriminated more than 2017. In CTT values of item discrimination fall usually within the interval $0 \leq d \leq 1$, while in IRT d values go beyond that range. This is because the concept of discrimination differs across the theories. In CTT discrimination is ability of item to distinguish between the high and low scorers while for IRT it is item ability to distinguish on the item response function (curve) testees with ability above item location and those below it.

This finding corroborates Bichi (2015) who undertook a comparative analysis of classical test theory and item response framework using a sample of 530 students. The results of the study showed that there is a significant difference between the item discrimination of items derived from CTT and IRT models. However, Abedalazez (2011) carried out a study to compare the CTT and IRT approaches in analyzing item characteristics. The developed instrument was administered to a sample of 602 grade 1 students, the data gathered was analyzed and the results revealed that there was a significant difference between the item difficulty and item discrimination of the item parameters using the IRT and CTT models.

However, the finding of this study agrees with Hotni (2006) who averred that discrimination index is a useful measure of item quality whenever the purpose of the test is to produce a desirable score spread reflecting

differences in students achievement. This enables distinctions to be made among the performance of respondents, particularly as discrimination measures the extent to which item responses distinguish between individual examinees who have higher overall scores on test and those that have lower overall scores. The finding also agrees with the earlier findings of Abel and Frisbre (1991) who contended that a test with a high average discrimination index is always better as such a test will produce a more reliable zone than the other.

The finding of this study however disagrees with that of Ekpenyong (1991) in a study of influence of psychological variables on classical test item parameters by item interaction analysis, where he discovered that classical test item parameters and item discrimination do not differ significantly across examinees. This may be attributed to weaknesses in examination administration.

The higher the discrimination index, the better the items because such a value indicates that the item discriminates in favour of the upper group, but when more students in the lower groups than in the upper group select the right answer to an item, the item actually has negative validity, the item is not only useless but is actually serving to decrease the discrimination of the test. There is need therefore for instructors to ensure that when constructing tests, the items should not be ambiguously worded as this may confuse the test takers and items with ambiguity should be redressed to function adequately.

Hypothesis three: There is no significant association between the option distraction powers of the 2016 and 2017 NECO mathematics multiple choice items across content areas.

The independent variable in this hypothesis is the content areas which is categorical, while the dependent variable is option distraction powers which is also categorical (positive and negative). This is to check if the option discrimination of the items in 2016 & 2017 and the content areas have a strong association. To test the null hypothesis of significant association between content areas and the option distraction powers, the contingency chi-square independence test was computed, with results as shown in Table 4.

Table-4: Contingency chi-square coefficients test of the option distraction power of 2016 and 2017 multiple choice test. N = 1351)

Year		Frequencies		df	x^2	sig
	Content area	No of + distracting options	No of - distracting options			
2016	Numbers/numeration	67 (67.9)	2 (1.1)			
	Algebraic processes	39 (39.33)	1 (.63)	5	2.48*	0.020
	Geometry	48 (47.2)	0 (.76)			
	Mensuration	28 (27.53)	0 (.45)			
	Trigonometry	57 (57.13)	1 (.92)			
	Probability	8 (0.77)	0 (.13)			
2017	Numbers/numeration	67 (64.88)	1 (3.12)			
	Algebraic processes	66 (68.7)	6 (3.3)			
	Geometry	38 (38.2)	2 (1.83)	5	6.09*	0.014
	Mensuration	12 (11.45)	0 (0.55)			
	Trigonometry	34 (34.35)	2 (1.65)			
	Probability	12 (11.45)	0 (0.55)			

*P<.05; critical x^2 value = 2.39

In Table 8 since the calculated x^2 values of 2.48 and 6.09 are respectively greater than the critical x^2 value of 2.39 at 0.05 level of significance and 5 degree of freedom for the two tailed test, it follows that there are significant associations between the option distraction powers of 2016 and 2017 NECO mathematics test and the respective content areas.

Results in the analysis of data for hypothesis three showed significant relationship in the frequencies observed and expected of the option distraction powers across the mathematics content areas for both 2016 and 2017 tests. That is there is a significant association between the option distraction powers across content areas. The frequencies of positive and negative distractors across the mathematics content areas in the two examinations are significant (strong association). This means that there is a strong significant association between option distraction powers of NECO 2016 and 2017 across content areas. The finding of the study also revealed that in 2016, 236 (98.4%) options had positive option distraction power while 4 (1.6%) had negative distraction power while in 2017 229 (95.4%) options had positive option distraction power and 11(4.6%), had negative option distraction power. This implies that in 2016 only 4 options attracted more of bright students than dull students while in 2017, 11 options attracted more of bright students than dull students. The items with these options need to be reviewed or totally removed from the pool of options. Items 1, 33, 58 in

2016 NECO need to be revised and items No 6, 8, 24, 27, 30, 46, 47, 54 in 2017 need to be reviewed.

This finding corroborates Ifewulu, Onjaja & Ojeunder (2013) who investigated the empirical relationship between the CTT and IRT eras in JAMB, using a sample size of 100,000 students. The data collected was analyzed using ANOVA, chi-square, with the xcalibre and Microsoft excel. The study found a significant improvement in the performances of candidates who repeated the examination in 2013 over those of the 2012 and significant difference between the item parameters using CTT and IRT models.

This finding agrees with Rodriguez (2005) who carried out a meta-analysis to synthesize the results from 27 research studies and the researcher concluded that providing more options does little to improve item quality and test score statistics and typically results in implausible distracters. This finding also corroborates the work of Ramos and Stern (1973) that compared 4 and 5 option test of reading spanish as foreign languages.

Also, in line with the finding of this study Haladyna and Downing (1993) found that approximately two-thirds of all four-option items they reviewed had one or two functioning distractor and none of the five option items had four functioning distractors. This is because it is difficult for teachers to develop three or more equally plausible distractors, additional distractors are often added as fillers.

Analyzing the distractors (incorrect alternatives) is useful in determining the relative usefulness of the decoys in each item. Items should be modified if students consistently fail to select certain multiple-choice alternatives. The alternatives are totally implausible and therefore of little use as decoys in multiple choice items. Distractors should be carefully examined when items show large positive discrimination values as the discrimination values for distractors should be lower and preferably negative.

CONCLUSION AND RECOMMENDATIONS

Based on the findings, it was concluded that psychometric properties depend on the measurement theory employed. Measurement theory and year of test significantly influenced the item difficulty on the NECO mathematics test. Similarly, item discrimination was significantly influence by the measurement theory and the year of test. There was a significant association between option distraction across the content areas for the 2016 and 2017 NECO mathematics multiple choice items. The guessing parameters also differed significantly with the years of study. There was a significant influence of measurement theory and year of test on item reliability on NECO mathematics examination papers.

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

i. Psychometricians should make construct explication a priority in estimation of item properties. Psychometric properties of items and person should always be interpreted relative to the measurement theory used.
ii. Items should always be constructed to have optimum sensitivity and ability to distinguish between bright and less bright testees.
iii. Item writers and educators should be retrained through workshop, seminar and conferences to improve on their skills to enable them construct items with higher distraction ability. Plausibility of options improves item quality.

REFERENCES

- Abedalazez, N. (2011). The comparison of CTT and IRT approaches in analyzing item characteristics. *Statistical Journal*, *14*(2); 23 – 27.
- Abel, D. (1990). Comparison of the item discrimination and item difficulty of the quick mental aptitude test using CTT and IRT methods. *The International Journal of Educational and Psychological Assessment*, *1*(1), 12 – 18.
- Bandele, S. O. & Adewale, A. E. (2013). Comparative analysis of the item difficulty levels of WAEC, NECO and NABTEB mathematics achievement examinations. *Mediterranean Journal of Social Science*, 4(2), 761 – 705.
- Bichi, A. A. (2015). Evaluating the quality of chemistry achievement test. A comparative analysis of classical test theory and item response theory frameworks. *Education Journal*, *12*(2), 32 – 40.
- Ekpenyong, M. (1991). Psychometric properties of 1981 – 1989 WAEC mathematics instrument. An unpublished PGDE thesis submitted to the Institute of Education, University of Ilorin.
- Haladymna, T. M. & Downing, S. M. (1993). How many options is enough for a choice test item? *Educational psychological measurement*, *53*(4), 999 – 1010.
- Hotni, C. E. (2006). Personal elements in teachers' works. *Journal of Educational Research*, 12, 49 - 55.
- Ifewulu, C. B., Onoja, G. O. & Ojenude, D. E. (2013). Comparative analysis of candidates performance in the pre and post IRT eras in Jamb: case study is the use of English in 2012 and 2013 UTME. *International Journal of Education science*. 13(3). 7 – 19.
- Mehrens, C., & Lehman, P. (1996). Measurement and evaluation in education and psychology. *Journal of Educational Measurement*, *22*(4), 316 – 318.
- Micheal, C. (2005). "Correlation between attitudes and mathematics achievement". *Journal of Educational Psychology*, *51*(2), 16 – 18.

- Okeme, I. (2009). Development and validation of psycho-productive skill items for measuring performance of students in agricultural science in secondary schools in Kogi State. Unpublished Ph.D Thesis of University of Nigeria, Nsuka.
- Ramos, L., & Stern, M. B. (1973). Web-base assessment. Two UK initiatives, Australia. Retrieved from http://www.auscreb scu.edu/aw2k/papers/index.html.
- Rodriquiz, M. C. (2005). Three options are optimal for choice items: A meta-analysis of 80 years of research. *Educational measurements*: issues and practice, 24(2), 3-13.
- Sandelowski, M. (1995). Sample size in qualitative research. *Research in nursing and health*, 18, 179 – 183.
- Sotaridona, M., Pornel, B. & Valleyo, M. (2003). Effect of examinee ability on test equating in variance. *Applied Psychological Measurement*, *12*, 69 – 82.
- Stanley, E., & Hopkins, L. (1972). *Data collection design and linking procedures:* New York – NY Springer Verlag.