

TITLE PAGE

**DETECTION OF DIFFERENTIAL ITEM FUNCTIONING
IN MATHEMATICS MULTIPLE CHOICE EXAMINATION
ADMINISTERED BY THE NATIONAL EXAMINATION
COUNCIL (NECO) IN ABUJA**

BY

**HOGAN-BASSEY, VICTORIA INYANG
REG. NO: PG/M.ED/07/43861**

**AN M.ED PROJECT SUBMITTED TO THE DEPARTMENT
OF SCIENCE EDUCATION, FACULTY OF EDUCATION.
UNIVERSITY OF NIGERIA, NSUKKA IN FULFILLMENT
OF THE REQUIREMENT FOR THE AWARD OF MASTER
OF EDUCATION (MEASUREMENT AND EVALUATION)**

MAY, 2011

APPROVAL PAGE

THIS PROJECT HAS BEEN APPROVED FOR THE FACULTY
OF EDUCATION, UNIVERSITY OF NIGERIA, NSUKKA.

BY

DR. B. C. MADU
SUPERVISOR

INTERNAL EXAMINAR

EXTERNAL EXMINAR

DR.(MRS) C.R.NWAGBO
HEAD OF DEPARTMENT

PROF. S. A. EZEUDU
DEAN FAULTYOF EDUCATION

CERTIFICATION

HOGAN-BASSEY, VICTORIA INYANG, A post graduate master in education student in the Department of Science Education, Faculty of Education, University of Nigeria, Nsukka with Registration number PG/M.ED/07/43861 has satisfactorily completed the requirement for research work for the degree of master in education (MEASUREMENT AND EVALUATION).

The work embodied in this project is original and has not been submitted in part or full for any diploma or degree of this or any other university.

VICTORIA HOGAN-BASSEY
CANDIDATE

DR. B.C. MADU
SUPERVISOR

DEDICATION

This work is dedicated to the Holy Trinity:

God the father

God the son and

God the Holy Spirit

ACKNOWLEDGEMENTS

The researcher is in indeed very grateful to the Almighty God her Father in heaven whose love, protection, encouragement and provision has led her through to this point.

Special thanks go to her supervisor, Dr. B.C. Madu whose fatherly concern; prompt supervision was a major force that propels the completion of this study.

Expression of sincere thanks to prof. B.G. Nworgu, who at the beginning of this task advice me on how to get materials and the path to follow. I am also grateful to Dr. (Mrs.) J.U. Eze, Mr. U.J. Utibe, and Mr. John Agah all of faculty of education, university of Nigeria, Nsukka, who read through the work and make corrections and suggestions accordingly.

The researcher is deeply grateful to the registrar National Examination Council (NECO), the Head of Department, Quality Assurance, Dr. Nwankpa, and Dr. Dayo of the computer unit of NECO for the release of data for this study.

Particularly the researcher is indebted Dr. Umoinyang of the University of Calabar for supplying her with material to back up the literature. I appreciate the contribution of My Head of Department Mrs Fajuyi, Government Secondary School Garki-Abuja, Engr. Olaitan Solarin, Ms Mbachirin Mbaumbur.

Special appreciation to Federal Capital Territory (FCT) Secondary Education Board (SEB) for giving me the approval by way of study leave to embark on this study. I appreciate Mr. Peter Olokun the secretary of (SEB) for his fatherly advice and love. My thanks also go to Education Resource Centre (ERC) Abuja, for supplying me with information on the population of students in the FCT. Many thanks to both the academic and non-academic staffs of the department of science education university of Nigeria, Nsukka for their assistant during the course of my study.

Finally I appreciate my mother and siblings for showing me love and concern in the course of this program and to all who, time and space has not permitted me to mention. Above all may the name of the Lord be glorified both now and for evermore. Amen.

TABLE OF CONTENTS

Title page	i
Approval page	ii
Certification	iii
Dedication	iv
Acknowledgement	v
Table of Contents	vi
Appendices	ix
List of Tables	x
List of Figures	xi
Abstract	xii

CHAPTER ONE: INTRODUCTION

Background to the problem.....	1
Statement of problem.....	8
Purpose of study.....	9
Significance of the study.....	9
Scope of the study.....	10
Research questions	11
Hypotheses.....	11

CHAPTER TWO: LITERATURE REVIEW

Conceptual Framework.....	13
The Concept of mathematics.....	13
Gender.....	14
Location.....	14

Academic Functioning/achievement.....	15
Item bias Item fairness and Item Impact	17
Detection of Differential Item Functioning.....	21
Difficulty index.....	25
Item discrimination.....	26
Multiple Choice Items.....	26
Statistical methods use for detecting differential item functioning...	27
Uses of Differential Item Functioning.....	35
Theoretical Framework.....	36
Classical Test Theory.....	36
Item Response Theory.....	38
Empirical Studies.....	42
Test and Differential item functioning.....	42
Gender and differential item functioning.....	46
Location and differential item functioning.....	52
Summary of Literature	53

CHAPTER THREE: RESEARCH METHOD

Research design	55
Area of study	55
Population of study.....	56
Sample and sampling technique.....	56
Instrument for data collection.....	57
Validity of Instrument.....	57

Reliability of Instrument.....	57
Method of Data Collection.....	58
Data analysis.....	58

CHAPTER FOUR: RESULTS

Research Question One.....	59
Hypothesis One.....	68
Research Question Two.....	68
Hypothesis Two.....	77
Summary of Results.....	78

CHAPTER FIVE: DISCUSSIONS, CONCLUSIONS, AND SUMMARY

Gender and Differential Item Functioning.....	79
Location and Differential Item Functioning	81
Conclusion.....	82
Implications.....	83
Recommendations.....	83
Limitation of the study.....	84
Suggestions for Future Research.....	84
Summary of Study.....	85
REFERENCES.....	86

APPENDICES

APPENDIX A: National Examination Council Senior Certificate
Examination Mathematics Objective Test 2008.....
95

APPENDIX B: Data Table for the calculation of Reliability.. 108

APPENDIX C: Tables showing the distribution of ability by gender and
location..... 112

APPENDIX D: Letter of Request for person-by-item response pattern of
examinees in the Federal Capital Territory, Abuja on mathematics multiple
choice examination administered by the National Examination Council
(NECO) 114

LIST OF TABLES

TABLES	PAGES
1. Illustration of analysis of bias with M-H statistics	28
2. Scheuneman modified chi-square test for computation of Differential Item Functioning	30
3. Computation of Differential Item Functioning using Distracter method	32
4. Distribution of schools and number of examinees	56
5. Summary of M-H procedure for boys and girls group at the NECO 2008June/July multiple choice mathematics examination	59
6. Summary of the analysis of gender DIF in NECO examinations using the Scheuneman chi-square statistics	61
7. Summary of Transform Difficulty Method for male and female Examinees	64
8. Items identified as exhibiting gender DIF by the three detection methods	68
9. M-H procedure for NECO mathematics by school location	69
10. Scheuneman chi-square analysis by school location	70
11.Transform Item Difficulty analysis by school location	74
12. Summary of the three methods of detecting DIF by location	77

LIST OF FIGURES

FIGURES	PAGE
1. Item plot of the Transform Item Difficulty for boys and girls	67
2. Item plot of the Transform Item Difficulty for urban and rural schools	76

ABSTRACT

This study sought to detect Differential Item Functioning in National Examination Council (NECO) Senior School Certificate Examination (SSCE) in the 2008 June/July multiple choice mathematics examination papers 1. The research was carried out on one thousand seven-hundred and twenty (1720) examinees. A random sample of 14 out of 43 schools were used. 9 Urban schools with 1119 examinees and 5 Rural schools with 849 examinees, also 849 boys and 871 girls comprised the sample. The Mantel-Heanszel procedure, Scheuneman chi-square test and the Transform Item Difficulty methods were used for the analyses. Mantel-Heanszel procedure did not significantly function differentially in terms of gender, but it however exhibited 14 items as functioning differentially with respect to location. Scheuneman modified chi-square test flagged up 46 items as functioning differentially in terms of gender and 41 items differentially function with respect to location. Transform Item Difficulty methods detected 2 items as differentiating between gender and 11 items significantly differentiating with respect to location. The three methods exhibit significant different with respect to location, converging at 7 items, that is item numbers 15, 22, 24, 25, 45, 53, and 57. From the result of the analyses there is slight gender differential tending towards the girls while location significantly function and tended towards urban schools. Thus National Examination Council (NECO) Senior School Certificate Examination (SSCE) should always be analysed to check for location differential item functioning for possible elimination.

CHAPTER ONE

INTRODUCTION

Background of the Study

Mathematics is one of the most important subjects used in educating a man. Hence it is one of the compulsory core subjects in the primary and secondary schools system in Nigeria. A credit level pass in mathematics at the West African Examination Council (WAEC) or National Examination Council (NECO) in the Senior Secondary School Certificate Examination (SSCE) is a pre-requisite for getting admission into the tertiary institutions. A failure in it poses a major obstacle to getting registered into these institutions and even getting some kind of jobs.

The Federal Republic of Nigeria (FRN, 2004) in the National Policy on Education recognised the role of mathematics towards technological and industrial development, gave it a special place in Primary, Secondary and Tertiary levels of education. Alio (1997) supported this fact when he said that, there can be no real technological development without mathematics. Mathematics test is usually administered to learners, this help to measure their abilities to enumerate, calculate, collate, group, analyse and relate data (Onugwu, 1991 in Odo 1998). Odo noted that mathematics when properly conceived is a model for thinking, developing scientific structures, solving problems and for drawing conclusions.

Many countries of the world have tests at the end of education, and test for entrance into higher education after upper school. The students take the same test, but have taken mathematics courses of different lengths; that is, in some schools the syllabus/scheme of work would have been covered while some would not cover these. Conscientious teachers used the prescribed textbook in addition to other texts which help to simplify and

explained concepts clearly for the students while other teachers just restrict themselves to the recommended text only. In some cases the parents employ the services of mathematics teacher for their wards for extra lessons outside of the school. These privileges might bring about inequality in students ability to tackle mathematics problems.

In Sweden, the national course tests are given at the end of the course because, the aims of the test are to support the teacher's grading of the students and make the ground for assessment across the country as uniform as possible for each course in mathematics (Gunilla 2003). North America use large-scale achievement test pervasively by school districts to make decision about individual student, to track students; classes and schools as well as determine their promotion from one grade level to another, (Vandenberghe & Gierl, 2001). The Canadian province of Alberta for example, tests students at Grades 3, 6 and 9 in the content areas of Language Art, Mathematics, Science and Social Studies. These tests are administered to approximately 40,000 students annually and the test results are used by many teachers to evaluate student, (Anderson and Postl, 2000).

In Nigeria, there are several bodies of examination which administer test, among which are: the West African Examination Council (WAEC), National Examination Council (NECO), National Business and Technical Examination Board (NABTEB), National Teacher's Institute (NTI), Joint Admission and Matriculation Board (JAMB), and so on. Their characteristics according to Nwana (2007:133) are to:

- i. Present several thousands of candidates for examinations which are based on the existing national curriculum and the examiners are outside the authority of the educational institution.

- ii. Issue Results slip, Certificates, Diplomas to examinees who has attempted the examination and accreditation is by or on behalf of the Federal Ministry of Education to which the business of minimum academic standard belongs.
- iii. To developed tests measuring common skills and bodies of knowledge taught in school without introducing any extraneous characteristics that would favour some groups of candidates to perform better than the others, other than the skills on which the measurement was based.

National Examination Council (NECO), and West Africa Examination Council (WAEC) conduct Senior School Certificate Examination (SSCE) at the completion of Senior Secondary School to determine the extent of knowledge acquired by examinees during their secondary school program, as a pre-requisite for entrance or admission into tertiary institution and as a basic to those seeking for employment after their secondary education.

The efforts and policy directives that brought NECO into being ensured that the academic standard of the NECO senior school certificate was to be the same as that of WAEC at that level, including such examining features as; the style of developing the examination papers and the marking and grading of papers etc.. WAEC and NECO syllabuses are in broad outline similar to each other. In addition, virtually the same school teachers and lecturers who contribute trial test items and mark for WAEC also mark the NECO examination (Nwana 2007).

Public examinations are conducted by or on behalf of the state or nation and open to all those who meet certain defined entry criteria. It includes examinations used to assess students in or wishing to gain access to state and federal schools, polytechnics, colleges of education, universities

and other educational institutions. The examination is typically formal (that is, the examination has duly undergone the process of validation and standardisation of the items by the examination bodies), summative, and controlled by an agency external to the school where the student has studied. The purpose of examinations are; selection of candidates for educational opportunities and/or employment. The results determine future educational, vocational and life chances. Some are use for certification of achievement. For any examination to be adjudged good it has to possess the following qualities, these include;

- i examination questions and marking schemes should produce scores that are both reliable and valid.
- ii The conduct of the examination and the published results should be deemed fair and achieve a high level of acceptance by the public.
- iii The examination should ensure that no particular candidate or group of candidates has an unfair advantage over others (Ekoja, Happiness, Ekoja & Oluwu, 2006).

Sax (1980:227) states that ñno test item is perfect no matter how much time and consideration it receivesò. Sax noted that, even the most carefully prepared items are susceptible to human error and on analysis may prove to be ambiguous, too simple, overly difficult, or none discriminating. Knowing how students responded to items can also help teachers diagnose studentsô errors and develop more effective tests. Many of these problems can be detected and at least partially resolved by item analysis procedures which are particularly useful for multiple choice tests. This is in consonance to what the researcher intends to do with the multiple-choice section of NECO examination, to analyse and detect for possible Differential Item Functioning (DIF).

Brown (1983) attests that, a test is only as good as the items it contains. A complete evaluation of test quality must include an evaluation of each question. Questions should assess only knowledge or skill that are identified as part of the domain being tested and should avoid assessing irrelevant factors. They should be unambiguous and free of grammatical errors, potentially insensitive content or language and other confounding characteristics.

There is a general concept that as much as possible, test constructed should represent the best test one could obtain for each group given the overriding constraint of content validity, (Green, 1980). If a test so constructed is unfair or yields different scores from subgroups for instance (gender, location, religion, culture, and socio economic background) with the same ability, then such a test is said to be biased or functioning differentially. This implies that it is possible for a test item to measure different things in a subject for two students who are of the same subject matter ability, but of different subgroups of the population of test takers; this phenomenon is known as Differential Item Functioning (DIF), (Jensen.1980; Nenty, 1986).

Differential Item Functioning is an item behaviour used to identify differential item response pattern between groups of test takers. That is, male versus female students, urban versus rural school, socioeconomic status, region, ethnicity and so on. This aid in identifying potentially biased test item. Questions regarding group difference in scores of standardised tests have been present for nearly a century according to (Jensen 1980). Binet became concerned when the test, which was develop in 1905 to measure innate ability, demonstrated significant group difference in overall scores among social status groups. These differences were found across multiple countries, including US. Binet began to speculate that certain items were

measuring cultural training, such as schooling, home environment, verbal fluency, cultural climate, educational level, rather than innate ability. The study eliminated categories of items that were believed to be dependent upon social standing. However, this attempt yielded little difference in the between group discrepancies of overall scores. Since this time, numerous researchers have also reported group differences in mental test scores and identify possible sources of group differences.

The issue of Differential Item Functioning or bias in testing came to public attention in the late 1960s, (Candall & Coffman, 1964; Cleary & Hilton, 1968; Green, 1980). In this study Differential Item Functioning and test bias shall be used interchangeably. Several statistical methods of estimating the index of item/test bias which led to different definitions by different researchers and authors has been proliferated. Many reviews, criticisms, recommendations have consequently been carried out on the different methods of DIF, (Jensen, 1980; Petersen, 1977; Ruder, Getson & Knight, 1980; Nenty, 1986). Previously, variety of methods had been proposed for detecting Differential Item Functioning including but not limited to the following: The Mantel-Haenszel (M-H) Procedures, Transform Item Difficulty (TID-45⁰), Scheuneman Modified Chi-square (SSG²), Distractor Analysis (DA) Method, Expert Judges, Regression Method, Cochran's Chi-square Test Method (CTG²), and Item Characteristics Curve (ICC) Method.

To evaluate whether a study item displays Differential Item Functioning, examinees are first separated into reference and focal group (for example males versus females, urban versus rural examinees). The reference group is the group expected to perform at higher ability while the focal group perform at lower ability. The reference and focal groups are

matched on ability, ideally d . Because d is an unobservable variable, DIF statistics must approximate matching on d with a matching based on the observable data at hand. In assessing DIF the M-H statistics matches examinees on the basis of total test score, including the score on the studied item and then evaluate the response pattern to the test items, (Rousons, Schnipke, & Pashley 2000).

Studies exist that provide evidence to show that test used in some public examination in Nigeria contains items that are differentially functioning. For instance Umoinyang (1991), study on incidence of Differential Item Functioning (DIF) in mathematics objective tests used in the WACE/GCE, Nov/Dec, 1990, showed that test items in the subject differentially functioned for test takers described in terms of gender, educational level of development and region (Northern and Southern Nigeria). The recorded studies exhibited significant difference in functioning of test takers which is not supported by the FGN in the National Policy on Education.

Every year in Nigeria and in the Federal Capital Territory (FCT), examinees write the NECO and WAEC SSCE examinations of which mathematics is one of the tests, but there has not been any analysis to show how boys and girls functioned in this examinations and how school location contributes to the functioning of examinees (ERC, 2008). Thus there is the dire need to carry out Differential Item Functioning DIF on mathematics to determine which group boys or girls and which school location urban or rural functioned differentially in the examinations in the FCT.

The Federal Government of Nigeria, in the National Policy on Education (FGN 2004), indicates that any existing ambiguities and lack of uniformity in educational practices should be removed, instead the provision

of equal access to educational opportunity for all citizen of the country at the primary, secondary, and tertiary levels both inside and outside the formal school system should be encouraged. This entailed that test should measure the knowledge and skills taught in schools and this should be to all students irrespective of gender, socio-economic background, location, school type, state and ethnic group.

This study hinges on the measurement theory which is divided into Classical Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory CTT considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored). This success rate of a particular pool of examinee on an item, well known as the p-value of an item is used as the index for the item difficulty.

Item Response Theory (IRT), postulates, that characteristic of a test item, such as its difficulty, can interact with a person's ability or trait to determine the probability of a correct response to that item, (Lord 1980). Exponents of IRT postulate that latent trait or mental trait is responsible for an individual performance in a given test (Warm, 1978). According to the unidimensionality assumption in IRT, a test item should be constructed such that performance in the item is sustained only by the subject matter ability demand, and nothing more.

Concern about equity with respect to men and women has generated a considerable interest in educational achievement. Differences in the educational background and achievement of the two groups are likely to contribute to disparity in the allocation of cognitively demanding roles in our society, for example there seem to be more women in the nursing profession and more men in engineering, more women teaching English Language and

more men teaching mathematics. Consequently, group difference in relevant test score are cause for concern.

The focus of this study is to find out whether there exist differential functioning items between male versus female students, rural versus urban schools in multiple choice mathematics achievement test set and administered by National Examination Council, NECO June/July 2008 SSCE in government coeducation secondary schools in the Federal Capital Territory (FCT), Abuja.

Statement of the Problem

There are cases of Differential Item Functioning in performance of students in some form of examination, test or assessment. In some cases the differential favour one group of test taker, that is, female may functioned significantly better than the male, also sometimes urban school may function better than their rural counterparts and vice versa. Test with differentially functioning item distorts the meaning of test results and decision that is based on it for some groups of test taker. In the National Policy on Education (FRN, 2004) education should be used as a tool for development of individual and building a nation where every citizen has equal opportunity. Test with differentially functioning items cannot be used to achieve the above goals.

Moreover from the Education Resource Centre (ERC 2008) Abuja, there has been no record of study on differential item functioning on the examination items of National Examination Council NECO in the Senior School Certificate Examination (SSCE) in Abuja; hence the differential status of our examinees in the FCT is unknown. The researcher intends to use NECO SSCE 2008 June/July multiple choice mathematics questions

because as at the time of her admission, the examination was the recent to determine the differential composition of examinees. Thus the researcher has undertaken to study the response pattern of examinees, and uses the methods of detecting DIF to establish whether gender of examinee and school location functioned differentially in mathematics multiple choice examination administered by NECO in the June/July 2008 SSCE examination in all the coeducation secondary school in the Federal Capital Territory, Abuja.

Purpose of the Study

The main purpose of the study is to determine: Differential Item Functioning in mathematics multiple choice questions administered by NECO in June/July 2008 Senior School Certificate examination SSCE.

Specifically this study tends to:

1. Identify if NECO items in mathematics multiple choice question of SSCE examination function differentially in terms of gender.
2. Identify if school location is a factor to differential item functioning in multiple choice questions administered by NECO in Senior School Certificate Examination.

Significance of the Study

Result of the study is useful to examination bodies like the: West African Examination Council (WAEC), National Examination Council (NECO), Joint Admission and Matriculation Board (JAMB), National Business and Technical Examination Board (NABTEB), National Teacher's Institute (NTI), also to Policy Makers, Teachers and Curriculum Planners.

- i The examination bodies test/item writers would be able to write appropriate test that will be devoid of Differential Functioning Items.
- ii Test developers for examination bodies will be better able to construct items that appropriately measure test takers knowledge of the examination items. That is the items will only portray the differential status of male versus female examinees, urban versus rural schools when exposed to answering multiple choice questions in NECO mathematics achievement test.
- iii Policy-makers and school administrator will use the result of the study in decision-making since the test is useful in identifying subgroups of test takers who require additional help and support. Thus the results from this study will shed light on the effectiveness of a well-established test review used to screen for gender and location bias, and identify areas where possible refinements and research are needed.
- iv Teachers, schools and district (or zone) will benefit from the results since the test are also being used as part of an educational accountability system that assesses them. In order words the absence of difference in functioning of students infer that they were taught effectively and are at the same ability
- v To curriculum planners: it provides information needed to improve instruction and student learning in selected areas of the curriculum in relation to national standards.

Scope of the Study

The study was carried out in all the public senior secondary schools in the Federal Capital Territory Abuja, preferably coeducation schools. Coeducation schools have both boys and girls attending, and are found in

both urban and rural areas. The researcher used the NECO June/July 2008 SSCE multiple choice mathematics examination score for the analyses. The choice of mathematics is that the researcher is a mathematics teacher, also WAEC and NECO examiner. The study is concerned with only Differential Item Functioning DIF in mathematics multiple choice examination. The tests measure a unidimensional construct, the situation under which the Mantel-Haenszel (M-H) procedures, Scheuneman chi-square (SSG^2) statistics and the Transform Item Difficulty ($TID-45^0$) are intended to be used for the analyses and the three methods shall be calculated manually.

Research Questions

The following questions have been posed to guide the study.

1. How do the items in NECO 2008 June/July multiple choice mathematics questions differentially function in terms of gender?
2. Are there differential functioned items in NECO mathematics multiple choice examinations in terms of school location?

Hypotheses

In order to test further the answers to the above research questions the following hypotheses will be tested at the 0.05 level of significance:

Hypothesis (H_{01})

Items in the multiple choice questions set and administered by NECO does not significantly function differentially among students based on gender.

Hypothesis (H_{02})

Items in the multiple choice questions set and administered by NECO does not significantly function differentially among students based on location.

CHAPTER TWO

LITERATURE REVIEW

This chapter is organised around the following major headings and sub headings as follows:

Conceptual Framework:

The concept of mathematics

Gender

Location

Academic functioning/achievement

Item bias, Item fairness and item Impact

Difficulty index

Item discrimination

Multiple choice items

Statistical methods use for detecting differential item functioning

Uses of Differential Item Functioning

Theoretical Framework:

Classical Test Theory

Item Response Theory

Empirical Studies:

Detection of differential item functioning

Test and Differential Item Functioning

Gender and Differential Item Functioning

Location and differential item functioning

Summary of Literature

Conceptual Framework:

The concept of mathematics

Globally, mathematics is made up of a set of concepts, facts, principles, and operations that are fundamental to the existence of every individual. The position mathematics occupy in the National Policy on Education (FRN, 2004), and its role towards technological and industrial development put mathematics in a special place in primary, secondary and tertiary levels of education. It is a science of quantity and space, the corner stone of every field of education Ezeameny and Alio (2004) reported in (Nwabueze 2006). The abstract nature of mathematics is that it demands perseverance, constant practice and a lot of thinking both critically and analytically from its learners, but unfortunately, most students lack the patience and required time to think properly in solving mathematical problems Iheanyi (2006), which might make them to function differentially in total test score. Also Obioma, (1986), considered mathematics as a science and a way of thinking that has rules, goals and players just like a game. Its learning calls for imagination, participation and skills. From the wordnetweb, mathematics is a science (or group of related science) dealing

with the logic of quantity and shape and arrangement. It is the study of quantity, structure, space, relation, change and various topics of pattern, form and entity, (www.en.cidipedia.org/wiki/mathematics) as already described by Ezeameny and Alio (2004). Mathematics is an abstract representational system used in the study of numbers, shapes, structure and change and the relationship between these concepts. Also a person's ability to count, calculates, and uses different system of mathematics at differing levels. Thus mathematics is a science which deals with the study of concepts, space, quantity, shape, form, pattern and numbers which every child must learn in the primary and secondary school systems in Nigeria and also a pre-requisite for getting admission in the tertiary institution. The broad mathematics concept covered in the senior secondary school include: Number and Numeration, Algebraic Processes, Menstruations, Plane Geometry, Trigonometry, Statistics and Probability Utibe (2009). These must be properly represented in the questions, otherwise, there will be bias that will culminate into Differential Item functioning DIF in mathematics achievement test.

Gender

Gender is the result of socially constructed ideas about the behavior, actions, and roles a particular sex performs. The beliefs, values and attitude taken up and exhibited by them is as per the agreeable norms of the society and the personal opinions of the person is not taken into the primary consideration of assignment of gender and imposition of gender roles as per the assigned gender (wikipedia). Gender may be defined as a socially ascribed attribute which differentiate feminine from masculine (Espino, 1992). It refers to the socially constructed roles of and relations between men and women. This

social attribute draws the perception that some academic programs are normal for one gender and abnormal for the other. Gender also refers to the biologically determined characteristics of men and women. Gender relations are characterised by unequal power. Gender norms assign specific entitlements and responsibilities to men and women - for example, women might be expected to take on caring or domestic duties and remain close to home, while men may be expected to be the main breadwinner, working outside the home, with greater freedom to move around in public places. But within this study, gender shall refer to male and female students.

Location

Students who attend the same school are grouped together, for example, for each student in the same school; their family income is based on the average family income of the school location. And also, all students who attend the same school are assumed to live in the same rural or urban location (Zumbo, & Gelin, 2005). Lesser (1976) stated that individual who share a common cultural background also share common patterns of cognitive styles. Cognitive style refers to a habitual mode of processing information that is adopted by a group (Das, Manos & Kanungo, 1975). Hence location in this study shall refer to schools located in the urban and rural areas that share a common background and cognitive style.

Academic Functioning/Achievement

Academic functioning/achievement mean the comprehensive, basic and important ability of the learner, who was systematically instructed for the duration of the senior secondary school period. (In this study academic achievement and functioning will be used interchangeably). The

achievement in this examination determines the process of admission into tertiary institution for those who wish to continue schooling. An achievement test is that which focuses on assessing the extent to which a given content area or task has been mastered after teaching and how much of what is taught has been learned by the individual. Hassan (2001) in Daman, wakjissa (2006), maintained that standardise achievement tests are used for the purpose of certification. This is most pronounced in Nigeria secondary school where a successful and meaningful completion of secondary school education is only determine through the passing of the NECO and WAEC/SSCE examination. The NECO mathematics achievement tests have these traits according to Daman (2006):-

- i It is a standardise test.
- ii Measurement experts and experts from every subject of the curriculum designed the tests.
- iii It has a specific blue print. The blueprint designed follows by the objectives of instruction, the criteria of the curriculum.
- iv The test is an item bank test, which is supported by IRT.
- v The scores from different tests can be compared because the test is in line with the requirement of IRT.

With the above traits listed, the functioning of the subgroups will be compared to ascertain whether NECO examination items in the senior school certificate examination differentially functioned/or not on our examinees in the Federal Capital Territory (FCT) Abuja.

Unlike measurement in the physical sciences functionality testing requires the cooperation of the object of measurement (that is examinees). This means that whenever we administer a functional test to an examinee, we implicitly assume that the examinee will try to do his or her best on the

item. Despite our assumption to the contrary, however, non effortful responding by examinee is not uncommon. Consequently because the item response provide the basis for estimating the examinee's functionality level, any response that do not reflect that examinee's functionality level will distort functionality estimation thereby diminishing the validity of the examinee's test score. Moreover because non effortful response tend to be correct less frequently than those for which the examinee exhibits good efforts, their presence tends to have a systematic negative bias on functionality estimation. Achievement may also be regarded as a behaviour exhibited at the end of a given period of time or within a given time range. Glaswer, (1963; 449) in Iheanyi (2005), defined achievement measure as "the assessment of terminal or critical behaviour involving the determination of characteristic of students performance with respect to specific standard". Achievement test result enables us to obtain information on the extent to which a student has attained the criterion performance. It also enables us to determine the relative position or rank of individual students with respect to their test performance. Thus academic functioning/achievement here are the attainment of a particular standard of performance in the course of study at the National Examination Council NECO Senior School Certificate Examination SSCE which will determine the examinee getting admission into tertiary institution.

Item Bias, Item Fairness, and Item Impact

The research on item bias can be traced back to 1905. A. Binet and T. Simon administered the original version of the intelligent test. They found there is a significant difference between children of working class background compare to those from middle class families. From that time on,

cultural bias has always become a factor in research. Item bias is an interaction between items and group, reasoned (Cleary and Hilton 1968) while Angoff and Ford (1973), thought item bias was the difficulty parameters difference between two groups. Anantasi, (1998), wrote that item bias seeks to identify items for which equally able person from different cultural groups have different probability of success. Equal ability means equal with regard to the construct that the test is design to assess or the criterion behaviour it is design to predict. In the language of Item Response Theory (IRT), the presence or absence of differential item functioning (DIF) indicates the degree to which items are biased or not. Differential Item Functioning (DIF) exists if an item is relatively more difficult, discriminating, or easily guessed for one group than for another Nunnally & Bernstein, (1994) in (Gunilla 2003). Test item is considered to be bias or functioned differentially when a dimension on the examination is deemed to be irrelevant to the construct that is being measured, placing one group of examinees at a disadvantage in taking the examination (Hambleton & Rodgers, 1995). Alternatively, item bias is defined as invalidity or systematic error in how a test item measures a construct for the members of a particular group

(Camilli & Shepard, 1994:8). Item bias is of particular concern on tests of mathematics achievement where differences between females and males are commonly found (Kimball, 1989; Scheuneman & Grima, 1997). In order to further ensure equal access to educational and career opportunities, careful analysis of the items that compose the test in question is necessary to ensure their fairness to all of the examinees, not just the majority, so much so that the objectives of the Federal Government in the National Policy on Education (FGN:7) be accomplished.

Despite an already large and growing literature on the subject of item bias in educational and psychological tests, it seems that no universally accepted definition yet exists (Keeves 1990). Keeves also wrote in an International Handbook, that the identification of item bias is important, even for tests that are mainly dependent on predictive validity (e.g. selection tests) since internal indications of bias make it increasingly likely that the test will also show predictive bias. Hence, techniques for identifying test item bias are useful in test construction, since potentially biased items, in the predictive sense, can be rejected during the initial item-selection process and prior to any attempts to evaluate a test's predictive validity in populations.

Hambleton & Rodgers, (1995) further stated that test item bias can presents itself in a variety of ways in items through the language, the format, the content or the behaviour required. It can result from membership in a specific subpopulation or from factors correlated to the subpopulation, it can affect all members of the subpopulation or it may affect only those in specific ranges of ability. An item might also be considered biased if it contains language or content that is differentially difficult for different subgroups of test-taker. In addition an item might demonstrate item structure and format bias if there are ambiguities or inadequacies in the item stem, test instruction or distracters continued Hambleton & Rodger, (1995). A biased item measures attributes irrelevant to the tested construct, (William, 1997). Again examination item is considered biased if it functions differentially for a specified subgroup of test-takers, in such a case, students who are equally able do not have an equal chance of success (Zumbo, 1999). Bias or differential item functioning is said to exist when a test or an item causes systematic errors in the measurement. For example if test scores indicate that boys performed better on a certain test or item than girls do, when in fact

both sexes have equal ability, the test or item provides a biased measure. Something else than what was intended is measured in biased items. Zumbo, (1999) cautioned that "DIF" is a necessary, but not sufficient condition for item bias.

A fair test is one that is comparably valid for all groups and individuals and that affords all examinee an equal opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the test purpose, Roever, (2005). This is stated in the National Agency for Education (NAE) and the National Policy on Education (FRN, 2004). Roever, (2005); Willingham and Cole, (1997), defined a fair test as a test that is comparably valid for all individuals and groups. According to them, fair tests design should provide examinee comparable opportunity as far as possible to demonstrate knowledge and skills they have acquired that are relevant to the purpose of the test. Fairness among examinee implies fairness among all possible subgroups of examinees including those from different ethnic groups, socio-economic status groups, and those who live in different neighbourhood e.g., rural and urban areas, adopted from (Zumbo, & Gelin, 2005). They establish that fairness among examinees is important because standardised functioning tests for elementary and secondary school students are commonly used to provide information on student learning in selected areas of the curriculum in relation to national standards.

When Differential Item Functioning DIF is not evident for an item, then there is no item bias. Conversely, DIF is required but not sufficient for item bias. An item might show DIF, but not be considered biased if the difference is as a result of the actual difference in the group's ability to respond to the item, that is, if one group of test-takers is at a high level and the other group of test-takers is at a lower level, the lower group would

perform significantly lower, (Roever, 2005). If test-takers differ in knowledge, a difference in item response would be evident. Consequently, a difference in the performance of groups of examinees with difference abilities on specific items is not indicative of test bias but rather of item impact (Schumacker, 2005).

Item impact is evident when one group of examinee is found to endorse the item more than the other group of examinees because the two groups truly differ on the underlying ability or factor being measured by the test i.e. item impact occurs when the item measures a relevant characteristics of the test, and a real difference between the two groups of interest are found. For example if female were less likely to endorse an item from an achievement test of ability, and men and women truly differed on mathematical aptitude, item impact is present. Item impact can be described as any group disparity in item performance that reflects actual knowledge and experience differences on the construct of interest (Clauser & Mazor, 1998). Item impact and item bias differ in terms of whether group differences are based on relevant or irrelevant characteristics of the test.

Only when difference in a group's ability to respond to a test item is caused by construct-irrelevant factors can DIF be considered as bias. Thus Roever, concluded logically that to detect test bias or DIF is to locate examination item on which one group of test-taker performed significantly better than another group. Therefore, there should be a level playing field where, for example male and female students from both urban and rural schools have equal opportunity to do well in large-scale assessment like WAEC, NECO, & JAMB examination and be treated equitable in terms of test score functioning. DIF requires that members of the two groups be matched on the relevant underlying ability before determining whether

members of the two groups differ in their probability for success. Hence the researcher will use the overall test scores of the examinees on NECO examination to match the ability of the difference subgroups and then apply the M-H procedures, Schuneman statistics and the TID to determine the differential functioning of the examinees.

Detection of Differential Item Functioning (DIF)

DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure. In item bias study, the method of detecting DIF is variable. The earliest method of detecting DIF was "Transformed Item Difficulty Method" provided by Angoff, (1972). The method is use to compare the correct answer probability of two groups because transformed item difficulty method only manipulate the item difficulty, it cannot observe the relationship between item bias and discrimination, (Merz & Grossen, 1979). Researchers have found the transformed item difficulty method was inferior to the item characteristics curve method and G^2 test procedure, (Ruder, Getson, & Knight, 1980: Shepard, Camilli & Averill, 1981: Shepard, Camilli and Williams, 1985: Scheuneman, 1979). It is similar to the method of using IRT as a base measure. Chi-square (G^2) test procedure usually separate examinees into many subgroups by the total test scores and assume the ability of all examinee approximately equal. Again researchers have found that statistical detection of item bias is at best an inexact science that is to say there has been a variety of methods proposed for detecting bias but no one statistic to date can be considered either necessary or sufficient. Different methods are more or less successful depending on the situation and circumstances, (Subkoviak, Mack, Ironson, & Craig, 1984). A problem exists whenever

examinees from two different populations are matched on the same level of the observed test score, but actually have quite different levels of the unobserved ability (Spray & Miller 1992). This imbalance is always seen in the observed total test score of the examinees.

Item Characteristic Curve ICC method is based on item response theory; it can be divided into 3 categories, if we use IRT to detect DIF. (Camilli & Shepard, 1994) referred to the raw or uninterrupted relative difficulty as Differential Item Functioning (DIF). Differential Item functioning statistics would be used to identify all items that functioned differently for different groups and establish the index of the differential. Tinsley & Dawis, (1995) in Joseph (2007), stated that when item estimate vary from one subgroup to another or from one administration to another then the conditions of DIF or Item Parameter Drift (IPD) are considered to be present. The statistical assessment of DIF falls into the second state or the structural state of construct validation as described by Benson, (1988), and Benson and Hagtvet (1996). In Benson's first state of construct validation, the theoretical and empirical domains of the construct are identified, following this stage, the items that are written to measure the construct are themselves studied in what Benson term's "internal domain" studies. Often this study involves the use of factor analysis to identify or confirm the dimensionality of the items. Internal domain study also include the examination of DIF to determine if at the item level, subgroups differ on a second extraneous construct being measured by the item.

The prominent statistical method used by psychometricians to detect bias is generally described as Differential Item Functioning (DIF) analysis, (Clauser & Mazor, 1998). DIF first involves dividing examinees into two separate groups; the reference group and focal group. Typically, the

reference group examinees serve as the standard for comparison while examinees from the focal group are viewed as being disadvantageous on the test. (In other words, the reference group is the group that is functioning at higher ability and the focal group is the group functioning at lower ability). Examinees from each group are then matched according to their ability from the test and statistical procedures are used to estimate group differences. An item is deemed to exhibit DIF when examinees from two groups have a different probability of answering an item correctly after controlling for overall ability. It is only when two persons from different ethnic groups, but with the same level of the construct being measured, perform differently on an item that the item is flagged as functioning differently for different ethnic groups confirmed (Clauser and Mazor, 1998). Zumbo (1999), supported others when he assert that DIF occurs when a statistically significant difference is evident in a problem that test takers from the two distinct groups (e.g., males and females) who have the same underlying ability on the measured construct, demonstrate differing probability of correctly answering the item. Further DIF can be attributed to impact bias.

Oftentimes, DIF technique are used to statistically assess whether item are measuring materials extraneous to the construct of interest, yet relate to group membership. Zumbo iterated; DIF statistical techniques are based on the principle that if difference groups of test-takers have roughly the same level of something example knowledge, they should perform similarly on individual test items regardless of group membership. In their essence, all DIF techniques match test takers from different groups according to their total test scores and then investigate how the difference groups perform on individual test items to determine whether the test items are creating problem for a particular group. Roussos, Schnipke & Pashley (2000), agree

with previous researchers that when examinees from two different subgroup have the same ability distribution (or are "matched" on ability but are not equally likely to answer a particular item correctly, the item is said to exhibit Differential Item Functioning, that is the item functions differently in the two groups.

Zumbo (1999), define two types of scores that can arise from tests and measures. Binary scores which are also referred to as dichotomous item responses and Ordinal item responses. But we are concern with Binary scores. Items that are scored in a binary format are;

- (a) Items (e.g, multiple-choice) that are scored correct/incorrect in aptitude or achievement test or
- (b) Items (e.g true/false) that are dichotomously scored according to a scoring key in a personality scale.

With the idea of Zumbo, the researcher using the multiple choice question of NECO items has collapsed the options into dichotomous scores items. That is the correct option (key) is given the value of 1 and the incorrect options (distracters) are given the value 0.

Roever, (2005), also join previous researcher and state that logically the first step in detecting test bias is to locate examination items on which one group of test-takers perform significantly better than the other group. Schumacker, (2005), also supported previous researcher by stating that DIF is a collection of statistical method utilized to determine if examination items are appropriate and fair for testing the knowledge of difference groups of examinees. (e.g, male versus female or Caucasian versus African-America). In order for an item to exhibit DIF Ackerman & Evans 1992 in Metcalf, (2008), added that three conditions must be present;

- (a) More than one dimension is being measured

(b) one group of examinees has a different ability distribution than another group on the dimensions that are not intended to be measured, conditioned on the ability intended to be measured;

(c) A single score is reported.

Traditionally an item is said to show DIF if conditionally on the ability, the probability of correctly answering the item depends on the group the person belongs to and models and techniques for DIF treat both the item and the groups as fixed Wim & Boeck (2005). If for example a mathematics test item has boys answering correctly more often than girls of equal ability, because the subject of the item is on a topic more familiar to boys (e.g., sports), then the item is said to exhibit DIF and should be considered for modification or removal from the test. Thus operationally Differential Item Functioning is the difference in functioning of two subgroups of examinees with the same subject matter ability but from different population of test takers.

The researcher will use Mantel-Haenszel (M-H) procedure, Scheumman χ^2 method and the Transform Item Difficulty method to detect DIF of gender and location in the mathematics multiple choice question administered by NECO in the 2008 June/July SSCE examination in the Federal Capital Territory, Abuja.

The researcher believes more on these statistics in detecting DIF items correctly because they are relatively inexpensive yet statistically powerful techniques for identifying differentially functioning test items, and in addition present enough information about the direction of relative ease of the item even if it is not biased.

Difficulty Index (Item Facility)

All tasks are evaluated in terms of difficulty according to standard Classical Test Theory (CTT) practices. (Sax, 1980), see difficulty level (p) of an item as the proportion of students responding correctly to it. The higher the proportion is, the easier the item is. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item. Similarly, difficulty is defined as the average proportion of point achieve on an item, and is measured by obtaining the average score on an item and dividing by the maximum score for the item (Zumbo & Gelin, 2005). It is properly interpreted as an "easiness index" because larger values indicate easier questions. Ideally, the items on an assessment will have a range of difficulties between 0.25 and 0.9 with most items falling between 0.4 and 0.7. Items that have either a very high or very low difficulty index are considered to be potentially problematic because they are either so easy that nearly all students get them right or so difficult that few students get them right. If an assessment were comprise entirely of very easy or very hard items all student would receive nearly the same scores and the assessment would not be able to differentiate high ability students from low ability students. Hence a minimal discrimination is needed to differentiate ability level.

Item Discrimination (Item-Test Correlation)

Sax (1980), states that discrimination indices measure the extent to which items differentiate between those persons with the highest and lowest scores on the total test. This view is also shared by the Montana Comprehensive Assessment System Technical Manual (2004) that a desirable feature of an item is that the higher ability students perform better on the item than lower ability student. Within Classical Test Theory (CTT),

the item test correlation is referred to as the item's discrimination. It indicates the extent to which successful performance on an item discriminate between high and low score on the test. The theoretical range of this statistic is -1 to +1, with a typical range from .3 to .6. Discrimination indices can be thought of as measures of how closely a question assesses the same knowledge and skills assessed by other questions contributing to the criterion total score.

Multiple choice items

Multiple choice items according to Keeves (1990) consist of two parts; a stem and a number of options or alternatives. The stem is a question or statements that is answered or completed by one of the alternatives. All incorrect or less appropriate alternatives are called distracters or foils, and the student's task is to select the correct or best alternative from all the options. Multiple choice items is by far the most frequently used in educational achievement testing. The number of alternative answers offered varies but is usually four (4) or five (5). As a rule only one of the alternatives is correct, the others (the distracters) are being constructed so as to provide plausible examples of common errors. If the distracters are carefully written a wrong-answer analysis can yield valuable diagnostic information about the types of error being made by students.

Advantages of multiple choice items are as follows:-

- i It has great versatility in measuring objectives from rote knowledge level to the most complex level.
- ii The teacher can sample a substantial amount of course material in a relatively short time since students writing is minimised.

- iii Scoring is objective because little interpretation is needed to count the number of correct responses.
- iv Teachers can construct multiple choice items that require students to discriminate among options that vary in degree of correctness. This allows students to select the best alternative and avoids the absolute judgements required in true-false tests
- v Because multiple choice items have from 3 to 5 options they reduce the effects of guessing.
- vi The multiple choice format is particularly amenable to item analyses to detect areas of students weakness, evidence of item ambiguity, item difficulty, and the extent to which the item can measure individual differences.

The major problem with multiple choice items are that they are time-consuming to write and students sometimes complain that there is more than one defensible correct answer. This difficulty can be remedied by careful editing and pretesting of items (Sax 1980)

Statistical Methods of Detecting Differential Item Functioning

Statistical detection of item bias/Differential Item Functioning (DIF) is at best an inexact science. There has been a variety of methods proposed for detecting DIF but no one statistics to date can be considered either necessary or sufficient, (Demars, Wise & Madison, 2007). That is no single method has been found to identify DIF accurately hence researchers should use methods that are best for their study. Different methods are more or less successful depending on the situation and circumstances. If a test overall is unbiased but with one or two biased items most method will locate the problems. If however the test itself is biased against one or more

subpopulation, a statistical analysis of the items will not be able to separate bias effects from true difference in achievement (Camilli & Shepard 1994). Some of the methods used in detecting DIF are:

Mantel-Haenszel (M-H) Procedure.

The Mantel-Haenszel (M-H) procedure for detecting DIF is the most commonly used technique in educational testing. It is a chi-square technique with a number of distinct advantages over its predecessors (Keeves 1990). It does not depend on the application or the fit of any specific measurement model. The method, simple chi-square G^2 statistics was proposed by Holland & Thayer (1988) and recommended for use as a very effective method for detecting bias items especially for items that are dichotomously score (0 or 1).

To use the method, the two groups being examined are divided into matched subgroups and a $k \times 2$ contingency tables for each item, where k is the number of distinct number-correct scores observed for the test, is formed. The simplest method of matching is to use the total test scores and for each matched group a table is calculated as in table 1 below. Using the notation of Holland and Thayer, T_j is the total number of students in the matched set; N_{rj} is the number of examinees from reference group and A_j is the number from the reference group who got the item correct.

Table 1 Illustration of analysis of bias with M-H statistics.

Group/score on item	1	0	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{fj}
Total	M_{1j}	M_{0j}	T_j

For the k 2 x 2 tables for a given item the following statistics is computed: MHCHISQ =

$$\frac{\left[\left| \sum_j A_j - \sum_j E(A_j) \right| - 0.5 \right]^2}{\sum_j \text{var}(A_j)}$$

$$\text{Where } E(A_j) = \frac{N_{rj} M_{lj}}{T_j}$$

$$\text{Var}(A_j) = \frac{N_{rj} N_{fj} M_{lj} M_{oj}}{(T_j)^2 (T_j - 1)}$$

$$\text{M-H Alpha} = \frac{\sum_j \frac{A_j D_j}{T_j}}{\sum_j \frac{B_j C_j}{T_j}}$$

$$\text{M-H Delta} = - (2.35) \log_e (\text{M-H Alpha})$$

Where M-H Alpha is the common odds-ratio across the k 2 x 2 table for a given item. It shall be used to show the direction of relative ease of the test items between the two comparable groups and it ranges from 0 to ∞. A value of 1.0 signifies no bias. Value less than 1.0 indicates that the item is relatively easier for the focal group (after controlling for total score), while M-H Alpha values greater than 1.0 indicates that the item is relatively easier for the reference group. The index of DIF utilise in this study is the M-H Delta statistics. M-H Delta indicates the average amount by which test-takers in the focal group found the items in question more difficult than did a comparable test-takers in the reference group (Holland & Thayer, 1986).

For the hypotheses, an item shall be considered biased or function differentially if the absolute value of the M-H delta exceeds 1.5 (Shapard, Camilli & William 1984). M-H delta is simply the transformation of M-H Alpha statistic to obtain a measure of bias expressed in terms of differences

on the Educational Testing Service (ETS) delta scale (Holland & Thayer, 1986).

Scheuneman's Modified Chi-Square Method (SSG^2)

With this method, an item is unbiased if for all persons of equal ability, the probability of a correct response is the same regardless of group membership (Scheuneman, 1987). With the method, each major comparative group is divided into various groups based on the ability level on the basis of observed total test scores. The p-values for each score group are then computed for comparison using the Chi-square (G^2) statistics. With this procedure, each item is separately tested for bias. A statistical test of item biases in the test as a whole is simply the sum of all the G^2 value for the single item, which is G^2 with $n-2$ degree of freedom, where n is the number of items. An item shall be considered biased if the calculated G^2 is statistically significant in a G^2 distribution at .05 level of significant. Chi-square with one degree of freedom must equal or exceed 3.84 to be significant at the .05 level of which is the conventional level of rejection of the null hypothesis.

Table 2 Scheuneman Modified Chi-square statistics for computation of DIF

Example of the computation of Chi-square for one item

Total score on sub-	No. With scores in each range			Obtained Frequencies (No. With item correct)			Proportion correct	Expected Frequencies Black White	
	Black	White	Total	Black	White	Total	(T_o/T)	(p.B)	(p.W)

test	B	W	T	B _o	W _o	T _o	P	B _e	W _e
13-14	25	315	340	22	300	322	.9471	23.68	298.32
12	24	110	134	18	99	117	.8731	20.96	96.04
10-11	48	118	166	23	93	116	.6988	33.54	82.46
1-9	65	92	157	14	33	47	.2994	19.46	27.54
$\chi^2 = \sum \frac{(B_e - B_o)^2}{B_e} + \sum \frac{(W_e - W_o)^2}{W_e} = 7.912, d.f = 3$									

Source; Scheuneman (1979) p.146. Adopted from Odili (2004).

Transform Item Difficulty (TID-45⁰) Method:

This method is classified by many researchers and authors as a major method of detecting DIF (Jensen, 1980; Rudner, Getson & Knight, 1980; Ironson & Subkoviak 1979; Joshua, 1990). The need for the transformation of the p-values before using it as a measure of detecting bias arose because equal intervals on the scale of p-value do not represent equal intervals on a true scale of item difficulty. The items p-value is the proportion of the total area under the normal curve that passed the item and $q = 1 - p$, (proportion of the area under the normal curve that failed the item). The vertical line that cuts the total area under the normal curve into areas p & q intersects the baseline of the curve, which is the z-scale of deviation of the normal distribution with mean of 0 and standard deviation 1. The z-scale of the normal curve thus represents an interval scale of item difficulty. The value of z on the baseline of the normal curve that is intersected by the vertical line that divides the curve into areas p and q where p is always the upper tail of the curve is the difficulty of any given item. The p-value for each item are calculated separately for each group and are transformed to a scale that is more likely to be interval than the p-value scale. The usual transformation involves converting the p-values to z-values where z is the $(1 - p)$ th

percentile in a normal distribution and then producing a delta scale where for item i of group j ($\Delta y = 4z_{ij} + 13$). Item with average ability corresponds to a z scale value of 0. Easier items have z -value less than 0, while more difficult items have z values greater than 0. This transformation controls for the negative values and ranges from 0 to 26 since the mean is 13 and standard deviation is 4. Transformation removes any group difference in the mean or standard deviation of the delta value and any group difference in delta value that remain represent group difference in the item relative difficulties which is a criterion to test bias. In many instances, the TID method is adequate for identifying item bias; it is conceptually simple, easy to compute and provides a visual indication of the amount of bias. Comparative studies of item bias detection techniques have shown that the TID method performs well as an approximation to more sophisticated methods, especially when short tests are being analysed or sample sizes are small (Keeves 1990). An item shall be said to function differentially for a group if its standard deviation is greater than 1.5 units from the 45° line (Ruder 1977) while using the TID method.

Distractor Analysis (DA) Method:

Distractor Analysis is a technique that inspected multiple-choice item by determining which distractors (incorrect choices) were more attractive to particular subgroups, (Scheuneman, 1979). According to Veale and Foreman (1983), this method was based on the idea that more can be learned from incorrect than correct responses. The assumption was that incorrect responses on a multiple-choice exam were not picked randomly. Instead, there was some sort of logic applied by the examinee in attempting to pick the correct answer. If a test is unbiased with respect to two groups, we

should expect the distracters to have the same relative degree of attraction within each group, in which case the distracters for any given item would have the same relative frequencies of takers in both groups, Jensen (1980). Jensen further commented that the simplest statistical test of the null hypothesis in this is chi square. The correct response is omitted from the error analysis, and the frequency of the major and minor groups are compared across distracters. For example;

Table 3 Computation of DIF using distracter method

Test item no 10	1	2	3	4
Group A	60	25	41	16
Group B	76	32	58	35

$\chi^2 = 2.66$, $df = 3$, $p = .50$

If a significant chi squared is found, then the item is biased.

The benefits of these approaches are: it does not require the majority of the remaining items to be unbiased (Veal & Forman, 1983). It gives more insight into why an item was biased, these insights allowed the item to be modified rather than discarded completely. A major concern with distracter analysis as a method for detecting bias was that distracter attractiveness was also a function of examinee ability (Scheuneman, 1979).

Regression Method:

This method stresses that a regression analysis for an item in two groups should yield equal intercepts, slopes and standard error of estimate in the two groups if the item is unbiased. A weighted-least-square, stepwise regression analysis involving goodness-of-fit statistics (chi-square, χ^2) is employed to test the hypothesis of equal parameters in the groups. If any of

the hypotheses is rejected, the item is flagged as potentially being biased (Marascuilo & Slaughter, 1981; Jensen, 1980; Reschly & Sabers, 1979; and Joshua, 1990).

Item Discrimination Method:

This method referred to as item correlation method. Jensen, (1980) defines an item as unbiased if the item total-score correlation (which is the discrimination) is the same in all subgroups of the population taking the test. For this to be determined, the point biserial (which Jensen prefers the biserial) correlation of item and total test scores is computed separately for each subgroup. The absolute value of the difference between the subgroups biserial coefficient represents the index of bias. The direction of bias is indicated by the sign attached to the difference (Ironson & Subkoviak, 1979).

Cochran's Chi-Square Test Method (CTg^2):

Cochran's test method is based on the use of g^2 statistics in detecting biased items. It was recommended for use by Marascuilo and Slaughter (1981) as they believe that the differences between the probabilities of success for subjects in each group are equal for all ability intervals. As a watch for the use of this method, the null hypothesis associated is that there is no difference in the probabilities of success across the entire ability group. A rejection of this hypothesis is an indication of differential difficulty across the compared groups.

The Item Characteristic Curve (ICC):

The item characteristic curve (ICC), otherwise called item response function (IRF) describes the relationship between an examinee's ability level and the probability of his answering an item correctly. It is also a graph of the percentage passing an item as a function of total raw score on the test. It is commonly used to reveal generally defective items in a test, but it is also one of the most direct and sensitive methods for bias items, (Warm, 1978; Jensen, 1980; Nenty, 1986).

Rank Order of Item Difficulties and Delta Decrements:

This is another variation of the transformed item difficulty method. Within it, all the test items are simply ranked for difficulty within the groups, and the Spearman's ranked order correlation, ρ or J_{AB} , is computed between the two sets of ranks. It does not matter whether the ranking is based on the p-values or Δ . Next is that, within the major group, the test items are ranked in order of the decreasing Δ value as in the major group (the item's actual order notwithstanding).

The third step is that the Δ decrement is obtained within each group. A decrement is the difference in values between adjacent items when the items are ranked. Example, $\Delta_1 - \Delta_2$, $\Delta_2 - \Delta_3$ and so on. In the major group all the Δ decrements are necessarily zero or greater than zero. The inter-item variance due to the rank order of item difficulties is completely eliminated by this procedure, leaving only the variance due to the unequal difficulty decrements between items.

The final step is the computation of Pearson r between the Δ decrements of the two groups. This index, r , indicates group resemblance in relative item

difficulties when the rank order of item difficulties is eliminated, and r^2 is the proportion of variance in the minor group's Δ decrements that is predictable from the item's Δ decrements in the major group. The value $1 - r^2$ is the proportion of the groups-by-item interaction variance that is due to group differences in the relative difficulties of items over and above that which is attributable to the item's differences in rank order.

Expert Judges

The method of using expert judges involved selecting individuals who represented a gender group and/or a different ethnic minority group or had special expertise about a particular group. These individuals were asked to rate items on their offensiveness or likelihood of being biased against a particular group. The use of expert judges has been useful in identifying items containing sensitive content and in increasing popular acceptability of a measure.

Uses of Differential Item Functioning

There are three general uses of DIF according to Zumbo and Gelin (2005):

1. Fairness and equity in testing. This purpose of DIF is often because of policy and legislation in which the group are defined ahead of time.
2. Dealing with a possible threat to internal validity. In this case DIF is often investigated so that one can make group comparisons and rule-out measurement artifact as an explanation for the group difference. The group are identified ahead of time and are often driven by the investigators research question (e.g., gender difference in mathematics achievement)

3. Trying to understand the (cognitive and/or psycho-social) processes of item responding and test functioning and investigating whether these processes are the same for different groups of individuals. In this context the groups are not identified ahead of time and instead latent class or other such methods are used to identify or create groups and learn about the process of responding to the item (Zumbo & Gelin 2005).

Theoretical Framework

Classical Test Theory (CTT)

Classical Test Theory (CTT) is one of the measurement theories that underlie the functioning of test-takers during examination. It is regarded as practical test theory based on the decomposition of observed scores (which are ordinal, but analysed as interval) to true and error scores. Classical test theory evaluates the examinee's true scores primarily according to the examinee's response observation scores which is examinee's response for certain item. That is, CTT attempts to evaluate the association between observation scores and true scores. Therefore, CTT is build on the true score model which is based on mathematical model. The assumptions of CTT are to define the relationship between observation score, true score and error score. According to (Croker &Algina, 1986), the three main assumptions are:

- i Observation scores are equal to true score plus error scores.

$$X = T + E.$$

Where X = score of an individual, T = true score component and E = extraneous variable error component.

The true score is the average of the observed scores obtained over an infinite number of repeated testing with the same test.

- ii There is no relationship between error score and true score. $\rho_{TE} = 0$

- iii There is no relationship between two error scores from two tests. $J_{EE} = 0$

The classical test model and procedures for constructing tests and interpreting test scores have served measurement specialist and other test users well for a long time. However, there are many shortcomings in the way in which educational and psychological tests are usually constructed, evaluated and used (Hambleton & Swaminathan, 1985).

The values of commonly used item statistics in test development such as item difficulty and item discrimination depend on the particular examinee samples where they are obtained.

Comparison of examinees on an ability measured by a set of test item comprising a test is limited to situation where examinees are administered the same test items. However, the problem is that because many achievement and aptitude tests are most suitable for middle ability students, the tests do not provide very precise estimate of ability for either high or low ability examinees.

Test reliability is one of the fundamental concepts defined in terms of parallel forms. The concept of parallel measures is difficult to achieve in practice.

Classical test theory provides no basis for determining how an examinee might perform when confronted with a test item. It presumes that the conditional standard error of measurement is the same for all examinee.

The major advantage of CTT is its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993).

The limitations of CTT can be summarise as circular dependency which poses some theoretical difficulties in CTT's application in some

measurement situations (e.g., test equating, computerised adaptive testing) as follows;

- i. The person statistics (i.e., observed score) is item sample dependent and
- ii. Item statistics (i.e., item difficulty and item discrimination) are examinee sample dependent, (Xitao Aan, 1998).

In the analyses of NECO examination items, the researcher will make use of person statistic, which is the observe total score of each examinee and item statistics, which is the difficulty indices of each of the items in the examination as stated in the classical test theory to calculate for differential item functioning.

CTT is by far the most influential theory of test scores in social sciences. In psychometrics, the theory has been superseded by the more sophisticated model in Item Response Theory (IRT). Item Response Theory (IRT) model however is catching on very slowly in mainstream research.

Item Response Theory (IRT)

The performance of examinees on a test item can be predicted or explained by a single factor, because the factor cannot be observed, it is called latent trait or ability which is the desired measurement objective. For instance if an examiner gives several tests in the same subject matter area to a group of examinees, he will find that in general the same examinees score high on the test and the same examinee score low. In other words we find consistency in the performance of the examinees on the different tests. To explain this consistency, we assume that there is something inside the examinees that causes them to score consistently; Jensen (1980) calls that something a mental trait. No one has found a physical reference for a mental

trait, and it's never observed directly. The scale of this trait is traditionally given the name of the Greek letter theta (θ). Here theta connotes the terms; ability level, amount on trait, and amount of subject matter knowledge interchangeably. The values of theta usually vary from -3 to +3 but may be observed outside this range Warm (1978). When an examinee walks into a testing room, he brings with him his theta; the purpose of the test is to measure the relative position of the examinees on the theta scale. The test as a measuring instrument interprets the examinee's theta and produces a measurement of ability, which is often the raw score.

The relationship between performance and ability can be expressed by monotonically increasing mathematical function which is called item characteristic Function (ICF). Item characteristic function provides the right response probability of various kinds of examinee ability level. If we use graph to express the relationship, it is called an item Characteristic Curve (ICC). Thus the probability of a correct response to an item by examinee will not depend on how many other examinees are located at the same level. In other words, the shape of an item characteristics curve does not depend on the distribution of ability in the examinees' population.

Every response data has corresponding item response measurement model and ICC because of its different types and assumption. Every ICC usually includes one or many parameters to describe item trait and examinee's ability. Therefore the shapes of ICC are different if the number of parameter is different. The most often seen is a non-straight regression line. Because the relationship between examinee's item performance and ability can be expressed by mathematical function, item response model is also called mathematical model.

The basic assumptions of IRT are:

i Unidimensionality: All of the items in the test measure the same ability or latent trait. In here, test can be expressed as item composition of measuring the same ability. The meaning of unidimensionality is simple, but it is not easy to correspond with unidimensionality for general measurement data. In general, unidimensionality means also that the items measure one and only one area of knowledge or ability Warm, (1978). However unidimensionality does not mean that the items correlate positively with each other. In fact, it is conceivable for all items to correlate negatively with each other and still be unidimensional.

ii Local Independence: The examinee's response in each item is independent in statistics. That is, his response to a particular item will not influence his or her response to any other item no matter whether examinee's response is correct or not correct on a particular item. Usually, local independence will be true if the assumption of unidimensionality can be proved successfully. The two concepts are interlinked on the definition of local independence Lord (1980). Local independence also presumes the probability for an examinee that an item score will fall in a pattern is simply the product of occurrence of the scores on each item of the test. Assume the probability of occurrence of a five item response pattern is v ;

1 = correct response:

0 = incorrect response:

P_1 = the probability of a correct response:

$1 - p_1$ = the probability of incorrect response.

Suppose the pattern of the occurrence of a five item response is:

$v = (1, 0, 1, 1, 0)$, then

$v_1 = p_1(1 - p_2) p_3 p_4(1 - p_5)$.

Thus the total probability is the product of the independent item response rates. Also p_1 is specific to a particular ability level.

iii Non-Speed Test: The bad test performance of examinee is because of lack of ability rather than of time confinement.

iv Know-right Assumption: If examinee knows the right answer on a certain item, he or she will certainly answer correct the item.

IRT was established in order to improve the above shortcoming of classical test theory, thus IRT possess the advantages that CTT doesn't have (Hambleton & Swaminathan 1985; Hambleton, 1989; Lord. 1991). These advantages and disadvantages are as follows:

Advantages of Item Response Theory IRT

i Item parameters adopted by IRT are different from the item parameters adopted by CTT. Item parameters adopted by IRT are free from influence by the samples.

ii IRT can provide the exact index of standard error for the ability estimation of each student rather than use the identical standard error to represent all the examinees.

iii Examinee ability estimates are independent of the particular choice of test items used from the calibrated population of items. It can be significant when comparing the scores coming from different tests or examinees.

iv Item information or test information of IRT can be the index of accuracy estimation of the item or of the test.

v IRT considers simultaneously the traits of examinee response type and item parameters when it calibrates the individual ability of examinee.

vi Statistics of goodness of fit can detect whether the model is appropriate to the data, whether the examinees response is unusual and whether test or item is biased.

Disadvantages of IRT

- i The assumptions of IRT build on the strict mathematical probability model that is difficult to be understood by the public; thus, it is limited in promotion and application.
- ii IRT is limited by strict basic assumptions and usually it is therefore necessary to have a big sample size. Therefore, the application area of IRT data is limited. It cannot be supported by general test users in promotion and application.
- iii IRT has to combine with computer technology in order to estimate it. It would be more widely accessible if the price of software is cheap and time consumption is more economical.

Looking critically at examinees response in National Examination Council (NECO) Senior School Certificate Examination (SSCE) which has been subsumed in the measurement theory of CTT and IRT, student's achievement in NECO may or may not function differentially in mathematic between gender of students and between locations of school. Measurement theory reveals why some student responds better than others considering the inherent or latent trait possess by the individual. This trait made it possible for high ability students to consistently function higher and low ability students to also function consistently lower in an achievement test.

When examinees preparing for NECO examination walk into the testing hall, they bring along with them, their theta (θ), which according to IRT, is ability level, an amount of subject matter knowledge

interchangeable. The NECO examination measures the relative position of the examinee on the theta scale, interpret the examinees theta and produce a measurement of ability in the form of raw scores. It is these raw scores that the researcher uses to calculate the psychometric properties of the examination and in turn used these properties in the difference methods of calculation DIF to determine the differential status of the examinees.

Empirical Studies

Tests and Differential Item Functioning

Differential Item Functioning DIF is a collection of statistical method that gives indication of items that are functioning differently for different groups of students (Gnnilla, 2003). Typically, DIF analyses are conducted after the tests are administered using large samples of examinee data.

The presence of bias or differential functioning item is a cause for concern because tests are used as a door opening for educational opportunities. It is a very important issue whether or not the test items are fair for every examinee. For instance, Ironson and Subkoviak (1979) found that, some items in the National Longitudinal Study (NLS) test battery differentially functioned in favour of white test takers. Van (1980) applied the logit model approach to a Word Exclusion and a Word Analogy test administered to 500 Tanzanian and 500 Kenyan students. The statistical method was able to exclude 8 out of the 29 items in the test. In their study of DIF between White and African American students on mathematics test items, Linn and Hirnisch, (1981:16) speculated that the "DIF associated with the racial groups may in fact masked the difference in the amount of instruction that the two groups received". The dissimilarity in instruction patterns could be attributed to "attendance patterns and school-to-school variability in content

coverage and emphasis. Further, Van, Mellengergh, Ader and Wijn (1984) applied the iterative logit procedure to the same group of 500 Tanzanian and 500 Kenyan on 29 items used by Van (1980). The total scores were split into five categories, and used 15 iterations to test the χ^2 statistics at .05 and .01 levels of significance. Without iteration, no item was identified as exhibiting DIF at .01 levels, but only two items were identified as exhibiting DIF when the significant level was .05. With iteration, 15 out of the 29 items were flagged biased. Rogers and Kulick (1987) in particular, with the item difficulty approach reported 14, 11 and 11 (out of 85) SAT-verbal items of the three forms of SAT as performing differentially in favour of city candidates. For Scholastic Aptitude Test SAT-Mathematics, Rogers and Kulick reported 7, 7 and 4 (out of 60) items on the three forms of SAT respectively as exhibiting the same bias. According to Scheuneman, (1987), Graduate Record Examination (GRE) general test was found to contain test items that measured different things for white and black examinees of the same subject matter ability. Schmitt (1988), used the item difficulty approach to study bias on the Scholastic Aptitude Test (SAT) for Hispanics, in particular, Mexican-Americans and Puerto-Ricans who reside in the continental United States and who speak English as their best language. Two forms of the SAT were studied. In the first form, 278, 166 city whites were used as the reference group, 2,963 Mexican-American and 3,230 Puerto-Rican candidates were respectively used as the focal group. In the second form, the reference group were 285, 885 whites with 3,456 Mexican-Americans, and 3,384 Puerto-Rican candidates as the focal groups. Schmitt observed no much bias in the SAT-Mathematics test. For the verbal test, he observed 6 and 12 items out of 85 biased for Mexican-Americans and Puerto-Rican groups respectively. Steve and Hillary (1994) conducted a

study to try to explain differences in the performance of black and white students on the Maryland School Performance Assessment Program (MSPAP) using a measure of Opportunity To Learn (OTL) rather than school socioeconomic status or race and ethnicity. Analyses of differential item functioning (DIF) used an item response theory-based approach focusing on the difference between observed versus expected response probabilities for a race-ethnic examinee subsample and each of the deciles groups in that subsample. 1 of 28 science items was flagged in favour of whites and 2 of 26 mathematics content items and 3 of 12 mathematics process items were flagged, all in favour of blacks. Abiam (1996) also reported research study in which he found that tests item used in the First School Leaving Certificate (FSLC) mathematics examination in 1992 in Cross River State differentially functioned in favour of some group of examinees. A total of 10 out of the 24 items that exhibited DIF were in favour of males while 14 items were in favour of females. Freedle and Kostin, (1988), researched whether the items of test exhibited item differential functioning for different groups, their findings established that the items were advantageous to whites.

Also Elliot & McGregor, (2001) assessed the presence of DIF for African-American and Anglo-American students in the Achievement Goal Questionnaire (AGO) using the ordinal logistic regression (OLR) and Logistic discriminant function analysis (LDFA). Both procedures flagged the same two items as having statistically significant DIF. The differences between Performance Avoidance, Performance Approach, and Task (Mastery) goals were evaluated in a sample of 525 African-American and Anglo-American sixth grade students taking a mathematics course. Their results indicated that, for females, African-Americans were more Mastery

focused than Anglo-Americans. DIF therefore means the performance difference between two groups of comparable ability or performance.

Under the Item Response Theory (IRT) framework, a test item is showing DIF if the Item Response Curve (IRC) is not the same for the boys and girls group who are equal in level on the latent trait, and who do not have the same probability of endorsing a test item, (Embretson & Reise, 2002). Cheong in International Journal of Testing (2006) investigated 2,076 U.S. ninth-grade students in 92 schools as part of the Civic Education Study of the International Association for Evaluation of Educational Achievement. The sample was composed of White Americans 67%, African Americans 15%, Hispanic Americans 11%, and Asian Americans, Native Americans, Pacific Islanders, and other 7%, using item response procedure found on the average, White students ($\bar{X} = 118.4$) scored higher than Hispanic ($\bar{X} = 106.0$) and Black students ($\bar{X} = 100.2$).

DIF matches the group on the latent variable of interest. Matching groups on the variable measured by the test is important for determining whether item responses are equally valid for different groups. Another investigation of the incidence of DIF in WAEC/SSCE Biology multiple choice tests used in 1999, 2000, and 2001 in Delta state was by Odili (2004). The analyses using Schuneman modified chi-square statistics shows that 10 out of 60 items of the multiple choice test used in WACE/SSCE biology examination in 1999 functioned differentially for test-tees from high and low socio-economic status (SES), none of the items in the test used in 2000 differentially functioned for this group of test-tees, while only one item differentially functioned in the test used in 2001 examination. 30 items in 1999 differentially functioned for candidates from urban and rural areas. 8 items in 2000 differentially functioned for test-tees from different location

while 14 items showed evidence of differential functioning for test-takers from different locations in 2001. 8 items in 1999 WACE/SSCE biology multiple choice test significantly functioned differentially for test-takers on the basis of gender, 2 items significantly functioned differentially for test-takers on the basis of gender in similar test in 2000 exams and 3 items in 2001. The results revealed that there are instances of DIF in biology multiple choice questions used in WACE/SSCE. Such items functioned differentially for test-takers described by socioeconomic status (SES), geographical location and gender. However, DIF is a statistical method engaged by the researcher to determine which subpopulation of test takers are at advantageous ability and which are not as would be decided by the analysis of NECO examination result response of students in the FCT Abuja. Therefore it is the first step in determining whether there is item bias or item impact as earlier discussed in the literature:

Gender and Differential Item Functioning

In the past few decades, considerable attention has been paid to gender differences on standardised achievement test. Robert, (1958), stated that gender difference in arithmetic reasoning and spatial relation, have been found. Men are superior to women in this test while women excel in tasks requiring verbal ability and memory. Similarly in a research conducted in Ghana, Mitchelmore, (1973), found a standard deviation, ($p < 0.01$) in mathematics achievement in favour of Ghanaian boys. When a sample of academically talented students in primary schools was given a test of mathematical reasoning ability, boys perform better than girls, Burnett and Lane, (1980) in Keeves (1990). In line with the above, Obioma and Ohuche, (1982) administered a teacher-made-test to a sample of 360 class five

students and analysis showed that male achieved significantly, ($p < 0.05$) higher than female. In another instance Obioma and Ohuche, (1982), in a longitudinal study found that there is a significant difference ($p < 0.05$), between male and female students in mathematics achievement. Abonyi, (1982), in a comparative study analysed school certificate result of 3000 students in Imo and Cross River States for the years 1978, 1980, and 1982. The purpose was to investigate any gender difference in students' achievement in mathematics. They made use of both t -test and F -test. They found that gender difference exists in students' achievement in school certificate mathematics in Cross River State, which supported the findings of Obioma and Ohuche, and non in Imo State which contradict previous findings. Obioma, (1984), use a Diagnostic Mathematic Achievement Test (DAMAT) as instrument; found that girls are more deficient in mathematics than boys. Obioma, (1986) again use 'DAMAT' as instrument, discovered that gender is a significant predictor of one's level of achievement in mathematics. Schmitt (1988) used M-H procedure to analyse for DIF in the Graduate Record Examination (GRE) general test and the Graduate Management Admission (GMA) test respectively, taken in 1985. The result showed that the test contains items with gender differential functioning. Buchanan, (1978), use third grade gifted boys and girls, fifth grade average boys and girls for an 8-week study, found that boys achieved higher than girls. To accurately interpret gender difference in performance assessment DIF is essential because it aids us in ruling out item and test bias in explanation for the observed gender difference in test performance. Therefore, standardised assessment must be fair so that examinees with equal ability levels have an equal probability of correctly answering each task. Miller, Dolittle and Ackerman, (1988), demonstrated that Mathematics

and English usage Test of the American College test contains items that differentially functioned against some group of testees. Again Doolittle and Welch (1989), investigated gender differences in performance on five tests of the Collegiate Assessment of Academic Proficiency (CAAP), including Reading, Writing Skills, Mathematics, and Critical Thinking and a direct measure of writing proficiency. Random samples of 1,000 males and 1,000 females were used for the study. From the analyses of results using the mean, standard deviations, t-statistics, and effect sizes found for each test ; the results indicate that females tended to perform better than males on the multiple choice Writing Skills Test and on the essay test; male tended to outperform females on the Mathematics Test. Ogu, (1989), found in a study carried out in Etiti local government area of Imo State that gender and school location have significant effect on the performance of Senior Secondary Students (SSS) in the SSCE chemistry examination. Likewise in the investigation of Differential Item Performance (DIP) of basic competence test for junior high school students between males and females in the three categories of Algorithms, Algebra and Geometry, Friedman & Lamon, (1994), found that females scored slightly better than males during elementary and middle school years, but males score substantially better than females in the high school and college years.

According to Harbor-Peters (2001), the gender issues in mathematics have been a source of aversion. Mathematics has been male stereotyped, since the subject is viewed by students as abstract, difficult and only reserved for boys continue Harbor-Peters. Speculation and research efforts concerning gender difference in cognitive abilities appear to be a national pre-occupation. Research results showed that boys performed better in mathematics reasoning abilities than girls in primary schools (Mills, Abland

& Stumpt, 1993). Agwagah (1993), found that female students performed significantly higher than their male counterparts in mathematics reading. One conclusion from several studies is that men have a better spatial ability than women, Geary, (1994). Studies have indicated that women are better than men in verbal skills and this give them advantage on items involving verbal communication. They also score relatively higher than the men on tests in mathematics that better match coursework (Willingham & Cole 1997). Men tend to outperform women in geometry and in arithmetic and algebraic reasoning questions. The researchers also stated that women tend to be better at intermediate algebra, arithmetic and algebraic operations. Gallagher, De Lisi, Holst, McGillicuddy, Morely, & Cahalan, (2000), studied gender differences in problem solving of conventional and unconventional problems. According to them conventional problems were defined as routine textbook problems with method that are clearly defined. Unconventional problems are seldom presented in the textbook and either requires an unusual use of familiar algorithm or can readily be solved by used of logical estimation or insight. The unconventional problems are more likely to be called problem solving according to the definition which is used in the National Test Bank for mathematics. In the test bank problem solving is defined as follows;

“Problems solving” items in this category should contain a more evident element of independent thinking or require the demonstration of creative ability. Items that require multiple steps but that can be judged to be complex in character should be assigned to this category even if the demand for independent thinking and creativity are small. In such cases the item should require insight of a more technical nature.

Gallagher et al, (2000), confirmed that men out performed women in all kinds of problems but that the difference were greater for problems requiring spatial skills or multiple solution path than for problem requiring verbal skills or containing classroom-based content. The differences between the sexes are smaller than between schools, social groups and ethnical groups according to Kimball, (1994). The study stated; women underestimation of their ability, the teacher's depreciation of the women's results, the men's better strategies for speed tests and women's tendency not to solve items that seem too difficult to them are some other reasons why men perform better on problem solving items in test. Obioma (1984) and Obodo (1990) findings recorded that, students in urban schools are better achievers in mathematics than their rural counterparts.

Umoinyang, (1991) similarly carried out indigenous investigation on the incidence of DIF in test used in some public examination. He analysed mathematics multiple-choice test used by WAEC in the 1990 GCE, using three DIF detection techniques: M-H statistics, Scheuneman chi-square statistics and Transformed Item Difficulty. The result showed that 29 test item differentially functioned in favour of candidates from educationally developed states, while 5 test items showed evidence of gender DIF. Joshua (1991) also used the Transformed Item Difficulty (TID-45) to investigate gender bias in 200 female and 200 male candidates taking the polytechnic mathematics entrance examination. Out of the 33 item in the test, 2 were identified as being biased in favour of the females. The same results were observed using the Schuneman chi-square procedure. In his conclusion, he supports the hypothesis that there is gender bias in mathematics achievement tests. Stevenson, Chen & Lee, (1993) claimed that mathematics achievement is significant between male and female when the students are in

high school or higher level. This means that DIF set in when students progresses from junior to senior secondary school and to higher level of schooling. Whenever male demonstrate higher test score than female, question arises concerning whether those disparities are due to actual achievement differences or bias in the test, or combination of both, Linn, (1993). Maccoby & Jacklin, (1994), maintain that males have a small advantage on mathematics and science examination whereas female tends to perform slightly better on verbal ability measures well. Erinosh, (1994), found that between 1985 and 1990, the most significant disparity between boys and girls in the WAEC result was in mathematics, with 55.5% boys scoring various grade up to the pass grade and for girls it is 44.5%. Wei, (1996), investigated the 482 fifth grade students of elementary school in the central of Taiwan. The study found that there was no significant relationship between the mathematics achievements of gender. The findings were close to the research of Randhawa, Beamer, & Lundberg, (1993), who indicated that mathematic achievement was not significant in elementary school but the mathematic achievement would be significant when the students enter the middle school. These confirmed the findings of Stevenson, Chen & Lee (1993). Ezenwa (1996) in Nwabueze (2006) determined the effect of gender and school type on the mathematics achievement of students. The researcher used 134 males and 106 females randomly selected from three schools in Minna, Niger State for the study. The data was collected using 30 multiple choice items on three mathematics concepts areas. The analysis of data revealed that male students were superior to the female in mathematics achievement. Wu, (1997) investigated the 464 fourth grade students of elementary school in the south of Taiwan. The study found there was no significant difference between the mathematics achievement of gender in

multiple choice, completion, calculation and application categories. The findings corresponded with the research results of Lee, (1983); Wei, (1996); Samules, (1991); and Tan, (1992). But their findings are different from the results of Lee, (1993) and Tsai & Walberg, (1983), whose research indicated that mathematics achievement of high school students, are significant between males and females. Males and females were equally capable in mathematics during elementary school which indicated no difference in performance.

Akukwu (1990) in Ozofo (2001) used a sample of 450 students in a study. The analyses of data collected using mean, standard deviation and two-way ANOVA, found that females are superior to males in achievement. But interaction between school type and location in achievement was not significant. When Abiam investigated DIF in mathematics test use in the First School Leaving Certificate (FSLC) examination 1992 in cross River State, the study revealed that some items in this test functioned differentially for candidates grouped by gender. Ozofo (2001), is consistent with the study of Akukwu (1990) in his study of the effects of two modes of Computer Aided Instruction (CAI) using 10 intact classes in Enugu State, the data collected were analyzed using mean, standard deviation and ANCOVA, found females superior to males in achievement, but interaction was not significant between location.

In general equality among examinees is commonly investigated in relation to gender and ethnicity; however, equality among examinees also implies comparisons among other contextual factors such as those at the community level as well as those at the individual level, (Zumbo and Gelin, 2005). These findings raised the question of whether a similar relationship might exist between male and female examinees in NECO multiple choice

examinations in the FCT. That is if there is going to be a differential in functioning of our examinees, which gender will it favour? Therefore to accurately interpret gender difference in performance assessment, DIF is essential because, it aids us in ruling out item and test bias in explanations for the observed gender difference in test functioning.

Location and Differential Item Functioning (DIF)

Rose and Rose (1962) in Odo (1998) and Barracks (1980), carried out studies to find out if the location of a school has anything to do with students difficulty in mathematics. They found location difference as a factor affecting achievement in mathematic. They concluded that differences in mathematics achievement were due to difference in socio-cultural setting and socio-economic status. They found that students from uniform socio-cultural setting and higher socio-economic status achieved higher in mathematics than those who are from poor cultural setting and low socio-economic status. In a similar study, King, (1985), used 58 secondary schools from London and outskirts of London. After collecting, collating and analysing for DIF in mathematics, found that there is significant difference between sub-urban and urban school in favour of sub-urban school. In addition Roger and Kulick, (1987), used Mantel-Haenszel technique to analyse the mathematical and verbal component of SAT test item for DIF. The result shows that some of the items differentially functioned against white test-takers from the rural settlement. This finding was supported by the result of a similar research carried out by Schmitt, (1988). He found that some test items in mathematical subtest of SAT were differentially functioned against white test-takers from the rural settlement.

In the first indigenous study of DIF, Inyang (1991), studies location bias on 522 rural and 512 urban Akwa Ibom and Cross River States

examinees at the 1980 Common Entrance Examination in Mathematics. The study used three detection methods. The modified Scheuneman chi-square (SSG^2) procedure, transformed item difficulties ($TID-45^0$) and item discrimination methods. The (SSG^2) method identified 13 items out of the 33 multiple choice test as being biased, 5 items were flagged biased by the transformed item difficulty, while 9 out of the 33 items in the test were spotted as exhibiting DIF by the discrimination procedure. Further in the results, 5 items already identified by the $TID-45^0$ were identified by the SSG^2 , six out of the nine identified by the SSG^2 method were also identified by the discrimination method, her findings supports the hypothesis that there is location bias in Mathematics achievement tests. Considering location again Akukwe (1991), used 450 senior secondary two (SS2) students with Continuous Assessment Cumulative Average Score (CACAS) as instrument, found that school location was a significant factor in student's achievement in mathematics. Students in urban school achieved higher than their counterpart in the rural setting which countered the findings of Kings (1985). Obodo, (1990), conducted a study involving three teaching models on performance of JSS 3 students in algebra his finding showed that urban students achieve better than their rural counterparts in algebra.

It is worthy of note that the International Association for the Evaluation of Educational Achievement (IEA:193) have pointed out that large schools in terms of size were located in urban areas and smaller schools in rural areas, and such school location could affect the nature of students and the quality of educational experience they receive. In the literature reviewed thus far, none of the findings in differential item functioning has been carried out in Federal Capital Territory Abuja to ascertain whether there exist DIF in mathematics in NECO SSCE

examination and to determine which school location (urban versus rural) exhibit the differential, hence the researcher wish to study, analyse and establish whether a similar set of findings as in above would hold in the FCT.

Summary of Literature

Functioning was highlighted to mean the comprehensive, basic and important ability of the learner, who was systematically instructed for a certain period of time, say three years in the case of senior secondary school. The test items is presume to be fair in which case, it should be comparably valid for every group of individual and that it should afford all examinee an equal opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the test purpose.

Sometimes bias or Differential Item Functioning set in. Differential Item Functioning DIF occurs when examinees with the same underlying ability on the construct measured by the test, but who are from different group have a difference probability of correctly answering the item. It is said to exist when a test or an item causes systematic error in the measurement. Test item bias can present itself in a variety of ways in items through the language, the format, the content or the behaviour required. Thus a desirable feature of an item is that higher ability students function significantly better on an item than lower ability students. In the study of gender differential, some findings shows significant in functioning in favour of males while some favours females, yet in some no significant in functioning was establish at all. But in location some of the findings favour urban schools. DIF and students functioning has been subsumed in the measurement theories; The Classical Test Theory (CTT) and the Item

Response Theory (IRT). The CTT evaluate the examinee's true score which the examinee response to certain item while in Item Response Theory IRT the performance of examinee on a test item can be predicted or explained by a single factor or trait called theta (θ). Because the factor cannot be observed, it is called latent trait or ability which is the desired measurement objective.

In the Federal Capital Territory Abuja, from the Education Resource Centre (ERC 2008), there is no record available to the researcher on DIF, thus the status of test-takers in Abuja is not known that is, if there is a differential in functioning or not and if there is, which gender is it favouring likewise location of schools. That is why the researcher is prompted by these inadequacies to undertake the study of DIF to detect differential functioned items in NECO SSCE examination and to identify the differential status of examinees in the FCT.

CHAPTER THREE

RESEARCH METHOD

This chapter specifically present the method that was used in carrying out the study under the following sub-headings: Research Design, Area of study, Population of study, Sample and Sampling Technique, Instrument for Data Collection, Validity of Instruments, Trial-testing and Reliability of Instrument, Method of Data Collection and Data Analysis.

Research Design

The design for this study is the Ex-Post-Facto research design. In this design the researcher has no control over the variables of interest and therefore cannot manipulate them Nworgu (2006). Data are collected after the event or phenomenon under investigation has taken place hence the name ex-post-facto. At the commencement of the study, the researcher finds that the subjects are already assigned to or classified into various levels of the variables whose effects are being investigated and cannot alter these. Thus the researcher deals with non-manipulable independent variables of which in this study, are gender and location. The researcher is surveying the differential status of the already taken examination on examinees in the FCT. It shall therefore describe the existence or non-existence of DIF in the test to infer from the result of the data analyses.

Area of the Study

The area of study is the Federal Capital Territory (FCT), Abuja. It is made up of the following area council: Abaji, Bwari, Gwagwalada, Kuje, Kwali, and Abuja Municipal area council (AMAC). Abuja is the seat of government and the administrative head quarters of all ministries, parastatals and other agencies in Nigeria. It houses the Three Arm zone viz: The Executive, The Judiciary, and The National Assembly. It is bounded by Niger State, Kogi State and Nassarawa State. Abuja as emerging capital City has rural areas mostly inhabited by the indigenes who are low income earners and others who cannot afford the high standard of living in the city. These areas have inadequate basic amenities like good roads, health services, water supply, electricity and schools. They are mostly farmers and traders, majority of which are subsistent. Abuja is chosen for the study because there is no record of research work on differential item functioning

of NECO examination carried out before now to ascertain the quality of the examination or its differential status on the test-takers.

Population of the Study

The population of the study is made up of all the forty-three (43) government coeducation senior secondary schools with a population of four thousand five hundred and sixty-eight (4,568) SS3 students in 2008 in the FCT, Abuja. These examinees were prepared by way of instruction of the mathematics curriculum from SS1 to SS3 and were expected to be at the same ability level to write NECO SSCE examination. According to statistics from Education Resource Centre (ERC 2009) Abuja, table 4 shows the total number and the distribution of students and schools.

Table 4 Distribution of schools and number of examinees

Location	No. of School	No. of students	Percentage (%)
Urban	32	3724	82
Rural	11	844	18
Total	43	4568	100

ERC (2009)

Sample and Sampling Technique

A sample of fourteen schools was used for the study. Proportionate Stratified Random sampling was employed to select the sample schools. The population was first stratified in terms of two subgroups, the reference group and the focal group. Samples were drawn randomly from each stratum (subgroup) in such a way that the relative proportion of the subgroup in the

samples is the same as in the parent population. In each of the sample school selected all the students who wrote the June/July 2008 NECO SSCE examination were studied. The reference group sample for the analyses was 28% of the total number of urban schools which is 9 schools, and the focal group sample is 45% of the total number of rural schools which is 5 schools, both adding up to 14 schools.

Instrument for Data Collection

The instrument for this study, developed by the National Examination Council NECO was the multiple choice mathematics test questions which comprises of 60 items of paper 1 in SSCE 2008 in FCT, Abuja. It was used to collect data on students' response to the items/questions in order to identify the presence of DIF. The option of the multiple choice items is due to the objectivity involved in scoring. The options/alternatives were collapse into dichotomously scores items which was 1 for the correct alternative and 0 for the distracters. Thus DIF would be attributed to examinee's ability and not the scoring bias of the examiner.

Validity of Instrument

The items covered the existing syllabus and were constructed and validated by experts in the Department of Examinations and Quality Control of National Examination Council NECO.

Reliability of the Instrument

The instrument was administered to 52 students outside the FCT who were not part of the study. The responses of the students were used to

determine the reliability coefficient of the items using Kuder-Richardson formula 20 (KR_{20}) and was found to be 0.86. See table 2.

Method of Data Collection

The data for this study were gathered from the multiple choice answer sheets of candidates who took the 2008 June/July NECO SSCE multiple choice mathematics examination. The researcher applied for the data from head office of NECO Minna. Person-by-Item response matrix which was given by NECO head office Minna was used by the researcher to map out the ability groups for each of the subgroups for the analysis of DIF. All the candidates from both the reference and the focal group were grouped into five score groups based on the ability level on the basis of observed total test scores. For example: excellent candidates (41 ÷ 60), very good (31 ÷ 40), good (21 ÷ 30), average (11 ÷ 20), and poor (1 ÷ 10) candidates. The p-values for each score-group was computed for comparison using the χ^2 statistics.

Data Analysis

Mantel-Haenszel delta was used to test the hypotheses that will enabled the researcher provide answers to the question of existence of items with significant DIF in the mathematics multiple choice items used in NECO (SSCE). The chi-square was tested at the .05 level of significance. The Schuneman chi-square statistics and Transformed Item Difficulty were used to detect DIF between the sub groups.

The M-H alpha indicates the directional relative ease in the test items.

CHAPTER FOUR

RESULTS

This chapter presents the analyses and interpretations of the data collected and prepared for the study. Mantel-Haenszel procedure was used to analyse the data and test the hypotheses for the purpose of identifying differentially functioning test items, while Scheuneman modified chi-square (SSG2) test and Transform Item difficulty (TID-45⁰) methods were used to identify items numbers that functioned differentially.

Research Question 1

How do the items in NECO 2008 June/July mathematics multiple choice questions differentially functioned in terms of gender?

To answer the research question the frequency of male and female examinees that succeeded or failed at each of the 60 items were subjected to three statistical analyses. Mantel-Haenszel procedure, Scheuneman modified chi-square statistics (SSG2) and the transform item Difficulty (TID-45⁰). Results of these statistical analyses to flag items that differentiate between boys and girls are presented in the tables 7, 8, and 9 below

Table 7 Summary of Mantel-Haenszel DIF Procedure for male (reference) and female (focal) groups at the NECO 2008 June/July multiple choice mathematics examination.

Item Number	M-H CHISQ	M-H α	M-H Δ
1	23.68	1.63	-1.15
2	0.02	0.98	.04
3	27.82	1.68	-1.23
4	24.90	1.62	-1.14
5	11.67	1.40	-.79
6	3.52	1.23	-.48
7	0.01	1.00	-.00
8	7.89	1.32	-.67
9	5.17	1.26	-.55

10	5.82	1.27	-.56
11	5.16	1.25	-.53
12	5.49	1.27	-.56
13	1.80	0.85	.38
14	27.37	1.78	-1.35
15	4.92	0.80	.52
16	1.52	1.05	-.12
17	0.47	1.59	-1.09
18	0.16	1.36	-.72
19	24.31	1.62	-1.13
20	0.32	0.93	.16
21	0.10	0.97	.08
22	2.38	0.81	.48
23	25.78	0.61	1.17
24	0.74	0.92	.21
25	18.31	0.63	1.09
26	38.77	1.86	-1.45
27	36.12	1.80	-1.39
28	0.25	1.05	-.13
29	0.97	1.11	-.23
30	3.63	1.21	-.46
31	3.72	1.33	-.68
32	0.02	1.02	-.05
33	3.25	0.83	.43
34	0.24	1.06	-.14
35	0.92	1.11	-.24
36	0.12	0.95	.12
37	0.14	0.94	.15
38	3.12	0.94	.14
39	3.43	1.23	-.49
40	1.40	0.87	.33
41	0.27	1.06	-.19
42	4.78	1.26	-.54
43	3.74	1.21	-.45
44	0.01	1.02	-.05
45	6.52	1.40	-.80
46	10.90	1.38	-.76
47	0.00	0.99	.03
48	28.96	1.74	-1.30
49	0.52	0.90	.24

50	13.16	0.70	.84
51	1.48	1.13	-.29
52	2.11	1.17	-.37
53	0.10	0.96	.09
54	10.01	1.40	-.79
55	2.03	0.86	.36
56	0.08	0.99	.03
57	8.72	1.37	-.74
58	0.00	0.99	.03
59	9.18	0.74	-.69.
60	0.42	0.93	.16

$P < .05 \quad |\Delta| = 1.5$

Looking at table 7 above none of the items from the index of DIF considered, significantly differentiate between boys and girls. That is the M-Hq did not exceed 1.5. Thus boys and girls did not significantly function differentially with this procedure.

Table 8 Summary of analysis of gender DIF in NECO 2008 June/July multiple choice mathematics examination using the Scheuneman Modified chi-square statistics (SS χ^2)

Items	Group	Number of correct responses by score group					Chi-sq
		1-10	11-20	21-30	31-40	41-60	
1	Male	4	162	252	109	41	36.64*
	Female	17	145	198	76	34	
2	Male	6	66	113	95	49	4.49
	Female	9	106	133	60	28	
3	Male	9	139	241	104	42	35.53*
	Female	9	138	179	80	32	
4	Male	8	103	228	50	37	21.36*
	Female	4	112	127	62	30	
5	Male	9	116	188	97	44	13.25*
	Female	7	127	150	74	35	
6	Male	6	53	124	42	42	38.25*

	Female	4	56	84	65	28	
7	Male	7	46	25	04	00	25.73*
	Female	5	55	20	44	00	
8	Male	7	49	131	90	35	29.74*
	Female	5	83	83	63	31	
9	Male	4	101	121	88	45	13.99*
	Female	7	130	95	56	31	
10	Male	5	79	193	101	29	28.96*
	Female	10	115	140	69	32	
11	Male	6	67	170	120	41	12.81*
	Female	9	101	153	72	31	
12	Male	11	191	216	98	44	20.72*
	Female	28	213	188	68	29	
13	Male	3	63	23	52	35	34.38*
	Female	1	45	82	52	25	
14	Male	9	84	100	57	34	41.84*
	Female	12	65	49	40	26	
15	Male	13	140	176	84	39	12.03*
	Female	18	214	169	75	35	
16	Male	3	81	110	97	39	21.68*
	Female	7	96	159	72	31	
17	Male	1	54	151	65	29	91.41*
	Female	3	50	120	39	11	
18	Male	7	67	79	34	15	15.46*
	Female	5	62	71	21	03	
19	Male	5	124	219	93	44	34.52*
	Female	9	117	164	74	29	
20	Male	4	53	70	47	39	4.83
	Female	5	66	79	52	29	
21	Male	11	89	177	39	44	53.24*
	Female	7	129	135	77	29	
22	Male	2	31	31	37	33	15.81*
	Female	5	65	39	27	27	
23	Male	10	82	141	90	38	19.30*
	Female	20	183	172	73	30	
24	Male	4	87	160	98	37	12.44*
	Female	8	134	168	71	34	
25	Male	1	50	60	61	34	23.12*
	Female	8	120	85	51	30	

26	Male Female	10 7	250 124	250 191	112 81	41 34	39.30*
27	Male Female	4 5	226 97	226 181	112 81	44 29	35.61*
28	Male Female	3 6	125 126	125 130	87 62	35 32	3.53
29	Male Female	8 3	110 99	110 92	85 65	41 29	14.86*
30	Male Female	10 8	127 107	127 101	92 69	41 30	15.30*
31	Male Female	3 2	69 40	69 37	16 15	08 05	20.67*
32	Male Female	3 7	107 126	107 85	79 57	39 31	15.17*
33	Male Female	5 9	101 134	131 138	60 69	37 31	9.02
34	Male Female	4 9	68 80	76 66	61 56	37 31	4.21
35	Male Female	10 5	95 123	109 111	75 48	38 28	7.54
36	Male Female	7 11	51 75	57 59	33 24	16 06	4.65
37	Male Female	3 5	42 59	39 37	16 11	07 04	2.26
38	Male Female	3 3	44 51	42 50	65 28	34 30	6.41
39	Male Female	5 5	61 75	129 102	37 23	07 05	13.92*
40	Male Female	0 7	90 115	65 60	26 31	08 03	15.68*
41	Male Female	3 5	85 106	100 98	59 39	35 30	5.07
42	Male Female	6 6	68 86	113 82	60 52	44 29	17.23*
43	Male Female	17 4	147 163	195 182	82 73	41 31	33.27*
44	Male Female	7 5	64 98	81 62	37 12	07 09	19.26*
45	Male	6	56	67	27	08	16.05*

	Female	2	53	53	11	08	
46	Male	7	121	201	102	44	16.15*
	Female	5	136	177	68	31	
47	Male	2	41	50	29	16	9.39
	Female	6	69	34	25	09	
48	Male	9	70	194	62	20	55.28*
	Female	3	87	116	27	22	
49	Male	2	32	61	27	16	17.37*
	Female	7	68	49	16	14	
50	Male	8	30	239	115	44	102.36*
	Female	15	195	216	74	24	
51	Male	6	132	142	86	39	10.95*
	Female	13	120	151	69	36	
52	Male	6	72	135	40	19	20.45*
	Female	6	89	106	42	07	
53	Male	3	55	91	75	39	3.21
	Female	8	83	98	58	30	
54	Male	0	75	153	60	35	19.38*
	Female	3	52	121	44	22	
55	Male	11	75	83	46	31	4.22
	Female	10	97	103	41	30	
56	Male	10	106	198	82	30	28.77*
	Female	13	170	159	67	31	
57	Male	9	67	77	58	37	15.18*
	Female	9	71	53	34	32	
58	Male	9	122	215	97	44	39.46*
	Female	20	228	164	62	28	
59	Male	4	91	145	83	42	6.14
	Female	8	149	170	76	36	
60	Male	1	64	121	75	42	22.21*
	Female	8	115	101	69	32	

*Item that differentially functions for the group. $P < .05$; $df = 4$; $G^2_{critical} = 9.488$

With the Scheuneman's modified chi-square test, forty-six (46) items in the test indicated by the asterisk (*) were identified as significantly exhibiting Differential Item Functioning between boys and girls at the .05 level of significant.

Table 9 Summary of Transformed item Difficulty (TID-45⁰) analysis for male and female examinees

Table 9

	Female			Male		
Items	P	Z	ϕ	P	Z	ϕ
1	0.55	-.13	12.46	0.67	-.44	11.24
2	0.39	.29	14.16	0.39	.29	14.14
3	0.50	-.00	13.00	0.63	-.33	11.67
4	0.38	.29	14.17	0.50	-.00	13.00
5	0.45	.13	13.50	0.54	-.09	12.64
6	0.28	.57	15.29	0.31	.48	14.93
7	0.10	1.30	18.21	0.10	.13	13.52
8	0.30	.51	15.05	0.37	.34	14.35
9	0.37	.33	14.30	0.42	.19	13.78
10	0.42	.20	13.80	0.47	.07	13.26
11	0.42	.20	13.80	0.48	.06	13.24
12	0.60	-.26	11.94	0.66	-.42	11.32
13	0.24	.72	15.88	0.21	.82	16.26
14	0.22	.77	16.08	0.33	-.43	14.71
15	0.59	-.22	12.12	0.53	-.08	12.68
16	0.42	.20	13.80	0.39	.28	14.13
17	0.26	.66	15.62	0.35	.38	14.50
18	0.19	.88	16.53	0.24	.71	15.85
19	0.45	.12	13.48	0.57	-.18	12.28*
20	0.27	.63	15.51	0.25	.67	15.68
21	0.43	.17	13.68	0.43	.18	13.70
22	0.19	.89	16.55	0.16	1.00	17.02
23	0.43	.17	13.66	0.43	.19	13.75
24	0.48	.06	13.24	0.45	.11	13.45
25	0.34	.42	14.68	0.24	.70	15.80
26	0.50	-.00	13.00	0.65	-.39	11.44*
27	0.45	.12	13.48	0.61	-.28	11.90
28	0.41	.23	13.92	0.42	.20	13.79
29	0.33	.44	14.76	0.37	.35	14.38
30	0.36	.36	14.42	0.41	.23	13.94
31	0.11	1.20	17.82	0.15	1.05	17.21
32	0.35	.38	14.52	0.36	.37	14.48
33	0.44	.13	13.53	0.39	.27	14.08

34	0.28	.59	15.36	0.29	.55	15.22
35	0.36	.35	14.42	0.39	.29	14.16
36	0.20	.74	15.94	0.19	.86	16.45
37	0.13	1.11	17.43	0.13	1.15	17.58
38	0.19	.89	16.57	0.22	.77	16.07
39	0.24	.70	15.81	0.28	.58	15.31
40	0.25	.68	15.72	0.22	.76	16.05
41	0.32	.47	14.88	0.33	.43	14.74
42	-.29	.55	15.18	0.34	.41	14.62
43	0.52	-.05	12.80	0.57	-.17	12.32
44	0.21	.79	16.18	0.23	.74	15.94
45	0.15	1.09	17.36	0.19	.86	16.45
46	0.48	.06	13.25	0.56	-.16	12.36
47	0.16	.98	16.91	0.16	.98	16.94
48	0.29	.55	15.18	0.42	.21	13.82
49	0.18	.93	16.71	0.16	.98	16.94
50	0.60	-.26	11.97	0.51	-.04	12.86
51	0.45	.13	13.54	0.48	.06	13.23
52	0.29	.56	15.25	0.32	.47	14.86
53	0.32	.47	14.98	0.31	.50	14.98
54	0.28	.59	15.36	0.35	.39	14.54
55	0.32	.46	14.84	0.29	.55	15.22
56	0.51	-.01	12.95	0.50	-.00	13.00
57	0.23	.74	15.97	0.29	.56	16.19
58	0.58	-.19	12.23	0.58	-.19	12.24
59	0.50	-.01	12.96	0.47	.09	13.35
60	0.37	.32	14.29	0.36	.37	14.46

(*) Items exhibiting gender DIF

From the table it is observed that two (2) items exhibit significant DIF. Item number 19 and 26 and are in favour of girls. These items are seen as outliers in the item point plot shown in figure1. That is they are 1.5 units below the 45⁰ line.

Figure 1 Item plot of the TID for boys and girls

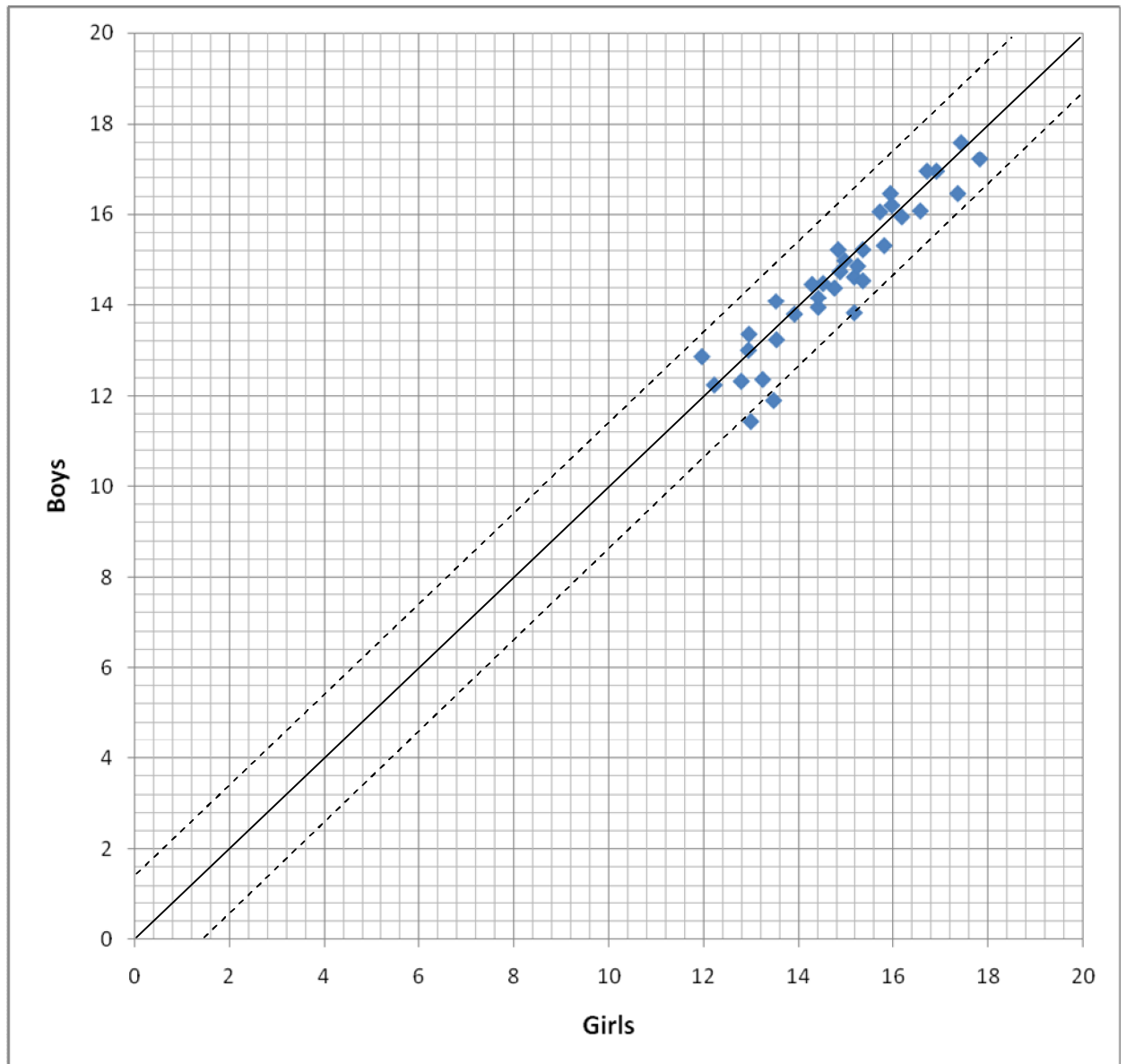


Table 10 Items identified as exhibiting gender DIF by the three detection methods

Methods	Items	Percentage (%)
M-H Procedure	Nil	00
(SSg2)	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,21,22,24,25,26,27,29,30,31,32,39,40,42,43,44,45,46,48,49,50,51,52,54,56,57,58,60	76.67
TID-45 ⁰	19,26	3.33

This table shows that none of the items show DIF with M-H procedure. 46 items representing 76.67% were spotted as showing DIF with (SSg2) statistics and only 2 item representing 3.33% are identified as showing DIF by the TID-45⁰.

(SSg2) statistics and TID-45⁰ converges items 19 and 26.

Hypothesis 1

Items in the multiple choice questions set and administered by NECO does not significantly function differentially among students based on gender.

From table 10, items 19 and 26 are common to SSg2 and TID-45⁰ methods, indicating that there is a slight difference in the functioning of gender in favour of female in NECO/SSCE multiple choice mathematics examinations.

Research Question 2

Are there differential functioned items in NECO mathematics multiple choice examinations in terms of school location?

The data gathered on urban and rural schools were subjected to three (3) detection methods: Mantel-Haeanszel procedure, Scheuneman modified chi-

square statistics and the Transform Item Difficulty method. The results of the analyses are presented in tables 11, 12, and 13.

Table 11 M-H procedure for NECO 2008 June/July SSCE multiple choice mathematics examination identified by school location

Item Number	M-H CHISQ	M-H $\bar{U}$	M-H ϕ
1	16.02	0.65	1.01
2	8.98	1.38	-.75
3	0.21	0.61	1.15
4	68.56	2.41	-2.07
5	16.74	1.52	-.99
6	36.79	2.06	-1.70*
7	2.64	0.75	.67
8	18.45	1.62	-1.12
9	9.31	1.38	-.76
10	13.09	1.46	-.89
11	7.29	1.33	-.66
12	2.49	1.19	-.40
13	24.47	1.93	-1.54*
14	15.49	1.60	-1.10
15	191.83	4.30	-3.43*
16	0.34	0.94	.15
17	8.07	0.73	.73
18	23.86	0.55	1.40
19	12.32	0.70	.85
20	0.18	1.06	-.13
21	35.99	1.86	-1.46
22	80.98	4.77	-3.67*
23	1.35	1.13	-.28
24	63.30	2.30	-1.96*
25	62.86	2.66	-2.30*
26	1.63	1.14	-.32
27	0.01	0.99	.03
28	0.01	0.98	.04
29	14.43	1.52	-.99
30	33.22	1.87	-1.47
31	23.00	2.27	-1.93*

32	30.18	1.84	-1.43
33	10.33	1.40	-7.97*
34	11.83	1.54	-1.02
35	8.46	1.37	-.74
36	0.41	0.91	.21
37	6.12	1.50	-.95
38	43.96	2.59	-2.23*
39	16.57	0.63	1.09
40	2.32	0.82	.48
41	3.98	1.25	-.52
42	7.67	1.36	-.73
43	0.02	1.02	-.04
44	1.93	0.87	.33
45	31.80	0.48	1.72*
46	0.91	0.86	.35
47	13.13	1.69	-1.23
48	20.17	0.62	1.14
49	0.38	0.90	.24
50	1.32	1.22	-.27
51	9.84	1.30	-.62
52	9.55	0.71	.80
53	110.51	3.69	-3.07*
54	1.86	0.86	.36
55	4.05	0.80	.53
56	6.64	0.77	.61
57	82.29	3.36	-2.85*
58	0.40	1.07	-.16
59	0.95	0.90	.24
60	34.50	1.91	-1.52*

$P < .05$ $|\Delta| = 1.5$ * items that exhibited location DIF

From the table, fourteen (14) items were identified to exhibit significant DIF at .05 level of significant. They are item numbers: 4, 6, 13, 15, 22, 24, 25, 31, 33, 38, 45, 53, 57, and 60. Out of which 13 items were in favour of the urban (reference) schools while 1 item was in favour of rural (focal) schools that is item 45.

Table 12 Scheuneman chi-square statistics for the analysis of DIF in the NECO 2008 June/July multiple choice mathematics examination identified by school location.

Items	Group	Number of correct responses by score group					Chi-sq
		1-10	11-20	21-30	31-40	41-60	
1	Urban	12	186	230	144	72	15.52*
	Rural	09	133	220	41	03	
2	Urban	11	98	143	136	74	19.29*
	Rural	04	74	103	19	03	
3	Urban	08	185	229	145	71	1.24
	Rural	10	92	191	39	03	
4	Urban	11	148	261	92	65	55.78*
	Rural	01	67	94	20	02	
5	Urban	07	148	221	140	76	17.23*
	Rural	09	95	117	31	03	
6	Urban	04	76	149	86	68	21.71*
	Rural	06	33	59	21	02	
7	Urban	08	57	27	06	00	6.03
	Rural	04	44	18	02	00	
8	Urban	03	74	146	130	63	28.35*
	Rural	09	58	68	23	03	
9	Urban	06	144	133	115	73	5.09
	Rural	05	87	83	29	03	
10	Urban	08	131	208	134	58	5.75
	Rural	07	63	125	36	03	
11	Urban	09	99	196	155	69	8.26
	Rural	06	69	127	37	03	
12	Urban	21	263	232	135	73	3.59
	Rural	18	141	172	31	03	
13	Urban	03	58	72	99	57	36.01*
	Rural	01	50	33	05	03	
14	Urban	09	96	93	87	60	14.50*
	Rural	12	53	56	10	00	
15	Urban	18	281	243	147	74	78.51*
	Rural	13	73	102	12	00	
16	Urban	04	132	111	132	67	30.42*
	Rural	06	45	158	37	03	
17	Urban	02	68	131	73	37	11.61*

	Rural	02	36	140	31	03	
18	Urban	06	68	63	43	17	26.01*
	Rural	06	63	87	12	00	
19	Urban	08	139	191	128	70	15.38*
	Rural	06	102	192	39	03	
20	Urban	07	78	62	80	66	16.44*
	Rural	02	41	87	19	02	
21	Urban	10	160	210	87	75	20.14*
	Rural	08	58	102	29	03	
22	Urban	05	79	53	64	60	43.76*
	Rural	02	17	17	00	00	
23	Urban	19	169	174	137	65	7.21
	Rural	11	96	139	26	03	
24	Urban	08	161	223	140	68	26.32*
	Rural	04	60	105	29	03	
25	Urban	05	135	100	96	61	27.89*
	Rural	04	35	45	16	03	
26	Urban	07	170	255	153	72	3.49
	Rural	10	94	186	40	03	
27	Urban	05	133	225	152	73	4.78
	Rural	04	88	182	41	03	
28	Urban	04	147	133	115	64	3.66
	Rural	05	87	122	34	03	
29	Urban	06	102	127	123	67	7.49
	Rural	05	63	75	27	03	
30	Urban	05	128	145	137	71	18.86*
	Rural	13	55	83	24	00	
31	Urban	03	43	86	36	13	38.32*
	Rural	02	25	20	00	00	
32	Urban	07	156	106	112	67	15.18*
	Rural	03	44	86	24	03	
33	Urban	07	163	155	105	67	3.35
	Rural	01	72	114	24	01	
34	Urban	04	84	98	104	65	27.71*
	Rural	09	64	44	13	03	
35	Urban	09	152	117	105	63	12.61*
	Rural	06	66	103	18	03	
36	Urban	09	74	62	48	22	6.23
	Rural	09	52	54	09	00	

37	Urban Rural	03 05	69 32	57 19	22 05	11 00	12.71*
38	Urban Rural	02 04	62 33	69 23	86 07	62 03	27.02*
39	Urban Rural	03 07	92 44	119 112	31 29	09 03	40.93*
40	Urban Rural	03 04	114 67	66 59	40 17	11 00	3.90
41	Urban Rural	03 05	128 63	106 92	84 14	62 03	5.65
42	Urban Rural	05 07	101 53	110 85	95 17	70 03	4.52
43	Urban Rural	12 09	210 100	196 181	123 32	69 03	3.74
44	Urban Rural	03 09	101 61	78 65	44 05	13 03	19.62*
45	Urban Rural	02 06	54 55	48 72	30 08	13 03	39.78*
46	Urban Rural	06 06	145 112	205 173	141 29	76 03	16.23*
47	Urban Rural	04 04	77 33	58 26	47 07	23 02	10.63*
48	Urban Rural	04 08	83 74	158 152	66 23	42 00	20.42*
49	Urban Rural	03 06	61 39	50 60	37 06	27 03	13.23*
50	Urban Rural	16 07	221 104	243 212	155 34	66 03	7.34
51	Urban Rural	10 09	167 85	158 135	135 20	72 03	9.64*
52	Urban Rural	09 03	87 74	133 108	58 24	24 02	17.30*
53	Urban Rural	07 04	105 33	153 36	116 17	67 02	61.09*
54	Urban Rural	00 02	66 35	128 146	86 18	57 00	16.64*
55	Urban Rural	13 08	109 63	78 108	63 24	61 00	20.64*
56	Urban	07	163	189	121	58	14.93*

	Rural	16	113	168	28	03	
57	Urban	08	107	105	84	66	50.12*
	Rural	10	31	25	08	03	
58	Urban	22	227	204	126	71	9.98*
	Rural	07	123	175	33	03	
59	Urban	04	153	153	126	75	10.75*
	Rural	08	87	162	33	03	
60	Urban	02	134	134	124	71	16.92*
	Rural	07	45	88	20	03	

*Items that differentiate between locations. $P < .05$; $df = 4$; $G^2_{critical} = 9.488$

A total of 41 items exhibited significant DIF at $p < .05$ level of significant. They are: 1, 2, 5, 6, 8, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 30, 31, 32, 34, 35, 37, 38, 39, 44, 45, 46, 47, 48, 49, 51, 52, 53, 54, 55, 56, 57, 58, 49, and 60.

Table 13 Transform Item Difficulty (TID-45⁰) analyses of NECO 2008June/July multiple choice mathematics examination resulting from school location.

	Urban Schools (ref. Group)			Rural schools (focal group)		
Items	P	Z	ϕ	P	Z	ϕ
1	0.56	-.19	12.24	0.68	-.46	11.18
2	0.41	.22	13.88	0.34	.42	14.68
3	0.57	-.18	12.30	0.56	-.14	12.44
4	0.52	-.05	12.81	0.31	.51	15.03*
5	0.53	-.07	12.72	0.42	.19	13.76
6	0.34	.42	14.66	0.20	.84	16.35
7	0.01	2.37	22.48*	0.11	1.21	17.84
8	0.37	.33	14.31	0.27	.62	15.48
9	0.42	.20	13.80	0.34	.40	14.60
10	0.48	.02	13.06	0.39	.28	14.12
11	0.47	.07	13.28	0.40	.26	14.02
12	0.65	-.38	11.49	0.61	-.27	11.92
13	0.26	.65	15.59	0.15	1.20	17.80*
14	0.31	.51	15.04	0.21	.80	16.20
15	0.68	-.47	11.12	0.30	.53	15.12*

16	0.40	.26	14.03	0.41	.22	13.86
17	0.28	.59	15.36	0.35	.38	14.52
18	0.18	.63	16.72	0.28	.58	15.32
19	0.48	.06	13.24	0.57	-.17	12.30
20	0.26	.63	15.52	0.25	.67	15.68
21	0.48	.04	13.16	0.33	.43	14.73
22	0.23	.73	15.92	0.06	1.56	19.22*
23	0.50	-.01	12.96	0.46	.02	13.06
24	0.53	-.09	12.64	0.33	.43	14.71*
25	0.35	.37	14.48	0.17	.95	16.80*
26	0.59	-.21	12.16	0.55	-.14	12.46
27	0.53	-.06	12.74	0.53	-.07	12.71
28	0.41	.22	13.87	0.42	.21	13.83
29	0.38	.31	14.22	0.29	.56	15.24
30	0.43	.17	13.66	0.29	.55	15.18
31	0.16	.99	16.94	0.08	1.42	18.67
32	0.40	.26	14.02	0.27	.62	15.50
33	0.44	.14	13.56	0.36	.35	14.40
34	0.32	.48	14.90	0.22	.77	16.07
35	0.40	.26	14.03	0.33	.45	14.80
36	0.12	.87	16.48	0.21	.82	16.28
37	0.14	1.06	17.24	0.10	1.27	18.09
38	0.25	.67	15.68	0.11	1.20	17.81
39	0.23	.75	16.00	0.32	.46	14.82
40	0.21	.81	16.24	0.24	.69	15.76
41	0.34	.41	14.62	0.29	.54	15.16
42	0.55	.41	14.64	0.27	.60	15.39
43	0.55	-.14	12.44	0.54	-.01	12.95
44	0.21	.79	16.18	0.24	.71	15.85
45	0.13	1.12	17.48*	0.24	.70	15.80
46	0.51	-.03	12.88	0.54	-.09	12.63
47	0.19	.89	16.56	0.11	1.18	17.70
48	0.32	.48	14.92	0.43	.18	13.73
49	0.16	.99	16.96	0.19	.88	16.52
50	0.63	-.32	11.71	0.60	-.25	12.00
51	0.48	.03	13.16	0.42	.20	13.81
52	0.28	.59	15.36	0.35	.39	14.57
53	0.40	.25	14.01	0.15	1.02	17.09*
54	0.30	.52	15.08	0.33	.43	14.71
55	0.29	.56	15.22	0.34	.42	14.67

56	0.48	.06	13.23	0.55	-.11	12.54
57	0.33	.44	14.75	0.13	1.14	17.54*
58	0.58	-.20	12.18	0.57	-.17	12.32
59	0.46	.19	13.76	0.49	.03	13.12
60	0.42	.21	13.86	0.27	.61	15.44

*Items that differentially function for the group

From table 13, 11 items were identified as functioning differentially. They are; 4, 7, 13, 15, 18, 22, 24, 25, 45, 53, and 57. Three items; 7, 18, and 45 were in favour of urban (reference) school while eight items; 4, 13, 15, 22, 24, 25, 53, and 57 were in favour of rural (focal) school. This information is shown on the item point scattergram on figure 2. The items that are functioning differentially are those outliers that lie above or below 1.5 units from the 45^0 line on the scattergram.

Figure 2 Item plot of the Transform Item Difficulty for urban and rural schools

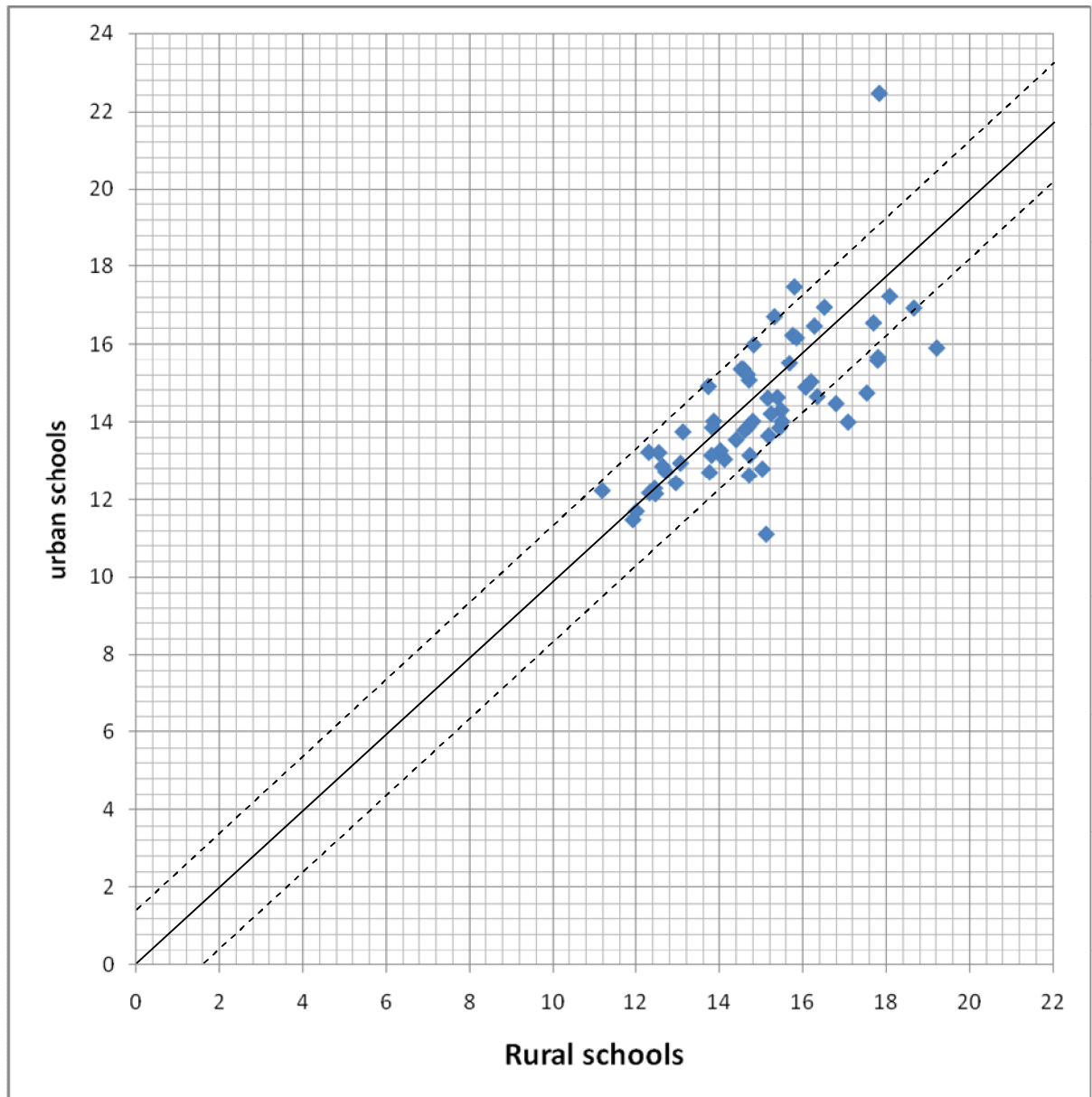


Table 14 A summary of the items identified as exhibiting significant DIF resulting from location in the FCT using the Mantel-Haeanszel procedure, Scheuneman chi-square statistics and the Transform Item Difficulty methods are presented below.

Method	Items	%
M-H procedure	4,6,13,15,22,24,25,31,33,38,45,53,57, and 60	23.33
SS χ^2	1,2,5,6,8,14,15,16,17,18,19,20,21,22,24,25,30,31,32,34,35, 37,38,39,44,45,46,47,48,49,51,52,53,54,55,56,57,58,49, and 60.	68.3
TID-45 ⁰	4,7,13,15,18,22,24,25,45,53,and 57	18.33

The three methods of detecting DIF have identified item variously in their analyses. From table 14 the highest number of items identified as functioning differentially is done by SS χ^2 . Items identified and common to the three methods are 15, 22, 24, 25, 45, 53, and 57.

Hypothesis 2

Items in the multiple choice questions set and administered by NECO does not significantly function differentially among students based on location. From table 14 above the three methods converged on 7 items (15, 22, 24, 25, 45, 53, and 57) and others that are significantly function differentially between locations. This is evident that school location significantly functions differentially for examinees.

Summary of Results

From the statistical analyses while answering the research questions for this study the following findings were made;

The person by item response matrix of the NECO 2008 June/July SSCE mathematics multiple choice examination was the instrument used by the three methods of detecting DIF.

M-H procedure at $p < .05$, $|\Delta| = 1.5$ did not detect any item as functioning differentially between gender of examinees in the FCT. However, it spotted 14 items as having significant different between school location.

SSG2 identified 46 items functioning differentially with respect to gender and 41 items exhibit significant DIF with respect to school location.

In TID-45⁰, 2 items were spotted as exhibiting DIF with respect to gender and both items favoured the girls. By school location, 11 items functioned differentially with 3 items favouring urban school and 8 items favouring rural schools.

But some items converged at the three methods. These items common to the three methods appears as exhibiting significant DIF. Thus the discussion shall be based on those items which are identified as exhibiting DIF by at least two methods.

CHAPTER 5

DISCUSSION, CONCLUSION, AND SUMMARY

This research was carried out solely to ascertain the differential status of examinees in the Federal Capital Territory (FCT) Abuja on the National Examination Council (NECO) Senior School Certificate Examination (SSCE) multiple choice mathematics test. The discussion is organised according to the hypotheses that were tested. The conclusions of the study, implications of the findings were also highlighted. The chapter will also presents recommendations and summary of the study.

Applying the three methods of detecting Differential Item Functioning (DIF): Mantel-Haenszel (M-H) procedure, Scheuneman Modified Chi-square (SSG2) statistics and the Transform Item Difficulty (TID-45⁰) methods, many items were identified as exhibiting significant DIF. However the discussion shall be based on items on which at least two of the detection methods agree.

Gender and Differential Item Functioning (DIF)

Results from testing hypothesis 1 using the M-H procedure shows that both boys and girls are performing at par. That is the results of the analysis did not exhibit differential in functioning of the examinees in terms of gender. But the SSG2 statistics is the reverse of this method, as it exhibited

46 items as functioning differentially with respect to gender. The SSg2 tests, spots out items which exhibit significant DIF at different ability levels. This is partly responsible for the high number of items identified as being differentially functioned. This means that the items identified as functioning differentially by the SSg2 tests, include all those items that are functioning differentially throughout all the ability levels, and those which may not function differentially at higher ability groups, but functioned differentially for the low ability examinees and vice versa Umoinyang (1991). In the TID-45⁰ only 2 items (19 and 26) exhibited DIF. The items also appear in Scheuneman chi-square test and both are in favour of girls. This results of using SSg² test and TID-45⁰ method is a contradiction to the research findings of Umoinyang (1991), Habor-Peters (2001), Abiam (1996) and Odili (2004), whose studies revealed that test item in the subject mathematics significantly function differentially for test takers described in terms of gender. That mathematics has been male stereotyped since the subject is viewed by students as abstract, difficult and only reserved for boys.

Under the Item Response Theory (IRT) framework, a test item show DIF if the Item Response Curve (IRC) is not the same for the boys and girls group who are equal in ability on the latent trait, and who do not have the same probability of endorsing a test item Embretson and Reise, (2002). Zumbo (1999) iterate that DIF statistics techniques are based on the principle that if different group of test takers have roughly the same level of something example knowledge, they should function similarly on individual test items regardless of group membership. Even though there are only 2 items in NECO/SSCE multiple choice mathematics examination that significantly function differentially in terms of gender, it indicate that

something else than what was intended was measured. But according to Zumbo (1991) again, DIF is a necessary but not sufficient condition for bias.

Therefore NECO should continuously subject their items to DIF analyses to ensure the fairness of the item. Since a fair test design provide examinees comparable opportunity as far as possible to demonstrate knowledge and skill they have acquired that are relevant to the purpose of the test. Thus considering SSG² test and TID-45⁰ method one can evidently say that there is a significant difference in the functioning of gender in the FCT. Girls perform significantly better than boys in the NECO 2008 June/July multiple choice mathematics examination.

Location and Differential Item Functioning (DIF)

The M-H procedure exhibited 14 items in the 2008 June/July NECO SSCE multiple choice mathematics examination in the FCT as functioning differentially with respect to location. 13 items were in favour of the urban school while only 1 item favours the rural school. The TID exhibited 11 items as functioning differentially in the examination of which 3 were in favour of urban schools while 8 were in favour of the rural schools. The analysis of SSG2 statistics identified 41 items exhibiting significant DIF at the .05 level of significant.

Out of the 14 items spotted by the M-H procedure as exhibiting DIF, 11 of the items were also detected by the SSG2 statistics and 9 items were again detected by TID-45⁰; and out of the 11 items exhibited as functioning differentially by the TID-45⁰ method, 8 of the items were also exhibited by

the SSc2 statistics. The three methods of detecting Differential Item Functioning (DIF) converge at 7 items which are items (15, 22, 24, 25, 45, 53, and 57). The findings of these 3 statistical methods of detecting DIF is in consonance to several research studies such as that of Obioma (1984); Obodo (1990); Inyang (1991) and Odo (1998), whose finding recorded that students in urban schools are better achievers in mathematics than their rural counterparts. Odili (2004) ascertains the different in functioning of examinees with respect to location in the multiple choice tests used by WAEC in the SSCE examination of 1999, 2000 and 2001. The findings of Akukwe (1991), using students continuous Assessment Cumulative Average Score found that school location was a significant factor in student achievement in mathematics. Students in urban school achieved higher than their rural set up.

The results of these study revealed that there are instances of DIF in terms of location in the mathematics multiple choice examinations used in the NECO/SSCE certificate examination. Such items functioned differently for examinees described by location. From table 14 we can see that the different in functioning tends to favour the urban schools, so we can confidently say that urban school function significantly better than the rural school in the NECO/SSCE 2008 June/July multiple choice mathematics examination in the FCT.

Conclusion

The aim of this study was to analysis the NECO/SSCE 2008 June/July multiple choice mathematics examination for possible DIF. From literature Sax (1980), asserts that no item is perfect no matter how much time and consideration it receives. Even the most carefully prepared items are

susceptible to human error and on analysis may prove to be ambiguous, too simple, overly difficult, or none discriminating. Brown (1983) attests that a test is only as good as the items it contains. Hambleton and Rodgers (1995) state that DIF can present itself in a variety of ways in items through the language, the format, the content or the behaviour required. Though we are not ruling out the effect of item impact of which one group of examinee is found to endorse the items more than the other group of examinees because the two groups truly differ on the underlying ability or factor being measured by the test. Others may be associated with difference in the amount of instruction received by the two groups which could be attributed to attendance patterns and school-to-school variability in content coverage and emphasis.

A complete evaluation of test quality must include an evaluation of each question. Therefore question should assess only knowledge or skill that are identified as part of the domain being tested and should avoid assessing irrelevant factors and examination items should be fair among examinees from all possible subgroup of test takers. The results of the analysis shows that SSg2 and TID exhibit difference in functioning of gender, thus we can state categorically that there is a significant difference in functioning of gender at the NECO 2008 June/July multiple choice examinations administered in the FCT in favour of girls. However the analysis using the 3 methods reveal significant DIF with respect to location of schools which indicate that school variability differentiate in the functioning of examinees and this favour the urban schools.

Implications

The following implications are noted;

Teachers should be aware that DIF may masked itself in actual instructions receive and school-to-school variability; They should endeavour to teach selflessly to cover the content or the recommended scheme of work to ensure that examinees function at their best during standardised examinations.

The issue of location variability in functioning of examinees during examinations is a recurring decimal, thus items writers for standardise bodies of examination like NECO, WAEC, NTI, JAMB, NABTEB and so on should note this and seek modalities of writing items that will not discriminate against location.

Students should be attentive in class and put in their best in the course of their study to reduce the effect of location differential in functioning during examination.

The results of research findings in education should be made available to stake holders as information to guide the process of development and accountability in the industry.

Standardised assessment must be fair so that examinees with equal ability levels have an equal probability of correctly answering each item.

Recommendations

From the foregoing the following recommendation can be made.

Examination bodies such as the NECO, WAEC, NTI, JAMB, NABTEB, and so on should readily release information to researcher who seek such for research studies which will expose the strength and weakness of examinations to enable them improve in their items writing and conduct of examination.

- They should make use of DIF for pilot study using gender, location and other possible groupings to ensure fairness in examination items.
- Selection of items from the result of pilot study should be done such that the test is made up of equal attractions to the groups used for the study.
- Examinees from all subgroups of test takers should work hard to achieve good results because DIF detection depends on their ability (person statistics).
- Research grant should be made available to researcher in the field of education.
- Researcher in education should be encourage to research more on differential item function, since literature is scanty on this topic in Nigeria.

Limitations

In the data given by NECO, all examinees in the schools considered were observed to response to all the 60 items in the multiple choice examination.

The analyses of the person-by-item response matrix for this study took a long time to complete.

Suggestions for further research

Further research study should be carried out to establish factors that bring about location differential in functioning of examinees in the FCT.

Factors that made it possible for gender to function at par in standard examination in the FCT.

Find out whether other school subject could function differentially for gender, location and other groupings in the FCT.

Other methods of analyses of detecting differential item functioning should be used other than the 3 used in these study.

Summary of Study

This study was to ascertain whether or not examinees in the FCT significantly functioned differentially in the NECO/SSCE 2008 multiple choice mathematics examination. The results of the analyses using the 3 detection methods of M-H procedure, SSg2 tests, and the TID show that there is a slight significant different in the functionality of gender in the NECO examination which can be observed in the agreement between the SSg2 tests, and the TID method. However SSg2 tests, exhibited 46 items as having functioned differentially with respect to gender. Also there is a significant different in terms of location. The M-H procedure shows 14 items functioning differentially with respect to location of which 13 favours the urban schools and 1 favour the rural schools. SSg2 tests located 41 items which functioned differentially and TID show 11 items on which 3 are in favour of urban schools and 8 items in favour of rural schools. Thus even though tests may be highly reliable, some items are prone to DIF that is to say that no mathematics items used in standardise examination could be entirely free from DIF, a position taken by other researcher from the results of other studies in their countries as recorded in the literature.

REFERENCES

- Abiam, P.O. (1996). An Analysis of DIF of 1992 First School Leaving Certificate Examination (FSLCE), in Cross River State. In Odili, J (2004). *Effect of Language Manipulation on Differential Item Functioning of Biology Multiple Choice Tests*. Publish P.hd Thesis. University of Nigeria, Nsukka
- Abonyi, E.N., (1982). *A Comparism Study of the performance of male and female student in W.A.S.C Mathematics Examination Cross River and Imo State*. Unpublished BSc Thesis UNN,
- Agwagah, U.N.V. (1993). *Instruction in Mathematics Reading as a Factor of Student Achievement and Interest in Word Problem Solving*. Unpublised Ph.D thesis UNN.
- Akukwu, A.C. (1990). *School Location and School Type as Factors in Senior Secondary School mathematics achievement in Imo state*. Unpublished M.Ed. Theses. UNN.
- Alio, B.C. (1997). *Polya's Problem Solving Strategy in Secondary School Students Acievement and Interest in Mathematics*. An Unpublished Ph.D Thesis University of Nigeria, Nsukka.
- Allen, M.J. & Yen, W.M. (2002). Introduction to Measurement Theory. www.measurement/theory.wikipedia.pdf
- Amoo, S.A. (2002). Analysis of Problems Encountered in Teaching and Learning ABACUS. *The Journal of Mathematics Association of Nigeria*. 27(1);30-34.
- Anastasi A. And Urbina S. (1998). Psychological Testing. Department of Psychology Fordham Univ. USA. Mac-Millan Publishing Company.
- Anderson, B. and Postl, B. (2000). Assessing the Performance of School: A Case Study with respect to Aboriginal Students; *Paper presented at the meeting of the Canadian Society for the Study of Education*. Edmoniton, A.B.

- Angoff, W.H. & Ford, S.F. (1973). Item-race Interaction on a Test of Scholastic Aptitude. *Journal of Educational Measurement*. 10;95-105.
- Angoff, W.H. & Sharon, A. T. (1974). The evaluation of Different in Test Performance of Two or More groups. *Educational and Psychological Measurement*. 34(807-816).
- Angoff, W.H. (1972). A Technique for the Investigation of Cultural Differences. *A Paper presented at the Annual Meeting of the ASA. Honolulu*. (ERIC Document Reproduction Service No. Ed 069686).
- Barracks, S.W. (1980). Achievement in and Attitude towards High School Mathematics with respect to sex and socio-economic status. *Dissertation Abstract International*. 14(5), 129-133.
- Benson, J (1998). Developing a Strong Program of Construct Validation: A Test Anxiety Example. *Educational Measurement Issues and Practice* 17,10-17
- Benson, J. & Hagtvet, K.A. (1996). The Interplay Among Design, Data Analysis, and Theory in the Measurement of Coping. In M. Ziedner & N.S. Endler (eds.). *Handbook of coping. Theory, Research and Applications*. New York: John Wiley & Sons, Inc.
- Brown, F.G. (1983). *Principles of Educational and Psychological Testing* (3ed.) New York: Holt, Rinehart and Winston.
- Buchanan K.N. (1987). Factors Contributing to Mathematical Problems Solving Performance. An Expository Study *Educational Studies in Mathematics*. 18(4), 59-66
- Camilli, G. & Shepard, L.A. (1994). *Method for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.
- Candall, C., & Coffman, W.E. (1964). A method for Comparing the Performance of different groups on the Item in a Test. *Research Bulletin* 4-661. Princeton, N.J.: Educational Testing Services.
- Chaimongkol, S., Huffer, W.F., & Kamata, A. (2007). *An Explanatory Differential Item Functioning Model by the WinBUG 1.4*. <http://math.bu.edu/people/mg/papers/dif.pdf>
- Chen, C., Lee, S.Y. & Setverson, H.W. (1993). *Mathematics Achievement of Chinses, Japanese, and American Children*.
- Cheong, Y.F. (2006). Analysis of School Context Effects on Differential Item Functioning Using Hierarchical Generalized Models. *International Journal of Testing*, 6(1), 57-79.

- Clauser, B.E., & Mazor, K.M. (1998). Using Statistical Procedures to Identify Differential Item Functioning Test Item. *Educational measurement Issues and Practice*, 17, 31-44.
- Cleary, T.A. & Hilton, T.L. (1968). An Investigation of Item Bias. *Educational and Psychological Measurement*. 28;61-75.
- Daman, D.N., & Wakjissa S. (2006). Issues in the use of Standardise Achievement Tests in Nigeria. University of Jos. Jos.
- Das, J.P.; Manos, J. & Kanungo, R.N. (1975). Performance of Canadian Native, Black and White Children on some Cognitive and Psesonality Tests. *The Alberta Journal of Educational Research*. 21;183-195.
- Demars, C.E., Wise, S.L., Madison J. (2007). Can Differential Rapid-Guessing Behaviour Lead to Differential Item Functioning? *A paper presented at the Annual Meeting of the American Educational Research Association, Chicago*.
- Doolittle, A. & Welch, E. (1989). Gender-based Differential Item Performance in Mathematics Achievement Items. *Journal of Educational Measurement*, 24, 157-166.
- Echols, P.S. (1992). *A study of relationships among students' attitudes toward mathematics and the variables of teacher attitude, parental attitudes, achievement, ability, sex of the student and grade level of the student*. Mich: UMI.
- Education Resource Centre (2009). Abuja.
- Ekoja, C.O., Happiness, I. & Ekoja, A.A. (2006). *Challenges of Public Examinations in Nigeria Society for Educational Psychologist*.
- Elliot & McGrejour, (2001). A 2X2 Achievement Goal Framwork. *Journal of Personality & Social Psychology*. 50,(501-512)
- Embretson, S.E. & Reise, S.P. (2002). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Association. www.questia.com
- Erinosho, S.Y. (1994). *Girls and Science Education in Nigeria*. Abeokuta: Ango Int
- Espino, M. (1992). Gender Partnership Development. *A paper presented at a conference of CAPA on policy seminar on gender equity in Technology Education and Location Training, Strategies for 1990 and beyond*. Daris Salem.
- Federal Republic of Nigeria FRN, (2004). National Policy on Education. (4th ed.) Lagos. NERDC Press.

- Freedle, R. & Kostin, I. (1988). *Relationship between Item Characteristics and Index of Differential Item functioning for the fore GRE Verbal item Types*.
- Friedman, L. and Lemon (1994). Meta-analytic Contributions to the Study of gender Difference in Mathematics-The Relationship of mathematical and Spatial Skills. *International Journal of Educational Research*. 21,361-371
- Gallagher, A.M., De Lisi, R., Holst, P.C., McGillicuddy-De Lisi, A.V., Morely, M. & Cahalan, C. (2000). Gender Differences in Advanced Mathematical Problem Solving. *Journal of Experimental Child Psychology*, 75(165-190).
- Geary, D.C. (1994). Children's Mathematical Development. USA. *American Psychological Association*.
- Gierl, M., Nylakhahq, S., & Boughton, K. (1999). *Gender Differential Item Functioning in Mathematics and Science Prevalence and Policy Implication*. www.education.ualberta.ca/educ/psych/crame/files/dif-csse99
- Glaswer, J. (1963). *Group in Guidance* (2nd ed.). Boston: Allyn and Becon.
- Gunilla Nasstrom (2003). Differential Item Functioning for Items in the Swedish National Test in Mathematics. Course B.
- Glickman, E.M., Seal, P. & Eisen V.S. (2009). *A non-Parametric Bayesian Diagnostic for Detecting DIF in IRT*. <http://tijepabooks.officelive.com/documents/article4v1>
- Green, D.R. (1980). Procedures Used for Reducing Bias in Tests at CTB/McGraw-Hill, 1966-1980. *A Paper Presented at the third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C.*
- Gunilla Nasstrom, (2003). *Differential Item Functioning for Items in the Swedish National Test in Mathematics course B*. Umea University
- Hambleton, R. & Rodgers, J. (1995). *Item Bias Review Practical Assessment, Research, and Evaluation*, 4(6).
- Hambleton, R.K. (1989). Principles and Selected Publications of Item Response Theory. In R.L. Linn. (ed.). *Educational Measurement*. Pp.147-200. New York. Macmillan Publishing Company.
- Hambleton, R.K., & Swaminthan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer. Nijhoff.
- Hambleton, R.K., Swamnathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. USA: SAGE Publications, Inc.

- Harbor-Peters, V.F.A. (2001). Unmasking Some Aversive Aspect of School Mathematics and Strategies for averting them. *Inaugural Lecture*. Enugu. Snaap Press.
- Holland, P.W. & Thayer, D.T (1988). *Differential Item Functioning and the M-H Procedure*. In H. Wainer & H.I. Brown (eds.). *Test Validity: 129-145*. Hillsdale, N: Lawrence. Erlbaum.
- Holland, P.W. & Thayer, D.T. (1986). *Differential Item Performance and Mantel-Haenszel Procedure*. Paper Presented at the AERA Annanual Meeting, San Francisco.
- Iheanyi, S.J. (2005). Effect of Peer Assessment on Academic Achievement and Interest of Students in Geometry. M.ed. Project. University of Nigeria.
- Inyang, M.I. (1991). Location Bias Analysis of Students Item/Test Performance *Paper Presentation at the 7th Annual Conference of the National Association of Educational Psychologist*
- Ironson, G.H. (1977). A comparative study of Several Methods of Assessing Item Bias. *Unpublished Doctoral Dissertation*, University of Wisconsin.
- Ironson, H.H. & Subkoviak M. (1979). A Comparative of Several Methods of Assessing Item Bias. *Journal of Educational Measurement*. 1:209-225.
- Jensen, A.R (1980). *Bias in Mental Testing*. New York; The Free Press
- Joseph, A.C. (2007). Testing the Assumption of Sample Invariance of Item Difficulty Parameters in the Rasch Rating Scale Model. www.pareonline.net/pdf/v137.pdf
- Joshua, M.T. (1990). Test Item Bias in Psychological Testing. *Unpublished Term Paper. Faculty of Education. University of Calabar. Nigeria*.
- Keeves, P.J. (1990). *Educational Research Methodology and Measurement. An International Handbook*.
- Kimball, M.M. (1994). It Is Only in Myth that Girls are Poorer in Mathematics. *Kvinnovetenskaphg tidskrift* 15(4)39-53
- King, W.H. (1985). Difference in Mathematics Achievement Related to Type of School and their geographical location. *Journal of Educational Research*. 8(1), 27-31.
- Lane, S., Wang, N. & Magone, M. (1996). *Gender Related Differential Item Functioning on a Middle-School Mathematics Performance Assessment*. <http://educ.ubc.ca/faculty/zumbo/dif> handbook.pdf.

- Lee, Mou-Eng (1983). The Relationship between Gender, Grades, Mathematics Attitude, Role and Mathematics achievement. *Unpublished Dissertation of Master. National cheng-chi University. Taipei.*
- Lesser, G. (1976). Cultural Difference in Learning and Thinking styles. In S. Messick (ed.) *Individuality in Learning*. San Francisco, CA; Jossey-B.
- Linn, R.L. & Harnisch, D.L. (1981). Interaction Between Item Content and group Membership on Achievement test Items. *Journal of Educational measurement*. 18:109-118
- Linn, R.L. (1993). The use of differential item functioning statistics: a discussion of current practice and future implications. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (349-364). Hillsdale, NJ: Erlbaum. ass,
- Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ; Lawrence Erlbaum Association inc.
- Maccoby, E.E. & Jacklin, C.N. (1994). The Psychology of Sex Difference. Standfrod. C.A. Stanford University Press.
- Marascuilo, L.A. & Slaughter, R.E. (1981). Statistical Procedures for Identifying Possible Sources of Item Bias Based on χ^2 Statistics. *Journal of Educational Measurement*. 18,229-248.
- Merz, W.R. & Grossen, N.F. (1979). An Empirical Investigation on Six Mathematics for Examining Test Item Bias. *Paper presented at the Annual Meeting of National Council on Measurement in Education. San Francisco.*
- Metcalf, A. (2008). Differential Item Functioning: Evidence for Instructionally Sensitive Items. Illinois State Board of Education
- Miller, S.K., Doolittle, A.E., & Ackerman, T.A. (1988). Differential Item Performance for Mexican-American ESL Students and White non ESL on Mathematics and English Achievement Tests. *Paper presented at the NCME Annual Meeting, New Orleans, LA.*
- Mills, Abland and Stumpt, (1993). In ñThe Mathematics Teacherò ed. By William, D.R. 34(3), 95-101.
- Mitchelmore, M.C. (1973). Performance in a Modern Mathematics Curriculum. W.A.J.E. 8(2)63-69.
- Nenty, H.J. (1986). Cross-Cultural Bias Analysis of Cattle Culture-fair Intelligence Test. *Paper presented at the Annual Meeting of the*

American Educational Research Association. San Francisco. California.

- Noortgate, V.D. and Boeck, P. (2005). Assessing and Explaining Differential Item Functioning using Logistic Mixed Models. *Journal of Educational and Behavioural Statistics. Winter, 30(4):443-464.*
- Nwabueze, M.U. (2006). Effects of Area Tiles Approach on Students Achievement and Interest in Quadratic Equation at senior Secondary School level. M.ed Project University of Nigeria.
- Nwana. O.C. (2007). Educational Measurement an Evaluation. Imo State University. Owerri, Nigeria. Boma Way Publisher.
- Nworgu, B.G. (2006). Educational Research. Basic Issues and Methodology.(2nd ed.)
- O'Neill, K.A., & McPeck, W.M. (1993). Item and Test Characteristics that are Associated with DIF. In P.W. Holland & H. Wainer (eds.), DIF:255-279. Hillsdale, N.J.; Erlbaum.
- Obioma, G.O. & Ohuche, R.O. (1982). Sex and Environmental Factors in Mathematics Achievement. *Abacus, 16(1)39-51.*
- Obioma, G.O.(1986). Differential Performance of Students in Secondary Mathematics: *A Paper presented at the 27th Annual Conference of STAN at FUTO 25-30th August.*
- Obodo, G.C. (1990). The Differential Effect of the Three Teaching Models on the Performance of Junior Secondary School Students in Some Algebraic Concepts: Unpublished Ph.D Thesis. University of Nigeria, Nsukka.
- Odo, I.O. (1998). Gender and School Location as a Factor of Students Difficulty in Secondary School Geometry. M.ED Thesis. University of Nigeria. Nsukka.
- Ogu, C.E. (1989). Evaluating the Performance of Students in Senior Secondary Chemistry in Ekiti L.G.A. B.Sc. (ed.) *degree Project. University of Nigeria, Nsukka.*
- Oluwu, A.A. (2006). Examinations in Nigeria. Appraisal, Issues and Problems.
- Ozofor, M.N. (2001). Effect of Two Models of Computer Aided Instruction on Students Achievement and Interest in Statistical and Probability. *Unpublished Ph.D Thesis.* University of Nigeria, Nsukka
- Perrone, M. (2006). Differential Item Functioning and Item Bias: Critical considerations in Test Fairness. *Teachers College, Columbia University*

Working Papers in TESOL & Applied Linguistics, 2006. 6(2).
<http://www.tc.columbia.edu/tesolalwebjournal>.

- Petersen, N.S. (1977). Bias in the Selection rule-bias in the test. *A paper presented at the 3rd International symposium on Educational Testing, University of London*. The Netherlands.
- Randhawa, B.S., Beamer, J.E., & Lundberg, I. (1993). Role of Mathematics Achievement. *Journal of Educational Psychology*. 85:41-48.
- Reschly, D.J., & Sabers, D.L. (1979). Analysis of Test Bias in four groups with the regression definition. *Journal of educational Measurement*. 16:1-9
- Roever, C. (2005). "That's non Fair!" *Fairness, Bias, and Differential Item Functioning in Language Testing Retrieved. November 18, 2006, from the University of Hawai*. www2.hawaii.edu/~rover/brownbag.pdf
- Rogers, H.J. & Kulick, E. (1987). An Investigation of Unexpected difference in Item Performance between Black and White taking the SAT. In A.P. Schmitt and N.J. Dorans (eds.) *DIF on the scholastic Aptitude Test*. Princeton, N.J., Educational Testing Services.
- Roussos, L.A., Schnipke, D.L. & Pashley, P.J. (2000). Formulation of the M-H Differential Item Functioning Parameter with Practical Implications. www.jstor.org/pss/2351824
- Ruder, L.M., & Convey, J.J. (1978). An Evaluation of Selected Approaches for Biased Item Identification. *A Paper presented at the Annual Meeting of the American Educational Research Association, New York*.
- Ruder, L.M., Getson, P.R., & Knight, D.L. (1980). *A Monte Carlo Comparison of Seven Biased Item Detection Technique*. *Journal of Educational Measurement*. 17:1-10.
- Samules, W.D. (1991). *Mathematics Achievement and Attitude in Grade Six Through Eight in Lebanon, Oregon*. Mech: UMI.
- Sax ,G. (1980). *Priciples of Educational and Psychological Measurement and Evaluation (3ed)*. University of Washington. Belmont, California. Wadsworth Publishing Company.
- Scheueman. J.D. (1979). A Method of Assessing Bias in Test Items. *Journal of Educational Measurement*. 16(3);143-152.
- Scheuneman, J.D. (1987). An Experimental, Exploratory Study of Causes of Bias in Test Items. *Journal of Educational Measurement*. 24(2);97-118.

- Scheuneman, J.D., & Grima, A. (1997). Characteristics of Quantitative Word Items Associated with Differential Performance of Female and Black Examinees. *Applied Measurement in Education*, 10,229-319.
- Schmitt, A. (1988). Language and Cultural Characteristics that Explain Differential Item Functioning for Hispanic Examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 25(1)1-13
- Schumacker, R. (2005). Test Bias and Differential Item Functioning. www.appliedmeasurementassociates.com.
- Shepard, L.A., Camill, G. & Williams, D.M. (1985). Validity of Approximation techniques for Detecting Item Bias. *Journal of Educational Measurement*. 22:77-105.
- Shepard, L.A., Camilli, G., Averill, M. (1981). Comparison of Procedures for Detecting Test-Item Bias with both Internal and External Ability Criteria. *Journal of Educational Statistics*. 6:317-375.
- Spray, A.J.& Miller, R.T. (1992). Performance of a M-H Statistics and the Standardised Difference in Proportion Correct when Population Ability Distribution are incongruent. <http://pareonline.net/pdf/v13n7.pdf>
- Steve. F. & Hillary. M. (1994). *A Preliminary Study of Black-White Performance Differences In the Maryland School Performance Assessment Program*. New Mexico
- Stevenson, H.W., Chen, C., & Lee, S.Y. (1993). *Mathematics Achievement of Chinese, Japanese, and American Children. Ten Years Later*, *Science*, 259:53-59
- Subkoviak, M.J., Mack, J.S. Ironson, G.H. & Craig, R.D. (1984). Empirical Comparison of Selected Item Bias Detection Procedures with Bias Manipulation. *Journal of Educational Measurement*. 21:49-58.
- Tan, L.C. (1992). To Analyse the Mathematics Attitude and Solving Ability of Children. *Journal of Taipei. City teachers college*, 5:621-677.
- Tatsuoka, K.K., Linn, R.L., Tatsuoka, M.M., & Yamamoto, K. (1988). Differential Item Functioning Resulting from the use of Different Solution Strategies. *Journal of Educational Measurement*. 25:301-319
- Tsai, S.L. & Walberg, H.J. (1983). Mathematics Achievement and Attitude Productivity in High School. *Journal of Educational Research*, 76(5):265-272.
- Umoinyang I.E. (2002). Regional Differential Item functioning (DIF) in Mathematics Achievement Test Using the Mantel-Haenszel (M-H) Statistics. *Global Journal of Educational Research* 1(1)23-30.

- Umoinyang, I.E. (1991). Item Bias in Mathematics Achievement Test; *Published M.ed. Thesis, University of Calabar.*
- Usman, K.O. (2002). The Need to Retain In-services Mathematics teachers for the Attainment of the Objective of Universal Basic Education. *ABACUS. The Journal of Mathematics Association of Nigeria.*(27-44)
- Utibe, U.J. (2009). Creating Secondary Education for Petrotism, Self-Reliance and Productivity. *Journal of Qualitative Education.*5(30)38-44
- Van, D. F. (1980). Vergelijkbaarheid Van Individual Test Prestaties. Lisse, The Neither Land Swet & Zeptlinger
- Van. D.F., Mellenbergh, G., Ader, A. & Wijn, M. (1984). An Iterative item bias Detection Method. *Journal of Educational Measurement*, 2(2)131-145.
- Vandenberghe, C.N. Gierl, M.J. (2001). Differential Bundle Functioning on three Achievement Tests. A Comparison of Aboriginal and Non-aboriginal Examinees. *A paper presented at the annual Meeting of the American Educational Research Association. (AERA). Seattle, Washington*, <http://www.education.ualberta.ca/educ/psych/crame>
- Veale, J.R. & Foreman, D.I. (1983). Assessing Cultural Using Foil Response Data: Cultural variation. *Journal of Educational Measurement*. 20: 244-258.
- Wang, N., Magone, M., & Lane, S. (1996). Gender Related Differential Item Functioning on a Middle School Mathematics Performance Assessment. *Educational Measurement Issues and Practice*, 15(4):21-27.
- Warm, T.A. (1978). *A Primer of Item Response Theory*. Spring Field, V A: National Technical Information services, US Department of Commerce.
- Wei, L.M. (1996). *Learning Motivation, Mathematics Anxiety and Mathematics Achievement of Primary School*. *Bulletin of Research on Elementary Education*, Graduate Institute of Elementary Education, Taichung Teachers College, 4;133-155.
- Williams, V. (1997). The “Unbiased” anchor. *Bridging the gap between Differential Item Functioning and Item Bias*. *Applied Measurement in Educaion*. 10:253-267
- Willingham, W.W., & Cole, N.S. (1997). *Gender and Fair Assessment*. Mahwah, N.J.; Lawrence Erlbaum Associates.

- Wim, V.D.N. & Boeck, P.D. (2005). Assessing and Explaining Differential Item Functioning using Logistic Mixed Models. Belgium. *Journal of Educational and Behavioral Statistics*. 30(4)443-464
- Wu, Y.L. (1997). *A Study Comparing the Differences on Mathematics Attitude and Achievement with Regard to Program Gender and Social Economics Status. Research on Elementary Education Peng-Doung Teachers college*. 1:163-200.
- www.journals.tc-library.org/templates/about/ed...**
- www.proedinc.com/customer/productview.aspx?ID=2045**
- www.vxu.se/msi/picne10/L2NG.pds**
- Xitao, F. (1998). Item Respond Theory and Classical Test Theory: An Empirical Comparison of Their Item/Person Statistics: *Journal of Educational and Psycal Measurement*. 58(357)25
- Yamamoto, K. (1988). Differential Item Functioning Resulting from the use of Different Solution Strategies. *Journal of Educational Measurement* 25:301-319.
- Yu, M.L. (1997). The Development Trend of Test Theory. In China Test Institute (ed.). *The Development and Application of Psychology Test* :23-62. Psychology Publisher: Taipei.
- Zumbo, B.D. & Hubley, A.M. (2003). *Differential Item Funtioning and Item Bias in R*. Fernandex-Ballesteros (eds). *Encyclopedia of Psychological Assessment*.:505-509. Thousand Oaks, CA.: Sage Press.
- Zumbo, B.D. (1999). *A Handbook on the Theory and Method of Differential Item Functioning; Logistic Regression Modeling as a Untary Framework of Binary and Liker-type (ordinal) Item Scores*. Ohawa. ON: Directorate of Human Resources Research and Evaluation, Department of National Defence. www.educ.ubc.ca/faculty/zumbo/dif
- Zumbo, B.D. and Gelin, M.N. (2005). *A Matter of Test Bias in Educational Policy Research: Journal of Educational Research and Policy Studies* 5(1). Roussonss, L.A., Schnipke, D.L. & Pashley, P.J. (2000). *A Formulation of The M-H DIF Parameters with Practical Implications*.
- Warm, T.A. (1978). *A Primer of Item Response Theory*. Spring Field, V A: National Technical Information services, US Department of Commerce.
- Warm, T.A. (1978). *A Primer of Item Response Theory*. Spring Field, V A: National Technical Information services, US Department of Commerce.

,

APPENDIX A

NATIONAL EXAMINATION COUNCIL (NECO)

**SENIOR SCHOOL CERTIFICATE
EXAMINATION (SSCE)**

GENERAL MATHEMATICS OBJECTIVE
EXAMINATION 2008

TIME 1HR; 45 MINUTES

2

PAPER I

Answer **ALL** the questions

Each question is followed by **five** options lettered A - E. Find out the correct option for each question and shade **in pencil** on your answer sheet the answer space which bears the same letter as the option you have chosen. Give only **ONE** answer to each question. An example is given below.

The product of three numbers is 3876. Two of the numbers are 17 and 19. What is the third number?

- A. 57
- B. 12
- C. 6
- D. 3
- E. 2

The correct answer is '12' which is lettered B. Therefore answer space B would be shaded.

[A] ~~[B]~~ [C] [D] [E]

Erase completely any answer you wish to change.

Do all rough work on this question paper.

1. Simplify $125^{1/3} \times 49^{1/2} \times 10^{-1}$

- A. $\frac{1}{35}$
- B. $\frac{2}{35}$
- C. $\frac{1}{14}$
- D. $3\frac{1}{2}$
- E. $4\frac{1}{3}$

2. Simplify $16^{3/4} \times 2^4 \times 32^{-1}$

- A. 0
- B. 2
- C. 4
- D. 10
- E. 20

3. Evaluate $27 \times \frac{2^0 \times 3^{-2}}{4^{-1/2}}$

- A. 48
- B. 12
- C. 6
- D. $\frac{3}{2}$
- E. $\frac{1}{3}$

4. Evaluate $\frac{6.42 + 2.13}{4.1 - 2.85}$, correct to two significant figures.
- A. 6.9
B. 6.8
C. 5.9
D. 5.8
E. 5.4
5. Without using tables, evaluate $2 + \log_{10} 20 - \log_{10} 2$.
- A. 2
B. 3
C. 4
D. 5
E. 6
6. Find the value of X if $213_{\text{four}} = X_{\text{six}}$.
- A. 54_{six}
B. 103_{six}
C. 124_{six}
D. 345_{six}
E. 1110_{six}
7. If $\left(\frac{4}{3}\right)^{m+2} = \frac{9}{16}$, find the value of m.
- A. -4
B. -2
C. 0
D. 2
E. 4
8. Find the 32nd term of the sequence 482, 480, 478, 476, ...
- A. -420
B. 420
C. 450
D. 514
E. 544
9. The sum to infinity of a G.P. is 80. If the first term of the series is 20, find the second term.
- A. 15
B. $11\frac{1}{4}$
C. 5
D. $1\frac{1}{4}$
E. $\frac{3}{4}$
10. If $x:y = 4:3$ and $y:z = 5:6$, find $x:z$.
- A. 8:5
B. 8:9
C. 9:8
D. 9:10
E. 10:9
11. A rope which is 4.85m long was measured by a boy as 4.95m. What is his percentage error?
- A. $2\frac{1}{50}\%$
B. $2\frac{3}{50}\%$
C. $2\frac{6}{97}\%$
D. $20\frac{3}{50}\%$
E. $20\frac{6}{97}\%$

4

12. A civil servant spent $\frac{2}{7}$ of his monthly salary on clothes, $\frac{7}{20}$ of the remainder on feeding and save the rest. If he spent ₦ 7,000.00 on feeding, find his monthly salary.
- A. ₦ 28,000.00
B. ₦ 32,000.00
C. ₦ 40,000.00
D. ₦ 48,000.00
E. ₦ 49,000.00
13. Given a Universal Set $U = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and the subsets $A = \{3, 5, 6, 7, 10\}$ and $B = \{2, 5, 8, 9, 10\}$. Find $A^1 \cap B^1$.
- A. $\{4\}$
B. $\{5\}$
C. $\{3, 4\}$
D. $\{4, 5\}$
E. $\{5, 10\}$
14. Simplify $56a^{-6} \div 7a^{-4}$.
- A. $8a^{-10}$
B. $8a^{-8}$
C. $8a^{-2}$
D. $8a^2$
E. $8a^{10}$
15. Given that $x = -\frac{3}{2}$, and $y = 4$, evaluate $3x^2y + xy^2$.
- A. -51
B. -3
C. 3
D. 57
E. 105
16. Evaluate $2 + \frac{2}{x} - \frac{x+4}{2}$.
- A. $\frac{2-x^2}{2x}$
B. $\frac{8-x^2}{2x}$
C. $\frac{4-x^2}{2x}$
D. $\frac{4+8x-x^2}{2x}$
E. $\frac{8-4x-x^2}{2x}$
17. If $x = 7$ and $y = -3$, evaluate $3x^2 - 4xy + 8y$.
- A. 255
B. 207
C. 87
D. 39
E. -87

5

18. Find the smallest number by which 1029 must be multiplied to make it a perfect square.

A. 3
B. 7
C. 9
D. 14
E. 21

19. Make S the subject of the formula $\frac{v}{S+t} = \frac{t}{S}$.

A. $S = \frac{t^2}{v+t}$
B. $S = \frac{t}{v-t}$
C. $S = \frac{t}{v+t}$
D. $S = \frac{t^2}{v-t}$
E. $S = \frac{vt^2}{v+t}$

20. If $t = \frac{3}{4}U^2 - 2$ and $U = \frac{4}{3}V^2 + 1$, find the value of t when $V = \frac{3}{2}$.

A. 5
B. 6
C. 9
D. 10
E. 12

21. Find the values of x for which $\frac{2x+5}{4x^2-9}$ is not defined.

A. $x = \frac{3}{2}$ or $-\frac{3}{2}$
B. $x = \frac{2}{3}$ or $-\frac{2}{3}$
C. $x = \frac{2}{5}$ or $-\frac{2}{5}$
D. $x = \frac{5}{2}$ or $-\frac{3}{2}$
E. $x = \frac{5}{2}$ or $-\frac{3}{2}$

22. The energy E possessed by a moving object varies directly as its mass m and the square of its velocity V. When $m = 8\text{kg}$, $V = 5\text{ms}^{-1}$ and $E = 100\text{ joules}$. Find E when $m = 6\text{kg}$ and $V = 2\text{ms}^{-1}$.

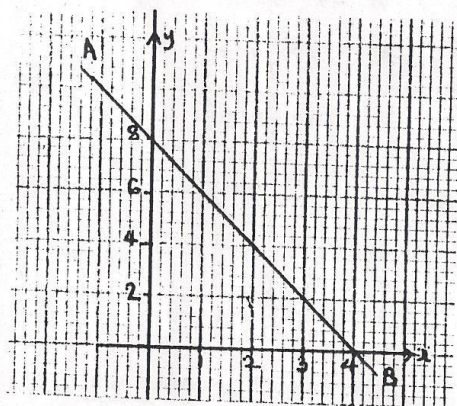
A. 6J
B. 12J
C. 24J
D. 48J
E. 96J

23. P varies directly as the cube root of Q. When $P = 20$, $Q = 125$. Find the value of P when $Q = 27$.

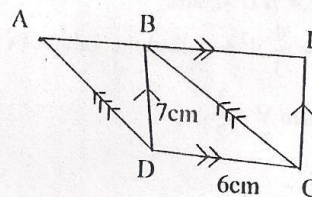
A. 12
B. 15
C. 24
D. 25
E. 30

6

24. What is the gradient of line AB in the following graph?



- A. -4
B. -2
C. 2
D. 4
E. 8
25. What must be added to $x^2 + 5x$ to make it a perfect square?
- A. $10x$
B. $\frac{5}{2}x$
C. $\frac{5}{2}$
D. $\frac{25}{4}$
E. 25
26. Simplify $\left(\frac{\sqrt{2}-\sqrt{3}}{\sqrt{2}+\sqrt{3}}\right)^2$
- A. $-19 - 20\sqrt{6}$
B. $-19 + 20\sqrt{6}$
C. $19 - 20\sqrt{6}$
D. $19 + 20\sqrt{6}$
E. $49 - 20\sqrt{6}$
27. Solve the inequality $(y-3) < \frac{y}{3}$.
- A. $y < 2$
B. $y < 3.5$
C. $y < 4.5$
D. $y > 4.5$
E. $y > 6$
28. Solve the equation
- $$\frac{3}{2(x-2)} - \frac{2}{3(2-x)} = 0$$
- A. 2
B. 3
C. 6
D. 13
E. 26
29. In the diagram, ABCD and BDCE are parallelograms. Calculate the area of the trapezium ADCE.



- A. 93 cm^2
B. 83 cm^2
C. 73 cm^2
D. 63 cm^2
E. 53 cm^2

30. The area of a rectangle is 30cm^2 and its perimeter is 22cm . State the dimensions of the rectangle.
- A. $10\text{cm} \times 3\text{cm}$
 B. $6\text{cm} \times 5\text{cm}$
 C. $15\text{cm} \times 2\text{cm}$
 D. $12\text{cm} \times 2.5\text{cm}$
 E. $25\text{cm} \times 1.2\text{cm}$
31. Calculate the total surface area of a solid cone of slant height 12cm and base radius 5cm in terms of π .
- A. $85\pi\text{cm}^2$
 B. $90\pi\text{cm}^2$
 C. $120\pi\text{cm}^2$
 D. $150\pi\text{cm}^2$
 E. $180\pi\text{cm}^2$
32. The angle of a sector of a circle of radius 4cm is 120° . If this sector is folded to form a cone, find the radius of the base of the cone. $\left(\text{Take } \pi = \frac{22}{7}\right)$
- A. $\frac{4}{3}$
 B. $\frac{16}{5}$
 C. $\frac{10}{3}$
 D. $\frac{14}{3}$
 E. $\frac{16}{3}$
33. Find the bearing of a point X from Y , if the bearing of Y from X is 110° .
- A. 070°
 B. 110°
 C. 250°
 D. 290°
 E. 310°
34. Calculate, in terms of π , the length of an arc of a circle of radius 12cm which subtends angle 240° at the centre of the circle.
- A. $16\pi\text{cm}$
 B. $24\pi\text{cm}$
 C. $32\pi\text{cm}$
 D. $48\pi\text{cm}$
 E. $96\pi\text{cm}$
35. Which of the following triangles are congruent?
-
- A. I and III only
 B. II and IV only
 C. I and II only
 D. II and III only
 E. I and IV only

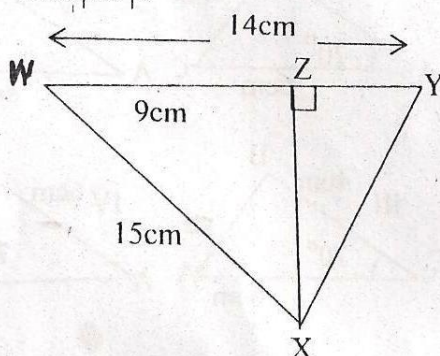
36. The sides of a right-angled triangle in ascending order of magnitude are 12cm, $(x + 2)$ cm, and $(x + 4)$ cm respectively. Find the value of x .

A. 23
B. 24
C. 33
D. 34
E. 43

37. If $(y - 20^\circ)$, $(x + 70^\circ)$ and 100° are three angles at a point, find the value of $x + y$.

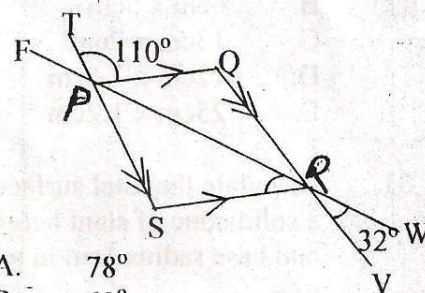
A. 30°
B. 90°
C. 120°
D. 180°
E. 210°

38. In the figure below,
 $\angle XZW = 90^\circ$, $|WZ| = 9\text{cm}$,
 $|XW| = 15\text{cm}$ and $|WY| = 14\text{cm}$.
Find $|XY|$.



A. 9cm
B. 10cm
C. 12cm
D. 13cm
E. 20cm

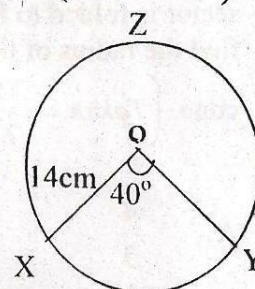
39. In the figure below, TPS, FRW and QRV are straight lines. If $\angle TPQ = 110^\circ$ and $\angle VRW = 32^\circ$, find $\angle PRS$.



A. 78°
B. 60°
C. 55°
D. 38°
E. 32°

In the diagram below, O is the centre of the circle of radius 14cm, $\angle XOY = 40^\circ$. Use this information to answer questions

40 and 41. (Take $\pi = \frac{22}{7}$).



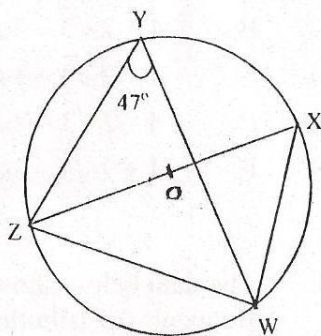
40. What is the length of the major arc XZY? (correct to 2 d.p.)

A. 4.89cm
B. 9.76cm
C. 9.78cm
D. 19.54cm
E. 78.22cm

41. What is the area of the minor sector XOY? (correct to 2 d.p).

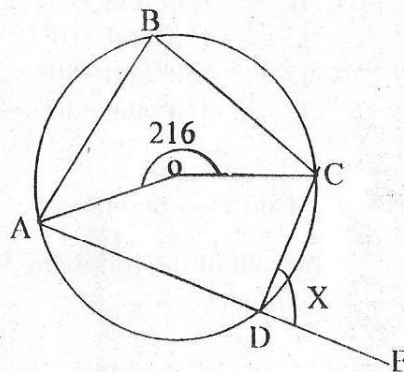
- A. 547.56cm^2
 B. 273.77cm^2
 C. 78.22cm^2
 D. 68.44cm^2
 E. 58.66cm^2

42. In the diagram below, O is the centre of the circle. If ZOY is a diameter and $\angle WYZ = 47^\circ$ find $\angle WZX$.



- A. 137°
 B. 117°
 C. 55°
 D. 47°
 E. 43°

43. Given a cyclic quadrilateral ABCD centre O with $\overline{AD}$ produced to E, if $\angle AOC$ is 216° . Find the value of $\angle CDE$.



- A. 72°
 B. 108°
 C. 144°
 D. 180°
 E. 216°

44. If $\tan x = \frac{a}{b}$, $0 < x < 90^\circ$, find $1 + \cos^2 x$.

- A. $\frac{a^2 - b^2}{a^2 + b^2}$
 B. $\frac{2a^2 + b^2}{a^2 + b^2}$
 C. $\frac{a^2 + 2b^2}{a^2 + b^2}$
 D. $\frac{a^2}{a^2 + b^2}$
 E. $\frac{b^2}{a^2 + b^2}$

45. If $\cos 30^\circ = 0.866$, find the value of the angles whose cosine is -0.866 .
- A. 120° and 330°
 B. 120° and 210°
 C. 150° and 210°
 D. 150° and 330°
 E. 210° and 330°
46. If $\sin x = \frac{3}{5}$ and $\tan x = \frac{3}{4}$, which of the following is correct?
- A. $0^\circ < x < 90^\circ$
 B. $90^\circ < x < 180^\circ$
 C. $180^\circ < x < 270^\circ$
 D. $270^\circ < x < 360^\circ$
 E. The value of x cannot be determined from the given information.
47. The angle of elevation of a point P on a tower from a point Q on the horizontal ground is 60° . If $|PQ| = 74\text{m}$, how high is P above the ground?
- A. $\frac{37\sqrt{3}\text{m}}{2}$
 B. $\frac{74\sqrt{3}\text{m}}{3}$
 C. $37\sqrt{3}\text{m}$
 D. $74\sqrt{3}\text{m}$
 E. 37m
48. The bearing of a point X from another point Y is 120° . Find the bearing of Y from X.
- A. 300°
 B. 210°
 C. 150°
 D. 120°
 E. 060°
49. Evaluate without using tables
- $$\frac{1 - \sin 45^\circ}{1 + \cos 30^\circ}$$
- A. $4 - 2\sqrt{3} - 4\sqrt{2} - \sqrt{6}$
 B. $4 + 2\sqrt{3} - 4\sqrt{2} + \sqrt{6}$
 C. $4 + 2\sqrt{3} - 3\sqrt{2} + \sqrt{6}$
 D. $4 - 2\sqrt{3} - 2\sqrt{2} - \sqrt{6}$
 E. $4 + 2\sqrt{3} - 5\sqrt{2} + \sqrt{6}$
50. The data below shows the frequency distribution of various categories of shoes purchased (in thousand) in the market by a trader
- | Categories | A | B | C | D | E |
|-----------------------|---|---|---|---|---|
| No of Shoes purchased | 2 | 3 | 4 | 5 | 6 |
| Amount (N'000) | 2 | 4 | 5 | 3 | 1 |
- Calculate the total amount spent on the shoes.
- A. 13
 B. 14
 C. 15
 D. 18
 E. 20

51. On which category of shoes is the modal amount spent?

A. E
B. D
C. A
D. C
E. B

52. Find the mean amount spent on the shoes to the nearest thousand.

A. 1
B. 2
C. 3
D. 4
E. 8

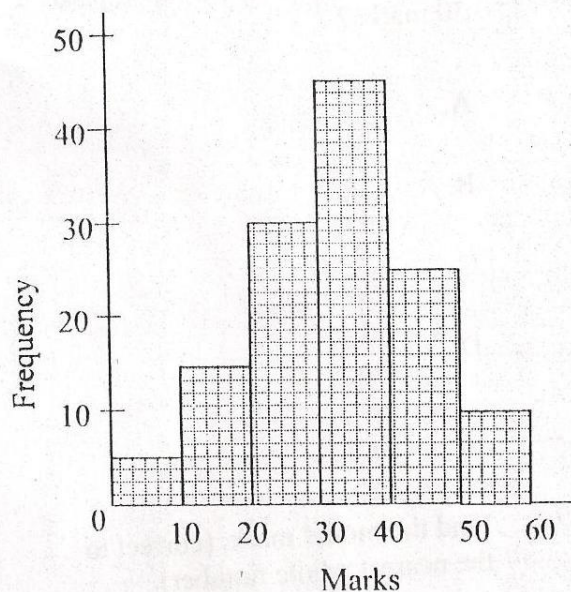
53. Considering this set of numbers, 3,3,4,5,6,6,6,7.
The mode, median and mean are respectively

A. 5,5,5.6
B. 5.5,6,5
C. 6,5,5.5
D. 5.5,5.6
E. 6,5,5.5

54. A number is chosen at random from the set $\{1, 2, \dots, 10\}$. What is the probability that it is a prime number?

A. 0.3
B. 0.4
C. 0.5
D. 0.6
E. 0.7

The histogram below shows the marks scored by the students in a class.



Use the graph to answer questions 55 to 57

55. What fraction of the students scored between 20 and 50 marks?

A. $\frac{7}{13}$
B. $\frac{7}{12}$
C. $\frac{4}{5}$
D. $\frac{5}{6}$
E. $\frac{10}{13}$

56. A student is selected at random from the class. What is the probability that he scored at least 30 marks?

- A. $\frac{8}{13}$
 B. $\frac{1}{2}$
 C. $\frac{9}{26}$
 D. $\frac{3}{14}$
 E. $\frac{1}{7}$

57. Find the modal mark. (correct to the nearest whole number)

- A. 33
 B. 34
 C. 35
 D. 36
 E. 37

50 students sat for NECO SSCE Mathematics examination, where the maximum mark obtainable was 100. and the following results were obtained:

Mark	Freq.
10-19	1
20-29	7
30-39	8
40-49	15
50-59	10
60-69	5
70-79	3
80-89	1

Use the information above to answer questions 58 and 59.

58. What is the mean mark?

- A. 14.5
 B. 29.5
 C. 33.5
 D. 46.1
 E. 57.3

59. What is the class interval?

- A. 9
 B. 10
 C. 14
 D. 19
 E. 20

60. A die and a coin were thrown once, what is the probability of getting a head and six?

- A. $\frac{1}{12}$
 B. $\frac{1}{6}$
 C. $\frac{1}{2}$
 D. $\frac{2}{3}$
 E. 1

APPENDIX B

Table 3: shows the number right correct answer for each of the question and the p-value for the calculation of reliability

Question Number	X=(no. correct answer)	p-value	q = 1-p	Pq
1	29	.5576	.4423	.2466
2	24	.4615	.5385	.2485
3	18	.3462	.6538	.2264
4	30	.5769	.4231	.2441
5	07	.1346	.8654	.1165
6	14	.2692	.7308	.1967
7	14	.2692	.7308	.1967
8	22	.4231	.5769	.2441
9	13	.2500	.7500	.1875
10	10	.1923	.8077	.1553
11	16	.3077	.6923	.2130
12	10	.1923	.8077	.1553
13	27	.5192	.4808	.2496
14	24	.4615	.5385	.2485
15	21	.4038	.5962	.2407
16	17	.3261	.6731	.2195
17	26	.5	.5000	.25
18	10	.1923	.8077	.1553
19	20	.3846	.6154	.2367
20	13	.2500	.7500	.1875
21	13	.2500	.7500	.1875
22	08	.1538	.8462	.1301
23	31	.5962	.4038	.2408
24	09	.1731	.8269	.1431
25	10	.1923	.8077	.1553
26	05	.0962	.9038	.0870
27	13	.2500	.7500	.1875
28	17	.3261	.6731	.2195
29	14	.2692	.7308	.1967
30	17	.3261	.6731	.2195
31	08	.1538	.8462	.1301
32	13	.2500	.7500	.1875
33	14	.2692	.7308	.1967
34	20	.3846	.6154	.2367

35	09	.1731	.8269	.1431
36	52	1	0	0
37	21	.4038	.5962	.2407
38	18	.3462	.6538	.2264
39	08	.1538	.8462	.1301
40	08	.1538	.8462	.1301
41	14	.2692	.7308	.1967
42	13	.2500	.7500	.1875
43	06	.1154	.8846	.1021
44	11	.2115	.7885	.1668
45	07	.1346	.8657	.1165
46	12	.2308	.7692	.1775
47	15	.2885	.7115	.2053
48	11	.2115	.7885	.1668
49	52	1	0	0
50	34	.6538	.3462	.2263
51	22	.4231	.5769	.2441
52	16	.3076	.6923	.213
53	20	.3846	.6154	.2367
54	10	.1923	.8077	.1553
55	52	1	0	0
56	52	1	0	0
57	10	.1923	.8077	.1553
58	09	.1731	.8269	.1431
59	13	.25	.7500	.1875
60	14	.2692	.7308	.1967
Total	1,066			10.6841

Table 3: shows the total score of each student in the achievement test use in the calculation of the variance and the reliability coefficient

S/N	X=(total score)	X ²
1	40	1600
2	38	1444
3	36	1296
4	33	1089
5	32	1024
6	32	1024
7	31	961
8	28	784

9	30	900
10	27	729
11	27	729
12	27	729
13	27	729
14	27	729
15	25	625
16	24	576
17	24	576
18	23	529
19	23	529
20	22	484
21	21	441
22	19	361
23	19	361
24	18	324
25	18	324
26	17	289
27	17	289
28	17	289
29	17	289
30	17	289
31	17	289
32	16	256
33	16	256
34	16	256
35	15	225
36	15	225
37	15	225
38	14	196
39	14	196
40	14	196
41	14	196
42	14	196
43	13	169
44	12	144
45	10	100
46	8	64
47	6	36
48	6	36
49	4	16
50	6	36
51	24	576
52	19	361
Total	1044	24592

$$S^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N} = \frac{24592 - \frac{(1044)^2}{52}}{52}$$

$$\frac{24592 - 20960.31}{52} = \frac{3631.69}{52} = 69.84$$

$$KR_{20} = \frac{N}{N-1} \left(1 - \frac{\sum pq}{S_t^2} \right) = \frac{52}{52-1} \left(1 - \frac{106841}{69.84} \right)$$

$$= 1.0196(1 - 0.1529)$$

$$= 0.847 \pm 0.85$$

APPENDIX C

Table 5 shows the distribution of ability by gender

Gender	41-60	31-40	21-30	11-20	1-10	Total
Male	44	130	318	320	37	849
Female	36	96	277	409	53	871
Total	80	226	595	729	80	1720

Table 6 shows the categories of the ability of the examinees with respect to location

Location	41-60	31-40	21-30	11-20	1-10	Total
Urban schools	77	174	335	486	47	1119
Rural schools	3	52	260	243	43	601
Total	80	226	595	729	80	1720