

TITLE PAGE

**RELATIVE EFFICIENCY OF TEST SCORES EQUATING
METHODS IN THE COMPARISON OF STUDENTS
CONTINUOUS ASSESSMENT MEASURES**

**A Ph.D THESIS SUBMITTED TO THE
DEPARTMENT OF SCIENCE EDUCATION
UNIVERSITY OF NIGERIA, NSUKKA**

BY

**AGAH, JOHN JOSEPH
PG/Ph.D/06/40715**

OCTOBER, 2013

APPROVAL PAGE

**THIS THESIS HAS BEEN APPROVED FOR THE DEPARTMENT OF
SCIENCE EDUCATION, UNIVERSITY OF NIGERIA, NSUKKA.**

BY

**PROF. B. G. NWORGU
SUPERVISOR**

INTERNAL EXAMINER

EXTERNAL EXAMINER

**PROF. D. N. EZEH
HEAD OF DEPARTMENT**

**PROF. I. C. IFELUNNI
DEAN OF FACULTY**

CERTIFICATION

This is to certify that AGAH, JOHN JOSEPH, a postgraduate student in the Department of science education with registration number PG/Ph.D/06/40715 has satisfactorily completed the requirements for the award of the Degree of Doctor of Philosophy in Measurement and Evaluation. The work embodied in this Thesis is original and has not been submitted in part or full for any other diploma or degree of this or any other university.

AGAH, JOHN JOSEPH
Student

PROF. B. G. NWORGU
Supervisor

DEDICATION

This work is dedicated to my late father who struggled to see this dream turned reality.

ACKNOWLEDGEMENT

The researcher is immensely grateful to God Almighty for good health, divine favour, provision and protection throughout the period of this research work. Making this work a reality was the dedicated effort of so many persons, institutions and Cross River State Government.

The researcher's appreciation goes to his supervisor Prof. B.G. Nworgu whose patience, tolerance, kindness, ever-preparedness to help, wealthy experiences and financial support propelled the researcher from the beginning of this study to the end. The researcher's fervent thanks go to Dr. B.C. Madu, Prof. K.O. Usman, Prof. U.N.V. Agwagah, Dr. A.O. Ovute and Dr. J.J. Ezeugwu who validated the instrument used for data collection and also read through the entire work with critical suggestions made. The path to this study was difficult to find until Dr. E.K.N. Nwagu read through the directionless concept paper and made some significant input that scholarly strengthen the researcher to commence the work in full. Dr. Frank Akubulo's effort in procuring the software (BILOG-MG and PARSCALE) that were used for data analysis from United State is highly appreciated.

The researcher also appreciates the former director of CUDIMAC Prof. E. Umeano, former Head of Department Science Education Dr. C. R. Nwagbo and the authority of University of Nigeria Nsukka for showing the researcher some measures of magnanimity to pursue this Ph.D. programme. The researcher thanks Mr. David Agah and Prof. Joseph Asor whose encouragement and invaluable advice greatly empowered the researcher to pursue this project more vigorously to the end. The effort of all the

teachers in Cross River and Rivers states that were used for administration of research instrument is highly appreciated. The researcher thanks Dr. (Mrs.) Ezeh and Mrs. C. Obi for organizing the students of Girls Secondary School Owerreĩ Ezeoba and University of Nigeria Secondary School to mark the answer scripts used for this study. Bro Patrick Asor is highly appreciated for helping the researcher to prepare the person by item matrix table for all the schools used in this study.

The researcher also wishes to express his profound gratitude to staff of WAEC and NECO particularly the quality Control Departments for discussing and showing the researcher how continuous assessment scores from various schools/states are standardized before they are incorporated into final assessment score. The researcher highly appreciates the financial support received from the government of Cross River State.

Equally appreciated are the following Ph.D students in measurement and Evaluation; Alabeke Christian, Ene Kate, Onyeabor Innocent, Dadughun Saurayi, Obika Grace, Ezeanya Loveline, Okofor Rita and Ayawei E. The researcher learnt a lot of things from their presentations during the internal seminars. The prayers of Bishop Ken Uloh, Pastor Okere (Ph.D), Evangelist. & Pastor Bethel Uloh, Pastor Francis and other Brethrens empowered the researcher to complete this work. Special thanks go to Sister Chika Sylvanus who painstakingly typed most of the work and Akobi Thomas who carefully formatted the work. I pray that God will bless you all.

The researcher immensely thank the chairman - Prof. D. N. Eze, the design reader - Prof. Z. C. Njoku, the content reader - Dr B. C. Madu and all panel members during the

proposal defence. Their criticisms and suggestions have helped in putting this work into a more refined and scholarly status it assumes now. Finally, the researcher appreciates the love, understanding and prayers of his wife - Priscillia, children - Vorda, David Ineneni, Etefolo and Eneda. To my dear mother, brothers and sisters thank you for tolerating my absence from home in the process of pursuing this study.

May God bless you all.

Agah, John Joseph

TABLE OF CONTENTS

Title page	-	-	-	-	-	-	-	-	-	i
Approval page	-	-	-	-	-	-	-	-	-	ii
Certification page	-	-	-	-	-	-	-	-	-	iii
Dedication page	-	-	-	-	-	-	-	-	-	iv
Acknowledgement	-	-	-	-	-	-	-	-	-	v
List of Appendices	-	-	-	-	-	-	-	-	-	x
List of Tables	-	-	-	-	-	-	-	-	-	xi
List of Figures	-	-	-	-	-	-	-	-	-	xii
Abstract	-	-	-	-	-	-	-	-	-	xiii
CHAPTER ONE: INTRODUCTION						-	-	-	-	1
Background of the Study			-	-	-	-	-	-	-	1
Statement of the Problem			-	-	-	-	-	-	-	12
Purpose of the Study			-	-	-	-	-	-	-	15
Significance of the Study	-	-	-	-	-	-	-	-	-	16
Scope of the Study	-	-	-	-	-	-	-	-	-	20
Research Questions		-	-	-	-	-	-	-	-	20
Hypotheses	-	-	-	-	-	-	-	-	-	21
CHAPTER TWO: LITERATURE REVIEW						-	-	-	-	23
Origin and Issues in Continuous Assessment (CA) in Nigeria	-							-	-	23
An overview of test score equating	-		-	-	-	-	-	-	-	34
Test scores equating methods/Models	-	-	-	-	-	-	-	-	-	62
Examinees's Ability, Population-Invariance/Differential Item Functioning (DIF)										78
Theoretical Framework	-	-	-	-	-	-	-	-	-	84
Classical Test Theory (CTT)	-		-	-	-	-	-	-	-	84
Item Response Theory (IRT)	-		-	-	-	-	-	-	-	87
Related Empirical Studies	-	-	-	-	-	-	-	-	-	91
Summary										111
CHAPTER THREE: METHODS						-	-	-	-	114
Design of the study	-	-	-	-	--	-	-	-	-	114
Area of the Study	-	-	-	-	-	-	-	-	-	115
Population of the Study			-	-	-	-	-	-	-	116
Sample and Sampling Techniques	-		-	-	-	-	-	-	-	117
Instrument for Data Collection	-		-	-	-	-	-	-	-	119
Validation of the Instrument	-		-	-	-	-	-	-	-	121
Reliability of the Instrument	-		-	-	-	-	-	-	-	122
Procedure for data collection	-		-	-	-	-	-	-	-	123

LIST OF APPENDICES

APPENDIX A; Request for validation of Instrument	-	-	-	-	181
APPENDIX B Letter of Introduction	-	-	-	-	182
APPENDIX C: Mathematics Achievement Tests (MAT) Test A ¹	-	-	-	-	183
APPENDIX D: Mathematics Achievement Tests (MAT) Test B ¹	-	-	-	-	188
APPENDIX E: Key to Test A ¹ and B ¹	-	-	-	-	193
APPENDIX F: Validator's Comment	-	-	-	-	194
APPENDIX G: Table of Specification	-	-	-	-	195
APPENDIX H: Item Facility Index and discrimination Index for Paper A ¹	-	-	-	-	196
APPENDIX I: Item Facility Index and discrimination Index for Paper B ¹	-	-	-	-	197
APPENDIX J: Cross River State Population Distribution for Twenty Three Schools	-	-	-	-	198
APPENDIX K: Rivers State Population Distribution for Twenty Two Schools	-	-	-	-	199
APPENDIX L: Sample Distribution of the Examinees in twenty three Schools from Cross River State	-	-	-	-	200
APPENDIX M: Sample Distribution of Examinees in Twenty Two Schools from Rivers State	-	-	-	-	201
APPENDIX N: Computation of the Reliability of Test A ¹	-	-	-	-	202
APPENDIX O: Computation of the Reliability of Test B ¹	-	-	-	-	204
APPENDIX P: Item Parameter estimates for test A ¹	-	-	-	-	206
APPENDIX Q: Item Parameter estimates for test B ¹	-	-	-	-	211
APPENDIX R: Differential Item Functioning for State-	-	-	-	-	216
APPENDIX S: Differential Item Functioning for Gender	-	-	-	-	221
APPENDIX T: Differential Item Functioning for Ability	-	-	-	-	226
APPENDIX U: Separate Calibration of ability for state A	-	-	-	-	231
APPENDIX V: Separate Calibration of ability for state B	-	-	-	-	234
APPENDIX W: Concurrent Calibration of ability for state A and B	-	-	-	-	237
APPENDIX X: Item Characteristic Curve for Test A1 and B1	-	-	-	-	241
APPENDIX Y: Linear Equating output	-	-	-	-	255
APPENDIX Z: Independent t-test analysis of student ability estimate when scores are standardize through linear equating-	-	-	-	-	278

LIST OF TABLES

Tables	Page
1: Item Parameter Indices of two parallel test - - - -	126
2: Independent t-test analysis of item parameter indices - - -	128
3: Item parameter consistency (differential Item Functioning) using Threshold (difficulty or b-parameter) Differences Values for State	130
4: Item parameter consistency (differential Item Functioning) using threshold (difficulty or b-parameter) Differences Values for Gender -	132
5: Item parameter consistency (differential Item Functioning) using threshold (difficulty or b-parameter) Differences Values for Ability -	134
6: Result of Chi-Square Goodness of Fit for 3PL IRT Model for A ¹	135
7: Result of Chi-Square Goodness of Fit for 3PL IRT Model for B ¹	136
8: Mean ability estimates of students in state A and B for scores equated through separate calibration - - - - -	140
9: Independent t-test analysis of students' performance when scores are standardized through Separate calibration equating - - -	141
10: Mean ability estimates of students in state A and B for scores equated through concurrent calibration - - - - -	142
11: Independent t-test analysis of students' performance when scores are standardized through Concurrent calibration equating - - -	143
12: Mean estimates of students score scaled through linear equating -	144
13: Independent t-test analysis of students' ability estimate when scores are standardized through linear equating - - - - -	145
14: Average Root mean square error - - - - -	146
15: Pearson's Product moment Correlation Analysis of MCA and MAT -	147

LIST OF FIGURES

Figures	Page
1: Schematic Representation of Test Score Equating Methods and Design for Test Scores - - - - -	38
2: Two tests with common item link - - - - -	42
3: A Chain of Two Links - - - - -	43
4: A Loop of Three Links - - - - -	44
5: Linking Five Different Test Forms - - - - -	44
6; Linking Five Different Test Forms using Box pattern - - - - -	45
7; Linking of continuous assessment scores with MAT - - - - -	45
8: Single-Group Design - - - - -	49
9; Counterbalanced Design - - - - -	49
10: Equivalent- Group Design - - - - -	50
11: Non-Equivalent Groups with Anchor Test design - - - - -	52
12: Summary of sample size - - - - -	119
13; Item Characteristic Curves of Test Form A ¹ and Test Form B ¹ -	138

ABSTRACT

The purpose of this study was to ascertain the relative efficiency of test score equating methods in the comparison of students' continuous assessment measures. Three equating methods (linear equating, separate calibration and concurrent calibration) based on classical test theory (CTT) and item response theory (IRT) frameworks were studied. The design of the study was "Non-Equivalent Anchor Test (NEAT) group Design" and the area of the study was Cross River (state A) and Rivers states (state B). The population of the study comprised of all senior secondary three (SSIII) students of 2010/2011 academic session in both states, and a sample of 2,905 students was drawn through multi-stage sampling procedure from the population. The instrument used for data collection were two parallel forms of Mathematics Achievement Test (MAT) with reliability of 0.83 and 0.89 respectively. The instruments were designed in such a way that students' scores from mathematics continuous assessment (MCA) were also obtained with it. Eleven research questions and six hypotheses guided the study. The research questions were answered using descriptive statistics, IRT differential item functioning likelihood ratio (IRTDIFLR) for three parameter logistic model and item characteristic curve (ICC). The six hypotheses were tested at 0.05 level of significance using independent t-test, chi-square statistic and Pearson product moment correlation coefficient. The data collected with MAT were analysed using BILOG-MG and SPSS. Major findings of the study showed that; (1) the average root mean square error (ARMSE) obtained were 0.09, 0.05 and 0.04 for separate calibration, concurrent calibration and linear equating respectively. This ARMSED values indicate that linear equating yielded the least error and therefore seems to be more efficient in this study, (2) there was no significant difference in the ability estimates of students in state A and B when their scores are scaled through separate calibration. (3) there was no significant difference in the ability estimates of students in state A and B when their scores are scaled through concurrent calibration. (4) there was a significant difference in the ability estimates of students in state A and B for MCA and MAT scores equated through linear equating. Other findings of the study include; (5) there was no significant difference in the item parameter estimates of the two forms of mathematics achievement test (MAT) used for test equating, (6) The items show negligible differential item functioning across states, sex and ability. (7) six items representing 15% did not fit the three parameter logistic model (3PLM) whereas 34 items representing 85% of the total test were not statistically significant and therefore fit the three parameter logistic model in both tests, (8) all the item characteristics curve (ICC) of the items in both tests were steep and vertically shifted towards the right corner of the curves except for item 1 and 2 in test A¹ whose ICC were slightly flat, (9) there was a significant relationship between the performances of students in mathematics continuous assessment test and mathematics achievement test. Based on these findings, it was recommended among others that linear equating method should be used to standardize students' continuous assessment (CA) scores. Score assigned to students' responses for every cognitive based continuous assessment should be reported in person-by-item response pattern. This will permit better CTT or IRT analysis to be performed.

CHAPTER ONE

INTRODUCTION

Background of the Study

In continuous assessment practice in Nigeria, different tests are developed by teachers and are used to determine the ability, proficiency or curriculum related achievements of students. A test according to Nworgu (2003), is a structured situation comprising a set of questions to which an individual is expected to respond, and each question in the test has a preferred answer. Nworgu further noted that the behaviour of an individual is quantified based on his responses to the questions. Onunkwo (2002) also described a test as an instrument which can be utilized in detecting some qualities, traits, characteristics, attributes, etc possessed by a person, an object or a thing.

Based on the above explanations, it can be deduced that, a test is a single occasion unidimensional timed exercise, usually in structured response item format or free response item format. It is used to quantitatively and qualitatively ascertain the magnitude of construct one possesses. It is one of the instruments used for the measurement and evaluation of students' educational achievements.

Testing in education helps in determining the learning difficulties or weakness, strength and level of mastery of examinees in a given task. One major goal of testing is to reveal the latent ability of examinees. The latent ability is determined from the number of correct answers made by an examinee and reported as raw test score or some norm-based transformation of it. When raw scores are standardized through any transformation, scaling or equating process, the resulting scores are known as measures (Moulton, 2004; Altonji, 2009).

The statistical characteristics of the tests used for continuous assessment in Nigeria vary across schools and depends on the characteristics of the population they were designed for. The characteristics of the schools and the teachers in these schools also vary. Given these variations, measurement errors are bound to occur from tests developed by individual teachers and the scores generated from them.

Scores obtained from different tests or examinations are often added up by teacher to get an examinee total score. When scores are treated in this manner, they are assumed to be interchangeable or comparable, even when they are not. The urge to use raw scores to make comparisons compel teachers to add up students' scores from different subjects and divide by total number of subjects to get the relative positions of students in the class. This act of adding up raw scores may lead to misinterpretation of marks because each assessment tool is crafted for a specific purpose and may not have the same mean and standard deviation. Therefore, for any comparison to be made over the achievement of students in tests, their test score should be standardized through an appropriate method.

In the classroom, students are assessed in a variety of ways by the teacher. As such, assessment is viewed from different perspectives by different teachers and authors. For instance, assessment according to Ajuonuma (2006), is a process of gathering data and fashioning them into interpretable form for decision-making. It involves the collection of data with the view to making valued judgment about the quality of a person, object or event. Anikweze (2005) refers to assessment as the process of investigating the status or standard of learners attainment, with reference to expected outcomes, that must have been specified as objectives when it concerns learners output.

The assessment system, the standard on which it is based, and all its parts must treat students equally. Assessment task should be sensitive to cultural, ethnic, class and gender differences, and to disabilities, and must be valid for and not penalize any group (Northwest Regional Educational Laboratory, 2001). To ensure fairness, students should have multiple opportunities to meet standards and should be able to meet them in different ways. No students' fate should depend upon a single test score. It therefore means that, several multiple forms of tests should be used to obtain students' test scores. This is exactly what continuous assessment (CA) practice in Nigeria is intended to achieve.

However, there exists an inherent problem in the implementation of continuous assessment (CA). This problem is in the quality of tests and comparability of test scores produced through CA. The problem arises probably because there are no nationally standardized tests used for a student's continuous assessment (CA). Rather each teacher in every school in the country prepares his/her own continuous assessment test (CAT). The CAT are used to obtain students continuous assessment scores (CAS) (National Teachers Institute, 2005).

Onjewu (2007) in her presentation on assuring fairness in the continuous assessment component of school based assessment practice in Nigeria observed that, there are inherent problems with the derivation of CA scores and the situation deserves the focus of academics. Onjewu further argued that, the entire practice of CA is surrounded by laxity and they are; laxity in timing, laxity in terms of the mode that the CA exercise takes, and laxity in relation to the content of CA. The continuous assessment scores (CAS) per se, has generated a lot of controversies between schools and

examination bodies like West African Examinations Council (WAEC) and National Examinations Council (NECO). Schools are accusing examination bodies of not using the CAS forwarded to them while the examination bodies are in turn accusing the schools of inflating the CAS sent by them.

The continuous assessment scores (CAS) takes 30% in Cross River, 60% in Rivers state, 40% in Enugu, Akwa Ibom, Anambra and Lagos states, of every examination at the secondary school level. A careful look at the CA scores assigned to students by schools in Cross River state and Rivers state shows a wide difference in the scores. This attest to the fact that, there is no uniformity in the practices of CA and hence the need to device a means of making the CA scores comparable.

At the primary, secondary and tertiary levels of education in Nigeria, CAS is incorporated or expected to be incorporated into the terminal assessment score (TAS) and there is growing concern from students, teachers, parents and some stakeholders on how effectively this is being done (Afemikhe, 2007). The terminal assessment, particularly those conducted by West African Examinations Council (WAEC) and National Examinations Council (NECO) has uniform items for each student across schools in Nigeria. This enhances comparability of raw scores. But in the continuous assessment situation, the comparison of raw scores from various schools is difficult and no reasonable measures have been taken towards this direction. The implementation of CA in Nigeria is one of the major functions of the teachers. The teachers's inability to effectively put CA into action may have caused some lapses in the grading, interpreting and reporting of students's performance. Pupils/students progression from primary to secondary level in some schools is now seen as automatic, and as such, they now move at

will from one school to another without their CA profiles which ought to be part of such movement. Many schools that admit such students fake their CAS which is forwarded to examination bodies (Afemikhe, 2007). This act to some extent constitutes pre-examination malpractices. One of the rationales for the introduction of continuous assessment was reduction of the incidence of examination malpractice, but this underlying reason seems elusive because the intensity of examination malpractice appears to increase astronomically.

Also the dismal performance of students particularly in Mathematics and English Language WAEC/NECO external examination calls for national concern. The continuous assessment scores forwarded to WAEC and NECO in all subjects and Mathematics and English Language in particular appear to indicate that most students' performance are above average. But surprisingly, the failure rate of students in public examinations that use CA is still high. How do the examination bodies standardize and compare CA scores from various states in Nigeria?

Comparability research conducted by Makiney, Rosen, & Davis, (2003); Merten, (1996); Pinsoneault, (1996); Mead & Drasgow, (1993) focused on the differences in means and standard deviations of test scores. The above authors placed little emphasis on underlying measurement issues like item parameters (Donovan, Drasgow, & Probst, 2000; King & Miles, 1995). Raju, Laffitte, and Byrne (2002) state that "without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully." It is therefore imperative not to only use mean and standard deviation in the comparability of student CA in Cross River and Rivers state but include other measurement issues such as item discrimination, difficulty and guessing parameters.

One of the core issue in comparing individuals and groups is to ensure that items of the tests used for assessment are consistent across all subgroups. Consistency is investigated through the determination of item bias or differential item functioning (DIF) between the groups in a study. This is often ignored by researchers /examination bodies despite the fact that it helps to minimize inappropriate interpretations arising from the use of tests. Examinees are often from different background/state, gender or ethnicity; this may systematically affect their performances on an item, and may lead to differential item functioning. It is therefore necessary to determine the consistency of the items of the test used to conduct measurement of ability.

The continuous assessment score forwarded to WAEC/NECO are incorporated into Terminal Assessment Score (TAS). Bearing in mind that examination bodies and schools in Nigeria seem to accept 100% as maximum score obtainable in each subject, the 30%, 40% or 60% allotted to CA is very significant and serious attention should be accorded to it. The continuous assessment Test (CAT) used for obtaining continuous assessment scores (CAS) in Nigeria varies across schools. Yet the scores obtained from CAT are incorporated into WAEC/NECO Terminal Assessment Score (TAS) whose items are uniform. The method of incorporating CAS into TAS (particularly by examination bodies) has been a source of worry to most persons (Onjewu, 2007; Afemikhe, 2007)

Currently, examination bodies like NECO incorporate CAS into TAS through the use of T-score transformation. The T-score transformation approach used by NECO and WAEC for the standardizing of CAS does not consider how scores are obtained. Also the T-score does not ensure that scores emanating from CA are authentic and actual

reflection of the students' academic achievement. This encourages faking of scores and does not guarantee accurate comparison of students' achievement.

The issue of comparability of standards, which deals with uniformity and quality of assessment instruments, as well as honesty and integrity in reporting of assessment result among others, has been a problem right from the introduction of continuous assessment. Based on the above issues, enhancement strategies such as moderation, self assessment and test scores equating have been suggested as educational standards control mechanisms in Nigeria (Afemikhe, 2007). Moderation and to some extent self assessment has been practiced especially at the tertiary level and not at the secondary level of education in Nigeria, whereas test score equating is not practiced at all. It is therefore pertinent to carry out a study on test score equating in Nigeria using CA scores from teacher-made-tests and researcher's mathematics achievement test score.

The assessment result, in which test scores are paramount, has a consequential effect on students' future. Therefore, the need for the standardization and placing of test scores on common scale to enhance fairness and comparability is as important as given the test itself. Different approaches can be used in ensuring that test scores are standardized and placed on a common scale, and one of such approach which has not been properly explored in Nigeria is test score equating.

The statistical process of making test scores comparable is called test equating (Kolen & Brennan, 2004). A process related to equating is linking or scaling of test scores, to achieve comparability. Wendy (2002) described test equating as a statistical procedure for measuring and controlling for variations in the difficulty (and other statistical characteristics) of different tests so that scores from equated tests have

comparable meaning. Test equating is a process used in comparing the test scores of more than one test form administered to examinees or group of examinees. The process of equating enables the test users to interchange multiple forms of a test (Angoff & Cowell, 1985; Keeve, 1990; Chong & Sharon, 2005). Test equating is an empirical procedure used in establishing the relationship between the raw-scores of two or more test forms. It enables the scores from one form of a test to be expressed in terms of the scores from the other form (Dorans & Holland, 2000; Vinder Linden, 2006;). Equating is also seen as a statistical process that is used to adjust scores on test forms so that scores on the forms can be interchangeable (Kolen & Brennan, 2004). Based on the definition of various authors above, one can deduce that, test equating is aimed at putting the scores obtained by students from different forms of a test on a common scale. In carrying out test equating, it is actually the scores from the test that are used. This is why it is called test score equating in most literature.

In conducting test score equating, numerous methods are available to the researcher. Some of these methods include; Mean Equating, Linear equating, Levine equally reliable linear equating, Levine unequally reliable linear equating, Tucker linear equating, Chained linear equating, Equipercentile equating (Frequency estimation equipercentile equating, Chained equipercentile equating), One parameter logistic (Rasch) model equating (Concurrent calibration, Fixed based procedure, Equating constant procedure, Major axis procedure), Two parameter logistic model equating (2pl concurrent calibration, 2pl partial credit model, 2pl generalized partial credit model), and Three parameter logistic model (separate and concurrent calibration).

These test scores equating methods are anchored mainly on the framework of two test theories namely; classical test theory (CTT) and item response theory (IRT). Classical test theory is founded on a test score model which assumes that, there are no perfect measures of ability. Each examinee's observed score (X) is comprised of True Score (T) and random error component (E) (Bielinski, Thurlow, Minnama & Scott, 2000; Wiberg, 2004; Schumacker, 2005). Item Response Theory (IRT) is based on the assumption that, there is a mathematical function that describes the relationship between an examinee proficiency and probability that an examinee will answer an item correctly (Chong, 2007; Hambleton & Swaminathan, 1995). For dichotomous items, the probability of correctly answering an item can be modeled mathematically using the logistic model or the normal ogive model. This relationship can also be represented graphically through the item characteristic curve or item response function (ICC/IRF).

The logistic model which dictated the three parameters: the item discrimination parameter (a), the item difficulty parameter (b), and the lower asymptote parameter (c) is known as the three parameter logistic model (Lord, 1980). The item difficulty parameter provides an indication of the difficulty level of the item and primarily dictates the location of the ICC with respect to the ability (θ) scale. A larger difficulty parameter results in a more difficult item and shifts the ICC upscale in reference to the ability scale. The item discrimination parameter provides an indication of how well the item discriminates between examinees of similar ability level and primarily dictates the magnitude of the slope of the ICC. A larger discrimination parameter results in more discrimination power and yields a steeper ICC slope. Finally, the lower asymptote

parameter takes random responding (guessing) into account and provides an indication of how well an examinee of very low ability should perform on the item (Lord, 1980).

The ultimate aim of both classical test theory (CTT) and item response theory (IRT) is to test people. Hence, their primary interest is focused on establishing the position of the individual along some latent dimension (ability). This study specifically examined linear equating method, separate and concurrent calibration of IRT. These methods were chosen because of their potentials in handling the problem under investigation.

In linear equating, the means and standard deviations of two forms for a particular group of examinees are set equal. Specifically, raw total scores that are the same distance from the mean in standard deviation units are set to be equal. The method ensures that scores on two tests are equivalent if they correspond to equal standard-score deviate. This method is most appropriate when the groups taking the test forms are equivalent or have equal ability, but can also be used in a non-equivalent anchor test group design (Kolen & Brennan, 2004; Tanguma, 2000).

Item parameters are estimated by setting the mean of the examinees' ability levels to zero and the standard deviation to one. Item parameters can be estimated using data from a common-item linking design either separately for each form or concurrently across forms. When two groups of examinees differ in ability levels, and when item parameters are estimated separately for each form, the units of the item parameters are not on the same scale because the examinees' mean ability levels and standard deviations are not equal. Therefore the item parameter estimates need to be transformed onto the

same scale. This procedure leads to equating test forms by separate calibration using three parameter logistic model (Mayuko, 2008).

Three parameter logistic model (3PLM) concurrent calibration combined two or more test forms that share common items into a single data set and then calibrated simultaneously. Because all items are calibrated at the same time, all the item parameters are estimated on a common scale and no further equating is necessary. The process of calibrating automatically equates the test forms (Morrison and Filzpatrick, 1992).

The efficiency of the above three test scores equating methods were determined using average root mean square error difference (ARMSED). Thus any of the equating method with the least amount of error was taken to be more efficient.

Apart from studying the efficiency of these identified equating methods, using Nigeria educational setting, the way the items of MAT functioned among the various subgroups used to perform test scores equating was examined. Specifically, ability and gender are among the constructs that were used to ascertain the Differential item functioning/invariance property of test score equating. Examinee's Ability is the cognitive, affective and psychomotor control of a person over his/her environment. In this study, examinee's test performance was treated as their ability and used for test score equating. Finally, examinee's gender was also considered as a factor linked to test score equating. Gender as used here refers to male and female. Examinees in school are made up of this nature distinguished group. Gender and Examinees's ability was used by the researcher in determining the population-invariance or consistency (differential item functioning) of test items.

Before conducting any test score equating, data are expected to be collected using very specific designs. Four commonly used designs to collect data before performing equating are: (i) Single-group design, (ii) random-group design (iii) equivalent-group design and (iv) anchor-test design (Kolen & Brennan, 2004; Tanguma, 2004). Other designs include but not limited to; Non-equivalent Anchor test group (NEAT) design and pre-equating non-equivalent group design (Von Davier, Holland & Thayer, 2004). A researcher or examination body carrying out test score equating, is expected to adhere to all or some of the following guidelines. These guidelines according to Dorans (2004) and Von Davier, et al (2004) are; (a) Same construct, (b) Equal reliability, (c) Symmetry, (d) Equity and (e) Population invariance. It is also important to assess whether equating has achieved its purpose by using any of the following criteria: (a) same distribution property, (b) first order equity property, (c) second order equity property and others that will be examined in literature review.

The goal of equating is for scores on multiple test forms to be used interchangeably. Test scores equating also ensure that no examinee is disadvantaged or advantaged because of the form of test taken. Thus, the inherent problem in the derivation of CAS and comparability of CAS among schools or states are the gap this study intends to fill. This study was therefore designed to examine the relative efficiency of test scores equating methods in the comparison of students' continuous assessment measures.

Statement of the Problem

Continuous assessment practice in Nigeria requires that teachers generate students' continuous assessment scores (CAS) in cognitive, affective and psychomotor

domain. About 30% of the scores from cognitive domain in particular are incorporated into terminal assessment scores (TAS) by WAEC/NECO. At the school level, scores allotted to continuous assessment varies between 30% and 60% among states in Nigeria. It is obvious that the manners in which CAS are produced leave much to be desired. The non-uniformity of scores allotted to CA attest to the fact that there is inherent problem of comparability of standard. This inherent problem has bedeviled continuous assessment practice right from its introduction and it has placed stakeholders in a perplexed state for so many years. Test scores appear to be the best matching criteria available to psychometricians, researchers and practitioners in making comparison among examinees. However, most of the classroom teachers are deficient in the construction of such valid test which can be used to generate scores. Teachers' deficiency in the construction of test is also a problem in the practices of CA. These nagging problems call for an urgent step to be taken if uniformity of standard and academic excellence must be monitored and maintained in Nigeria.

In as much as it seems desirable to use one common standard in judging the quality of the human capital being formed by the different schools within and across states in Nigeria. It is however impossible to do so given the diversity of the teachers, school environment, students, tests, to mention but a few in the case of continuous assessment. The social, economic, political and even administrative considerations also tend to seriously undermine or reduce the viability or practicability of subjecting all students among states to the same continuous assessment tests. This, in turn, creates multiple assessment standards within the same state. Some conceptual solution had been

suggested regarding the issues of comparability of standard in CA, but these seem not to be working. This is probably due to the fact that they are not based on solid statistical background. In an attempt to handle this problem, WAEC/NECO uses T-score transformation method in standardizing the CAS from various schools across the country. This T-score method is not quite appropriate because it does not ensure that scores emanating from CA reflects the students' academic achievement and also do not guarantee accurate comparison of students' achievement. It is on this basis that the researcher used a statistical technique known as test equating to handle the problems highlighted above. From preliminary investigation by the researcher, WAEC/NECO seem not to use test score equating methods in the comparison/standardization of raw score.

In this study therefore, an attempt was made to examine how scores from multiple assessment standards can be made comparable using a statistical approach – test score equating methods. In order to ensure that the test scores used for comparison are themselves free of questions that may be unfair, the psychometric properties of tests used and DIF analysis need to be conducted using item response theory. This is not practiced in Nigeria now. However, in conducting test scores equating, numerous methods are available to the researcher, and these methods are not of equal efficacy. Based on the above context, the researcher designed this study to determine the relative efficiency of test scores equating methods in the comparison of students' continuous assessment measures. Granted that this measurement technique is adopted as a complementary methodology for educational standard control mechanism, which of the test score

equating methods (Linear equating, separate calibration and concurrent calibration) would result in the most efficient or accurate technique for the comparison of students' continuous assessment measure?

Purpose of the Study

The main purpose of the study is to determine the relative efficiency of test score equating methods in the comparison of students' continuous assessment measures. The study specifically seeks to:

- (1) Determine the item parameter estimates of the two forms of Mathematic Achievement Test (MAT) used for Test Score Equating
- (2) Estimate the item parameter consistency values (Differential Item Functioning) for state A and state B
- (3) Estimate the item parameter consistency values (Differential Item Functioning i DIF) for male and female students
- (4) Estimate the item parameter consistency values (Differential Item Functioning i DIF) for High and Low ability students
- (5) Determine how well the items of MAT fit the three parameter logistic model
- (6) Determine the item characteristic curves of equated MAT
- (7) Determine the mean ability estimates of students in MAT in states A and B when their scores are equated through separate calibration
- (8) Determine the mean ability estimates of students in MAT in states A and B when their scores are equated through concurrent calibration
- (9) Determine the mean ability estimates of students scores scaled through linear

equating in states A and B

- (10) Determine which of the equating methods that is more efficient using average root mean square error differenced
- (11) Determine the relationship between the performances of students in mathematics continuous assessment tests and mathematics achievement tests

Significance of the Study

The findings from this study were theoretically significant and of immense benefit to the examination bodies that use continuous assessment score, school administrators, parents, teachers and students, policy makers and educators, and even tertiary institutions.

The theoretical knowledge of this study is considered significant because there is need for empirical information that can be used in explaining the various test scores equating methods and its appropriateness in the comparability of examinees scores from different forms of test. The major thrust of both classical test theory and item response theory is the use of test to determine the latent ability of examinees. In CTT the examinees' group response is the unit of analysis while in IRT the individuals' item response is the unit of analysis. The use of CTT based test score equating is most appropriate when the examinees' total scores is what is available to the researcher. However, if the examinees' individual response to each item of the test is available, IRT based test scores equating is most appropriate. The two theories may help to reveal to the teacher, the level of ability of the examinees with regard to the trait being measured. The information provided by this study may help teachers to evaluate how much underlying ability examinees possess and the comparison of this ability among examinees.

This is because the study helped in the establishment of an equivalency scale that enabled tests to be compared to one another. There will be uniformity in the reporting of information about student achievement in the nation's diverse educational system. Score derived through equating are standardized, as such examination bodies need not face the hurdle of transforming CAS by any method before incorporating it to TAS.

The test score equating methods in this study could be effectively utilized by examination bodies like; National Examinations Council (NECO), West African Examinations Council (WAEC), Joint Admissions and Matriculation Board (JAMB) to mention but a few, in comparing and standardizing Examinees's score. It also helped them to ensure that all examinations with parallel forms are of equal difficulty. Every examination body has an item bank where questions from known content area and psychometric properties are kept. The forms are expected to be of similar difficulties so that if by peradventure the form to be administered is presumed to have been exposed, it can be replaced with another one. The only way to guarantee equal difficulty of the test forms in the item bank is through test score equating

The issue of misinterpretation of marks by some teachers may be laid to rest as all raw scores obtained by examinee could be standardized through equating. With the application of test score equating in obtaining CAS by schools, examination bodies no longer place any doubt on the authenticity of CAS. It also erase the worries of how CAS is incorporated with TAS from the minds of students, teachers and all stakeholders.

The findings of this study may establish the appropriate test scores equating methods which could enhance the comparability of students's scores within a particular

grade level or across grade level. The primary practical advantage of tests cores equating is the placing of students's scores on common or equivalent scale. This is of particular advantage to the implementation of continuous assessment, because the obstacle encountered in the comparison of students continuous assessment scores may possibly be handled through the application of test scores equating methods.

The findings provided parents, students and teachers with clear information about the performance of individual student as measured by a common/equivalent or national standard. The findings of this study are of immense benefit to students, who for one reason or the other could not write their test/examination at a stipulated time and are administered supplementary one. This is possible because the equating methods used in this study, are capable of adjusting for difficulties between the supplementary test/examination and the original one taken by students such that the examinees taking different test are neither advantaged nor disadvantaged.

The findings provided scientific based information on the possible sources and magnitude of imprecision in equating test score. This is going to enable policy makers and educators take the responsibility for determining the degree to which they can tolerate imprecision in testing. Through this approach, quality assurance will be guaranteed in Nigeria testing enterprise. In order to meet the expectations of the society in-terms of higher educational standard, educators need information that can be used to project how students are doing against the class-level standards throughout the course of schooling. This in turn enable them to determine what need to be done to accelerate student's academic progress. The findings acted as a guide for the execution of this task

through the use of equivalency/common scale. The findings helped schools; identify the accomplishment of desired outcomes, goals or standards and compare such outcomes, goals or standards with that of other schools. This spurred the regular reviews of the assessment strategy which in turn help guarantee that, the instruction, curriculum and assessment are consistent with each other.

Tertiary institutions especially universities that use Post University Matriculation Examination (PUME) found this work a worthy tool. The universities set cut-off scores which define passing level using PUME. Fairness requires that this standard should be held constant over time, so that the meaning of score in one year is the same for another year. The findings enlightened the institutions on how to interchangeably use alternate forms of PUME built to the same content and statistical specification over years to maintain the standard set. This required the institution to equate the test scores of students from different administration or years so that, the performance of students in a particular year may be compared with that of another year.

In practice, the item parameter estimates, item consistency values and model fit estimates of item response theory used in this study may help examination bodies, researchers, and teachers with sound knowledge of test scores equating methods based on IRT framework. It may also help them to diagnose the weakness and strength of examinees per item in a test. Such diagnoses may be used to improve teaching and learning. Finally, researchers can use this work as a worthy tool in the advancement of knowledge as it will likely spark further research.

The Scope of Study

There are many test-score equating methods. This study strives to empirically investigate the use of some of these methods in the comparison of students continuous assessment scores. The approach involved using linear equating methods based on classical test theory framework and two methods based on item response theory (three parameter logistic model using separate calibration and concurrent calibration). The choice of these methods is informed by their robustness in handling the comparison of examinees' scores. The IRT based methods were used to develop and equate the test forms appropriate for equating continuous assessment to enhance comparability. Other equating methods like; Levine equally reliable linear equating, Levine unequally reliable linear equating and Tucker linear equating, as well as those based on one-parameter and two-parameter models were not examined but reference are made to them where necessary.

Research Questions

In carrying out this study, the following research questions were formulated

1. What are the item parameter estimates of the two forms of Mathematic Achievement Test (MAT) used for Test Score Equating?
2. What are the item parameter consistency values (Item Functioning) for state A and state B?
3. What are the item parameter consistency values (Differential Item Functioning i DIF) for male and female students?

4. What are the item parameter consistency values (Differential Item Functioning $\hat{\delta}$ DIF) for High and Low ability students?
5. What are the items of MAT that fit the three parameter Logistic model?
6. What are the item characteristic curve of equated test A^1 and B^1 ?
7. What are the mean ability estimates of students in state A and B when their scores are equated through separate calibration?
8. What are the mean ability estimates of students in state A and B when their scores are equated through concurrent calibration?
9. What are the mean estimates of students scores scaled through linear equating in state A and B?
10. Which of the equating methods is more efficient?
11. What is the relationship between the performance of students in continuous assessment mathematics tests and mathematics achievement tests?

Hypotheses

- H_{01} : There is no significant difference in the item parameter estimates of the two forms of mathematics achievement test (MAT) used for test equating
- H_{02} : There is no significant fit between the estimates of item difficulty and three parameter logistic model
- H_{03} : There is no significant difference in the ability estimates of students in states A and B for scores equated through separate calibration
- H_{04} : There is no significant difference in the ability estimates of students in states A and B for scores equated through concurrent calibration

H₀₅: There is no significant difference in the ability estimates of students in states A and B for scores equated through linear equating

H₀₆: There is no significant relationship between the performances of examinees in continuous assessment test and mathematics achievement test.

CHAPTER TWO

LITERATURE REVIEW

This chapter deals with the review of relevant literature. The researcher reviewed conceptual, theoretical and empirical research work carried out by other authors relating to the equating methods and factors under study. They are discussed under the following sections.

Conceptual Framework

- Origin and Issues in Continuous Assessment (CA) in Nigeria
- An overview of test score equating
- Test scores equating methods/Models
- Examinees's Ability, Population-Invariance/Differential Item Functioning (DIF)

Theoretical Framework

- Classical Test Theory (CTT)
- Item Response Theory (IRT)

Related Empirical Studies

Conceptual Framework

Origin and Issues in Continuous Assessment (C.A) in Nigeria

The present evaluation procedures in Nigeria took root from two interacting situations: first, the 1969 curriculum conference which gave birth to the National policy on Education approved by the Federal Government of Nigeria in 1979. Second, the recognition by those involved in drawing the policy of the importance of evaluation in education. The education policy recommended the 6-3-3-4 system to replace the former

6-5-4 system. The new system was to commence in all States in 1982 (Owolabi, 2003). The national policy recommended continuous assessment for measuring educational outcomes. Continuous assessment is used as an all embracing concept which covers all aspects of the development of learners. These include the cognitive, affective and psychomotor domains. In addition to assessing gains in school subjects, continuous assessment also assesses pupils' values, beliefs, attitudes and appreciation, interest, social relations, habits, emotional adjustments and life styles as well as manipulative skills and body movements like writing, drawing, typing, dancing etc (Owolabi, 2003).

The Nigerian National Policy on Education (Federal Republic of Nigeria, FRN, 2004) states that educational assessment and evaluation will be liberalized by basing them in whole or in part on Continuous Assessment of the progress of the individual. Anikweze (2005: 2) refers to Assessment as "the process of investigating the status or standard of learners' attainment, with reference to expected outcomes that must have been specified as objectives" when it concerns learners' output. Assessment enables the school to achieve an overall objective of having as complete a record of the growth and progress of each pupil as possible in order to make unbiased judgments in the cognitive, affective, and psychomotor evaluation in the classroom.

Continuous assessment, according to Federal Ministry of Education, Science and Technology (FMEST, 1985), is defined as a mechanism whereby the final grading of a student in cognitive, affective and psychomotor domains of behaviour takes account, in a systematic way, all his performances during a given period of schooling; such an assessment involves the use of a great variety of modes of evaluation for the purposes of

guiding and improving learning and performance of the student. This mode of assessment is considered adequate for assessment of students' learning because it is comprehensive, cumulative, systematic, guidance oriented, diagnostic and prognostic oriented (Habour-Peters, 1999). In consequence, the results obtained, are more valid and more indicative of the overall ability of the learner. The extent to which the overall ability of the student is assessed in Nigerian schools depends on how well teachers implement CA (Ajuonuma, 2007).

Airasian (1991) described continuous assessment as an assessment approach which should depict the full range of sources and methods teachers use to gather, interpret and synthesize information about learners; information that is used to help teachers understand their learners, plan and monitor instruction and establish a viable classroom culture. On his own part, Stites (1991) opined that continuous assessment should involve a formal assessment of learners' affective characteristics and motivation, in which they will need to demonstrate their commitment to tasks over time, their work-force readiness and their competence in team or group performance contexts.

From these definitions, Onjewu (2007), inferred that continuous assessment is an assessment approach which involves the use of a variety of assessment instruments, assessing various components of learning, not only the thinking processes but including behaviours, personality traits and manual dexterity. Continuous assessment will also take place over a period of time. Such an approach would be more holistic, presenting the learner in his/her entirety. It begins with the decisions that the teachers takes on the first day of school and ends with the decisions that the teachers and administrators make on

the learners regarding end-of-year grading and promotion. Continuous Assessment is crucial to school based Assessment. It usually forms a substantial component of any School Based Assessment policy as it ranges from thirty to forty percent in a majority of cases.

School based assessment as cited by Griffith (2005:2) and NTI (2005) referred to the process where students, as candidates, undertake specified assignments during the course of the school year under the guidance of the teacher as part of a subject examination. It is therefore expected that the school environment in its totality provides favourable atmosphere to facilitate learning and subsequent assessment procedures. School Based Assessment brings Assessment and teaching together for the benefit of the students and provides the teacher with the opportunity to participate in a unique way in the assessment process that leads to the final grade obtained by students. For this reason, Njabili, Abedi, Magesse and Kalole (2005:2) add that "The fundamental role of Assessment is to provide authentic and meaningful feedback for improving student learning, instructional practice and educational options" which means that Assessment is not and so should not be seen as an end in itself but a means to a justifiable end of learning.

Onjewu (2007) in a presentation on assuring fairness in the continuous assessment component of school based assessment practice in Nigeria observed that, there are inherent problems with the derivation of CA scores and the situation deserves the focus of academics. The author further argued that, the entire practice of CA is surrounded by

laxity. This laxity were listed as; laxity in timing, laxity in terms of the mode that the CA exercise takes, and laxity in relation to the content of the CA.

Kayode (2003) observed that only a partial continuous assessment is practice in Kwara state of Nigeria and encouraged both the government and the National Association of Educational Researchers and Evaluators (NAERE) to intervene to ensure that the laudable goals of continuous assessment are achieved. Adeyemo (2003) also complained of the haphazard way in which continuous assessment is implemented in Osun state of Nigerian. Some Nigerian researchers have complained about how different components of the policy are handled (e.g. Ezeudu, 2005; Onuka, 2005; Akinlua & Ajayi, 2003).

Afolabi (1999) warned that the combination of raw continuous assessment scores with students' examination scores for the SSCE results (instead of standardized scores) cannot ensure fair comparability. Afolabi further pointed out that any comparison made with a combination at the SSCE level cannot be considered fair, equitable, and just. The author stated that, students who have more observed scores for continuous assessment are likely to obtain final scores which are closer to the mean scores of the group (smaller variances), than those who have fewer observed scores.

Adebowale and Alao (2008) described continuous assessment as an on-going process of gathering and interpreting information about student learning that is used in making decisions about what to teach and how well students have learned. They highlighted some merits of continuous assessment as:

- Promotion of frequent interactions between pupils and teachers that enable teachers to know the strengths and weaknesses of learners to identify which students need review and remediation.
- Pupils receive feedback from teachers based on performance that allows them to focus on topics they have not yet mastered

The Rationale for Continuous Assessment

Generally, teachers are known to be very creative and full of innovation, which they could introduce into their class teaching. The readiness of the teachers to introduce innovations into their teaching is quite-often frustrated by the fact that the final examination which students would take at the end of a particular programme, does not seem recognize such innovations. But in a continuous assessment situation, the teacher could be flexible enough to accommodate innovations in the teaching and learning process since he has a part to play in the final grading of each student through his periodic continuous assessment practice. The system gives the teacher greater involvement in the total assessment of their students.

Continuous assessment procedure provides information on all the domains of educational process, which includes the cognitive, affective and psychomotor domains, which could be described as a complete assessment. It provides a more valid and reliable assessment of the students overall ability and performance. There are needs for teachers to assess their instructional methods from time to time in order to improve their performances. Continuous assessment provides a basis for the teacher to improve their

instructional methodology. A feedback from continuous assessment can be useful to the classroom teacher for such self-evaluation.

Any assessment procedure, which takes into account' the student's performance throughout the entire period of schooling, is likely to be more valid and more indicative of the learners overall ability than a single examination. The continuous assessment helps in providing the required data that enables the teacher to offer the appropriate guidance to the learner. This guidance function of continuous assessment is unique and very useful in the educational process.

The high rate of examination malpractice and leakage particularly in external examinations such as senior secondary school certificate examination have been blamed on the fact that the future of a candidate has often been based on a single final examination. This is so crucial in deciding the future of a learner hence; there is that temptation to ensure passing the examination by all means, whether fair or foul. It is strongly argued that with the use of continuous assessment scores, there would be a reduction of emphasis on the final examination which would in turn reduce students' anxiety and makes' malpractice a thing of the past.

Assessment is an integral part of the teaching and learning process. It is therefore important that the teacher should be involved in the final assessment of the students taught, as compared to the previous dispensation where the teacher had no contribution in the final grading of the students. The system whereby the final assessment at the end of a particular level of education is done through a single examination set by an external

examination body tends to deny the teacher the opportunity to participate in the final assessment of his students which the teacher considers very unfair.

The Problems of Implementation of Continuous Assessment

There are so many problems that have been identified as bottlenecks in the implementation of continuous assessment practice in Nigerian schools. They include: Comparability of standard, problem of record keeping, continuity of records, problem of cheating, and misconception of the concept of continuous assessment (Emaikwu, 2004; Onjewu, 2007).

The presence of two national examination bodies (WAEC & NECO), whose examinations all schools are expected to register their candidates for, provides a basis for making comparison as regards the quality of student's performance across schools. But in continuous assessment programme such comparisons become very difficult. This problem arises because there are differences in the quality of tests and other assessments used in different schools (Emaikwu, 2004). A student who has an "E" grade in a particular school may be better than a student who has a "D" grade in another school within the country in their continuous assessment.

To have a meaningful continuous assessment, there is a great need for keeping accurate records for each learner. Since continuous assessments are expected to be cumulative from class to class and from school to school, the need for proper record keeping is imperative. Nevertheless, with an educational policy, which seems not to have a firm grip on the principles and practice of CA in schools, keeping accurate records with

a uniform format across the country poses a great difficulty to the whole system (Ugodulunwa, 1999).

Furthermore, learners even within the same level of schooling may be required to move from one school to the other due to many factors including parental transfer or wish of parents due to economic reasons. It therefore, demands that a mechanism needs to be evolved to ensure that the records of a child moving from one school to the other can be easily transferred to his new school with the old school still maintaining a copy of these records in their care. This process of maintaining continuity from one school to the other is indeed not an easy task and poses a great problem in the whole system.

Emaikwu (2004) noted that, continuous assessment is perceived to be susceptible to cheating. Parents, brothers, sisters, and friends of students give helping hands on projects and assignments given at school. Students sometimes copy other student's work thus negating the very essence of assessment in schools. Experiential evidence shows there is problem of favoritism in the award of marks to students. Some students are given higher scores because of their relationship with the teachers, while others are not. At times students' scores are inflated arbitrarily contrary to the principles of continuous assessment for reasons best known to teachers and some school authorities. It is therefore, common to see continuous assessment scores that do not correlate positively with actual examination performance. It is not unusual to find students scoring very high marks on teacher's continuous assessment and some of them scoring very low mark in external examination (Ugodulunwa, 1999).

Another problem of implementation of continuous assessment in Nigerian schools is that of misconception of the concept of continuous assessment. Many teachers perceive continuous assessment as continuous testing in cognitive domain. Often times, teachers base their judgments of student's affective and psychomotor domains on student's performances in different subjects. The assessment conducted by such teachers do not present a total picture of the student's behavior as proposed in the national policy on education (FME. 1981). Nenty (1991) summarizes the demands of continuous assessment as:

(i) A readily available collection of a large number of reliable items for frequent measurement of specific objectives (item banking). (ii) A collection of items that will enable valid repeated cross-sectional or longitudinal measurement with which the same standards can be maintained across the years. (iii) Ensuring that the test items be constructed according to local specifications. (iv) A collection of items, which measure validity given well-defined strand of a given subject matter. (v) Having many parallel forms of the same test (p.9).

These demands pose a lot of challenges to the teachers, in terms of acquiring the right techniques and skills for measurement and evaluation. With the introduction of the new policy on education, the Federal Ministry of Education, and all the state Ministries of Education mounted series of workshops for teachers to enable them acquire the skills in developing quality test items, collection and collation of test data, as well as present effective and accurate interpretations of the data (NTI,2005). This effort is however frustrated by the teachers' inability to cope with the number of tests required. The ratio of students to teachers, in almost all the schools and the series of complain against

continuous assessment practices in school, from both teachers, students and parents alike had created a difficult scenario for its implementation. Teachers are still very much deficient in test construction techniques and scores interpretation, yet the new assessment policy of universal basic education require pupils to transit to junior secondary school with CAS only. Be that as it may, teachers must be given continuous re-orientation in test construction and test scores equating in order to be able to cope with the new trend of events in Nigeria education system.

In spite of these problems, it would be hard to abandon a system or an assessment procedure, which takes into account, the learner's performance throughout the entire period of schooling and thus takes several samples of learners' behavior (Emaikwu, 2004). Such an assessment procedure promises to be more valid and reliable and thus more indicative of the learner's overall abilities at school. Consequently continuous assessment has been adapted in Nigerian educational system with 'continuous assessment scores forming a percentage of the overall students' scores in most of the externally conducted examinations, like the junior secondary school certificate and senior secondary school certificate examinations.

Among the many demands for effective implementation of continuous assessment in schools, the development and maintenance of item banks is one of the most important. So far, an important problem faced by a classroom teacher in the implementation of continuous assessment is the skill and time demanded for constructing instruments for every subject areas for her different classes. This is compounded by the demand for the

development of new but equivalent test for students that were absent during initial testing. These would not pose any problem if a pool of tested and calibrated items in each aspect of teachers'subject area were available. (Emaikwu, 2004).

Overview of Test Equating

The use of different forms of the same test (or different tests aiming to measure the same constructs from year to year, or school to school as is the case of CA practice in Nigeria), raises the issue of the comparability of test scores. Since the different CA tests used in Nigeria vary across schools, it is possible to claim that the test scores on different forms are not directly comparable; as such 'test equating' is needed.

After two tests are equated, pairs of equivalent scores become available. For example, such a pair of equivalent scores could be (17, 19) which indicates that a total score of 17 on the first paper is equivalent to a total score of 19 on the second paper. In order to compare achievement using two different tests, one simply needs to use a conversion table or graph to convert the scores of one test to the equivalent on the other test. The widespread use of high stakes public examinations across the world, and the pressure on psychometricians to be able to interpret results from administrations of different tests, have generated an increased interest in the area of test equating research and development (Lamprianou, 2007).

Large scale testing programmes often require multiple forms of test in order to maintain test security over time or to enable the measurement of change without repeating identical questions. The comparability of scores across the multiple forms may have some consequential effect on students'sreported scores. This is because, the students

are admitted to colleges based on their test scores, and the meaning of a given scale score in one year should be the same as for the previous year. Also, institutions set cut-off scores that define passing levels (for instance university matriculation examination, UME) and fairness requires that these standards be held constant over time. To allow interchangeable use of alternate forms of test built to the same content, and statistical specifications, scores based on different sets of items need to be placed on a common scale, through a process called test equating (Haertel, 2004).

An equating method is an empirical procedure for determining a transformation to be applied to the scores on one of two forms of a test. Its purpose is ideally to transform the scores in such a way that it makes no difference to the examinee which form of the test he or she takes. This idea can be reached only if the two forms of the test measure exactly the same latent trait (ability or skill) and yield scores that are equally reliable and if the equating transformation is invertible (Gray, Nancy and Stewart, 1979). Similarly, Hanick and Chi-Yu (2002) referred to the term "equating" as a statistical procedure that adjusts test scores on different form of the same examination so that scores can be interpreted interchangeably.

Dorans, Pommerich and Holland (2007) opined that the goal of equating is to produce interchangeable scores. The authors further explain that users of test scores often assume that scores are direct and unambiguous measures of students' achievement. Consequently, an increase in test score could be assumed as evidence that students are learning more. Scores on most achievement tests, however, are only limited measures of the latent construct of interest, which is an aspect of student proficiency. As measure of

these constructs, test scores are generally incomplete, and they are fallible because they include measurement errors, and are vulnerable to other vices such as inflation. Therefore scores on most achievement tests are not inherently meaningful or useful unless they are subjected to certain statistical treatment like; linking, scaling or equating (Koretz, 1999).

Whenever equating is conducted, equating errors are bound to occur. This is because examinees who actually take tests are considered to be samples from a population of examinees, therefore equating errors are present in estimating equating relationships. These equating errors are often categorized into types: systematic equating errors and random equating errors. Systematic equating errors are usually caused by assumption failures in the equating method, bias in the sample statistics, and so on. Random equating errors on the other hand are caused by sampling errors. Both systematic and random equating errors influence the interpretation of results. Therefore, caution needs to be taken when sampling subjects or test for equating, this may help to minimize equating errors. Thus error in equating observed scores on two versions of a test can be seen as the difference between the transformations that equate the quantities of their distribution in the sample and in the population of examinees.

Figure 1 below clearly delineates, test scores equating into vertical and horizontal equating. Vertical equating involves equating test of different grades or levels. It allows comparison to be made between students at different levels and also comparison of their growth over time. Vertical equating is also called across-grade-scaling. This method places students' scores on two tests of different levels, such as mathematics for SSI and SSII on the same scale, so that scores of students in both tests can be compared (Lee,

2003; Leugn, 2003; Lissitz & Huynh, 2003). Horizontal equating on the other hand involves equating test of different forms or at different time of a single grade of level. It is also called within-grade-scaling. Horizontal equating places students's scores on two test at the same level, for the same content area and for the same population so that their scores can be directly compared.

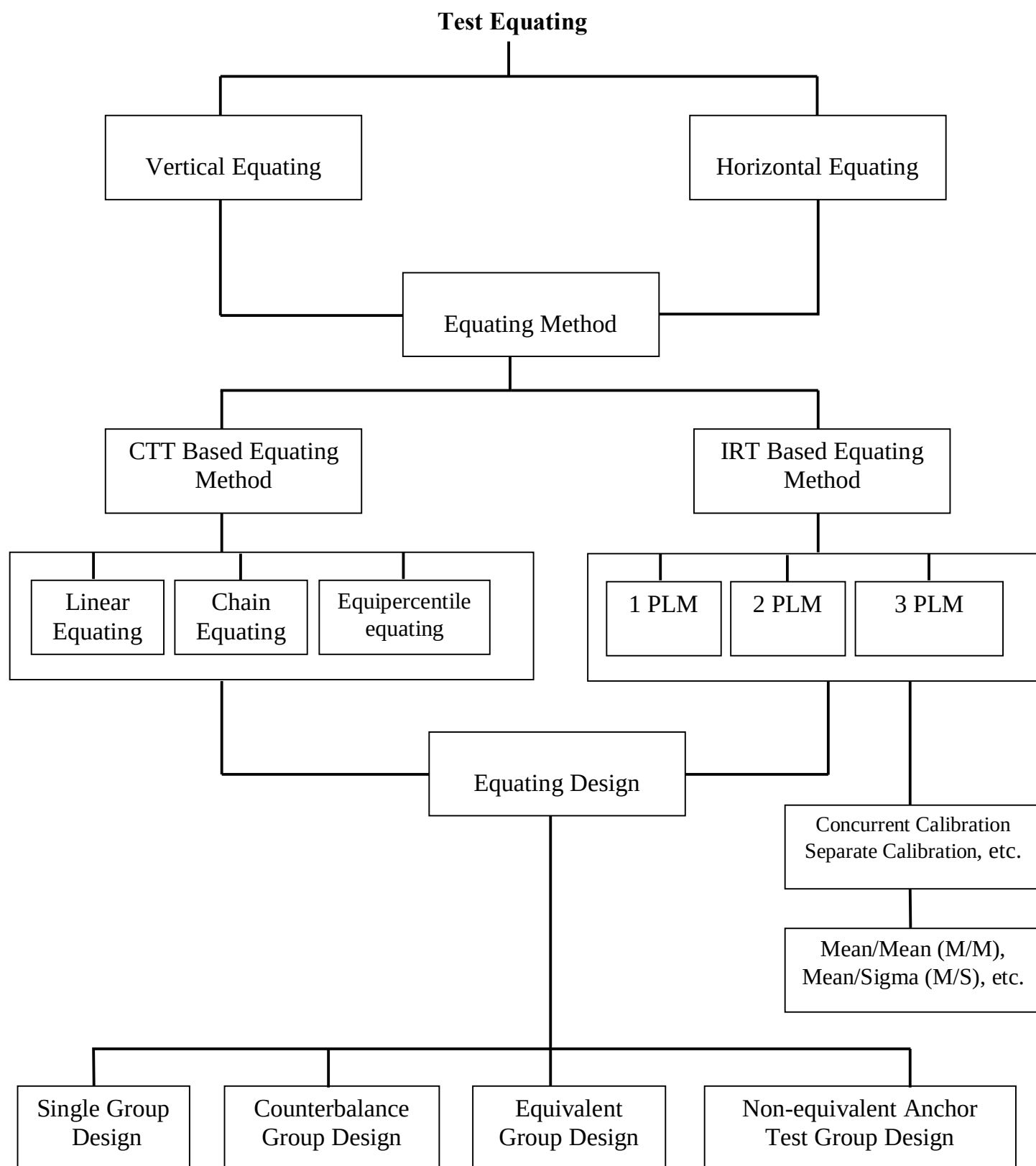


Fig. 1 Schematic Representation of Test Score Equating Methods and Design for Test Scores

The purpose of horizontal equating is to compare two or more groups of examinees of the same level of ability using two or more different test forms of the same content area and difficulty (Leung, 2003; Lissitz & Hunyh, 2003). Test score equating through vertical or horizontal approach use observed-score or true-score. When using the observed-score, raw total scores are calculated for each examinee and are used to create a score distribution for each test form. In true-score equating, item parameters are estimated using all the examinee. This method is to some degree based on latent trait variable such as proficiency parameter of Item Response Theory (IRT) or true score in Classical Test Theory (CTT) (Kolen & Brennan, 2004; Han, Kolen & Brennan, 1997).

There are many test scores equating methods subsumed into two major test theories: Classical Test Theory (CTT) equating and Item Response Theory (IRT) equating. The classical test theory is founded on a test score model which assumes that, there are no perfect measures of ability. Each examinee's observed score (X) is comprised of True Score (T) and random error component (E) (Bielinski, Thurlow, Minnama & Scott, 2000; Wiberg, 2004; Schumacker, 2005). Under classical test theory, the following test score equating methods are used; mean equating, linear equating, Levine-equally-reliable linear equating, Levine-unequally reliable linear equating, Tucker equating, equipercentile equating, frequency estimation equipercentile equating, chained equipercentile equating and chained linear equating (Von Davier, 2008; Von Davier & Kong, 2005; Chong & Sharon, 2005; Skaggs, 2005; Felan, 2002; Tanguma, 2000; Hanson, 1993; Petersen, Marco & Stewart, 1982).

Item Response Theory (IRT) methods, on the other hand, are based on the assumption that, there is a mathematical function that describes the relationship between an examinee proficiency and probability that an examinee will answer an item correctly (Chong, 2007; Hambleton & Swaminathan, 1995). Some of the equating methods under IRT are; Rasch model (one-parameter logistic model) based equating, two-parameter logistic model based equating and three-parameter logistic model based equating. Within the domain of the Rasch model (one-parameter logistic model), the equating methods used include but are not limited to concurrent/separate calibration, fixed based procedure, equating constant procedure and major axis procedure. Two-parameter and three parameter logistic models also have their respective equating methods, but this study was limited to some of the CTT and three parameter logistic model of IRT equating methods. Specifically, linear equating, frequency estimation equipercentile equating, chained equipercentile equating of CTT, and Three parameter logistic separate calibration and concurrent calibration, of IRT were examined. The ultimate aim of both classical test theory (CTT) and item response theory (IRT) is to test people. Hence, their primary interest is focused on establishing the position of the individual along some latent dimension. Because of the many educational applications the latent trait is often called ability. In conducting test score equating either based on CTT or IRT approach, the researcher needs an appropriate number of examinees (sample size).

Felan (2002) suggested that for linear and equipercentile equating, the sample size should be 400 and 1000-1500 respectively while Zen (1991) suggested that for Rasch model based IRT equating and three parameter logistic model based equating, the sample

size should be 400 and 1500 respectively. This information helped the researcher in drawing the appropriate number of sample size that was used for this study.

Felan (2002) and Kolen and Brennan (1995) outlined the following conditions as conducive to satisfactory equating; Firstly, the goal of equating, such as equating accuracy and the extent to which scores are to be comparable over long time periods are to be clearly specified. The design for data collection, the equating linkage plan, the statistical methods used and the procedure for choosing the goal in a particular practical context in which equating is conducted.

Secondly, in terms of test development using all designs, test content and statistical specification should be well defined and stable over time. When a test form is constructed, statistics on all or most of the items should be available from pretesting or previous use. The test should be reasonable in length (example 30 items or longer). Scoring key should be stable when items or forms are used on multiple occasions.

Thirdly, in terms of test development, when using common (anchor) item non equivalent group design, each common set should be representative of the total test in content and statistical characteristics. Each common (anchor) item should be of sufficient length (example 20% of the test for tests of 40 items or more). Each common item should be in approximately the same position in the old and New. Common item stems, alternatives and stimulus materials (if applicable) should be identical in the old and new forms.

Also, in terms of examinee groups, (a) these should be representative of operationally tested examinee, (b) stability over time (c) relatively large sample and (d) in

the common-item nonequivalent groups design, the groups taking the old and New Forms should not be extremely different.

Finally, in terms of administration, the test and test items should be secure and administered under carefully controlled standardized conditions that are the same each time the test is administered. In all, the curriculum and training materials and/or field of study (subject) should be stable.

Connecting/Linking Tests

The process of equating tests begins by understanding how to link two tests, then several tests and finally connecting all possible tests. The main aim of connecting tests intended to measure the same variable, is to ensure that the separate measures each test implies are expressed together on a single common scale.

The connecting of different tests forms can be done in several ways. Wright and Stone (1999) have suggested that when an easy test is to be linked or connected with a hard test, a set of common items are included in both tests, so that the common items become hard in the easy test and become easy in the hard test (Figure 2).

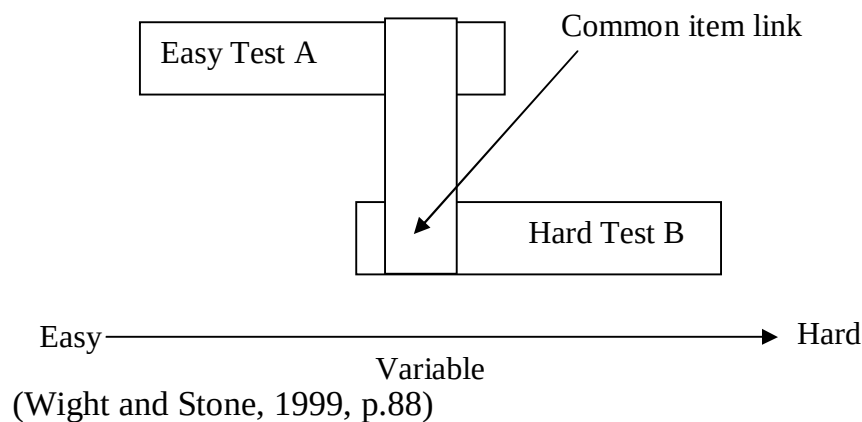
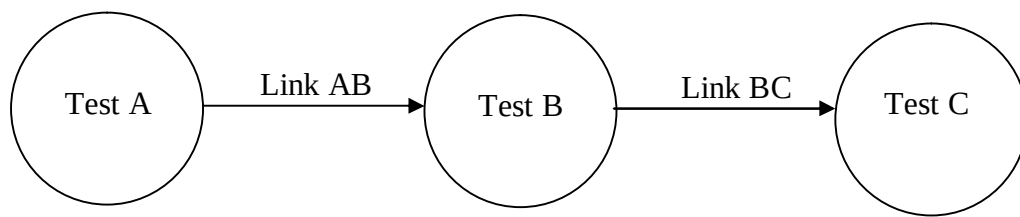


Figure 2: Two tests with common item link

When two or more test forms of approximately the same difficulty are to be linked, a set of common items is used to link the pairs of the different forms, as illustrated in Figure 3 below. In this design, Test A is linked to Test B by a set of items common to both Test A and Test B, and Test B is linked to Test C by a set of items common to both Test B and Test C. The design in Figure 3 is basically the extension of the simple design to link two forms of test illustrated in Figure 2.

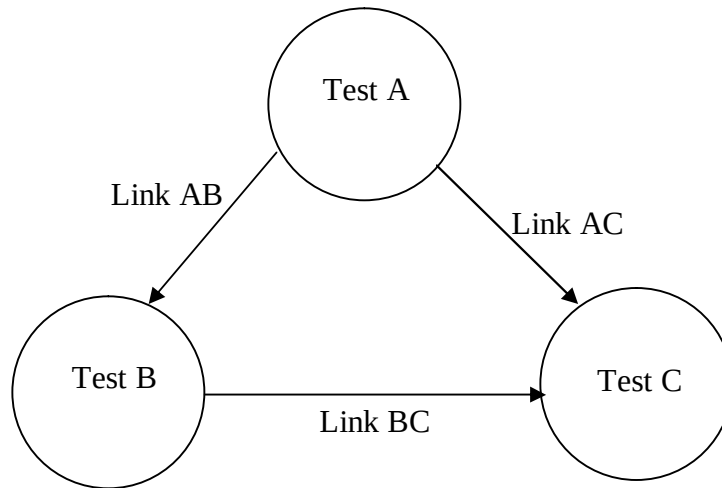


(Wright and Stone, 1999, p.89)

Figure 3: A Chain of Two Links

This pattern of linking can be extended by forming a loop (Wright and Stone, 1999, p.100) as presented in Figure 4. In this design, Test A and Test B are linked by a set of items common to both Test A and Test B, Test B and Test C are linked by a set of items common to both Test B and Test C, and Test C and Test A are linked by a set of items common to both Test C and Test A.

Wright and Stone (1999) have argued that for the different test forms of approximately the same levels of difficulties, the loop design is better than the chain design because in the loop design, the consistency of the equating can be checked through the equating of the last form back to the first form.



(Wright and Stone, 1999, p. 89)

Figure 4: A Loop of Three Links

If a researcher intends to use five different test forms in a study, Wright and Stone (1999) advise that the tests should be prepared with the same test specifications, so that it could be assumed that all five test forms had approximately the same level of difficulty and coverage of topics. The target population of the said tests is also students who are taught based on the same curriculum. In order to link all five different test forms, the chain design described earlier is applied, and the linking design of the five test forms is illustrated in Figure 5 below.

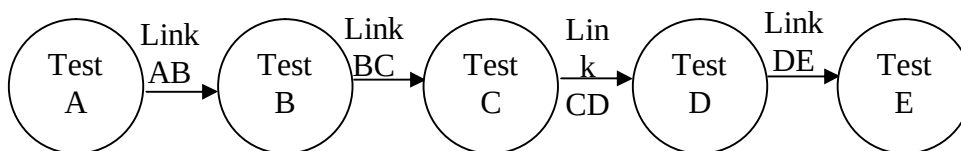


Figure 5: Linking Five Different Test Forms

In this connecting process, Test A and Test B are linked with a set of items common to both Test A and Test B, Test B and Test C are linked with a set of items

common to both Test B and Test C, Test C and Test D are linked with a set of items common to both Test C and Test D, and Test D and Test E are linked with a set of items common to both Test D and Test E. The representation of this process is also illustrated in Figure 6 below. AB represents a set of items common to Test A and Test B. BC represents a set of items common to Test B and Test C. CD represents a set of items common to Test C and Test D. and DE represents a set of items common to Test D and Test E.

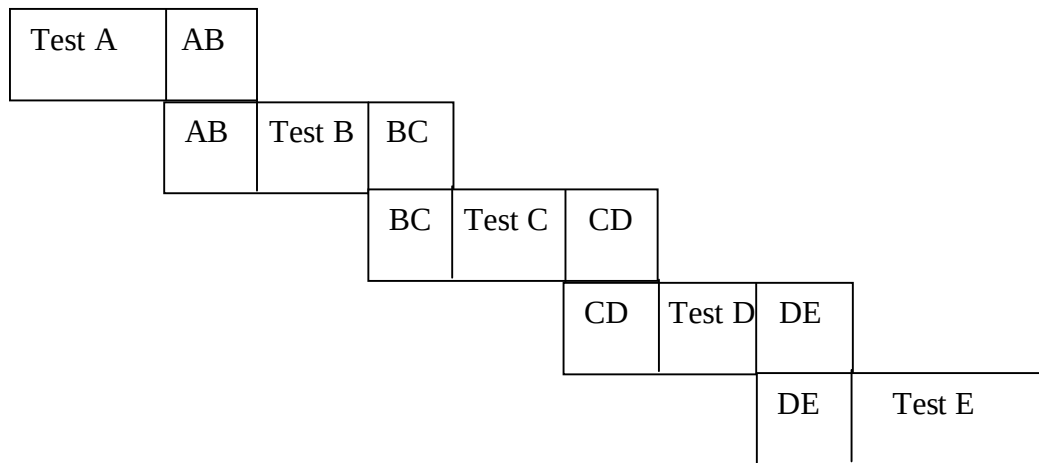


Figure 6: Linking Five Different Test Forms using Box pattern

The linking methods reviewed above, will help the researcher to link continuous assessment scores from various states with the researchers made test score. This is shown in Figure 7 below.

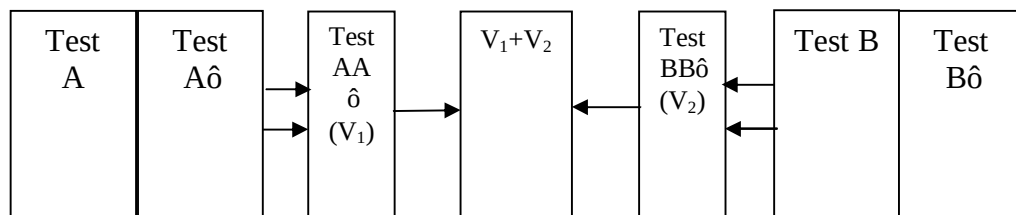


Figure 7: Linking of continuous assessment scores with Mathematics Achievement Test (MAT)

Where,

Test A = state A continuous assessment score

Test A¹ = scores from researchers made test (paper A¹)

Test B = state B continuous assessment score

Test B¹ = scores from researchers made test (paper B¹)

V₁ = equated scores from test A and A¹

V₂ = equated scores from test B and B¹

Equating Designs

An equating design is a plan for collecting the data one needs for equating. Data collection is crucial for successful equating or linking of scores from various instruments. It is very important to control for differences in distributions of response propensities when assessing differential instrument difficulty. In test equating or linking, as in most scientific research, this has always been accomplished through the use of special data collection design. Chong & Sharon (2005) opine that, the choice of equating method is influenced by the design function. And that some commonality must exist between the two test forms and examinee groups, whether that is in common items or common examinees (subjects).

Dorans (2004), Kolen & Brennan (2004), Livingston (2004) Tanguma (2000) and Von Davier, *et al.* (2004) noted that for equating scores on a new test form (test, X) to scores on a reference test form (test, Y), a number of different design or data collection procedure can be distinguished and these are:

- ◆ The single-group design

- ◆ The Random (counterbalance) design
- ◆ The equivalent design
- ◆ The anchor test design.

Other designs derived from the above are:

- Nonequivalent design
- Internal anchor test nonequivalent groups design
- External anchor test nonequivalent groups design
- Pre-equating nonequivalent groups design and
- Post-equating nonequivalent groups design

Single Group Design

The use of single (one) group design requires all candidates to take the two types of the test that are to be equated. The tests are administered to all the candidates in the same order. This design reliably control for variations in answer predisposition by using the same candidates for both tests. This has the merits of giving the most accurate statistical result. Also the merit of this design stems from the comparability of the groups which are very comparable, while the demerit is that there may be order effect.

The statistical connection between scores on the forms provides proof as to the comparability of the content and difficulty of tasks. However, because the tasks can take several hours or days to completed, the candidates may become tired and fatigue may set in. This is a serious matter in assessment and measurement error may be imminent (Gordon, Engelhard, Gabrielson, & Bernknopf, 1996). Hence, an extended rest period

might be required between administrations of the types of tests. The tasks might also be contingent upon each other – having a positive or negative relationship. This challenge becomes more likely when the tasks are intended to be relatively parallel and to evaluate the same construct.

Again, the assumption of Local independence in items of test may also affect data analysis using a given model, and violations of this assumption weaken any linkages. Item parameters and item weights have been found to sway the accuracy of linking in some equating situation. Yen, (1993) posits that, in optimal IRT scoring, item weights are a function of item parameters. Therefore, once the examinee is probable of guessing the response, the optimal weights will tend towards becoming proportional to the IRT item discrimination parameter. These parameters tend to be overstated by the presence of local item dependence; this can have significant effects on the scoring weights of such items, making the weights less than optimal in reality.

The one-group design may be used if all forms of evaluation instruments are present at the same time, and also if there are sufficient number of candidates. Otherwise, counterbalancing the test administration order was desirable if order effects are envisaged (Muraki, Hombo and Lee 2000). A lot of researchers believe that counterbalancing the forms could control order and fatigue effects (Crocker & Algina, 1986, Kolen & Brennan, 1995). Muraki et al noted that concerns about fatigue due to test session length can be alleviated if the same group of candidates is available for all forms at different times and locations. Over delays between forms administrations can however lead to changes in candidate's ability, and hence should be avoided. Although this design

appears simple, it may be practically difficult to implement in Africa. This is because the social, economic, political and even administrative considerations in Africa may reduce the viability of this design.

Population	First Administration	Second Administration
Group C	X	Y

Figure 8: Single-Group Design

Counterbalance Design (CB Design)

The order of administration of the tests has been taken into consideration by this design. Hence, half of the candidates in the group are given test X first and then test Y second. The other half of the candidate group takes the same two tests in reversed order. The time between the two forms should be very short- so that there will be no real change in their level of the knowledge of the skills that the test measures. A major merit of this design is similar to that of one-group design (OG). This design may be useful within a given education zone because result from this design is valid from relatively small sample.

Population	First Administration	Second Administration
Group C	X	Y
Group D	Y	X

Figure 9: Counterbalanced Design

Equivalent Groups Design (EGD)

When using the Equivalent Groups Design, two comparable samples are taken from a common population A, one is administered test X and the other test Y. The success of this design depends on taking large representative comparable groups.

Population	Sample	X	Y
Group C	1	X	
Group D	2		Y

Figure 10: Equivalent Groups Design

When this data collection design (DCD) is adopted, two types of tests are developed and administered on separate group of candidates who are randomly assigned to complete a test (Kolen & Brennan, 1995; Yen & Ferrara, 1997). This design reduces the time required for each candidate and therefore appropriate where there is time limitation. However, using randomly comparable groups should lead to outcome that is theoretically comparable. Differences in performance between the groups can be attributed to differences in the difficulty of the tests taken.

The result of randomly equivalent groups design is acceptable if the following conditions according to Muraki, Hombo and Lee (2000) can be met: ñ(1) a sufficient number of candidates can complete the assessment concurrently, (2) candidates can be randomly assigned to the forms, and (3) all forms to be equated are available simultaneously. This DCD is once in a while achieved in practice because the level of control required to randomly assign candidates is uncommon to meet, particularly if the tests are given over an extended time period.

Non-Equivalent Anchor Test (NEAT) Group Design

When using a NEAT design, two candidate groups, A and B take test form X and Y respectively. Because the two groups are not assumed equivalent and they are not taking the same test forms, they must be connected through an anchor test A, or an equivalent test to the form to be equated that is used to equate the test forms and account for group differences in ability.

Hence, in this DCD, the tests are developed to have some subset of test items in common. The forms are administered to different groups that are sampled in a way that the assumption of random equivalence does not hold. The common items in NEAT design are such that they are likely to contribute to the candidate's total score. When common items contribute to the total score, they are called *internal items* but when the contribution of the common item to the total score is not significant, and then the items are called *external items*. External related items are commonly administered as a separate, timed block of items; internal related items are often interspersed throughout a form (Kolen & Brennan, 1995). Similarly, when the related test selects common items from the two tests to be equated, the related test is called "Internal related test". But when different test measuring the same ability as the two test forms is developed and used as related test, it is called "external related test".

Population	Sample	X	A	Y
C	1	X	A	
D	2		A	Y

Figure 11: Non-Equivalent Groups with Related Test Design

From literature reviewed, the use of appropriate test equating and its designs, could settle the comparability problem of continuous assessment standards across states in Nigeria. Also, the application of test equating and its design by examination bodies may serve as a quality control mechanism in judging human capital being produce by different schools within Nigeria. The equating methods and designs reviewed are primarily important because they help the researcher to choose the equating methods and design that were used in the current study.

Test Scores Equating Guidelines

Most practitioners agree that there are five requirements for a linking between scores on two tests to be considered an equating (Dorans & Holland, 2000; Petersen, 2008):

- Å Same construct: The two tests must both be measures of the same characteristic (latent trait, ability, or skill).
- Å Equal reliability: Scores on the two tests are equally reliable.
- Å Symmetry: The transformation is invertible.
- Å Equity: It does not matter to examinees which test they take.

Å Population invariance: The transformation is the same regardless of the group from which it is derived.

In reality, equating is used to fine-tune the test construction process. We equate scores on tests because of our inability to construct multiple forms of a test that are strictly parallel. Thus, the same construct and equal reliability constraints simply imply that the two tests to be equated should be built to the same blueprint, that is, to the same content and statistical specifications. The study by Liu and Holland (2008) emphasize the importance of construct similarity in the tests to be equated. In line with construct similarity, the continuous assessment items used by teachers in Nigeria are from the same content (subject).

The population invariance and symmetry conditions follow from the purpose of equating: to produce an effective equivalence between scores. If scores on two tests are equivalent, then there is a one-to-one correspondence between the two sets of scores. This implies that the conversion is unique; that is, the transformation must be the same regardless of the group from which it is derived. It further requires that the transformation be invertible. That is, if score y_o on test Y is equated to score x_o on test X, then x_o must equate to y_o . Thus, regression methods cannot be used for equating.

The same ability and population invariance conditions go hand in hand, as do the same ability and equity conditions. If the two tests were measures of different abilities, then the conversions would certainly differ for different groups (see Liu & Holland, 2008). And if the two tests measure different skills, then examinees will prefer the test on which they will score higher.

However, Angoff (1971) suggested that the method of equating need to meet only two of the five conditions mentioned above. The conditions according to Angoff (1971) are; (1) the two instruments in question should yield measures of the same characteristics, (2) In order to be on a transformation across systems of units, the conversion must be unique, except for the random error associated with unreliability of the data and the method used for determining the transformation. The resulting conversion should be independent of the individual from whom data were drawn to develop the conversion and should be freely applicable to all situations.

Dorans (1990) argued that, in order to equate one test to another, both tests have to measure the same construct. The author further argued that, both tests do not have to contain unidimensional items, but they have to measure the same dimension. Going by Angoff and Dorans opinions, what is most paramount in equating test scores is for the tests involved to measure the same construct. Thus, a test which is to measure mathematics ability should not be equated to another test whose aim is to measure history ability.

Test Equating Criteria/ Equating Error

The goal of equating is for scores on multiple forms to be used interchangeably. Many methods have been developed to equate forms, and it is important, after equating, that the results be evaluated. Evaluation of equating results requires that the criterion be identified. Different methods may be preferred under each criterion. However, Gao (2004) highlighted some of the equating criteria to include;

1. Weak equity or tau equivalence is considered a special case of Lord's (1980) equity definition. It only requires the mean of the conditional distribution on each after equating to be equal. This special criterion includes the equating expected scores and conditional variable of the equating function. The advantage of this criterion over other equating criterion is that, it is directly aligned with a special case of Lord equity definition of equating. Therefore, whenever Lord's definition is adopted, it is suggested that the Lord's equity criterion be used. The disadvantage of equity criterion is that it is relatively difficult to compute and explained.
2. Summary indices are often used to compare two sets of equating conversions. The Root Mean Square Error (RMSE) is frequently used. The advantage of using these indices is that they are easy to interpret. The disadvantage is that the index may not specify the loss function or choice of standard.
3. Standard Error of Equating (SEE) is an analytical method to estimate the amount of equating error from sampling, which is one aspect of the accuracy of the equating. This method is easy to apply and interpret; however, it ignores systematic errors. The difficulty in using this criterion is that although smaller errors are preferable to larger errors, whether the magnitude of difference between the standard error is important or whether the size of the errors of equating themselves is larger is unanswered.
4. Estimated scaling constant can be compared to actual constants with generated data. Generated data means that the data are generated or simulated. The

advantage of this method is that the true equating relationship is known and can be used to evaluate the results. This method is most useful when the generated data closely resemble the real data of interest. The disadvantage of this method is the potential bias and question of how well generated data mimic real data remains unanswered.

5. Estimated scaling constants can be compared to actual constant of a test which is equated to itself. Equating a test to itself is also known as circular equating. A test is equated to itself either directly or through a chain of intervening forms. Traditionally, the circular equating criterion was intended to assess systematic error. The method has that advantage of knowing the true equating. However, the drawback is that no equating of sort ever works well.
6. A large sample criterion is used as an estimate of the population conversion to evaluate the equating result from smaller groups. This criterion is easy to interpret; however, a large sample is not always available.
7. Consistency criterion means that equating results are compared across methods. Usually, all that can be concluded from such a comparison is whether the method can provide similar or dissimilar result and then one method will be substituted for another practical reason. This method does not address accuracy.
8. A stability procedure compares new procedures to conventional equating methods to assess similarity but not necessary accuracy. Cross-validation is a

common example. This method is easy to apply; however, it does not address the accuracy.

Other criteria are the same distributions property (Kolen & Brennan, 2004), the first order equity property, and the second-order equity property (Morris, 1982). The same distributions property states that the distribution of scale scores in a common population should be the same on the old form and the new form after equating (Kolen & Brennan, 2004). The first-order equity property holds if, conditional on the true score, examinees have the same expected scale score on the two forms. The second-order equity property holds if, conditional on the true score, examinees have the same conditional standard error of measurement (SEM) on the two forms (Morris, 1982). The equity criteria are often used in the literature to assess the adequacy of equating and linking. Hanson (1989) used the equity criteria to evaluate the scaling of the PLAN and the American college test (ACT) Assessment, Harris (1991) also used the criteria to compare Angoff's designs for vertical scaling results, and Bolt (1999) equally used the criteria to investigate whether the IRT true score method is adversely affected by the presence of multidimensionality.

All three equating criteria and those highlighted by Gao (2004) as shown above, assess whether equating has achieved its purpose: the interchangeability of scores on alternate forms. To evaluate equating results in terms of the same distributions property, the nonparametric Kolmogorov-Smirnov T statistic is used. According to Conover (1999) the K-S T-statistic helps to quantify the difference between the score distributions. This statistic takes the maximum difference between two relative cumulative distribution

functions and is quite powerful in detecting distributional differences. To evaluate an equating result based on the first- and second-order equity properties, expected scale scores and their conditional SEMs are calculated assuming a psychometric model. Kolen, Zeng, and Hanson (1996) describe such procedures using dichotomous IRT models. Comparison between expected scale scores across proficiency levels on both forms help to quantify the extent to which the first-order equity holds. Similarly, the difference between conditional SEMs for scale scores on both forms can be used to investigate the extent to which second-order equity holds.

Kolen and Brennan (2004) indicate that each test-equating method is often designed to function optimally under at least one of the equating properties. The equipercentile method and the IRT observed score method both equate observed score distributions and therefore should be expected to perform relatively well under the same distributions property. The IRT true score method equates “true” scores and is expected to perform well under the first-order equity property. It is difficult to predict which equating method performs well to preserve the second-order equity property (Ye & Kolen, 2005).

Ideally, an equating method should be selected if it produces adequate results relative to all of the criteria. If two forms are nearly identical, the equating is expected to perform well relative to any of the criteria. When forms differ substantially in difficulty, Kim (2005) found that the equipercentile method performs well relative to the same distributions property and poorly relative to the first order equity property; the IRT true

score method performs well relative to the first-order equity property and poorly relative to the second-order equity property. Kim (2005) examined a small number of equatings; it was difficult to tell how large the difference in difficulty needs to be before equating performs poorly relative to each of the criteria. This study conducted by Kim (2005) has large number of separate equatings to investigate the impact of form-to-form difference on equating adequacy based on the three criteria.

There are many views in the equating literature about using statistical significance tests for selecting equating functions. Some views are ambivalent about using significance tests for selecting equating functions, possibly because in practice, equating function selection has been influenced by beliefs and heuristics and not necessarily by statistical criteria (Kolen & Brennan, 2004; Livingston, 2004). Other views encourage the use of significance tests because tests offer the potential to formalize equating decisions and reduce reliance on guesses, experience, and intuition (von Davier, Holland, & Thayer, 2004). Still other views encourage significance tests but note their limitations in addressing the practical implications of equating function selections, such as the implications of score rounding for score reporting (Dorans & Lawrence, 1990; Hanson, 1996). Several statistical significance tests have been proposed for equating function selection (Dorans & Lawrence, 1990; Hanson, 1996; Moses, Yang & Wilson, 2007; von Davier et al., 2004). Most of these proposals are based on demonstrating a limited number of tests on one or two data sets rather than on comparing the long-run accuracies of several tests to each other. The current investigation will focused on evaluating equating results using Root Mean Square Error (RMSE). This evaluation criteria is most

appropriate because it allows the researcher to use the summary indices to compare two sets of equating conversions. It also helps to determine the most efficient equating method.

When applying equating methods, different types of equating error influence the interpretation of the results. The first type - random equating error is always present because samples are used to estimate parameters such as means, standard deviations, and percentile ranks. However, random error can be reduced by using large samples of examinees and by the choice of equating design. Systematic error is more difficult because it results from the violation of assumptions and conditions unique to the particular equating methodology used. Unlike random error, which can be quantified using standard error calculations, systematic error is more difficult to estimate.

Divgi (2003) observed that only a sample of examinees, rather than the entire population, is available in practice during equating and that if the sample is small, the random error of equipercentile equating may be unacceptably large. A popular alternative is linear equating, which is more stable, that is, it has more less random error because it is based only on means and standard deviations of the two forms. This informed the inclusion of linear equating method in this study. However, when the score distributions of the forms have different shapes, linear equating suffers from bias, i.e., systematic errors, especially at very high and/or low scores. The choice between linear and equipercentile methods depends on one's judgment about the relative importance of random and systematic error. If the sample is very large, the bias of linear equating

exceeds its superiority in random error, and hence the equipercentile procedure is preferable. The opposite is true when the sample is small.

Equating error may occur as random or systematic. Random error presents itself whenever samples from populations of examiners are used to estimate parameters such as mean, standard deviation, and percentiles ranks (Barnard, 1996). This type of error can be reduced by using large samples and by choice of equating design (Felan, 2002). Systematic error on the other hand results from the violation of the assumptions and conditions of the particular equating methodology used (Zeng, 1991).

Felan, (2002) and Reng (1991) maintain that systematic error in equating will likely occur if:

- One fails to control for fatigue and practice effect in the single group design.
- The spiraling process is ineffective in achieving group comparability in the random group design
- Group differs substantially, or if the common items are not representative of the total test form in content and statistical characteristics
- Common item function differently from one administration to another in the nonequivalent groups design
- The new form and the old form differ in content, difficulty and reliability.

Felan and Zen however concluded that both random and systematic equating errors can be controlled through the use of an adequate sample size.

Test Scores Equating Methods/Models

Test Scores Equating Methods Based on CTT

Linear Equating: Linear equating assumes that, apart from differences in means and standard deviations, score distributions on two forms of a test are the same. This allows difficulty differences to vary along the score scale of the two assessments. Given this assumption, scores on the two forms can be matched using their z scores. The linear conversion is defined in terms of the mean and Standard deviation of the two scores (X and Y) and represented symbolically as;

$$\frac{x - (\bar{x})}{s_x} = \frac{y - (\bar{y})}{s_y} \text{ ----- (1)}$$

where, x_1 = any score on test x

$\bar{x}$ = mean of test x

s_x = standard deviation of test x

y_2 = any score on test y

$\bar{y}$ = mean of test y

s_y = standard deviation of test y

Solving for x, we have;

$$x = \frac{s_x}{s_y} (y - \bar{y}) + \bar{x} \text{ ----- (2)}$$

$$x = \frac{s_x}{s_y} y + \left(\bar{x} - \frac{s_x}{s_y} \bar{y} \right) \text{ ----- (3)}$$

$$x = Ay + B \text{ ----- (4)}$$

$$\text{Where, } A = \frac{s_x}{s_y} \text{ and } B = \bar{x} - \frac{s_x}{s_y} \bar{y}$$

In this case, A is referred to as the slope of the linear conversion and B the intercept. Linear equating is sometimes called linear conversion. It allows the relative difficulty of two or more forms of tests to vary along the score scale. When the standard deviation are equal, linear equating becomes the same as mean equating described by Kolen & Brennan (1995). Linear equating is appropriate if the score distributions differ only in mean and standard deviation. If there are differences beyond these first two moments, or if the shape of the distributions differ, then linear equating is inappropriate. Often the individual scales are of such limited range that score distributions are unstable and unsuitable for distribution-based equating methods. If multiple scales are aggregated to form a composite or total score, the score range might be large enough to permit linear equating. However, caution is needed when assuming that score distributions will be reasonably stable over multiple assessment populations. Large score distribution differences can invalidate a linear equating because the equating transformation would differ for another dataset (Eiji, Catherine & Yong-Won, 2000).

In linear equating, a transformation is found such that scores on X and Y are said to be equated if they correspond to the same number of standard deviation units above and below the mean in T, where T is the population in which the equating is performed. Individual score obtained from CA are aggregated to form a composite or total score. These composite score so obtained result in a large distribution which are stable and

suitable for linear equating. Hence the researcher decided to use linear equating as one of the equating method for this study.

Equipercentile equating. Equipercentile equating involves determining which scores in a distribution have the same percentile rank. Those scores are then declared equivalent. The percentile rank for each number-correct or scale score is determined on both test forms, and a percentile-rank score curve is created for each assessment. This graph is used to locate a corresponding base score for any equated score in the range. Because actual scores on assessments are discrete rather than continuous, a method is required to approximate a continuous distribution of scores (Holland & Thayer, 1989).

The equipercentile equating uses the cumulative percentages in order to equate two tests. The cumulative percentages (in other words the percentile values) are values of the total score that divide the data into two groups so that a certain percentage of the sample is above and the rest of the sample is below. For example, the 75th percentile indicates the value of the total score below which 75% of the students fall.

The aim of equipercentile equating method is to make the raw scores on two tests correspond to the same cumulative percentage for a certain group of examinees. The technique demands either the conversion of one set of scores to the other, or the conversion of the two sets of scores to a third (new) one. In other words, after successful equating, it should be indifferent to the examinees whether they sat for one paper or the other.

Therefore, for the equipercentile equating of two tests, it is enough to use the graphs of the cumulative percentages of their scores. If the two graphs are drawn on the

same axes, then the raw scores on each test that correspond to the same cumulative percentage can be identified. These scores form pairs of equivalent scores. Many such pairs can be plotted on a graph to form the "conversion line" which is the function that transforms the scores of the one test onto the score scale of the other test. A major advantage of equipercentile equating between two tests X and Y is the opportunity to set cutting scores (for groups similar to that used for the equating) on both tests and still be sure that the same percentage of examinees will succeed at each test. As mentioned by Kolen, the percentages and cumulative percentages are the only "statistics" you need to know in order to employ the equipercentile equation.

Equipercentile equating relies heavily on the presumption that scores will have sufficient variance to allow the formation of a stable statistical distribution. CAs could have so few items and/or such a limited score range that this might not be the case. When CA items are scored on multiple subscales, and a composite overall score is created, there might be sufficient variation. However, many CA items are scored on very limited, holistic scales, and the entire assessment can be composed of only one or two scored responses. In this situation, equipercentile equating might be inappropriate.

Most CA tests have free-response (FR) and multiple-choice, (MC) components, but the equating in this study was done using the students CA scores and MC tests developed by the researcher. Equating only on the MC items requires assuming that any trends over time in the candidates' achievement of the knowledge and skills measured by the multiple-choice section will be matched by similar trends in their achievement of the knowledge and skills measured by the free-response section. Performance on the free-

response section might be influenced by multidimensionality, and might not be well-represented by the conversion developed from the MC section. Free-response items might also have different reliabilities than MC items, because human judges rating the free responses may introduce another source of error variance not present in MC scoring.

Equipercentile equating is divided into frequency estimation equipercentile equating method and chained equipercentile equating method. The frequency estimation method uses the frequency distributions of Form X and Form Y for a common synthetic population and is estimated as follows:

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x), \text{-----} (5)$$

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y), \text{-----} (6)$$

where, $f(x)$ and $g(y)$ are the population distributions for X and Y scores, respectively. The subscripts s, 1, and 2 represent the synthetic population, Population 1, and Population 2, respectively. w_1 and w_2 are the weights for Population 1 and Population 2 that are used to define the synthetic population. However, $f_2(x)$ and $g_1(y)$ are not directly observable. The frequency estimation method assumes that the distributions of X and Y scores conditioned on the common item set V scores are population invariant; that is,

$$f_1(x|v) = f_2(x|v), \text{-----} (7)$$

$$g_1(y|v) = g_2(y|v), \text{-----} (8)$$

Then, it follows that

$$f_s(x) = w_1 f_1(x) + w_2 \times f_1(x/v)h_2(v), \text{-----} (9)$$

$$g_s(y) = w_1 \times g_2(y/v)h_1(v) + w_2 g_2(y), \text{-----} (10)$$

where $h_1(v)$ and $h_2(v)$ are the marginal distributions of the common item set scores in Populations 1 and 2. All the above quantities are directly observable with a common-item nonequivalent groups design. Equipercentile equating is then applied to $f_s(x)$ and $g_s(x)$. This equating method is also good for the equating of continuous assessment.

In *chained equipercentile* equating method, Form X scores are first equated to the common-item V scores in Population 1 using the equipercentile equating method:

$$e_y(v) = R_1^{-1}(p_1(x)), \text{-----} (11)$$

where, R_1^{-1} is the inverse of the percentile rank (PR) function for V in Population 1, and $P_1(x)$ is the PR function of X in Population 1. Then, the common-item set V scores are equated to Form Y scores in Population 2:

$$e_Y(v) = Q_2^{-1}(R_2(v)), \text{-----} (12)$$

where, $R_2(v)$ is the PR function of V in Population 2, and Q_2^{-1} is the inverse of the PR function of Y in Population 2. Finally, the Form X scores are equated to Form Y scores through a chain consisting of the two equipercentile equating functions:

$$e_Y(x) = Q_2^{-1}(R_2(R_1^{-1}(p_1(x)))), \text{-----} (13)$$

Both of these methods have been used in actual testing programmes, although they are seldom used in the same testing programmes.

Von Davier, Holland, and Thayer (2004) did some theoretical analyses of frequency estimation equipercentile equating and chained equipercentile equating methods and showed that they are both examples of what they termed “observed score equating.” The methods entail assumptions that are generally not testable in practice, and the methods produce essentially identical results under two extreme conditions: (a) the

two populations are very similar or (b) the anchor test is perfectly correlated with both tests. The theoretical work by von Davier et al., however, did not illuminate the comparative nature of the two methods under the realistic condition of group difference. This equating method is important to this study because it helps the researcher to report studies that compared linear and equipercentile equating.

Some IRT Models used for Test Scores Equating for Dichotomous Equating

Item response theory emerged, as early as the 1940s though the popularity came much later in the 1970s. As the name implies, IRT models consider examinee behavior at the item level, not at the test level. Modeling at the item level creates much more flexibility for applications to test development, study of differential item functioning, computer-adaptive testing, score reporting, etc. Early IRT models were developed to handle dichotomous responses (i.e., binary responses; for example, 0 (incorrect) and 1 (correct)) but today, models are available to handle just about all types of educational and psychological data (Linden & Hambleton, 1997).

Two of the fundamental assumptions with IRT models are unidimensionality and local independence. The assumption of unidimensionality means that a set of items and/or a test measure(s) only one latent trait (θ), and local independence refers to the assumption that there is no statistical relationship between examinees' responses to the pairs of items in a test, once the primary trait measured by the test is removed. The two assumptions are really just different ways to say the same thing about the data. The third main assumption concerns the modeling of the relationship between the trait measured by

the test and item responses. What follows are various models that make different assumptions about that relationship.

Normal Ogive Model;

The normal ogive model was the first IRT model for measuring psychological and/or educational latent traits. In the model, an item characteristic curve (ICC) is derived from the cumulative density function (CDF) of a normal distribution. A mathematical expression of the normal ogive model is as follows:

$$P_i(\theta) = \int_{-\infty}^{ai(\theta-b_i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, \quad - \quad - \quad - \quad - \quad (14)$$

where $P_i(\theta)$ is the probability of a randomly chosen examinee at ability level θ answering item i correctly, a_i is the discrimination parameter of item i , b_i is the difficulty parameter of item i , and z is a standardized score of the examinee involving trait score, and the two item parameters.

One-Parameter Logistic Model (1PLM- a.k.a. Rasch Model): A mathematician in Denmark, George Rasch, came up with a different approach to IRT in the 1950s. He used a logistic function to derive an ICC instead of the normal ogive function (though at the time he expressed his model differently), and his model contributed to simplifying the normal ogive model and the complexity of computation, he appeared to be unaware of the earlier work on the topic of item response theory. In the Rasch model, the probability of a randomly chosen examinee at a ability level θ obtaining a correct answer on item i can be expressed as

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}, \quad (15)$$

where e is an exponential constant whose value is about 2.718, and D is a scaling factor whose value is 1.7. The choice of this value for D , produces near equivalent values and interpretations between the item parameters in the normal ogive and two-parameter logistic models. Today, it is common to simply set $D=1$. Since the normal ogive model is rarely used in practice and so preserving consistent interpretations between the models is not important. But it is important to know when either studying item parameter estimates or generating them, that the value of D in the model be considered. It is still common, especially with the two and three parameter logistic models to retain D in the model with a value of 1.7. When $D = 1.0$ (it is common to say that the model parameters are placed on what is referred to as the "logistic metric") and with $D = 1.7$ (it is common to say that the model parameters are placed on what is referred to as the "normal metric".)

Two-Parameter Logistic Model (2PLM) is a generalization of the 1PLM. Instead of having a fixed discrimination of 1 across all items as in 1PLM, in the 2PLM, each item has its own discrimination parameter. Thus, the model is mathematically expressed as:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \quad (16)$$

The two-parameter logistic (2PL) model predicts the probability of a correct response to any test item from ability and two item parameters. The basic difference with respect to the 1PL model is that the expression $\exp(\theta - b_i)$ is replaced with $\exp[a_i(\theta - b_i)]$.

Just as in the 1PL model, b_i is the difficulty parameter. The new parameter a_i is called the discrimination parameter.

Three-Parameter Logistic Model (3PLM): The three-parameter logistic model (3PLM) allows an ICC to have non-zero lower asymptotes. This model is more suitable for response data to those items in which examinees at the extremely low proficiency level may get the items correct by chance; for example, a multiple choice item. In this model

$$P_i(\theta) = C_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \quad - \quad - \quad - \quad (17)$$

where c_i represents the probability that examinees at extremely low levels of the trait answer item i correctly. This third item parameter, c_i , is often called either the pseudo-chance-level parameter or the guessing parameter, although pseudo-chance-level parameter is theoretically more appropriate (Hambleton, Swaminathan, & Rogers, 1991). The 2PLM is a special case of 3PLM when $c=0$, and 1PLM is a special case of 2PLM when $a = 1$.

The viability of IRT is demonstrated based on the property of the item parameter invariance/consistency (differential item functioning). The 1PLM or 2PLM seems to be deficient in measuring all the parameters that are to be assessed or determine in a dichotomously scored instrument, but the 3PLM appears to support the determination of guessing factor, group invariance or item parameter consistency and was therefore appropriate for this study.

Nonparametric Item Response Model

Item Characteristic Curves (ICC) is always characterized by a single function in IRT models with parameters. However, assuming a single function for ICCs may not be appropriate to represent response data in some cases, Nonparametric item response models, in which a variety of shapes of ICCs are allowed, were developed in 1950s, even before parametric item response models were introduced.

Ramsay (1991) proposed a *kernel smoothing* approach for nonparametric item response models. In a kernel smoothing approach, $P(\theta)$ is estimated at g -th evaluation point, q_g , by a local averaging procedure. Thus,

$$\hat{P}(\theta) = \sum_{g=1}^G w_g(\theta) Y_i^{(g)},$$

where

$$w_g(\theta) = \frac{K[(q_g - \theta)/h]}{\sum_k K[(q_k - \theta)/h]}, \quad (18)$$

So,

$$\hat{P}(\theta) = \frac{\sum_g K[q_g - \theta] Y_i^{(g)}}{\sum_g K\left[\frac{q_k - \theta}{h}\right]},$$

where

h is the bandwidth parameter, which controls bias and sampling variance, and $K(u)$ is one of the kernel functions (Ramsay, 1991, p.617):

(a) $K(u) = 0.5, |u| \leq 1$, and 0 otherwise, for uniform,

(b) $K(u)=0.75 (1-u^2)$, $|u|\leq 1$, and 0 otherwise, for quadratic, and

(c) $K(u)=\exp(-u^2/2)$ for Gaussian.

Nonparametric item response models may not be as practically useful for operational uses as parametric models because nonparametric item response models do not provide informative, interpretable item parameters (for example, difficulty parameters), and it is hard to equate tests under nonparametric models. However, nonparametric models are frequently used for research purposes such as evaluating model fit for parametric models since nonparametric models produce item characteristic functions that are very close to the observed data.

Unidimensional IRT Models for Analyzing Polytomous Responses

In dichotomous item response models, the only type of response data is binary (i.e., 0 or 1). However, in some test situations, responses can be of more than two categories. For example, a questionnaire on attitude, using Likert-scale items, may result in 5 categorical responses (strongly disagree, disagree, neutral, agree, and strongly agree, which can be coded from 1 to 5, for a positively phrased item or 5 to 1 for negatively phrased item). Sometimes polytomous responses are dichotomized to be handled within dichotomous item response models, but it is very inappropriate in most cases because dichotomizing polytomous responses changes the nature of the scale of the measure and, as a result, validity of the measure could be seriously threatened. Several item response models were developed to enable uses of polytomous responses within an IRT framework. However many of polytomous item response models are basically generalizations of the dichotomous item response models.

Partial Credit Model (PCM)

The partial credit model is an extension of the 1PLM (Rasch model) (Dadughun, 2008).

Equation (15) for the 1PLM above can be rewritten as

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}} = \frac{\exp(D(\theta - b_i))}{1 + \exp(D(\theta - b_i))} = \frac{P_{i0}(\theta)}{P_{i0}(\theta) + P_{i1}(\theta)}, \quad (19)$$

where $P_{i0}(\theta)$ is the probability of a randomly chosen examinee, whose proficiency level is θ scoring 0 on item i , and $P_{i1}(\theta)$ is the probability of a randomly chosen examinee, whose proficiency level is θ , scoring 1 on item i . Thus, the probability of a person at θ , scoring x over $x-1$ can be computed as

$$\frac{P_{ix}(\theta)}{P_{ix-1}(\theta) + P_{ix}(\theta)} = \frac{\exp(D(\theta - b_{ix}))}{1 + \exp(D(\theta - b_{ix}))}, \quad x=1, 2, \dots, m_i, \quad (20)$$

where $P_{ix}(\theta)$ and $P_{ix-1}(\theta)$ refer to the probabilities of examinee at θ , scoring x and $x-1$, respectively. It should be noted that the number of item difficulty parameters are, now, m_i (one less than the number of response categories) in Equation (20). The probability of a randomly chosen examinees, who is at θ , scoring x on item i can be expressed as

$$P_{ix}(\theta) = \frac{\exp \sum_{k=0}^{x-1} (D(\theta - b_{ik}))}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h-1} (D(\theta - b_{ik}))}, \quad x=1, 2, \dots, m_i \quad (21)$$

The function of Equation (21) is often called the *score category response function* (SCRf).

Generalized Partial Credit Model (GPCM)

The generalized partial credit model is a modification of the PCM with a parameter for item discrimination added to the model. Muraki (1992) expressed the model mathematically as following:

$$P_{ix}(\theta) = \frac{\exp \sum_{k=0}^x (z_{ik}(\theta))}{\sum_h \exp \sum_{k=0}^h D(Z_{ik} - (\theta)),} \quad (21)$$

$$\text{where } Z_{ik}(\theta) = Da_i(\theta - b_i + d_{ix}), \quad (22)$$

where d_{ix} is the relative difficulty of score category x of item i . Although Muraki (1992) followed the same way of parameterization for item and score category difficulty as Andrich's (1978) rating scale model, the item difficulty parameters for each score category can be simply rewritten as

$$b_{ix} = b_i - d_{ix} \quad (23)$$

and so is Equation (22),

$$Z_{ik}(\theta) = Da_i(\theta - b_{ix}) \quad (24)$$

The only difference between the PCM and GPCM is the additional discrimination parameters for each item (a_i).

Rating Scale Model (RSM)

There are two different approaches to the rating scale model. Andersen (1983) proposed a response function, in which the values of the category scores are directly used as a part of the function:

$$P_{ix}(\theta) = \frac{e^{w_x \theta - a_{ix}}}{\sum_{x=1}^m e^{w_x \theta - a_{ix}}} \text{-----} (25)$$

where $w_1, w_2, \dots, w_m$ are the category scores, which prescribe how the m response categories are scored, and a_{ix} are item parameters connected with the items and categories. An important assumption of this model is that the category scores are equidistant.

Another form of the RSM was proposed by Andrich (1978a, 1978b), this can be seen as a modification of PCM. In Andrich's RSM, item response functions are computed via

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^x (\theta - (b_i + d_{ix}))}{\sum_{x=0}^m (\exp \sum_{j=0}^x (\theta - (b_i + d_{ij})))} \text{-----} (26)$$

where d_{ix} is the relative difficulty of score category x of item i . Andrich's RSM assumes that the category scores are fixed across all items in a testlet, and RSM should not be used if the scale of category scores varies across items in a testlet.

Graded Response Model (GRM)

The graded response model was introduced by Samejima (1995) to handle ordered polytomous categories such as letter grading. A, B, C, D and F, and polytomous responses to attitudinal statements (such as a Likert scale). The model is expressed as

$$P_{ix}^*(\theta) = \frac{\exp(Da_i(\theta - b_{ix}))}{1 + \exp(Da_i(\theta - b_{ik}))} \text{-----} (27)$$

where $P_{ix}^*(\theta)$ is the probability of an randomly chosen examinee with proficiency of θ scoring x or above on item i . This function is called the cumulative category response function (CCRF). Probability of each score category can be given by

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{ix+1}^*(\theta) \quad (28)$$

Thus, the score category response function (SCRF) of the GRM can be expressed as

$$P_{ix}(\theta) = \frac{\exp[-Da_i(\theta - b_{ix+1})] - \exp[-Da_i(\theta - b_{ix})]}{[1 + \exp[-Da_i(\theta - b_{ix})]] [1 + \exp[-Da_i(\theta - b_{ix+1})]]} \quad (29)$$

Unlike the PCM and GPCM, the interpretation of item parameters of the GRM should be based on the CCRF, not on the SCRF. Within the GRM, a value of A -parameter for each response category indicates where a probability that a randomly chosen examinee, whose proficiency level (θ) is exactly same as the value of b -parameter, scores x or higher is 50% on the CCRF.

Nominal Response Model (NRM)

The nominal response model (also called the *Nominal Categories Model*) is an IRT model propounded by Bock (1972). This model can be used in place of Graded response model, unlike the other polytomous IRT models introduced above; polytomous responses in NRM are unordered (or at least not assumed to be ordered). Even though responses are often coded numerically (for example, 0, 1, 2,..., w), the values of the responses do not represent some sort of scores on items, but just nominal indications for response categories. Some applications of the NRM are found in uses with multiple choice items. The category function of NRM can be expressed as:

$$P_{ix}(\theta) = \frac{e^{z_{ix}}}{\sum_{k=1}^{m_i} e^{z_{ix}}} \text{-----} (30)$$

where

$$Z_{ix} = a_{ix} \theta + C_{ix} \text{-----} (31)$$

In Equation (31), a_{ix} and c_{ix} are called the slope and intercept parameters, respectively, and they are related with item discrimination and location. For any successful test scores equating to be conducted, these models are essential tools which are always used in transforming the raw scores before equating.

Examinees' Ability, Population-Invariance/ Differential Item Functioning (DIF) and Test scores Equating

Forsyth (1987) observes that the major potential of item response theory (IRT) rests in the characteristics of parameter invariance. The author maintains that if the assumptions of IRT models are satisfied, item parameters are invariant across groups of examinees and ability parameters are invariant across groups of items. The invariance of the item parameters is particularly important in horizontal equating settings; while the invariance of ability parameters is essential in IRT applications in vertical equating and computerized adaptive testing. If the items in a test have been calibrated using IRT procedures, then, presumably, a subset of items could be used to estimate an examinee's achievement level and a school's mean achievement level.

In other words, Shoemaker (1980) as cited by Wilcox (1987) observes that if examinee's ability has been established through his or her responses to certain carefully

calibrated items, that it is possible to estimate the ability in a similar situation even though he or she is not present. The author concludes that the estimated test parameters on parallel test items are always statistically equivalent given the invariant nature of the ability. Shoemaker (1980) observes that from item response theory, it is possible to estimate scores that examinees make on items to which they do not respond from scores that they make on items to which they responded. The author used his result to illustrate that if examinee Z is administered item 1 and 2 but not 3, if one actually knows the three item characteristic curves, obviously one could estimate the probability of passing item 3 from the examinee Z's position on the attribute. Thus if sufficient information can be obtained from the subsets of a total collection of items, one can estimate scores on the underlying attributes, since performance is a monotonically increasing function of the latent trait of ability (Warm, 1987). The need to administer the same item to all subjects as it is done in conventional testing, very seldom constitutes a serious practical problem. The notion of item response theory encourages a situation of giving individuals a test in which not all subjects would be administered all the same items.

Shoemaker (1980) maintains that a good testing procedure is expected to produce results that do not depend heavily on the particular set of questions used, or the time at which the test is given or the person who scored the test. Factors such as these can cause examinees that are equal in the ability the test is intended to measure to receive different scores on the test. The influence of these factors-is referred to as measurement error. The term measurement error does not mean that someone has made a mistake in constructing, administering or scoring the test. It rather means that the examinee's test score is 'affected

to some extent by factors the test is not intended to measure. The author also observes that the term ability connotes the characteristics of examinees that the test is intended to measure. It includes factual knowledge and specific skills as well as more general abilities. For an ability to be tested, the questions or problems on a test are only a sample of all questions or problems that would have been used. A test score based on this sample of questions or problems will only be an approximate indicator of examinee's ability. Hence test calibrations are independent of the sample of persons used to estimate item parameters and person measurements, the transformation of test scores into estimates of person ability are independent of the selection of items used to obtain test scores (Wright and. Panchapakeson,1996).

MacDonald and Sampo (2002) observe that in theory, measures based on IRT overcome the principal limitation of measures based on classical test theory, that is, item parameter estimates are not dependent on the particular sample of examinees who have been administered the test items, and the person ability estimates are not dependent on the particular sample of test items administered. This invariance property of IRT models has been demonstrated extensively and has been widely accepted (Hambleton and Jones, 1997). Therefore meaningful comparisons of ability can be made even when the parameters of the items used to make the different measurement are not the same. The sample-free model assumes that all items have the same discrimination (for 1PLM), and that the effect of guessing is negligible.

It has been pointed out earlier that in the framework of IRT, item parameters are assumed to be invariant to group membership. In reality, this is not always the case as

some items in a test or set of tests seems to function differently across the subgroup within the population in a study or examination. This scenario results into item parameter variant or what is technically called differential item functioning (DIF) in IRT. Camilli & Shepard (1994) observed that, "DIF is said to occur whenever the conditional probability, $P(X)$, of a correct response or endorsement of the item for the same level on the latent variable differs for two groups". The authors advised that the "key decision that must be made for DIF analysis is the selection of the appropriate IRT model, noting that different models allow a different number of item parameters (i.e., b, a, c parameters) to be estimated from the data of item responses, and thus, allow for the evaluation of DIF for different item properties". Differential item functioning (DIF) is a condition when an item functions differently for respondents from one group to another. Items that show DIF is a serious threat to the validity of the instruments to measure the trait levels of members from different populations or groups. Instruments containing such items may reduce the validity for between-group comparisons, because their scores may be indicative of a variety of attributes other than those the scale is intended to measure (Thissen, Steinberg, & Wainer, 1988).

Zieky (2003) asserts that "the rules for assembling tests using DIF statistics must be followed within the context of the need to produce editions of the test that are parallel to one another". Apart from DIF statistic, there is no other way one can prove that a test question is either fair or unfair. Zieky (2003) also noted that "test may show DIF if it happens to be measuring a skill that is not well-represented in the test as a whole, or if a topic is of greater interest to some group(s) than to others, or if members of some

group(s) are more likely to be exposed to the information being tested. In those cases, judgment is required to determine whether or not the difference in difficulty shown by the DIF index is unfairly related to group membership.

IRT techniques provide a powerful means of testing items for bias or consistency, using what is known as differential item functioning (DIF) analysis. DIF has become an empirical method for investigating the interconnected ideas of (a) lack of invariance, (b) model-data fit, and (c) model appropriateness in model-based statistical measurement frameworks like IRT (Zumbo, 2007)

Differential Item Functioning (DIF) refers to the potential for items to behave differently for different groups. DIF is generally an undesirable characteristic of an examination because it means that the test is measuring both the construct it was designed to measure and some additional characteristic or characteristics of performance that depend on classification or membership in a group, usually gender or ethnic group classification. The principles of test fairness require that examinations undergo scrutiny to detect and remove items that behave in significantly different ways for different groups based solely on these types of demographic characteristics.

Zwick, Thayer, and Mazzeo, (1997) noted that the effects achievement tests have on different subpopulations that respond to it can be detected through DIF procedures. Zwick et al., (1997), Ackerman and Evans, (1994) in their researches have evaluated DIF analysis methods that involve matching examinees' test scores from two groups and then comparing the item's performance differences for the matched members using detection methods like Mantel-Haenszel procedure, and Shealy and Stout's simultaneous item bias

(SIBTEST) procedure. These procedures, however, lack the power to detect nonuniform DIF which may be even more important. Another procedure of detecting item bias which is not yet as popular as those mentioned above is, the Item Response Theory Likelihood-Ratio Test for Differential Item Function (IRTLRDIF). Of all of the procedures available for DIF detection and measurement, IRT-LR procedure posits several advantages over its rivals. IRT-LR procedures involve direct tests of hypotheses about parameters of item response models, they may detect DIF that arises from differential difficulty, differential relations with the construct being measured, or even differential guessing rates (Thissen, 2001). Given that the tests used in this study are parallel tests, and DIF based on sub-population (state A versus B; male versus female and high versus low ability students) was conducted, the use of IRTLRDIF is most appropriate in determining item parameter consistency.

In assessing the degree of DIF present, the odds-ratio estimator can be transformed onto Educational Testing Service (ETS) δ metric- D (Dorans and Holland, 1992). The D statistic represents the difference in item difficulty for the reference and focal groups after the total score has been taken into account (Scheuneman and Gerritz, 1990). The advantage of using the D statistic to classify degree of DIF present is that the ETS has defined the values into a classification scheme delineated by Dorans and Holland (1992). Following guidance proposed by ETS, items were classified into DIF levels as follows:

AA (no DIF), when the absolute value of δ was less than 1.0

ΔB (weak DIF), when the absolute value of delta was between 1.0 and 1.5

ΔC (strong DIF), when the absolute value of the delta was greater than 1.5

The difference in b parameters for the two groups conveys the difference rather than the statistical significance of the DIF (Camilli and Shepard, 1994). In the present study, the three-parameter logistic model was used to find out: the difficulty parameter for two states (state A and state B), sex (males and females) and ability group (high and low ability) by BILOG-MG program.

Theoretical Framework

A test can be studied from different angles and the items in the test can be evaluated and equated according to the framework of different theories. Two such theories upon which this study is anchored are: classical test theory (CTT) and item response theory (IRT).

Classical Test Theory (CTT)

Classical test theory (CTT) was originally the leading framework for analyzing and developing standard tests. It has dominated the area of standardized testing for several decades. This theory is based on the measurement concepts that, a test-taker has an observed score and a true score. The observed score of the test taker is usually seen as an estimate of the true scores of that test-taker plus or minus some unobservable measurement error. Thus CTT postulate linking the observed score (X) to true score (latent unobservable score, T) and error score: $X = T \pm E$. This is the starting point of

classical test theory. (Schumacker 2005; Richard, Donald, Bruno and Ross, 2003; Hambleton, Swaminathan and Rogers, 1991; Croker and Algina, 1986).

The following assumptions underlie CTT:

- True score and error score are uncorrelated
- The average error score in the population of examinee is zero
- Error score in parallel tests are uncorrelated

Classical test theory utilizes traditional item and sample dependent statistics. These includes: item difficulty and item discrimination estimates, distracter analysis, item-test inter correlation, and a variety of related statistics (Schumacker, 2005). Classical test theory also typically includes a measure of reliability of scores (e.g. Cronbach Alpha) and validity (concurrent, predictive, construct or content validity).

Several benefits are obtainable through the application of classical test theory. Firstly, when compared to item response theory models, analyses can be performed with smaller representative samples of examinees. This is particularly important when field-testing any measuring instrument. Secondly, classical test analyses employs relative simple mathematical procedures and models parameters estimations are conceptually straightforward. In addition, classical test analyses are often referred to as 'weak models' because the assumptions are easily met by traditional testing procedures.

While classical models have proven very useful in test development, they have several important limitations. The two statistics that form the cornerstones of most classical test theory, item difficulty and item discrimination, are both sample dependent.

Higher item difficulty values are obtained from examinee samples of lower-average knowledge, while lower item difficulty values occur from examinee sample of above-average knowledge. In terms of discrimination indices, higher values tend to be obtained from heterogeneous examinee samples, and lower values are associated with homogeneous samples. Such sample dependency relationship reduce the overall utility of these statistics.

Classical test theory applications are also test dependent or "test-based". Test difficulty directly affects the resultant test scores. Higher knowledge scores are associated with test composed of relatively easy items, and low knowledge score can be a function of tests composed of item that are more difficult. The true score model, upon which much of the classical test theory is based, permits no consideration of examinee responses to any specific item. Thus, no basis exists to predict how a given examinee will perform on a particular test item.

Despite the above disadvantages, CTT still plays a significant role in the field of psychometric particularly in test equating. All equating methods based on the framework of CTT serve similar purposes as that of IRT. It should be noted that, there exist specific situations in which only CTT based test equating methods are relevant. For instance, when two test forms are to be equated and it is only examinee's raw scores that are available to the researcher, the best equating method to apply are those based on classical test theory.

Item Response Theory (IRT)

Since the beginning of the 1980s, item response theory (IRT) has more or less replaced the role classical test theory used to play in the scientific field of measurement and evaluation (Crocker and Algina, 1986). Item response theory is a body of theory used in the field of psychometrics. Psychometrics is concerned with the theory and technique of educational and psychological measurement. In item response theory (IRT) mathematical models are applied to analyze data from questionnaire and tests as a basis for measuring things such as abilities and attitude studies in psychometrics.

IRT models are mathematical functions that specify the probability of discrete outcome, such as a correct response to an item, in terms of person and item parameters. Person parameters may, for example, represent the ability of a student or the length of a person's attitude. Item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptote). Items may be questions that have correct and incorrect responses or statement on questionnaires that allow respondents, to indicate level of agreement.

Among other things, as a body of theory, IRT provides a basis for evaluating how well assessment work and how well individual questions on assessment work. IRT is often referred to as latent trait theory, strong, true score theory or modern mental test theory. IRT models are used as a basis for statistical estimation of parameters that represent the "locations" of persons and items on a latent continuum or, more correctly, the magnitude of the latent trait attributable to the persons and items. For example, in attainment testing, estimates of the latent trait may be the magnitude of a person's ability

within a specific domain, such as reading comprehension. Once estimate of relevant parameters have been obtained, statistical test are usually conducted to gauge the extent to which the parameters predict item responses given the model used. Stated somewhat differently, such test are used to ascertain the degree to which the model and parameter estimates can account for the structure of and statistical pattern within the response data, either as a whole or by considering specific subsets of the data such as response vectors pertaining to individual items or persons. This approach permits the central hypothesis represented by a particular model to be subjected to empirical testing, as well as providing information about the psychometric properties of a given assessment, and therefore also the quality of estimates. Latent is used to emphasis that, discreet item response are taken to be observable manifestations of the trait or attribute, the existence of which is hypothesized and may be inferred from the manifest responses.

Certain assumptions underlie the use of IRT as a test theory and the first one is the unidimensionality assumption. This assumption postulates that only one ability is measured by the items that make up a test. What is required for the unidimensionality assumption to be met adequately is the presence of one dominant factor that influences test performance (Svend and Christensen, 2002). Most IRT models assume that it is only a single latent trait that underlies performance on an item, and that, responses to different items are independent given the latent trait.

The second assumption is that of local independence. This assumption states that, when the abilities influencing test performance are held constant and the examinee's

responses are equally held constant, then examinee's responses to any pair of items are statistically independent (Ponocny, 2002).

The relationship between examinee's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristics curve (ICC) (Hambleton *et al.*, 1991).

Advantages of IRT

Unlike CTT item statistics, which depend fundamentally on the subset of items and persons examined, IRT items and person parameters are invariant. This makes it possible to examine the contribution of items individually as they are added and removed from a test. IRT also allows researchers to calculate conditional standard errors of measurement based on a test information function, rather than assuming an average standard error across all trait level as in CTT. This allows researchers to select items that provide maximum measurement precision in a particular ability/trait range.

Again IRT allows researchers to conduct rigorous tests of measurement equivalence across experimental groups. This is particularly important in cross-cultural research where groups are expected to show mean differences on the attribute being measured. IRT methods can distinguish item bias from the differences on the attributes measured.

IRT also facilitates computer adaptive testing. Item can be selected that provide the most information for each examinee. This can dramatically reduce time and cost associated with test administration (Stephen, Sasha, David, Jyne and Patrick, 2001).

Comparison of Classical and Modern Test Theory

Classical test theory and modern test theory are largely concerned with the same problem but are different bodies of theory and therefore entail different methods. Although the two paradigms are generally consistent and complementary, there are a number of points of difference. These include:

- IRT makes stronger assumptions than CTT and in many cases provides correspondingly stronger findings; primarily, characterization of error. Of course, these results only hold when the assumptions of the IRT models are actually met.
- Although CTT results have allowed important practical results, the model-based nature of IRT afford many advantages over analogous CTT findings.
- CTT test scoring procedures have the advantages of being simple to complete (and to explain) whereas IRT scoring generally require relatively complex estimation procedures (note that in Rasch Model, the total score for a person is the sufficient statistic of the person parameter).
- IRT provide several improvement in scaling items and people. The specifics depend upon the IRT models, but most models scale the difficulty of items and the ability of people on the same matrix. Thus the difficulty of an item and the ability of a person can be meaningfully compared.
- Another improvement provided by IRT is that the parameters of IRT models are generally not sample or test-dependent whereas true-score is defined in CTT in the context of a specific test. Thus IRT provides significantly greater flexibility in situation where different samples or test forms are used.

Related Empirical Studies

Magno (2009) used the chemistry test data of junior secondary school students in Philipines to demonstrate the difference between CTT and IRT. Rasch model and Cronbach's alpha were used to analyse data and it was found among others that IRT estimates of item difficulty do not change across samples as compared with CTT which was inconsistent. Magno (2009) also found that the difficulty indices were more stable across forms of test in IRT than CTT approach. Silvestre-Tipay (2009) examined the behaviour of item and person statistics derived from two framework of a Biological Science test design for college fresh students. The findings of the study revealed that the degree of difference of item and person statistics across sample appears to be similar in CTT and IRT.

Edelen and Reeve (2007) applied IRT in modeling the responses of 6,504 adolescent respondents in the National Longitudinal Study of Adolescent Health public who completed the 19- item Feelings Scale for depression. The sample was split into a development and validation sample. Scale items were calibrated in the development sample with the Graded Response Model. The results obtained show that the 19 items varied in their discrimination (slope parameter range: .86 to 2.66), and item location parameters reflected a considerable range of depression (from -1.72, to 3.39). The authors concluded that when used appropriately, IRT can be a powerful tool for questionnaire development, evaluation, and refinement, resulting in precise, valid, and relatively brief instruments that minimize response burden.

In a study on the use of IRT to explore the psychometric properties of extended matching question examination (EMQ) in undergraduate medical education by Bhakta, Tennant, Horton, Lawson and Andrich (2005), it was reported that modern method (that is, IRT) provides a more

useful approach to the calibration and analysis of EMQ undergraduate medical assessment. The Bhakta et al (2005) study also revealed that IRT based metric calibration facilitates the establishment of item bank.

Tam, Griffith and Li (1997) conducted an investigation on the appropriateness of Item Response Theory (IRT) linking design using six (6) groups of students taking 6 forms of pilot test. A total of 8,357 students were used. They used a single set of anchor items with fixed common item parameter (FCIP) during the calibration process. The robustness of FCIP was examined under the situation of large standard error in the item difficulty and guessing parameters. Based on the study, item parameter estimates calibrated from IRT linking design are very consistent, except for guessing parameter under the characteristic curve method (CCM) of equating. It was concluded that IRT linking or equating procedures a very precise and stable parameter estimated.

Population Invariance/ DIF and Test Scores Equating

Given today's social and political climate, Petersen (2008) recommended that all testing programs with high-stakes outcomes should conduct population invariance equating studies for gender and major racial/ ethnic subgroups, especially because the results are likely to be comparable across these subgroups. Testing programs also need to conduct studies for major subgroups that could differ in ways related to the ability being measured and/or that comprise a varying proportion of the testing population at different administrations.

Determining population invariance can be done through differential item functioning (DIF). DIF is the statistical term that is used to simply describe the situation

in which persons from one group answered an item correctly more often than equally knowledgeable persons from another group. The introduction of the term *differential item functioning* allows one to distinguish item impact from item bias. Item impact described the situation in which DIF exists, because there were true differences between the groups in the underlying ability of interest being measured by the item.

Tim (2008) opined that equating functions are supposed to be population invariant, meaning that the choice of the subpopulation used to compute the equating function should not matter. The author further said that, the extent to which equating functions are population invariant is typically assessed in terms of practice difference criteria that do not account for equating functions sampling variability. In the same article, Tim pointed out that, the framework of kernel equating can be extended so that the standard errors of the root mean square difference (RMSD) and of the difference between two subpopulations' equated scores can be estimated using the kernel method of test equating for estimating the standard error of population invariance measures.

Yang (2004) examined whether the multiple-choice to composite linking functions of Advanced Placement Program (AP) exams remained invariant over subgroups by region. The study focused on two questions: (a) how invariant were cut scores across regions and (b) whether the small sample size for some regional groups presented particular problems for assessing linking invariance. In addition to using the subpopulation invariance indices to evaluate linking functions, Yang also evaluated the invariance of the composite score thresholds for determining final AP grades. Overall, linking across regions seemed to hold reasonably well. Males and females exhibit

differential mean score differences on the free-response and multiple-choice sections. Does this differential mean score differences affect the equatability of AP scores and the invariance of AP grade assignments across gender groups?

Based on Yang's unanswered question, von Davier and Wilson (2008) investigated population invariance for gender groups for the Advanced Placement Program Calculus AB exam. They used an internal anchor test data collection design to equate a multiple-choice (MC) test and a test composed of both MC and free-response questions. Item response theory, Tucker, and chained linear equating procedures were also used. Overall, the two administration groups did not differ much in ability, but the gender groups had large differences in ability. In general, they found that all equating methods produced acceptable and comparable results for both tests for equating based on men, women, or total administration groups.

Also, Liu and Holland (2008) used Law School Admission Test data from a single administration to investigate population invariance for highly reliable parallel tests, for less reliable parallel tests, and for nonparallel tests. The authors used subgroups based on gender, race/ethnicity, geographic location, application to law school status, and admission to law school status. As expected, Liu and Holland (2008) found that construct similarity between the two tests to be equated had much more effect on the population sensitivity of results than did differences in reliability.

Yang and Gao (2008) looked at population invariance for gender groups for forms of the College-Level Entrance Placement (CLEP) College Algebra exam. This was an equivalent-groups design. In general, the authors found that equating results based on

men, women, or total group were comparable overall and at the cut score. The article by Yi, Harris, and Gao (2008) also used an equivalent-groups design with a science achievement test. Unlike the other researchers, though, they looked at population invariance for subgroups that differed in ability, they created three sets of subgroups of dissimilar ability: (a) based on the average of four test scores that included the science test under study, (b) based on students' average grade point average (GPA) in all science courses taken, and (c) based on whether students had taken physics. Number-correct score difference for the composite subgroups were approximately 1 point, those for the GPA subgroups were approximately 4 points, and score differences for the physics subgroups were approximately 5 points. When the ability differences among the groups of interest were related to the construct being measured, they found the equating functions to be more population sensitive.

Dorans, Liu, and Hammond (2008) looked at the effect of anchor test, ability group differences, and equating method on population invariance for gender groups. Results of this study reconfirmed results of several earlier studies. The Tucker equating method did not work well when there were large ability differences between the groups. Also, use of an anchor composed of dissimilar content to the tests to be equated did not produce acceptable equating results.

Differential item functioning (DIF) effects of Biology examination items of WAEC and NECO for the years 2000 to 2002 were analyzed in terms of Gender and location by Obinne (2007). The result revealed that some items favoured girls while some others favoured boys. The same was observed in location; some items were easier for the

urban student than the rural students. The author concluded that this confirms the existence of DIF effects in the Biology test constructed by these two examination bodies in Nigeria.

All these articles were based on the premise that population invariance is a prerequisite for equating. In summary, these six studies found little sensitivity of equating results for subgroups formed on the basis of characteristics such as gender, race/ethnicity, and geographic location. They found that the use of anchor tests that were not miniatures of the tests to be equated did not yield sound equating. They also found that the Tucker equating (which is a type of linear equating) method did not work well when there were large differences in group ability. They found that construct similarity between the two tests to be equated had much more effect on the population sensitivity of results than did differences in test reliability. Finally, they found some sensitivity of equating results for subgroups that were selected on variables related to the construct being measured.

Equating Methods

Afressa and Keeves (1999) studied changes in students' mathematics achievement in five states of Australian lower secondary schools from 1964 to 1994. The purpose of the study was to investigate changes in Australian secondary school students' achievement between 1964, 1978 and 1994 using 13 years old students. The authors used first international mathematics studies (FIMS), second international mathematics studies (SIMS) and third international mathematics studies (TIMS) with samples of 5998, 3038 and 3786 respectively. The measurement procedure employed among others was Rasch model to scale students' responses and t-test statistic. According to Afressa and Keeves

(1999) the t-test statistic is appropriate because it takes into account (a) sampling error, (b) error in calibration and (c) equating error. Both QUEST and SPSS programs were used in analyzing the data collected. The result of the study revealed that, there was overall decline in the achievement of students in mathematics in the three years. It was concluded that a significant difference was found in one state. This study is very important to the current work because it provided a lead way to the use of t-test in analyzing data obtained from the field.

Fan (1998) examined IRT model fit of 60 multiple choice test items. The author used Texas Assessment of Academic skills (TAAS) test administered in October 1992. The sample used for the study was 6000 participants and model fit was examined through observing if each of the individual items fit or misfit a given IRT model. Likelihood ratio chi-square test was employed to detect item fit and BILOG software was used for data analysis. Fan's findings among others indicated that the data fitted the 2-and 3-PLM of IRT very well, whereas the fit of the data for IPLM was highly questionable.

Stage (2003) used IRT to determine how Swedish Scholastic Aptitude Test (SWESAT) quality can be improved upon. The SWESAT contained 122 multiple choice items in five subjects with 22 items measuring mathematical ability. The sample size used for the study was 2461, made up of 1349 female and 1112 males. The analysis conducted using BILOG chi-square test was to determine whether the items of SWESAT fitted the 3PLM of IRT. The findings of Stage showed that the 3PLM of IRT did not fit the SWESAT data. Stage (2003) asserted that the advantage of IRT models can only be maximized when the fit between the model and the test data of interest is satisfactory.

In working with two equating methods and determination of errors associated with them, Wang, Lee, Brennan, and Kolen (2006) used simulation to compare two test equating methods under the common-item nonequivalent groups design. These are frequency estimation method and the chained equipercentile method. Their study used data from four forms of a 60-item Mathematics test. Randomly equivalent groups of about 3000 examinees per form took the test. The item parameters were estimated using BILOG-MG assuming a three-parameter logistic (3PL) IRT model. Following a procedure used by Hanson and Beguin (2002), the item parameters for the four test forms were put on a common scale. The estimated item parameters were treated as "true" parameters that are used to simulate item responses. Parallel test forms with different test lengths are created from these four test forms. A common-item set is created by replacing some of the items from one test with items in the parallel form. Results for the aggregate equating error showed that the frequency estimation method tends to have larger bias than the chained equipercentile method when there are group differences. The frequency estimation method has smaller SEE's than the chained equipercentile method.

The effect of content mix and equating method on the accuracy of test equating using anchor-item design was studied by Yang (1997). Using anchor-item design of test equating, the effects of three equating methods (Tucker, two-and three-parameter logistic item response theory methods) and the content representativeness of anchor items on the accuracy of equating were examined. Data analyzed were test result from two (2) forms of professional competency examination with 197 and 203 items respectively. There were 145 anchor items embedded in both forms and the 2 examinees groups used in the study

were not randomly formed. From the 2 test forms, 4 pairs of shortened test forms were created to differ in the content representativeness of their anchor items. The total raw scores on the original items were regarded as "pseudo true score", which was used as a criterion for evaluating equating accuracy. Overall, the three equating methods appeared to yield moderately accurate equating result on every test, and the outcomes of the IRT methods seems to be more accurate from the outcomes of Tucker method.

Morrison and Fitzpatrick (1992) conducted an investigation on Direct and indirect equating. They compared four equating methods namely: (1) concurrent calibration; (2) equating constant procedure; (3) major axis procedure; and (4) fixed based procedure. An internal anchor test design was employed with five different test forms, each consisting of 30 items, 10 in common with the base test and 5 to 10 in common with one or more other forms. Simulated data were generated using Rasch model. Using one form as the base test, each of the other was equated directly to the base test and also to other forms. An attempt was made to determine which of the item response (IRT) equating method results in the least amount of equating error or scale drift when equating scores across one or more test forms. It was found that concurrent calibration resulted in the least amount of equating error. When concurrent calibration is not feasible, results indicate that major axis equating results in the least amount of equating error when equating across one or more forms.

Combinations of five methods of equating test forms and two methods of selecting samples of students for equating were compared for accuracy by Livingston (1989). The two sampling methods were representative sampling from the population and matching

samples on the anchor test score. The equating methods were: The Tucker method (2) The Levine method (2) the chained equipercentile method (4) the frequency estimation; and (5) the three parameter logistic model of IRT. The tests were the verbal and mathematics sections of the scholastic aptitude test. The criterion for accuracy was measures of agreement with an equivalent groups equating based on 115,000 students taking each form. Some inaccuracy in the equating was observed and attributed to overall bias. The results of all equating methods in the matched samples were similar to those of the Tucker and frequency estimation methods in the representative samples. These equating methods made too small an adjustment for the difference in the difficulty of the test forms. In the representative samples, the chained equipercentile method showed a much smaller bias. The IRT and Levine methods tended to agree with each other and were consistent in the direction of their bias.

Wang, Won-chan, Brennan & Kolen (2006) used simulation to compare two test equating methods under the common-item non-equivalent groups design; the frequency estimation method and the chained equipercentile method. An IRT model was used to define the true equating criterion, simulated group differences and generated response data. Three linear equating methods were also included for reference. The result showed that when there is substantial group difference, the frequency estimation method has larger bias than the chained equipercentile method. The frequency estimation method, however, has a smaller standard error of equating than the chained equipercentile method.

Ong and Sireci (2008) conducted a study on using bilingual students to link and evaluate different language versions of an examination. They used two language versions

of mathematics tests (English and Malay) which were administered on 505 students who are proficiency in both English and Malay from the state of Perak Malaysia for each intact class were use. The test consisted of 40 dichotomously scored multiple-choice items measuring topics such as numbers, algebra, measurement, geometry and statistics. Two separate booklets were prepared, one using only the English version of the items and the other using only the Malay version. The students were divided into two groups. The first group consisted of 255 students who took the English version first and the second group consisted of 250 who took the Malay version first. Three weeks later, the versions were reversed and administered on each group. Equating analyses were conducted using mean linear and equipercentile methods of CTT. They also conducted separate common-item equivalent group equating using BILO-MG and concurrent calibration, where common (Non DIF) and Unique (DIF) items were analyzed simultaneously using a one-parameter logistic model of IRT. Ong and Sireci (2008) among others concluded that both linear and unsmoothed equipercentile methods gave similar results. The result also indicated that bilingual examinees can be useful for evaluating different language versions of test and adjusting for differences in difficulty across test forms due to translation.

Wen-ling and Rui (2008) investigated whether the functions linking number-correct scores to the college-level examination program (CLEP) scaled score remain invariant over gender groups, using test data on the 16 testlet-based forms of the CLEP college algebra examination. In this study, linking of various test forms to a common reference form is based on the Rasch model. Overall, linking based on gender groups are

very similar to linking for the total group. At all levels, differences between subgroups and total group linking are smaller than the difference that will affect the pass / fail decision for CLEP candidates.

Tim (2008) carried out an investigation of population invariance for the equivalent group design. In the study, the accuracies of the derived standard errors were evaluated with respect to empirical standard errors. The evaluation showed that the accuracy of the standard error estimates for the equated scores difference is better than for the RMSD, and that accuracy for both standard error estimates is best when sample sizes are large.

Mikisley and Kingston (1987) conducted a study to investigate the feasibility of using IRT equating for the graduate Records Examinations (GRE) subject test in mathematics. Two forms of the test were equated using the three parameter logistic (3 pl) model and the results were compared to the result of the Tucker and equipercentile equating. It was found among others that, both 3pl IRT and equipercentile equating procedures yielded different results than the Tucker method. The results for the IRT procedure and the equipercentile procedure were quite similar to that of Rapp and Allalouf (2003).

Jodoin, Keller and Swaminathan (2003) conducted a study on three common items response theory equating approaches to capturing academic growth. Using data from a statewide testing program and (1) Linear transformation of separate calibrations, (2) Fixed common item parameter calibration and (3) Concurrent calibration, found that there exist differences in mean growth depending on ability estimate and equating procedure

Comparing traditional observed score methods and IRT methods are not a new concept in equating research. For instance Petersen, Cook, and Stocking (1983) used Scholastic Aptitude Test (SAT) data to investigate the differences between conventional observed score methods and IRT equating methods in the NEAT design. Comparing unsmoothed equipercentile, Tucker, Levine equally reliable and unequally reliable, concurrent calibration, fixed b's, and characteristic curve transformation method, the authors found that linear equating methods tend to work better when the tests are reasonably parallel and of equal lengths. When the items and lengths vary, IRT equating using the 3-PL model is more stable, with concurrent calibration being the most stable. The authors found that conventional and IRT methods are equally sufficient, with differing results between the two tests, Verbal and Mathematics. Based on these authors' opinions, the researcher developed two tests that are parallel.

Similarly, Han, Kolen, and Pohlmann (1997) explored the differences in IRT true-score equating, observed-score equating and traditional (unsmoothed) equipercentile equating methods in the NEAT design. Using multiple forms from the Mathematics and English portions of the ACT, the authors compared the results of the three methods to each other and investigated the relationship between the discrepancies in equating results and the difference in difficulty of the two equated test forms. They found that there is no significant difference in the equating stability of the two IRT methods, but that both methods are more stable than the traditional equipercentile equating. Han and colleagues conclude that there appears to be a positive relationship between the discrepancies in equating results and the difference in difficulty among the two test forms, and call for

farther investigation. Although Han and colleagues use two subjects in their study, for this work, the researcher used one subject only, mathematics.

Kim and Cohen (1998) explored three IRT methods of equation and linking in the NEAT design: concurrent calibration based on a posteriori estimation, characteristic curve transformation method, and concurrent calibration with marginal maximum likelihood estimation. The concurrent methods were calculated with IRT calibration software BILOG and MULTILOG. The authors found that when the anchor test length is short, the characteristic curve method worked better, delivering a smaller root mean square difference (RMSD) than the other methods. However, when the anchor test length was longer (i.e., more than 10 items), the three methods delivered similar results. The IRT model used was a 3-PL and data were simulated. This study helped the researcher to compare equating method based on classical test theory (linear) and Item Response Theory (separate and concurrent calibration methods).

Hanson and Beguin (1999) used simulation to investigate the performance of separate versus concurrent estimation in putting item parameter estimates for two forms of a test administered in a common item equating design on the same scale. This study uses two forms of a mathematics test with 60 dichotomous items. The mathematics tests were denoted forms A and Z. Randomly equivalent groups of 2696 and 2670 examinees took forms A and Z, respectively. The computer program BILOG was used with the data generated from the tests to estimate the item parameters for all items assuming a three parameter logistic IRT model. These estimated item parameters were treated as population item parameters for simulating data. The authors investigated five factors and

one of the factors was concurrent versus separate estimation using four item parameter scaling methods. The result of the study among others showed that, the differences among the item parameter scaling methods used in separate estimation were much larger than the differences between concurrent estimation and the better performing scaling methods in separate estimation. It was also found that, concurrent estimation resulted in less error than separate estimation. The authors concluded by saying that, "although concurrent estimation resulted in less error than separate estimation, more times than not, it is concluded that the results of this study, and other research performed on this topic, is not sufficient to recommend concurrent estimation should always be preferred to separate estimation".

Other research that compared separate and concurrent calibration have concluded that concurrent estimation performed somewhat better than separate estimation (Petersen, Cook, and Stocking, 1983; Wingersky, Cook, and Eignor, 1987), while Kim and Cohen (1998), using computer programs more commonly used today, concluded that the performance of separate estimation was equal to or better than concurrent estimation. From the above review, it appears that, there have been no consistent in the result of studied that that compared separate versus concurrent estimation hence the need for this study.

Hanson and Béguin (2002) published a report comparing various IRT equating methods under the NEAT design using simulated data. Specifically, they investigated the characteristic curve methods Stocking and Lord (1983) and Haebara (1980), the moment methods mean/mean and mean/sigma, and concurrent calibration in both BILOG-MG

and MULTILOG. The authors found that the moment methods (mean/mean and mean/sigma) produced much larger errors than the other methods, and that using BILOG-MG for concurrent calibration produced less error than for separate estimation, but that MULTILOG produced the opposite results.

Hanson and Beguin were not alone in investigating IRT equating methods. Jodoin, Keller and Swaminathan (2003) explored the differences between concurrent calibration, fixed common item parameter estimation (FCIP), and Stocking and Lord's transformation method. Unlike Hanson and Beguin, they used examinee data, and were one of the few that investigated FCIP against the other IRT methods. The authors found that although there was a lot of agreement between the proficiency classifications of the examinees using the three methods, there was sufficient disagreement to warrant further investigation using simulated data where truth is known.

In a study on unified approach to linear equating for the non-equivalence groups design, Von Davies and Kong (2005) used data which were collected following a NEAT design with an external anchor. The tests used consisted of two parallel tests, with 78 items tests and a 35 items external anchor in each test. The tests were given to two samples from a national population of examinees. The two sample sizes used for the study were 10,634 and 11,321. The sample correlations when the anchor test was correlated with the first and second samples were 0.88 and 0.87 respectively. The mean difference between the two sample groups was found to be 2.66. The authors concluded that, a mean difference of this magnitude indicates a fairly large difference between the two population used in their study.

In recent years, more studies have been conducted investigating the benefits and limitations of kernel equating versus the more traditional methods of observed score test equating (Kelly 2007). For instance Von Davier, Holland, Livingston, Casablanca- Grant and Martin (2005) used real data in an equivalent groups (EG) design to create pseudo-tests in the NEAT design to compare kernel equating results to those of other more traditional observed score equating techniques. Equipercentile equating and linear equating were investigated under the EG design, and Tucker, Levine observed-score, frequency estimation, and chained equipercentile were studied in the NEAT design, with linear and equipercentile equating both conducted in kernel equating. To do this, the authors created a criterion equating function with the EG designs with which the equating methods in the NEAT design could compare. Two smaller tests of 44 items each were created from a larger test, and an external parallel anchor of 24 items (54.5%) was used. Results indicated that kernel equating approximates an equating that is very close to the more traditional techniques, but that is actually closer to the criterion, and thus more accurate. The effects of differences in test form difficulty, length of anchor test, and sample sizes were not investigated.

In a related study, von Davier and Ricker (2006) examined the role that external anchor test length plays in equating in the NEAT design. Creating a criterion equating using the classical equipercentile method in the EG design, the authors compared the results of kernel equating with large bandwidths (linear equating), optimal bandwidths (equipercentile equating), and the traditional methods equipercentile and chained equipercentile with external anchor lengths of 24 (54.5%), 20 (45.4%), and 16 (36.4%)

items. They established guidelines for comparing results, using a score difference that matters (SDTM), which is any score difference 0.5 or larger. The authors found that kernel equating with optimal bandwidths results are very close to the other equipercentile observed score methods, but that kernel equating with large bandwidths did not closely approximate the other traditional linear methods, especially at the lower end of the score scale, and this method did not come as close to the criterion equating as well as the equipercentile method did. They concluded that the choice of equating function could determine the amount of error in the test scores, especially when the anchor length was shorter. The use of Score Difference That Matter (SDTM) as identified in Von Davier and Ricker work will help the researcher to know the score difference among equating methods.

The relationship that sample size plays in kernel equating results has been investigated to some extent. Grant, Zhang, Damiano and Lonstein (2006) studied small-sample equating with the NEAT design in kernel equating, investigating the effects on the standard error of equating. Previous studies had been conducted with large sample sizes only (over 2000 examinees per form), whereas Grant and colleague studied sought to compare the performance of kernel equating when the sample sizes are small: 1000, 500, 250, 125, and 75. With smaller sample sizes, there are more breaks in the score distributions, and results indicated that the equating accuracy increased with the increase of sample sizes. Their result also indicated that, increasing a smaller sample size improved the equating results much greater than increasing a larger sample size.

Continuous Assessment Practice in Nigerian Schools

Suffice to say that most works reviewed so far merely lay emphasis on Test Equating practice. However, in this section deliberate attempt was made to review some work on continuous assessment only. Thus, Adebowale and Alao (2008) examined the methods adopted by teachers in the implementation of the provisions of a continuous assessment policy in Ondo State in Nigeria. Data were collected from One hundred (100) primary school teachers who were drawn by simple random sampling from all the schools in two selected Local Government Education Authorities (LGEAs) in Ondo state. The Education Authorities were Akoko North West and Akoko South West of Ondo state in Nigeria. An instrument titled "Questionnaire on the handling of continuous assessment by primary school teachers" consisting of 15 items, were developed and used by the researchers. The reliability of their instrument was 0.81 and the data were analyzed using simple percentages, t-test, and ANOVA. Results indicated a non-uniform strategy of implementing continuous assessment policy provisions and are found to be independent of factors like gender, duty posts, teaching experience, and qualifications, as no significant difference were found in the score of respondents on all of these factors.

Obioma (no date) conducted a large scale-survey to examined the status, gaps and challenges of continuous assessment (CA) practices of primary and junior secondary school teachers in Nigeria. The survey sought information from the school teachers on their understanding of the CA as well as the appropriate application of the CA instruments, whether there are uniform CA guidelines across the country and how school teachers engage in CA practices. A random sample of 3,325 teachers (2,185 primary

school teachers and 1,140 junior secondary schools teachers) across the six (6) geopolitical zones of the country. The reliability of the instrument used for the study was 0.79 and the data were analyzed using simple frequency counts, mean and standard deviations.

Results from the Obioma's study showed that in general, school teachers demonstrated poor knowledge of the elementary concept of CA. Many teachers misapplied the CA instruments leading to more of continuous testing of learners instead of continuous assessment. CA guidelines not only varied across states and schools but were also different from the guidelines stipulated in the extant national CA handbook. School teachers are to be mandatorily and formally trained in CA principles and practice both at pre-service and in-service levels.

Ajuonuma (2007) designed a study to survey the implementation of continuous assessment (CA) in Nigerian universities. Two research questions and one hypothesis were formulated to guide the study. Eight universities in south-east zone of Nigeria were used. The sample for the study consisted of 1,340 respondents (940 males and 400 females) who were drawn through stratified random sampling technique. A 24 item self-report instrument called "Continuous Assessment Implementation (CAI) questionnaire" was used for the study and its reliability was 0.81. The data generated, were analyzed using mean and t-test. The result revealed that, out of the twenty-four continuous assessment implementation items, Nigerian University lecturers implement only eleven. Thirteen are not implemented, some of which include; setting questions using table of specification, assessment of students in affective and psychomotor domains, developing

and using valid instruments for assessment in the three domains. In addition, sex does not have any influence on the implementation of continuous assessment in Nigerian universities. Provision of adequate fund to schools and exposure of lecturers to conferences, seminars and workshops were some of the solutions proffered to remedy the ugly situation. Extensive and diligent search during this review of works done on equating methods and continuous assessment practices revealed that comparability of students' achievement scores is achievable. Therefore, it is pertinent to conduct a study on relative efficiency of test scores equating methods in the comparison of students' continuous assessment measures.

Summary

The literature review covered conceptual framework, theoretical framework and related empirical studies on test equating and the practice of continuous assessment in Nigerian schools. Within the conceptual framework, test equating was seen as a measurement process that involves test development, administration, analysis, scoring, reporting, and evaluation (Hattie, Jaeger & Bond 1999). Test score equating is a statistical procedure that adjusts test scores on different forms of the same examination so that scores can be interpreted interchangeably. Equating methods are empirical procedures for determining a transformation to be applied to the scores on one of two forms of a test. Its purpose is ideally to transform the scores in such a way that it makes no difference to the examinee which form of the test he or she takes. This idea can be reached only if the two forms of the test measure exactly the same latent trait (ability or skill) and yield scores that are equally reliable.

Each equating method uses a specific model which is based on classical test theory (CTT) or item response theory (IRT). The classical test equating or traditional test equating methods include mean equating, linear equating, and equipercentile equating while item response methods include one parameter logistic (Rasch) model equating, two parameter logistic model equating and three parameter logistic model equating. Specifically separate and concurrent calibration, fixed based procedure, equating constant procedure, and major axis procedure are the equating models used under IRT.

When a model, such as any IRT model, provides the conceptual measurement framework for test equating process, the results depend on how well the data fit that model. If the IRT model fits the data perfectly, then parameters will be invariant across administrations, except for sampling fluctuations that introduce random error in the responses of examinees. In that case, the changes in the behavior of item parameter estimates would follow a systematic pattern depending on the changes in the size and proficiency of the different examinee groups (example, state, gender or ability group). Therefore, the parameter invariance is the property of IRT which ensure that the item parameter estimates remain unchanged across various groups of examinees or the items in the instrument used for assessment do not show differential item functioning. For the applicability and usefulness of IRT, Hambleton, Swaminathan & Rogers (1991) assert that the ability estimates should also remain invariant across groups of items.

It is obvious from review of related empirical studies that, a lot of research has been conducted on test equating, and the comparison of test score equating methods is not a new concept to measurement and equating research. However, to date, no studies

within the reach of the researcher have compared students' continuous assessment scores using test scores equating in Nigeria. WAEC/NECO seems not to use test score equating methods in the comparison/standardization of raw score. Consequently, they use T-score transformation method in standardizing the CAS from various schools across the country. This T-score method is not quite appropriate in this particular situation because it does not ensure that scores emanating from CA reflects the students' academic achievement. There is therefore the need for WAEC/NECO to device a means to bridge the observed gap arising from the use of the T-score if the real essence of CA is to be realized. This thesis attempts to remedy the lack of information concerning the use of test scores equating in the comparison of students' scores. Most of the studies reviewed used simulated data while in this research, real data from the field (schools) was used. This will create situations in which truth is known about students' continuous assessment scores in Nigeria. Therefore, the study on relative efficiency of tests scores equating method in the comparison of students' continuous assessment measures is necessary.

Conclusively, a dangerous precedence and loop hole in the current National Policy on Education which seems not to have a firm grip on the principles and practice of CA across the country has created the problem of comparability of CA scores. This in turn has made examination bodies (WAEC, NECO, NABTEB, etc) to be at a cross road as to what the real CA score of Nigerian students should be. However, the concept of using Test score equating which is for now a uniquely "strange" phenomenon in our measurement and evaluation system might be a solution to comparability problem of CA.

CHAPTER THREE

METHODS

This chapter is concerned with the description of the procedure used for the study with regards to the research design, Area of the study, population, sample and sampling techniques, instrument for data collection, validation and reliability of instrument, and method of data analysis.

Design of the study

The equating design used in this study is non-equivalent anchor test (NEAT) group design. The design is also called common-item non-equivalent group design. In this design, a set of common or anchor items are included in both tests or forms, so that the difference between the two can be adjusted based on common item statistics (Zhu, 1998).

The design is also useful in measuring growth when the two groups are known to be non-equivalent, and is necessary when it is impossible to administer more than one test due to test security or other practical concerns like test adaptation. It is often used when developing an item bank in which test items are cumulated into a common scale. However, the use of an anchor item design requires stronger statistical assumptions of effect of groups and test difference; therefore, there should be enough items in the representative content to be measured.

In using Non-Equivalent Anchor group design, the researcher assumed that; (a) the continuous assessment tests scores from schools in State A are scores from test A, while the continuous assessment tests scores from schools in State B are scores from test B. (b)

each state is a group, (c) tests A and B are different tests from the same mathematics curriculum. Students in state A completed test A^1 whereas students in state B completed test B^1 , which is called Mathematics Achievement Test (MAT) developed by the researcher. MAT scores were used in equating students' continuous assessment scores in mathematics from both states. The researcher equated the CAS from schools in State A (test A score) with test A^1 scores to produce V_1 (that is, $A + A^1 = V_1$), and also equated the MCA scores from State B (test B scores) with test B^1 scores to produce V_2 (that is $B + B^1 = V_2$). The equated results were used in comparing students' continuous assessment scores.

The non-equivalent anchor test (NEAT) groups design was therefore very appropriate in the present study because, it helped the researcher to cumulate the items used for data collection into a common scale. The design also helped the researcher to use more than one test in this study and equally permitted the statistical invariance of each group and test difference to be ascertained.

Area of the Study

The area of the study was Cross River and Rivers States in the South-south geopolitical region of Nigeria. Cross River State is made of three senatorial zones, namely; Northern, Central and Southern senatorial zones; it has eighteen Local Government Areas. Rivers state is also made up of three senatorial zones namely; Rivers West, Rivers South-East and Rivers East, and has twenty three Local Government Areas. Cross River State was chosen as an area of this study because the state ministry of Education organizes general mock examination for all SSII students in Government owned

secondary school. The result of the mock examination is not only used in promoting the students to SSIII but also used to sponsor WAEC registration bill. Thus, Government pays the WAEC registration bill for students who pass the mock examination while parents pay for their wards who failed the mock exam. Parents also pay for NECO registration bill of their wards. All teachers in Cross Rivers do not go for first and second term holidays. They stay behind to prepare JSSIII, SSII and SSIII students for junior secondary III exam, state mock examination for SSII and external examinations for SSIII.

Rivers State was chosen for this study because the students are not given state organized holiday lesson. The state mock examination is decentralized in such a way that each school sets her own mock exam. The state government pays for NECO registration bill while parents pay WAEC registration bill. Another reason for the choice of the two states is that, in Cross River, 30% is assigned to CA and 70% to examination while in Rivers State, 60% is assigned to CA and 40% to examination. These characteristics enabled the researcher to categorize the population into two groups as demanded by the design of the study.

Population of the Study

The population of the study consisted of all senior secondary school III students in Government owned schools in Cross River and Rivers states. There are two hundred and thirty two (232) Government owned secondary schools in Cross River state (Cross River State Secondary Education Board Calabar, 2010), and two hundred and twenty four (224) Government owned secondary school in Rivers State (Rivers State Ministry of Education Port Harcourt, 2010). The number of SSIII students in Cross River and Rivers state are;

15,139 and 13,911 respectively and the population for this study was therefore 29,050 SSIII students in both states.

The study concentrated on senior secondary III students because they have taken mathematics for at least five years and has a cumulative continuous assessment scores in mathematics. The students have been taught almost all aspect of the WAEC/NECO syllabus in mathematics. They were equally preparing for senior secondary school certificate examination, and must have been exposed to the content covered on the test.

Sample and Sampling Techniques

The sample consisted of two thousand nine hundred and five (2905) students, such that 1,514 students were from Cross River and 1,391 were from Rivers State. The sample size was determined using 10% of the population from Cross River State and Rivers state respectively. The choice of 10% was based on Nwana (1981) opinion that, there is no fixed number and fixed percentage that is ideal in drawing sample size, rather, it is the circumstance of the study situation that determines what the percentage of the population should be in a study. Thus the use of 10% in drawing the sample size for this study was to get appropriate number of examinees that is adequate for performing any form of test scores equating. A total of 45 schools, (23 from Cross River and 22 from Rivers state) were used for this study. The school sample size was equally determined using 10% of the total number of schools in Cross River State and River State. Multi-stage sampling procedure was used for this study. Ali (2006) opined that multi-stage sampling required several stages of sampling the elements of the population. It may involve the use of two or more sampling techniques. The sampling stages involved in this study are; sampling of

school and sampling of research subjects (students). The sampling techniques that were used are simple random sampling and proportionate stratified random sampling.

In the first stage sampling, 23 secondary schools were drawn from Cross River state and 22 from Rivers State, through simple random sampling technique. The simple random sampling technique used here was slips of papers method. Under this method, the name of each school was written on slips of papers. The slips were folded and put into a bag. After thorough reshuffling, the researcher dips his hand into the bag and picks one slip. The slip was unfolded and the name of the school it contains recorded. The procedures continued until all the 23 schools and 22 schools were drawn from Cross River and Rivers States respectively. In the second stage sampling, 1,514 students from Cross River state and 1,391 students from River state were drawn, through proportionate stratified random sampling. The proportionate stratified random sampling according to Nworgu (2006) is used where the population is first categorized into groups that are distinctively different from each other on relevant variables and in which the elements are drawn at random from within stratum in such a way that the relative proportions of the strata in the resultant sample are the same as the parent population. This sampling technique is appropriate for this study because it helps to ensure that majority of the features of the parent population are included in the sample. The choice of this sampling technique is also justified on the ground that random equating errors that are caused by sampling errors is reduced because of sample typicality to the population. Finally, the number of students to come from the sampled schools were determined by proportionately sharing the total sample to the number of schools for each state. Students

were then randomly drawn from their schools based on their proportion to the total sample.

Figure 12: Summary of Sample Size of students used for data collection

STATE	SAMPLE SIZE OF STUDENTS		TOTAL
	MALE	FEMALE	
CROSS RIVER	693	821	1514
RIVERS	661	730	1391
TOTAL	1354	1551	2905

Instrument for Data Collection

The instruments for data collection are two parallel tests. They are multiple choice tests consisting of 40 items each and called Mathematics Achievement Test (MAT) developed by the researcher. Twenty (20) items are common to both tests while the other 20 items are unique to each test. The 40 items in the test were considered appropriate for this study because it is large enough to enhance high precision in the determination of the psychometric properties of the items as well as producing good ability estimates of the students. Parallel test was used because, single form of test allow students to copy answers from neighbouring colleagues if the test is not administered the same day. The scores of such dishonest students can be unreliably high and honest students may be disadvantaged by the act of such dishonest students. Given the above situation and the fact that the administration of MAT in the states was designed to take place on different days, the use of a single test was inappropriate. The use of parallel test form may

substantially reduce the dishonest and cheating problem anticipated in a single test. Honest students may not be disadvantaged by the act of dishonest students. Since test equating was conducted, some equity issues were not a problem to border about.

The parallel tests used in this study were designed in such a way that, the same latent trait, ability or skills are measured across the schools and states. The items in the instrument were drawn from the current West African Examinations Council (WAEC) Mathematics Syllabus. Each of the items has four (4) response options or alternatives. Only one of the alternatives is the correct option (key) for an item. The 40 items in each test form were spread over six major topics levels as follows: Number and numeration 20%; Algebraic process 18%; Mensuration 14%; Geometry 16%; Trigonometry 14%; Statistics and probability 18%. The weights along the topic dimension were based on how voluminous the content scopes are. Thus, a topic with larger content scope is assigned more weight than the one with lower volume of content scope. The cognitive levels were also weighed as follows: knowledge 8%, Comprehension 12%, Application 23%, Analysis 24%, Synthesis 18% and Evaluation 15%. The weights along the cognitive level were assigned based on the relative importance or emphasis attached to the different levels. In mathematics, emphasis is on whether the students can apply what had been learned and subsequently use such ideas in analyzing mathematical task. This is why higher weights were assigned to application and analysis. The number of questions for each topic in conjunction with each of the cognitive level were calculated and the result shown in appendix G. Items 1, 9 and 32 measure knowledge, items 2, 10, 17, 22, 29, and 35 measure comprehension, items 3, 4, 11, 12, 18, 23, 28, 34 and 36 measure application,

items 5, 6, 13, 14, 19, 24, 25, 30, 37 and 38 measure analysis, items 7, 15, 20, 26, 31 and 39 measure synthesis while items 8, 16, 21, 27, 33 and 40 measure evaluation in both tests. (See Appendix C and D for the instruments).

Validation of the Instrument

In order to ascertain the validity of the instruments, the instruments were subjected to both content and face validity. In the content validity; the researcher carefully prepared a test blue-print or table of specification where both the cognitive levels as well as the subject content were aligned on a two grid table. The test blueprint or table of specification shows the number of items per topic and cognitive level (see appendix G). This was enhanced by the assignment of percentages to both the cognitive and content dimensions. In the second stage of ensuring that the instrument is valid, the researcher consulted two experts in mathematics education/measurement and evaluation, one expert in mathematics education, and two experts in measurement and evaluation from University of Nigeria, Nsukka. The mathematics experts were specifically asked to undertake a careful systematic comparison on test content with mathematics content and to check if the test items adequately sampled the subject content. The experts in mathematics were also asked to solve the items of the test and indicate the correct answer among the options.

The experts in measurement and evaluation were required to assess the brevity and unambiguity of the statements used in phrasing each item. Each of the experts was asked to judge the adequacy of the items of the instrument and comment on the test blue print/table of specification. The experts were equally saddled with the responsibility of

determining how appropriate the items in the instrument are to the class level they were designed for (see appendix A for letter of validation of instruments).

The experts observed among others that, some of the keys to the items were not correct, some items were not addressing the cognitive level they intend to measure as shown by the table of specification/test blueprint. The validators also observed that, the issue of possibility and feasibility may not be determine quantitatively, as such research questions 3, 4 and 7 should be reconsidered. It was also suggested that the researcher should recast hypothesis 2, and rename the instrument. The issues raised by the experts were properly addressed by the researcher.

Reliability of the Instrument

To establish the reliability of the instrument, a sample of examinees was drawn from equivalent study sample. The instruments were trial tested on them. The essence of the trial testing was to find out how the respondents would react to the instrument and also to establish prior psychometric properties of the test items (particularly item facility index and item discrimination index). The trial testing was also aimed at helping the researcher determined the reliability of the test. The reliability of the instrument was computed using Kuder-Richardson 20 (K ï R 20).

Kuder-Richardson 20 attempts to show whether each of the test item measures the same characteristics as every other items (that is homogeneity of test items). Thus the reliability coefficient of paper A¹ was found as 0.83 and that of paper B¹ was found as 0.89. This indicates that the instruments are highly reliable, (see Appendix N and O for computation of the reliability of Test A¹ and B¹). The psychometric properties of the

items in the instruments were first determined using CTT approach and result shown in appendix H and I

Procedure for Data Collection

The instruments were administered to the stipulated respondents in their various schools/states by the researcher and some research assistants who were postgraduate students of the University of Nigeria, Nsukka. Some mathematics teachers in the sampled schools also helped in the supervision of the examinees. Prior to the administration of the instrument, the researcher moved round the sampled schools to establish rapport with the mathematics teachers and examinees. The examinees were intimated on the modalities and purpose of the test. They were employed to take the test as serious as possible.

After administering and retrieving the instrument from the examinees, the researcher scored the items according to a prepared scoring key. The correct option was scored 1 while the incorrect option was scored 0. The scores per examinee were recorded and possible raw score ranges from 0 to 40 in increment of 1. The instrument also contained a space where the mathematics continuous assessment scores of each student was recorded by their mathematics teacher.

Method of Data Analysis

The data collected using the instruments were analyzed in relation to the research questions and the hypotheses formulated to guide the study. The data analyses were carried out using BILOG-MG 3 and statistical packages for social sciences (SPSS). Research question 1 was descriptively answered using item parameters of test forms A¹ and B¹. Research question 2, 3 and 4 were answered using item response theory

likelihood ratio differential item functioning (IRTLRDIF). Research question 5 was answered using IRT Chi-square goodness of fit. Research question 6 was answered using item characteristic curve for test form A¹ and B¹. Mean and standard deviations as well as scores difference that matter (SDTM) were used to answer research question 7, 8 and 9. Root mean square error difference was used the answer research question 10 while correlation coefficient was used to answer research question 11. The hypotheses formulated were tested at 0.05 level of significance. Specifically, independent t-test, chi-square goodness-of-fit and Pearson product moment correlation coefficient were used in testing the hypotheses.

CHAPTER FOUR

RESULTS

This chapter deals with the presentation of results according to research questions and research hypotheses.

Research Question 1

What are the item parameter estimates of Mathematics Achievement Test (MAT) used for Test Score Equating?

This research question was answered using the item parameters indices obtained from the separate calibration of examinees responses to two parallel mathematics achievement test (MAT). The result of the item parameter indices from 1514 and 1391 students from state A and B respectively are shown in the table 1 below

Table 1: Item parameter indices of two parallel tests

STATE A				STATE B			
Item	Parameters			Item	Parameters		
	b	a	C		b	a	c
1.	-0.16	0.50	0.40	1.	0.22	0.75	0.38
2.	1.44	0.43	0.28	2.	0.61	0.88	0.19
3.	0.56	0.61	0.21	3.	0.66	0.95	0.19
4.	1.00	0.60	0.24	4.	0.63	1.34	0.18
5.	1.10	0.57	0.21	5.	0.87	1.02	0.17
6.	1.44	0.59	0.24	6.	1.13	0.72	0.15
7.	1.47	0.63	0.32	7.	1.35	0.70	0.26
8.	0.88	0.60	0.15	8.	1.01	0.91	0.14
9.	0.51	0.56	0.22	9.	1.11	0.85	0.16
10.	1.25	0.58	0.26	10.	1.08	0.74	0.24
11.	1.23	0.61	0.25	11.	1.37	0.93	0.24
12.	1.23	0.75	0.18	12.	1.27	0.81	0.14
13.	1.29	0.75	0.21	13.	1.45	0.82	0.16
14.	0.98	0.65	0.16	14.	1.09	0.90	0.13
15.	1.37	0.80	0.26	15.	1.42	0.06	0.22
16.	1.08	0.62	0.16	16.	1.11	0.86	0.14
17.	0.12	0.70	0.15	17.	1.26	0.88	0.13
18.	1.16	0.63	0.28	18.	1.37	0.81	0.28
19.	1.77	0.64	0.14	19.	1.01	0.78	0.15
20.	1.04	0.59	0.10	20.	1.25	0.89	0.12
21.	1.02	0.67	0.09	21.	1.33	1.06	0.11
22.	1.03	0.92	0.16	22.	1.52	1.54	0.18
23.	1.43	0.68	0.23	23.	2.00	0.83	0.30
24.	1.33	0.70	0.15	24.	1.52	1.57	0.21
25.	1.31	0.79	0.15	25.	1.94	1.15	0.21
26.	1.23	1.02	0.21	26.	1.71	2.38	0.29
27.	1.34	0.87	0.26	27.	1.64	3.30	0.36
28.	1.28	0.87	0.09	28.	1.50	1.55	0.13
29.	1.17	0.90	0.12	29.	1.57	2.38	0.20
30.	1.25	0.82	0.11	30.	1.67	1.71	0.18
31.	1.28	1.12	0.15	31.	1.73	3.34	0.21
32.	1.38	0.78	0.16	32.	1.56	2.67	0.22
33.	1.26	0.89	0.08	33.	1.73	2.58	0.14
34.	1.27	0.70	0.16	34.	1.62	2.57	0.24
35.	1.14	1.06	0.08	35.	1.67	3.44	0.14
36.	1.27	0.81	0.12	36.	1.70	2.90	0.23
37.	1.10	1.08	0.13	37.	1.58	3.13	0.21
38.	1.22	0.94	0.09	38.	1.64	4.02	0.16
39.	1.91	1.05	0.08	39.	1.70	2.59	0.18
40.	1.89	0.90	0.12	40.	1.66	3.85	0.22

b = item difficulty, a = item discrimination, and c = guessing factor

Table 1 shows that the discrimination values (a-parameter) of state A range from 0.43 to 1.12. Theoretically, the a-parameter values range from 0 to $+\infty$. The a-parameter values presented in table 8 for state A indicates that the items can separate examinees into different ability levels in the region of item difficulty level. The difficulty values (b-parameter) for state A range from -0.16 to 1.91. This shows that in state A, item 1 with b-value of -0.16 was the easiest item whereas item 7 with b-value of 1.47 was the most difficult item. Although b-value theoretically ranges from $-\infty$ to $+\infty$, typically b-value that range from -3 to +3 are used. The guessing factor (c-value) of state A range from 0.08 to 0.40. Theoretically, c-value range from 0.00 to 1.00 and any item with c-value above 0.50 was recommended not to be selected. Thus, in state A, the a-parameter values, b-parameter values and c-parameter values were appropriate for items in the instrument used in test equating.

Table 1 equally reveals that the a-value of state B range from 0.70 to 4.02. The a-value obtained for state B indicates that the items discriminate highly among students who took the test. The b-value range from 0.21 to 2.00, these b-values for state B indicate that the items are somewhat more difficulty for the students. Also, the c-value for state B range from 0.11 to 0.38. All the items in test B¹ are good enough to be used in conducting test equating. This is because the estimates of the item parameters are within acceptable range. (See appendix P and Q for BILOG-MG output of the item parameters of test form A¹ and B¹)

Hypothesis 1

H₀₁: There is no significant difference in the item parameter estimates of the two forms of mathematics achievement test (MAT) used for test equating.

In order to test this hypothesis, the item parameter indices obtained through BILOG MG analysis and presented in Table 1 were subjected to further statistical analysis using SPSS. The result is shown in table 2 below.

Table 2: Independent t-test analysis of item parameter indices

Parameter	Form	N	$\bar{X}$	SD	df	t	Sig	Decision
a-value	A ¹	40	1.34	0.61	78	-1.50	0.14	NS
	B ¹	40	1.64	1.06				
b-value	A ¹	40	1.26	0.44	78	0.92	0.36	NS
	B ¹	40	1.18	0.32				
c-value	A ¹	40	0.18	0.07	78	-1.18	0.24	NS
	B ¹	40	0.20	0.06				

$\bar{U}$ = 0.05, NS=Non-Significant

The result in Table 2 revealed that the t-values obtained for a-, b-, and c-parameter when test form A¹ and B¹ were compared are -1.50, 0.92, and -1.18 respectively. The associated probabilities to these t-values were 0.14, 0.36 and 0.24 respectively. Since the probability values of all the three parameter estimates are higher than 0.05, it means there is no significant difference in the item parameter indices of the two test forms used for test equating. The non-significance of the three parameters estimates implies that

invariance property the tests use for test scores equating holds and the test are equivalent. The null hypothesis was therefore upheld and either of the two states can use any of the tests. Hence, there is no significant difference in the item parameter estimates of the two forms of mathematics achievement test (MAT) used for test equating

Research Question 2

What are the item parameter consistency value (Differential Item Functioning) of state A and state B?

In order to answer this research question, the absolute adjusted threshold (difficulty) values for all common items (that is even items) in test A¹ and B¹ were used. Item difficulty parameter was used because it tends to yield more stable estimates than item discrimination parameters (Kolen & Brennan, 2004). Specially, item response theory likelihood ratio DIF (IRTLRDIF) was used. Differential item functioning (DIF) value of 1.0 criterion for no DIF set by Dorans and Holland (1992) was the index used to determine item consistency in this study. State A was used as reference group and state B was used as focal group.

Table 3: Item parameter consistency (differential Item Functioning) using threshold (difficulty or b-parameter) Differences Values for states

Item Group B-A		
S/N	DIF Value	SE
1.	-0.30	0.12
2.	-0.45	0.12
3.	-0.17	0.12
4.	0.08	0.12
5.	-0.32	0.12
6.	-0.15	0.12
7.	0.08	0.12
8.	-0.14	0.12
9.	-0.45	0.12
10.	-0.15	0.13
11.	0.24	0.13
12.	-0.16	0.13
13.	-0.33	0.12
14.	0.19	0.14
15.	0.12	0.13
16.	-0.07	0.13
17.	-0.08	0.13
18.	0.12	0.13
19.	0.25	0.14
20.	0.23	0.13

In Table 3, DIF values based on states range from -0.45 to 0.25. Eight (8) items had positive DIF whereas 12 items had negative. All items with negative DIF value imply that students from state A performed better in that item than their counterparts in state B. Also, items with positive DIF values indicated that, students in state B did better in those items than students in state A. Although all the items showed DIF, the /D/ values were not above 1.0 logit for an item to be flagged off. This implies that the items of the tests

are consistent across states and was used in this study to conduct test equating. (See appendix R for DIF output based on state)

Research Question 3

What are the item parameter consistency values (Differential Item Functioning $\hat{\theta}$ DIF) of male and female students?

This research question was answered with the aid of the absolute adjusted threshold (difficulty) values for all common items (that is even items) in test A¹ and B¹ based on male (M) and female (F) students. Specially, item response theory DIF likelihood ratio (IRTDIFLR) was used. Again, differential item functioning (DIF) value of 1.0 logit criterion set by Dorans and Holland (1992) was the index used to determine item consistency.

Table 4: Item parameters consistency (Differential Item Functioning ï DIF) using threshold (difficulty) difference values for sex

Item	Group F-M	
S/N	DIF Value	SE
1.	0.09	0.10
2.	-0.16	0.11
3.	0.20	0.11
4.	-0.33	0.12
5.	0.10	0.01
6.	-0.19	0.13
7.	-0.34	0.12
8.	0.18	0.13
9.	-0.11	0.01
10.	0.14	0.04
11.	-0.20	0.14
12.	-0.16	0.04
13.	0.41	0.13
14.	-0.34	0.12
15.	-0.40	0.26
16.	0.21	0.14
17.	0.38	0.23
18.	-0.27	0.15
19.	0.17	0.08
20.	-0.34	0.14

This research question was answered using the adjusted threshold (difficulty) value based on sex. The result on Table 4 reveals that the DIF based on sex range from -0.40 to 0.41. In this DIF analysis, male students represented by M were used as reference group while female student represented by F were focal group. 9 items had positive DIF whereas 11 items had negative. The result indicates that, female students appear to

outperform their male counterparts in 9 items whereas male students outperformed their female counterparts in 11 items. All the items showed some level of DIF which is not above 0.50 absolute value criterion. This result implies that the items of the tests are consistent across sex and was used for test equating. (See appendix S for DIF output based on gender)

Research Question 4

What are the Item Parameters Consistency values (Differential Item Functioning i DIF) of High and Low ability students?

This research question was also answered with the aid of the absolute adjusted threshold (difficulty) values for all common items (that is even items) in test A¹ and B¹ based on high (H) and low (L) ability students. Specially, item response theory DIF likelihood ratio (IRTDIFLR) was used. Again, differential item functioning (DIF) value of 1.0 logit criterion set by Dorans and Holland (1992) was the index used to determine item consistency in this study.

Table 5: Item Parameters Consistency (Differential Item Functioning ï DIF) using threshold (difficulty) difference values for ability

Item	Group L-H	
S/N	DIF Value	SE
1.	-0.06	0.06
2.	-0.19	0.11
3.	0.06	0.11
4.	-0.04	0.11
5.	-0.03	0.11
6.	-0.18	0.11
7.	-0.10	0.11
8.	0.01	0.11
9.	-0.07	0.11
10.	-0.03	0.11
11.	-0.09	0.12
12.	-0.05	0.11
13.	0.06	0.11
14.	-0.09	0.12
15.	0.01	0.12
16.	-0.08	0.12
17.	-0.13	0.11
18.	-0.03	0.12
19.	-0.13	0.12
20.	-0.12	0.12

The result in Table 5 shows that the DIF value based on students' ability ranged from -0.13 to 0.06. High ability students represented by H were used as reference group and low ability students represented by L were focal group. From the Table, 16 items had negative DIF value whereas 4 items had positive DIF value. The 16 items with negative DIF value were in favour of high ability student while 4 items with positive DIF value were in favour of low ability students. None of the DIF value as observe in the Table was above 1.0 logits absolute value. Again, the items of tests are consistent across ability. (See appendix T for DIF output based on ability)

Research Question 5

What are the items of MAT that fit the three parameter Logistic model?

Table 6: Result of Chi-Square Goodness of Fit for 3PL IRT Model for A¹

Items	Chi-Square	df	Prob	Items	Chi-Square	df	Prob
1	16.5	8	0.04*	21	12.4	8	0.16
2	15.3	8	0.05	22	32.8	8	0.00*
3	8.0	8	0.53	23	9.0	8	0.07
4	9.3	8	0.63	24	12.0	8	0.07
5	13.2	8	0.82	25	8.7	8	0.07
6	11.9	8	0.22	26	8.0	8	0.14
7	5.1	8	0.10	27	66.1	8	0.00*
8	10.8	8	0.22	28	9.2	8	0.21
9	7.0	8	0.20	29	13.5	8	0.15
10	14.6	8	0.10	30	10.0	8	0.08
11	9.3	8	0.20	31	15.4	8	0.30
12	8.2	8	0.11	32	8.2	8	0.21
13	9.5	8	0.42	33	14.0	8	0.06
14	11.4	8	0.18	34	87.8	8	0.00*
15	14.3	8	0.11	35	26.2	8	0.00*
16	15.6	8	0.15	36	6.4	8	0.17
17	6.4	8	0.09	37	51.5	8	0.00*
18	7.1	8	0.53	38	15.0	8	0.31
19	13.2	8	0.12	39	10.6	8	0.08
20	5.8	8	0.07	40	14.9	8	0.06

*Significant

Table 6 shows the result of the chi-square goodness-of-fit analysis for test A¹. From the chi-square values associated with the items in the test it is evident that six items representing 15% of the total items in the test were statistically significant and did not fit the three parameter logistic model. These items are: item 1, item 22, item 27, item 34, item 35 and item 37, and are all marked with asterisk. The Table also indicated that, the remaining 34 items representing 85% of the total test were not statistically significant and

therefore fit the three parameter logistic model. All item fit/misfit were determined at 0.05 level of significance.

Table 7: Result of Chi-Square Goodness of Fit for 3PL IRT Model for B¹

Items	Chi-Square	df	Prob	Items	Chi-Square	df	Prob
1	47.1	7	0.02*	21	40.6	7	0.01*
2	11.7	7	0.08	22	7.0	7	0.43
3	12.7	7	0.06	23	4.6	7	0.08
4	7.7	7	0.10	24	11.6	7	0.12
5	6.3	7	0.06	25	9.0	7	0.21
6	7.8	7	0.23	26	26.6	7	0.00*
7	10.0	7	0.10	27	59.9	7	0.01*
8	8.9	7	0.28	28	11.4	7	0.10
9	12.3	7	0.27	29	6.0	7	0.08
10	9.2	7	0.50	30	10.2	7	0.18
11	12.6	7	0.20	31	6.1	7	0.09
12	6.6	7	0.10	32	9.7	7	0.12
13	12.9	7	0.17	33	3.6	7	0.50
14	8.2	7	0.25	34	20.5	7	0.00*
15	6.0	7	0.11	35	4.7	7	0.20
16	6.4	7	0.10	36	9.7	7	0.47
17	5.5	7	0.30	37	23.4	7	0.00*
18	11.2	7	0.12	38	4.7	7	0.30
19	8.8	7	0.21	39	10.7	7	0.24
20	3.3	7	0.38	40	9.0	7	0.16

*Significant

Table 7 equally shows the result of the chi-square goodness-of-fit analysis for test B¹. From the chi-square values associated with the items in the test it is evident that six items representing 15% of the total items in the test were statistically significant and did not fit the three parameter logistic model. These items are: item 1, item 21, item 26, item 27, item 34 and item 37, and are all marked with asterisk. The table also indicated that, the remaining 34 items representing 85% of the total test were not statistically significant

and therefore fit the three parameter logistic model. (See appendix P and Q for BILOG-MG output of the item parameters of test form A¹ and B¹)

Hypothesis 2

H₀₂: There is no significant fit between the estimates of item difficulty and three parameter logistic model

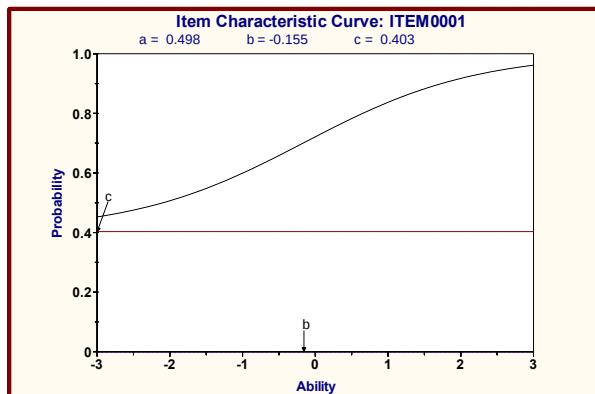
The chi-square goodness-of-fit was used to test whether there is fit between the items of MAT and three parameter logistic model. The data with respect to hypothesis 2 are presented in tables 6 and 7. From table 6, the probability values obtained range from 0.00 to 0.82 for items in test A¹. The result shows that in test A¹, 6 of the items had probability values that were less than 0.05. Since the probability values associated with the chi-square values of these items were all below 0.05, the 6 items representing 15% were statistically significant and did not fit the model. Thirty four (34) items representing 85% in test A¹ had probability values that were greater than 0.05. These items were statistically not significant and therefore fitted the model. Also in table 7 the probability values obtained range from 0.00 to 0.50 for items in test B¹. Six (6) of the items had probability values that were less than 0.05. Since the probability values associated with the chi-square values of these items were all below 0.05, the 6 items representing 15% were statistically significant and did not fit the model. Thirty four (34) items representing 85% in test A¹ had probability values that were greater than 0.05. These items were statistically not significant and therefore fitted the model. It means that the null hypothesis which states that there is no significant fit between the estimates of item difficulty and three parameter logistic model was upheld for 6 items and rejected for 34

items in both tests. Hence there is a significant fit between the estimates of 34 items and the three parameter logistic model.

Research Question 6

What are the item characteristic curve of equated test A¹ and B¹?

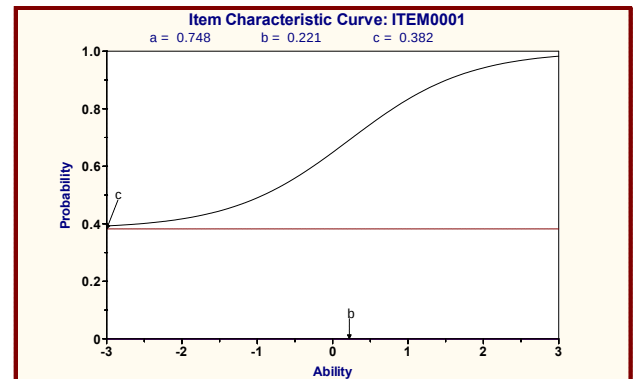
ICC of Test Form A¹



3-Parameter Model, Normal Metric Item: 1
 Subtest: TEST0001

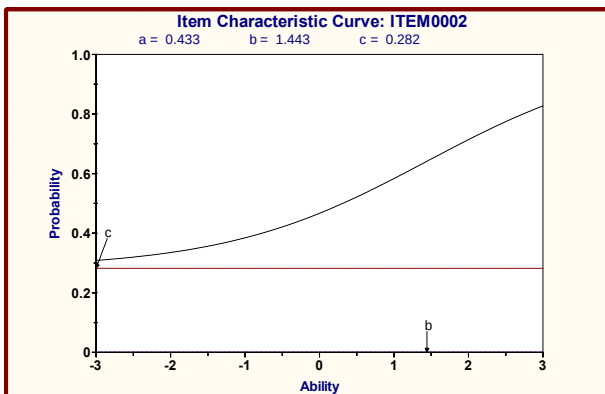
ITEM 1A¹

ICC of Test Form B¹



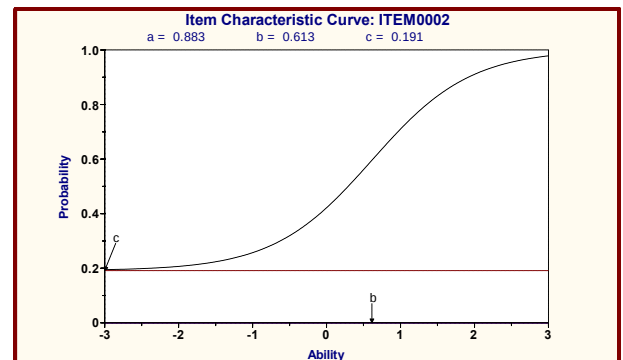
3-Parameter Model, Normal Metric Item: 1
 Subtest: TEST0001

ITEM 1B¹



3-Parameter Model, Normal Metric Item: 2
 Subtest: TEST0001

ITEM 2A¹



3-Parameter Model, Normal Metric Item: 2
 Subtest: TEST0001

ITEM 2B¹

Figure 13: Item Characteristic Curves of Test Form A¹ and Test Form B¹

In figure 13 item 1 for test form A¹ had a discrimination (a) value of 0.50, difficulty (b) value of -0.16 and guessing parameter (c) value of 0.40. The item characteristics curve (ICC) of this item 1 was more or less flat and is slanted upward. Item 1 for test form B¹ on the other hand had a discrimination (a) value of 0.75, difficulty (b) value of 0.22 and guessing parameter (c) value of 0.38. The item characteristics curve (ICC) of this item shifted to right as the probability of correct response is low for most of the students' ability and it increases only at the high ability levels. The dotted line in the ICC of item 1 for both tests show great similarity of student tendency to guess in state A¹ and B¹. Although Item 1 in test form A¹ appears to be less discriminating and difficult than in test form B¹, the a-, b-, and c-values show that they are good items.

Similarly in figure 3, Item 2 for test form A¹ had a discrimination (a) value of 0.43, difficulty (b) value of 1.44 and guessing parameter (c) value of 0.28. The item characteristics curve (ICC) of this item 2 also slanted upward. Item 2 for test form B¹ on the other hand had a discrimination (a) value of 0.88, difficulty (b) value of 0.61 and guessing parameter (c) value of 0.19. The item characteristics curve (ICC) of this item shifted to right in S like shape. The difficulty value of this item in form A¹ is above 1 theta (d) while that of form B¹ tends to 1 theta (d). The dotted lines in the ICC of item 1 and 2 show that student tendency to guess in state A was higher than that of B. Although Item 1 in test form A¹ appears to be less discriminating and difficult than in test form B¹, the a-, b-, and c-values show that there were good items. All the ICC of the items in both tests were steep and vertically shifted towards the right corner of the curves except for

item 1 and 2 in test A¹ whose ICC were slightly flat (See appendix X for the ICC of other items).

Research Question 7

What are the mean ability estimates of students in state A and B when their scores are equated through separate calibration?

Table 8: Mean ability estimates of students in state A and B for scores equated through separate calibration.

State	$\bar{X}$	SD	S^2	RMS	RMS(S^2)
A	1.95	0.89	0.80	0.45	0.20
B	1.89	0.86	0.75	0.55	0.30
Mean Diff.	0.06				

In table 8, the mean ability estimates of students in state A is 1.95 while that of state B is 1.89. The table also shows that, the standard deviation, variance, root mean square and root mean square variance posteriori of estimated ability for the two states are; State A [SD = 0.89, S^2 = 0.80, RMS= 0.45 RMS(S^2) = 0.21] and State B [SD = 0.86, S^2 = 0.75, RMS= 0.55 RMS(S^2) = 0.30]. There appears to be some slight difference in the mean ability estimates of students in state A and B. However, the difference in mean and other moments do not show score difference that matter (SDTM). (See appendix U and V)

Hypothesis 3

H₀₃: There is no significant difference in the ability estimates of students in state

A and B for scores equated through separate calibration

This hypothesis was tested at 0.05 level of significance using the summary of the theta (d) values from separate calibration

Table 9: Independent t-test analysis of students' performance when scores are standardized through Separate calibration equating.

State	N	$\bar{X}$	SD	df	t _{cal}	t _{crit}	Decision
A	1514	1.95	0.89	2903	0.53	1.96	NS
B	1491	1.89	0.86				

Ū = 0.05, NS=Non-Significant

The distribution moments (mean and standard deviation) of separate calibration equating were used in testing this hypothesis. In Table 9, the value of t-calculated was 0.53 and t-critical 1.96. The calculated t value was less than the critical t. Since the calculated t-value was less than the critical t, the null hypothesis (H₀₃) was upheld. Therefore, there was no significant difference in the ability estimates of students in states A and B when their scores are scaled through separate calibration. The observed difference in mean ability estimates may be due to sampling error. The observed mean ability difference was not up to 0.50 and was therefore negligible.

Research Question 8

What are the mean ability estimates of students in state A and B when their scores are equated through concurrent calibration?

Table 10: Mean ability estimates of students in state A and B for scores equated through concurrent calibration.

State	$\bar{X}$	SD	S^2	RMS	RMS(S^2)
A	1.92	0.52	0.27	0.81	0.66
B	1.89	0.71	0.51	0.76	0.58
Mean Diff.	0.03				

In Table 10, the mean ability estimates of students in state A is 1.92 while that of state B is 1.89. The Table also shows that, the standard deviation, variance, root mean square and root mean square variance posteriori of estimated ability for the two states are; State A [SD = 0.52, S^2 = 0.27, RMS= 0.81 RMS(S^2) = 0.66] and State B [SD = 0.71, S^2 = 0.51, RMS= 0.76 RMS(S^2) = 0.58]. In concurrent calibration, there also appears to be some slight difference in the mean ability estimates of students in state A and B. However, the difference in mean and other moments do not show score difference that matter (SDTM). (See appendix W)

Hypothesis 4

H₀₄: There is no significant difference in the ability estimates of students in state A and B for scores equated through concurrent calibration

This hypothesis was tested at 0.05 level of significance using the summary of the theta (d) values from concurrent calibration

Table 11: Independent t-test analysis of students' performance when scores are standardized through Concurrent calibration equating.

State	N	$\bar{X}$	SD	df	t_{cal}	t_{crit}	Decision
A	1514	1.92	0.52	2903	1.32	1.96	NS
B	1391	1.89	0.71				

$\alpha = 0.05$, NS= Non-Significant

The distribution moments (mean and standard deviation) of concurrent calibration equating were used in testing this hypothesis. In table 11 above, the value of t-calculated was 1.32 and t-critical 1.96. The calculated t-value was less than the critical t. Since the calculated t was less than the critical t, the null hypothesis H_{04} was upheld. Therefore, there is no significant difference in the ability estimates of students in state A and B when their scores are scaled through concurrent calibration. The observed difference in mean ability estimates may be due to sampling error. The observed mean ability difference was not up to 0.50 and was therefore negligible.

Research Question 9

What are the mean estimates of students scores scaled through linear equating in state A and B?

To answer this research question, the students' scores in both MCA and MAT were first converted to 100%. This was done in order to bring all the scores from

different state and test forms to a common denominator. The first two distribution moments (mean and standard deviation) in MCA and MAT for the two states were then used to standardize the scores through linear equating and the summary result presented in Table 12 below.

Table 12: Mean estimates of students score scaled through linear equating

State	$\bar{X}$	SD	S^2	S.E
A	66.07	8.53	72.76	0.22
B	61.65	8.13	66.10	0.22
Mean Diff.	4.42			

Table 12 revealed that, the mean of students' performance when standardized through linear equating method are 66.07 and 61.65 for state A and state B respectively. The difference in the mean value of the scaled score between the two states is 4.42 in favour of state A. This difference seems to indicate that, students in state A have higher ability estimate than those in state B when MCA scores and MAT scores are used to determine their ability (see Appendix Z).

Hypothesis 5

H₀₅: There is no significant difference in the ability estimates of students in state A and B for scores equated through linear equating

In order to test this hypothesis, the measures from linear equating (see appendix Y) were further subjected to statistical analysis using SPSS and the result obtained is shown in table 13.

Table 13: Independent t-test analysis of students' ability estimate when scores are standardized through linear equating.

State	N	$\bar{X}$	SD	df	t	Sig	Decision
A	1514	66.07	8.53	2903	14.28	0.00	S
B	1391	61.65	8.13				

$\bar{U} = 0.05$, S= Significant

Table 13 revealed that t-computed was 14.28 with associated probability value of 0.00. Since the associated probability (0.00) is less than 0.05, the null hypothesis H_{05} was rejected. Thus, there was a significant difference in the ability estimates of students in state A and B for scores equated through linear equating.

Research Question 10

Which of the equating methods produced the least Average root mean square error?

In order to answer this research question, the Average root mean square error (ARMSED) obtained from the result of separate calibration, concurrent calibration and linear equating were used. Root-mean-squared-error (RMSE) functions as the transformations which result in the measure of equating accuracy. In separate and

concurrent calibration, the RMSE is obtained from summary indices. But in linear equating, it is calculated by finding the square root of the sum of the squared average standard deviation and the squared absolute mean deviation.

Table 14: Average Root mean square error

Equating Methods	ARMSED
Separate Calibration	0.09
Concurrent Calibration	0.05
Linear Equating	0.04

From Table 14, the average root mean square error obtained were 0.09, 0.05 and 0.04 for separate calibration, concurrent calibration and linear equating respectively. The absolute ARMSED indicates that linear equating yielded the least error and therefore seems to be more efficient in this study.

Research Question 11

What is the relationship between students' mathematics continuous assessment (MCA) scores and mathematics achievement test (MAT) scores for state A and B?

Table 15: Pearson's Product moment Correlation Analysis of MCA and MAT

State	Test	N	$\bar{X}$	SD	r	sig	Decision
A	MCA	1514	66.07	8.53	0.32	0.00	S
	MAT	1514	45.34	16.62			
B	MCA	1391	61.65	8.13	0.33	0.00	S
	MAT	1391	32.53	15.90			

$\bar{U} = 0.05$, S=Significant

For this research question to be answered, MCA scores and MAT scores were first converted to 100%. In state A and B, the scores from MCA and MAT were correlated using Pearson's product moment correlation and the coefficient of 0.32 and 0.33 respectively were obtained. The result is shown on table 15 and the coefficients obtained show weak positive relationship between MCA and MAT in both states.

Hypothesis 6

H_{06} : There is no significant relationship between the performances of examinees in continuous assessment test and mathematics achievement test.

The result presented in Table 15 indicated that, there exist a weak positive relationship between MCA score and MAT score. The r-value for state A was 0.32 with associated probability of 0.00 and 0.33 with associated probability of 0.00 for state B. Since the associated probability in each case was less than 0.05 level of significance with 1,512 and 1,389 degree of freedom respectively. It means that the result was significant and the null hypothesis was not upheld. This implies there is a significant relationship

between the achievement of examinees in mathematics continuous assessment (MCA) test and mathematics achievement test (MAT).

Summary of Findings

Major Finding

- The average root mean square error (ARMSE) obtained were 0.09, 0.05 and 0.04 for separate calibration, concurrent calibration and linear equating respectively. The absolute ARMSED indicates that linear equating yielded the least error and therefore seems to be more efficient in this study.
- There was no significant difference in the ability estimates of students in state A and B when their scores are scaled through separate calibration. The observed difference in mean ability estimates may be due to sampling error.
- There is no significant difference in the ability estimates of students in state A and B when their scores are scaled through concurrent calibration. The observed difference in mean ability estimates may be due to sampling error.
- There was a significant difference in the ability estimates of students in state A and B for MCA and MAT scores equated through linear equating.

Other Findings

- There is no significant difference in the item parameter estimates of the two forms of mathematics achievement test (MAT) used for test equating.

- The items of tests are consistent across states, sex and ability and were used to conduct test equating.
- 6 items representing 15% did not fit the 3PLM whereas 34 items representing 85% of the total test were not statistically significant and therefore fit the three parameter logistic model in both tests.
- All the ICC of the items in both tests were steep and vertically shifted towards the right corner of the curves except for items 1 and 2 in test ^{A1} whose ICC were slightly flat. There appears to be some slight difference in the mean ability estimates of students in state A and B. However, the difference in mean and other moments do not show score difference that matter (SDTM).
- There was a significant relationship between the performances of examinees in mathematics continuous assessment test and mathematics achievement test.

CHAPTER FIVE

DISCUSSION OF THE FINDINGS, CONCLUSION, RECOMMENDATIONS AND SUMMARY

This chapter deals with the discussion of the findings of the study, the conclusion and the recommendations based on the findings. The implications of the study and suggestions for further research are also highlighted. Finally, limitations of the study as well as a brief summary of the entire work are presented

Discussion of findings

The use of three parameter logistic model in the calibration of 40-items mathematics multiple-choice test using BILOG-MG yielded item parameter and ability estimates/results that are very important to test developers and users as well as researchers. The item consistency/differential item functioning (DIF), fitting of items to three parameter logistic model, item characteristics curves of all items, separate calibration, concurrent calibration, linear equating and determination of equating error using root mean square were all explained based on findings made about them.

Item Parameter Estimates

The result from research question 1 and the corresponding hypothesis 1 for state A in table 1 revealed that the easiest item in MAT was item 1. This item had a difficulty, discrimination and guessing parameter estimate of $b = -0.16$, $a = 0.50$ and $c = 0.40$ respectively. The negative sign in front of item 1 shows that, this question was easy for most examinees (especially those below average ability) that were used in the calibration exercise and they did not have any difficulty answering it correctly. Item 1 had a

discrimination value which Baker (2001) classified as being low. This low discrimination value shows that this item did not adequately differentiate between the high and low ability examinees. The tendency of examinees answering item 1 correctly by guessing was 0.403. For item 1 being the easiest question to have the highest guessing estimate, it implies that the item needs further investigation. The most difficult item was item 40. This item had a difficulty, discrimination and guessing parameter estimate of $b = 1.89$, $a = 0.90$ and $c = 0.12$ respectively. Item 40 was so difficult that even students whose ability is above average but not up 1.89 cannot answer it correctly without guessing or cheating. The discrimination value of item 2 was very high and this shows that, the item adequately differentiate between the high and low ability examinees. The tendency of examinees answering item 1 correctly by guessing was 0.12. The parameter values of other items of MAT are interpreted in similar manner.

Similarly, in State B, item 1 was the easiest item with parameter estimates of: $b = 0.22$, $a = 0.75$ and $c = 0.38$. The value of a and b parameters in state B appear to be higher than that of state A, and the examinees in state B seem to have lower tendency towards guessing item 1 correctly. The most difficult item was item 25 with parameter estimates of; $b = 1.94$, $a = 1.15$ and $c = 0.21$. The result of the t-test statistic as presented on table 2 indicated that, there is no significant difference in the item parameter estimates of the two test forms used for test scores equating.

This finding is consistent with Forsyth's (1987) assertion that, if the items in a test have been calibrated using IRT procedures, then, presumably, a subset of items could be used to estimate an examinee's achievement level and a school's mean achievement level.

Forsyth maintained that if the assumptions of IRT models are satisfied, item parameters are invariant across groups of examinees and ability parameters are invariant across groups of items. The invariance of the item parameters is particularly important in horizontal equating settings. The finding of this study is also in line with that of Yang (2004) who evaluated the invariance of the composite score thresholds (item difficulty) and concluded that, linking across regions seemed to hold reasonably well. In the same vein, the results presented on table 2 appear to hold when the tests are used to conduct test equating across states.

Item consistency/Differential item functioning (DIF)

The analyses from the results of research question 2, 3, and 4 indicated the invariant nature of the items of MAT for state, gender and ability groups. In Table 3, the DIF analysis conducted using the 20 common items imbedded in both tests forms revealed that for the state, 12 items had a negative values indicating that the items are in favour of examinees in state A whereas 8 items had positive values indicating that they are in favour of state B. The DIF values obtained ranged from -0.45 to 0.45, this DIF values were not up to 1.0 logit and as such they were negligible or had no differential item functioning between examinees in state A and state B. The DIF analysis conducted based on gender in Table 4 shows that, 11 items had a negative values indicating that the items are in favour of male examinees whereas 9 items had positive values indicating that they are in favour of female. The DIF values based on gender ranged from -0.40 to 0.41, this DIF values were not up to 1.0 logit and as such there were negligible or no differential item functioning between male and female examinees was observed. In the

same vein, the DIF analysis based on high and low ability group as presented in Table 5 revealed that, 16 items had negative values and are in favour of high ability examinees whereas 4 items had positive values and are in favour of low ability examinees. The DIF values based on examinees' ability ranged from -0.26 to 0.12, this DIF values were not up to 1.0 logit and as such there were negligible or no differential item functioning existed between examinees based on ability.

The findings of this study is not consistent with that of Yang (2004) who examined whether the multiple-choice to composite linking functions of Advanced Placement Program (AP) examinations remained invariant over subgroups by region. The study focused on two questions: (a) how invariant were cut scores across regions and (b) whether the small sample size for some regional groups presented particular problems for assessing linking invariance. In addition to using the subpopulation invariance indices to evaluate linking functions, Yang also evaluated the invariance of the composite score thresholds for determining final AP grades. Overall, linking across regions seemed to hold reasonably well, and Males and females exhibit differential mean score differences on the free-response and multiple-choice sections. The finding of this study agrees with that of Obinne (2007) on Differential item functioning (DIF) effects of Biology examination items of WAEC and NECO for the years 2000 to 2002 analyzed in terms of Gender and location and reported that some items favoured girls while some others favoured boys. The finding is also in line with that of Won-ling and Rui (2008) who found that the difference across the groups used in their study show smaller difference that will not affect pass/fail decision.

Item Fit and Three Parameter Logistic Model

The result from the researchers' attempt to fit a three parameter logistic IRT model to the response data of MAT produced an encouraging outcome, as few items turned out to be significant. The separate estimates of examinees in state A and B corresponded very well. From Table 6, the chi-square goodness-of-fit statistic for state A revealed that, out of the 40 items used in this study, 6 items representing 15% did not fit 3PLM whereas 34 items representing 85% fitted the 3PLM. Also in Table 7, the chi-square goodness-of-fit statistic for state B equally revealed that, out of the 40 items responded to by examinees, 6 items representing 15% did not fit 3PLM whereas 34 items representing 85% fitted the 3PLM. From Gruijter and Kamp's (2002) suggestion, the 34 items that fitted the chosen model are selected as most appropriate items in measuring students' ability in mathematics. However, item 22 in state A and item 21 in state B that were some of the items with good parameter estimates and ICC shape but did not fit the 3PLM. The fact that item 22 in state A and item 21 in state B did not fit the 3PLM does not imply that they are bad item, rather it shows that the item require some other model for it to fit properly. The null hypothesis which states that there is no significant fit between the estimates of item difficulty and three parameter logistic model was upheld for 6 items and rejected for 34 items in both tests. Hence there was a significant fit between the estimates of 34 items and the three parameter logistic model. The findings of this study are in line with that of Adedoyin (2010) and Ene (2005) who used chi-square test with probability greater than the alpha level of 0.05 significant level to selected items that fit models they used in their respective studies.

Item Characteristics Curves of all Items

The result presented in figure 1 for research question 6 showed the item characteristic curve (ICC) of test form A¹ and B¹. In test form A¹, the ICC of item 1 was flat and is slanted upward. The horizontal axis represents ability (θ), with the vertical axis representing the observed probability of a correct response (PCR). The flat nature of item 1 ICC shows that it did not discriminate properly. Adedoyin (2010) had explained that the flatter the ICCs curve, the less the item is able to discriminate and the steeper the curve, the better the item can discriminate. Item 1 need to be revised for further use. The item characteristics curve (ICC) of item 2 in test form A¹ shifted to right as the probability of correct response was low for most of the students' ability and it increased only at the high ability levels. Also in test form B¹, the item characteristics curve (ICC) of item 1 shifted to the right as the probability of correct response was low for most of the students' ability and it increased only at the high ability levels. The item characteristics curve (ICC) of item 2 in test form B¹ equally shifted to the right as the probability of correct response was low for most of the students' ability and it increased only at the high ability levels. Judging from the shape of the item characteristics curves of item 1 and 2 in test form A¹, these two items did not possess good ICC shape but other items of Test form A¹ had good ICC shape. All the items in test form B¹ appears to be S-like shape and had good ICC shapes. The numbers of items with poor ICC shape are negligible indicating that the items of MAT are good ones.

Separate Calibration, Concurrent Calibration and Linear Equating

The analysis of results from research question 7 and the corresponding hypothesis 3 revealed that the mean ability estimates of students in state A and B did not show any significant difference. As observed from Table 7, the difference in the mean ability estimates and other moments (standard deviation and variance) did not show any score difference that matter (SDTM). The result of t-test statistic as presented in Table 8 indicated that the estimated mean ability from calibrated test form A¹ and B¹ through separate calibration were statistically equivalent.

The analysis of result from research question 8 and the corresponding hypothesis 4, also revealed that the mean ability estimates of students in state A and B did not show any significant difference. As observed from Table 9, the difference in the mean ability estimates and other moments (standard deviation and variance) did not show any score difference that matter (SDTM). The result of t-test statistic as presented in Table 10 indicated that the estimated mean ability from calibrated test form A¹ and B¹ through concurrent calibration were statistically equivalent.

These findings are not in consonant with Hanson and Beguin (1999) who investigated the performance of separate versus concurrent estimation in putting item parameter estimates for two forms of a test administered in a common item equating design on the same scale. Their results among others showed that, the differences among the item parameter scaling methods used in separate estimation were much larger than the differences between concurrent estimation and the better performing scaling methods in separate estimation. Other research that compare separate and concurrent calibration have

concluded that concurrent estimation performed somewhat better than separate estimation (Petersen, Cook, and Stocking, 1983; Wingersky, Cook, and Eignor, 1987), while Kim and Cohen (1998), concluded that the performance of separate estimation was equal to or better than concurrent estimation. The findings of this study have shown that both separate and concurrent calibration performed equally in the estimation of students' ability.

The analysis of results from research question 9 and the corresponding hypothesis 5, revealed that the mean ability estimates of students in state A and B did show a significant difference. As observed from Table 11, the difference in the mean ability estimates shows a score difference that matter (SDTM). The result of t-test statistic as presented in Table 12 indicated that the estimated mean ability from mathematics continuous assessment (MCA) and mathematics achievement test (MAT) scores using linear equating was statistically not equivalent. This finding is in consonant with Von Davies and Kong (2005) who in a study on unified approach to linear equating for the non-equivalent groups design used two parallel tests, with 78 items tests and a 35 items external anchor in each test found out that, the mean difference between the two sample groups was 2.66. Based on this result, the authors concluded that, a mean difference of this magnitude indicates a fairly large difference between the two population used in their study.

Determination of efficiency equating methods using root mean square (RMS)

From Table 13, the average root mean square error obtained were 0.09, 0.05 and 0.04 for separate calibration, concurrent calibration and linear equating respectively. The

absolute ARMSED indicates that linear equating yielded the least error and therefore appears to be more efficient. This finding seems to negate that of Morrison and Fitzpatrick (1992) who found that concurrent calibration resulted in the least amount of equating error among four equating methods considered in their study. The finding of this study also differs from that of Yang (1997) who found linear equating to produce the largest amount of error. Yang's (1997) result shows that IRT equating methods were better than the linear (Tucker) method. In all the three equating methods that were used in this study, the error estimates shows some level of similarity among them. This is because when error values were compared among equating methods, none shows score difference that matter.

Relationship between MCA and MAT

The result presented in Table 14 indicated that, there exist a weak positive relationship between MCA score and MAT score. The r-value for state A was 0.32 with associated probability of 0.00 and 0.33 with associated probability of 0.00 for state B. The associated probability in each case was less than 0.05 level of significance with 1,512 and 1,389 degree of freedom respectively. This means that the results were significant and the null hypothesis was not upheld. This implies there is a significant relationship between the achievement of examinees in mathematics continuous assessment (MCA) test and mathematics achievement test (MAT). This finding is not in consonant with the opinion of Ugodulunwa, (1999) who pointed out that, it is common to see continuous assessment scores that do not correlate positively with actual examination

performance. A careful look at MCA scores and MAT scores shows that, some students scored very high marks on teacher's mathematics continuous assessment and the same students surprisingly scored very low marks in researchers' developed mathematics achievement test. This may be indicative that the mathematics continuous assessment scores collected by the researcher from various schools may not have correctly reflected the students' achievement.

Conclusion

Based on the result, the following conclusions were drawn:

- The test equating methods used in this study had different root mean square errors, but linear equating was the most efficient as it produced the least amount of error.
- The three parameter logistic model was successfully applied in the calibration of a mathematics test.
- The item parameter estimates obtained show that, the items of MAT were good as they did not function differentially among examinees in states, and by gender and ability levels.
- On the issue of item fit, 34 items fitted the three parameter logistic model whereas 6 items did not fit it.
- The item characteristics curve of the items presented show that item 1 and 2 of test form A¹ had poor ICC shapes while that of the other items in test form A¹ and B¹ had good ICC shapes.
- There exists weak relationship between the score of students in MCA and MAT.

Educational Implication of the Findings of the Study

Developing assessment instruments that will be used in continuous assessment practices is very important to the teacher. Given the array of task before the teacher, it is difficult for him/her to develop assessment instrument that will produce result, which can be said to be comparable with that of other teachers. Therefore experts in psychometrics can be hired to produce/develop items and such items can be saved in school store or nation item bank and used when the need arises.

Items generated through the use of IRT may likely improve instructional delivery among teachers and understanding among students. With IRT, content area that students find difficult to learn can easily be detected from the beginning of the lesson. The teacher is in this case expected to bring such difficulty area to the understanding of the student. Teacher will have the opportunity of detecting individual student level of latent ability which may not be easily noticed in CTT approach only.

Determining how items of the instrument function among the various subgroups used in this study helped practitioners prepared valid instruments that are fair to every group to be tested under the same curriculum content. This study has shown that even when the psychometric properties of the item used to measure ability are good, item may treat some group unfairly. The extent to which this may occur will make some group to appear as underperformers.

It is not just enough to use items from an instrument just because its parameter attributes appears to be good when IRT is applied. Since models are used to calibrate this item, fitting the item into the model used for calibration will give a further credence to

the items from doubt that other practitioners would want to place on the instrument. Therefore this study has implication on the practitioner, students and users of such test scores.

Building test forms require that such forms should be of equal difficulty so that the result emanating from them can be comparable. Therefore equated test forms in test/item banks can save man time, energy and resources of producing items of test on regular bases.

The equating method employed in this study has some implication on student, teachers and score users. For students and teachers, the study can help in the development of instrument with similar difficulty. If for any reason a child/student is absent on the day of assessment, the student will be sure that the next test he/she will take is of equal difficulty with the one his/her colleague had written. With a partial knowledge of the students' ability, linear equating can be use to estimate his/her performance in test which was not taken. Teachers and scores users can easily compare performance or even assessment standard in the situation.

Recommendations

It is therefore recommended that:

- Test scores equating methods should be used to standardize students' continuous assessment (CA) scores. Particularly linear equating method should be used if students' CA scores are reported as aggregate of the CA test scores they have taken over a given period. This equating method may help in the comparison of students' scores within the same subject

- Score assigned to students' responses for every cognitive based continuous assessment should be reported in person-by-item response pattern. This will permit better CTT or IRT analysis to be performed.
- Instruments used for testing the cognitive, affective and psychomotor ability of the students should employ IRT based process in their development.
- The differential item functioning effect of items in instrument for measurement of latent trait should always be determined. This will help to ensure that sub-group of examinees are not unfairly treated.

Limitations

- Treating the CA scores from all the schools within a state as a single group when there is variation in their test forms may have created some loss of vital information.
- The content covered by the test may not have been taught by all schools as at the time of the examination as such some schools may have found the test more difficult than others.
- The continuous assessment scores collected by the researcher from various schools may not correctly reflect the students' achievement. The CA score was aggregate performance over several examinations by the teachers and there was a weak relationship between the MCA and MAT which probably indicated that something was wrong with the CA score.

Suggestion for further study

This study in itself is not exhaustive in terms of scope as there are other areas that could not be researched into under test score equating. It is rather an attempt to stimulate and facilitate further research in this direction of academic endeavour. It is therefore suggested that:

- A replica of this study should be carried out to cover more states.
- The standardization of students' score through test score equating can be study with other forms of equating designs
- The comparison of students' score in Essay test can be study using other test equating models like partial credit model and graded response model.
- The comparison of students' score across different grades using vertical equating can be research into.
- Research in test equating that takes into consideration uncontrolled environmental variables that influence the examinees' response is necessary.

Summary of the study

The study aimed at ascertaining the relative efficiency of three test scores equating methods (Linear equating, separate calibration and concurrent calibration) in the comparison of students' continuous assessment measures. To accomplish this task, eleven research questions and five hypotheses were formulated to guide the study. Review of literature was extensively utilized to unveil some of the researches conducted in this area

of study. The researchers review shows that several methods are available for test equating and each method is unique for a given purpose. Using test equating as a means for comparison of students performance requires standardized test, special design and software for data analysis. In this study therefore, two parallel tests consisting of 40 multiple choice items each were developed by the researcher. The items of the instruments were administered on a sample of 2905 SS III students for 2010/2011 academic session multistage sampling technique was used to randomly draw students from Cross River and Rivers States. The study adopted the Non-Equivalent Anchor Test (NEAT) group design.

In analyzing the data collected, item parameter estimates, item characteristics curve, chi-square (χ^2) goodness of fit test, descriptive statistics, score difference that matter (SDTM), equating Error (which was used as estimate of efficiency), t-test and Pearson product moment correlation coefficient were used to answer research question and test the stated hypotheses.

The results indicated that:

- The absolute Average Root Mean Square Error Difference (ARMSED) indicates that linear equating yielded the least error and therefore seems to be more efficient than the other methods.
- All the estimates of the item parameters in both tests were within acceptable range. There is no significant difference in the item parameter estimates of the two forms of mathematics achievement test (MAT) used for test equating

- The items of tests are consistent across states, sex and ability and were used to conduct test equating.
- 6 items representing 15% did not fit the 3PLM whereas 34 items representing 85% of the total test were not statistically significant and therefore fit the three parameter logistic model in both tests. There was a significant misfit between the estimates of 6 items and the three parameter logistic model. There was also a significant fit between the estimates of 34 items and the three parameter logistic model.
- All the ICC of the items in both tests were steep and vertically shifted towards the right corner of the curves except for item 1 and 2 in test A¹ whose ICC were slightly flat. There appears to be some slight difference in the mean ability estimates of students in state A and B. However, the difference in mean and other moments do not show score difference that matter (SDTM)
- There was no significant difference in the ability estimates of students in state A and B when their scores are scaled through separate calibration. The observed difference in mean ability estimates may be due to sampling error. The observed mean ability difference was not up to 0.50 and was therefore negligible.
- there is no significant difference in the ability estimates of students in state A and B when their scores are scaled through concurrent calibration. The observed difference in mean ability estimates may be due to sampling error. The observed mean ability difference was not up to 0.50 and was therefore negligible.
- There was a significant difference in the ability estimates of students in state A and B for scores equated through linear equating.

- There was a significant relationship between the performances of examinees in mathematics continuous assessment test and mathematics achievement test.

Based on these findings, it was recommended that item response theory (IRT) methods of item parametrization should be incorporated into all examinations conducted in Nigeria. The standardization of students' continuous assessment score should be done through test score equating. This will allow for the comparison of scores and test forms. If several equating methods are to be used at a time, their efficiency should be determined using the method that produces the least quantity of error. Score comparison is a necessary evaluation step which should be undertaken for valid judgment to be passed or made on students.

REFERENCES

- Adebowale, O. F. & Aluo, K. A. (2008). Continuous assessment policy implementation in selected local government areas of Ondo State (Nigeria): Implication for successful implementation of the UBE program. *KEDI Journal of Educational Policy* 5(1), 3-18.
- Adedoyin, O. O. (2010). Using IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics. *Educational Research and Reviews*, 5 (7), 385-399, Jul 2010. Retrieved January 10, 012 from <http://www.academicjournals.org/ERR2>
- Afemikhe, O. A. (2007). "Assessment and educational standard improvement: Reflections from Nigeria". A paper presented at the 33rd Annual conference of the International Association for Educational Assessment held at Baku, Azerbaijan. September 16th to 21st 2007.
- Afolabi, S. O. (1999). Six honest men for continuous assessment evaluating the equation of achievement scores in Nigeria Secondary Schools. *Ifè Journal of Behavioural Research*, 1(2), 7-15
- Afressa, T. M. & Keeves, J. P. (1999) Changes In Students' Mathematics Achievement In Five States Of Australian Lower Secondary Schools over Time. *International Education Journal*, 1 (1) 1-21.
- Airasian, P. W. (1991). *Classroom assessment*. New York. McGraw-Hill.
- Ajuonuma, J. O. (2006). *Competence Possess by Teachers in the Assessment of Students in the Universal Basic Education (UBE) Programme*. A paper presented on the 2nd Annual National Conference of the Department of Educational Foundations. Enugu State University of Science and Technology.
- Ajuonuma, J. O. (2007). *A Survey of the Implementation of Continuous Assessment in Nigeria Universities*. A paper presented at second regional conference of Higher Educational Research and Policy Network (HCRPNET). Held at Ibadan 13th to 16th August, 2007.
- Akinlua, A. A. & Ajayi, P. O. (2003). Evaluation of continuous Assessment practice in primary schools: *Nigeria Journal of Education Research and Evaluation* 4(2): 16-23.

- Alausa, Y. A. (2003). *Continuous Assessment in our schools: Advantages and problems*. Retrieved May 18, 2007, from <http://www.edne.nalResources/Reform%20Foriem/Journal9/Journal%209%Article%202.Pdf>.
- Alausa, Y. A. (ND). *Continuous Assessment in Our Schools: Advantages and Problems*
- Ali, A. (2005). *Conducting Research in Education and the Social Sciences*. Enugu: Tashiwa Networks Ltd.
- Altonji, J. G. (2009) Constructing AFQT Scores that are Comparable Across the NLSY79 and NSLY97. Retrieved November 10, 2009 from <http://www.econ.yale.edu/-F188/AFQTmatch.pdf>
- Andrich, D. (1978). A Binomial latent trait model for the study of likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, 31: 84-98.
- Angoff, W. H. & Cowell, W.R. (1985). An Examination of the assumption that the equating of parallel forms is population independent. (ETS Research Report 85-22). Princeton, JJ: Educational Testing Service.
- Angoff, W. H. (1971). Scale, norms and Equivalent scores. In R.L. Thorndike (ed.). *Educational Measurement* (2nd ed. 508-606) Washington DC. American Council of Education.
- Angoff, W. H. (1984). *Scales, Norms, and Equivalent scores*: Princeton, NJ: Educational Test Service.
- Anikweze, T. M. (2005, September). *Assessment and the Future of School and Learning*. A paper presented at the 31st Annual Conference of the International Association of Educational Assessment. Abuja, Nigeria 4th ï 9th.
- Baker, E. L. (1991). Trend in Testing in the United State of America. In S.H. Fuhrman and B. Malen (eds.) *The politics of curriculum and testing*. 139-159.
- Baker, F. B. (2001). *The basics of item response theory on Assessment and Evaluation*. Wisconsin: ERIC clearing house.
- Barnard, J. J. (1996). *In search of equity in Educational Measurement: Traditional versus modern equating methods*. Paper presented at ASEESA& National Conference at the HSRC Conference Centre, Pretoria, South Africa.
- Bhakta, B., Tennant, A., Horton, M., Lawson, G., & Andrich, D. (2005) Using Item response theory to explore the Psychometric Properties of extended matching question

- examination in undergraduate medical education. Retrieved from <http://www.biomedicalcentral.com/1472-6920/5/9>
- Bielinski, J., Thurlow, M., Minnema, J. & Scott, J. (2000). How out-of-level testing affect the psychometric quality of test scores. Out-of-level Testing Report.
- Bock, R. (1972). Estimating item parameter and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 27: 29-51.
- Bolt, D. M. (1999). Evaluating the Effect of Multiple dimensionality on IRT True Score Equating. *Applied Measurement in Education* 12: 383-407.
- Braud, H. L & Holland, P.W. (1982). Observed score equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland D.B. Rubin (Eds.). Test Equating (pp. 9-49). New York: Academic Press.
- Camilli, G., & Shepard, L. A. (1994). *Method for Identifying Biased Test Items*. Newbury Park, CA: Sage.
- Chong, H. Y. (2007). *A simple guide to the item Response theory (IRT) and Rash Modelling*: Retrieved March 27th, 2009 from <http://www.creative-wisdom.com>
- Chong, H. Y. & Sharon E. O. P. (2005) Test Equating by Common Items and Common Subjects: Concepts and Applications. *Practical Assessment, Research & Evaluation* 10 (4), 1-19 Retrieved March 21st, 2009 from <http://pareonline.net/getvn.asp?v=10&n=4>
- Conover, W. J. (1999). Practical Nonparametric statistics (3rd ed.) New York: John Wiley.
- Crooker, L. & Algina, J. (1986). *Introduction to classical and Modern Test theory*. New York, NY: Holt, Rinehart and Winston.
- Donovan, M. A. Drasgow, F., Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology*, 85(2), 305i 313.
- Doran, N. J. & Holland, P.W. (2000). Population Invariance and Equating of Tests: Basic Theory and the linear case. *Journal of Educational Measurement*, 37: 281-306.
- Dorans, N. J. (2000). *Linking scores from multiple instruments*. Center for Statistical Theory and Practice Educational Testing Service. ETS.

- Dorans, N. J., Pommerich, M. & Holland, P.W. (Eds.) (2007). *Linking and Aligning Scores and Scales*. New York: Springer.
- Dorans, N. J. & Lawrence I. M. (1990). Checking the Statistical Equivalent of nearly Identical Test Editions. *Applied Measurement in Education*, 3, 245-254.
- Dorans, N. J., Liu, J., & Hammond, S. (2008) Anchor Test Type and Population Invariance: An Exploration Across Subpopulations and Test Administrations. *Applied Psychological Measurement*, 32, 81-97.
- Dorsans, N. J. (2004). Equating, concordance and Expectation. *Applied Psychological Measurement*, 28(4), 227-246.
- Eiji, M., Catherine, M. H. & Yong-Won, L. (2008). Equating and Linking of Performance Assessment. *Applied Psychological Measurement*. 24(4): 325-337. Retrieved April 22, 2009 from <http://upm.sagepub.com/cgi/content/abstract/24/4/325>.
- Emaikwu, S. O. (2006). Relative efficiency of four multiple matrix sample models in estimating aggregate performance from partial knowledge of examinees ability levels. An Unpublished PhD Thesis. University of Nigeria, Nsukka.
- Ene, C. U. (2005). Application of Rasch Model in Assessing the Attitude of Students Towards Biology in Senior Secondary Schools in Enugu Education Zone. *Unpublished M.ED Project*. Department of Science Education, University of Nigeria, Nsukka.
- Ezeudu, S. A. (2005). Continuous assessment in Nigeria Senior Secondary School geography: Problems and Implementation strategies. Paper presented at the annual conference of the International Association for Educational Assessment, Abuja, Nigeria.
- Fan, X. (1998). Item Response Theory and classical test theory: an empirical comparism of their items/person statistics. *Educational and Psychological Measurement* 58(3) 1-17
- Federal Ministry of Education (1985). A Handbook on Continuous Assessment Lagos: Heinemann Educational books Ltd.
- Felan, G. D. (2002). *Test Equating: Mean, linear, Equipercentile and item Response theory*. Paper Presented at the Annual Meeting of the Southwest Educational Research Association. Austin TRXAS.

- Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, 14, 31-57.
- Gao, H. (2004). *The effect of Different Anchor Test on the Accuracy of Test Equating for Test Adaptation*. Retrieved May 12, 2009. <http://www.ohiolink.edu/etd/send-pdf.cgi/Gao%20itua%20pdf?accnum=ohiou/089917802>.
- Gordon, B., Engelhard, G., Gaberialson, S. & Bernknoff, S. (1996). Conceptual issues in Equating performance assessment: Lesson from writing assessment. *Journal of Research and development in Education*, 29: 81-88.
- Grant, M. C., Zhang, L., Damiano, M. & Lonstein, L. (2006). An Evaluation of the Kernel Equating Method: Small Sample Equating in non-equivalent groups. Paper presented at National conference of AERA/NCME, 2006.
- Grant, M. C., Zhang, L., Damiano, M. & Lonsterin, L.L. (2006). *An evaluating of the kernel equating method: Small sample equating in non-equivalent groups*. Paper presented at the national conference of AERA/NCME 2006.
- Gray, L. M., Nancy, S.P. & Stewart, E.E. (1979). *A Test of Adequacy of Curvilinear Score Equating models*. A paper presented at the Computerized Adaptive Testing Conference. Minneapolis. M.N, June 27-30, 1979.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144i 149.
- Haertel, E. H. (2004). The Behaviour of Linking Items in Test Equating. (Case Report. 630). Educational Testing Service, Princeton.
- Hambleton, R.K.Swaminathan, H.& Roger, H. J. (1991). Fundamentals of item response theory. Newbury Park. C.A: Sage
- Han, T., Kolen, M. & Pohlman, J. (1997). A comparison among IRT true i and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, 10(2), 105-121.
- Hanick, P. L. & Chi-Yu, H. (2002). Effect of decreasing the number of common items in Equating Link item sets.
- Hanson, B. A. & Beguin, A. A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in common item non

- equivalent groups equating design. *Applied Psychological Measurement*. 26(1), 3-24.
- Hanson, B. A. (1989). Scaling the P-ACT+ In R.L. Brennan (Ed.) *Methodology used in Scaling the ACT Assessment and P-ACT+ (Pp. 57-73)*. Iowacity, IA: American College Testing.
- Hanson, B. A. (1996). Testing for Difference in Test Scores Distributions using loglinear Models. *Applied Measurement in Education* 9(4): 305-321.
- Harbor-Peters, V. F. A. (1999). *Noteworthy points on Measurement and Evaluation*. Enugu: Snaap Pres Ltd.
- Harris, D. J. (1991). A Comparison of Angoff's Design I and Design II for Vertical Equating using Traditional and IRT Methodology. *Journal of Educational Measurement* 28, 221-235.
- Hennings, S. S. (1996). A Comparison of Equating Methods Applied to Performance-Based Assessment. Retrieved July 9, 2008 from <http://eric.ed.gov/ERICwebportal/homeportal?nfpb=true ERICeXGSEARCH> Description= Result
- Holland, P. W. & Thayer, D. T. (1989). *The Kernel Method of Equating Score Distributions (Technical Report No. 89-84)*. Princeton N.J: Educational testing Service.
- Holland, P. W., Von Davier, A. A., Jinharay, S. & Han, N. (2006). Testing the unobsenable assumption of the chain and post stratification equating methods for the NEAT design. Research report ETC, 2006.
- Ipaye, T. (1982). *Continuous Assessment in Schools with some Counseling Applications*. Ilorin: University of Ilorin Press.
- Jodoin, M.G., Keller, Lisa, A. & Swaminathan (2003). A comparison of Linear Equating, Fixed common item and concurrent parameter Estimation Equating procedure in Capturing Academic growth. *Journal of Experimental Education*. 71(3): 2
- Kim, D. L. & Cohen, A.S. (1998). A comparison of linking and concurrent calibration under item Response Theory. *Applied psychological Measurement* 22(2), 131-143.
- Kim, D. L. (2005). A comparison of IRT Equating and Beta 4 equating. *Journal of Educational Measurement*, 42(1), 77-99.

- Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling and Linking Methods and Practices*. New York: Springer-Verlag.
- Kolen, M. J. Zeng, L., & Hanson, B.A. (1996). Conditional Standard Error of Measurement for Scale Scores using IRT. *Journal of Education Measurement* 35(2), 129-140.
- Kolen, M. J., and Brennan, R. L. (1995). *Test Equating Methods and Practices*. New York: Springer-Verlag.
- Koretz, D. (1999). *Limitations in the use of Achievement Tests as Measures of Educator's Productivity*. Paper presented in the National Academic of Sciences Conference held at Beckman Centre Irvine, California December 18, 1999.
- Kyang, T. H. (ND) Item Response models used with Wingen: Undimensional IRT models for Dichotomous Responses. Retrieved July 11, 2008 from <http://www.Umass.edu/remf/software/wingen/modelsF.html>.
- Lamprianou, L. (2007). An investigation into the test equating methods used during 2006, and the potential for strengthening their validity and reliability. Cyprus Testing Service. Manchester.
- Lee, O. K. (2003). Rasch simultaneous vertical Equating for measuring Reading Growth. *Journal of Applied Measurement*, 4(1), 10-23.
- Lissitz, R. W. & Huynh, H. (2003). Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate yearly Progress and School Accountability. *Practical Assessment, Research & Evaluation* 8(10): 2003.
- Lissitz, R. W. & Huynh, H. (2003). Vertical Equating for the Arkansas ACTARD Assessment: Issues and Solutions in Determination of Adequate Yearly progress and school Accountability. A Report Submitted to the Arkansas Department of Education.
- Liu, M. & Holland, P.W. (2008). Exploring Population Sensitivity of Linking Functions across three School Admission Test Administrations. *Applied Psychological Measurement*, 32, 81-97.
- Livingston, S. A. (2004). *Equating Test Scores (without IRT)*. Princeton: Educational Testing Service.
- Livingston, S. A., Dorans, N. N. & Wright, N.K. (1990). What combination of sampling and equating methods work best? *Applied Psychological Measurement*, 3(1): 73-95.

- Lord, F. N. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ. Lawrence Erlbaum.
- MacDonald, P. & Shampo, V.P. (2002). A Monte Carlo Comparison of Item and Person Statistics based on Item Response Theory versus Classical Test Theory. *Journal of Educational and Psychological Measurement*, 48(5): 1040-1050.
- Magno, C. (2009) Demonstrating the difference between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment*, 1 (1), 1-11
- Makiney, J. D., Rosen, C., Davis, B.W., Tinios, K. & Young, P. (2003). *Examining the measurement equivalence of paper and computerized job analyses scales*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL
- Mao, X., Von Davier, A.A. & Rupp, S. (2005). *Comparative of kernel equating methods on PRAXIS data (ETS Research Report)*. Princeton, NJ: Educational Testing Service.
- Marco, G. L., Peterson, N.S. & Stewart, E.E. (1983). A Test of the Adequacy of Curvilinear Score equating models: In D. Weiss (Ed.), *New horizons in testing* 147-128. New York: Academic.
- McKinky, R. & Kingston, N. (1987). Exploring the use of IRT Equating for the GRE Subject Test in Mathematics. Educational Testing Service, Princeton, N.J. 08541.
- Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458. *measurement* 58(3) 1-17
- Merten, T. (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Personality and Individual Differences*, 20, 281-291.
- Millman, J. & Arter, J.A. (1984). Issues in Item Banking. *Journal of Educational Measurement*, 21 (4), 312-330.
- Morris, L.N. (1982). On the Foundation of Test Equating. In P.W., Holland & D.B. Rubin (Eds.) *Test Equating* (pp. 169-191). New York: Academic Press.

- Morrison, C. A. & Fitzpatrick, S.J. (1992). Direct and Indirect Equating: A Comparison of Four Methods using the Rasch Model. Retrieved July 9, 2008 from http://eric.ed.gov/ERICdoes/data/ericdocs2sgl/contentstorage_01/0000019b/80/13/5d/d3.pdf.
- Moses, T., Yang, W. & Wilson, C. (2007). Using Kernel Equating to Assess Item Order Effects on Test Scores. *Journal of Educational Measurement*, 44(2): 157-178.
- Moulton, M. H. (2004) Weighting and Calibration: Merging Rasch Reading and Maths Subscale Measures into a Composite Measure. Retrieved May 12, 2009, from <http://www.aobfoundation.org>.
- Moulton, M. H. (2004). Weighting and calibration: Merging Rasch Reading and Maths subscale Measures into a composite measure. Retrieved May 12, 2009 from www.aobfoundation.org.
- Muraki, E. (1992). A generalized partial credit model: Application of the EM algorithm. *Applied Psychological Measurement*, 16(2), 159-172.
- National Teachers' Institutes (2005). Manual for re-training of primary school teachers on school based assessment. NTI press, Kaduna-Nigeria.
- Nenty, H.J. (1991). *Item Banking for Continuous Assessment*. A paper presented at the 7th Annual Conference of the National association of educational researchers and evaluators. Ahmadu Bello University, Zaria from 24th to 28th March.
- Nwana, O.C. (1981). *Introduction to Education Research*. Ibadan: Heinemann Educational Books Ltd.
- Nworgu, B. G. (2003). *Educational Measurement and Evaluation: Theory and practice (Revised Edition)*. University Trust Publisher, Nsukka-Enugu.
- Nworgu, B. G. (2006). *Educational Research, Issues and Methodology (2nd ed)*. University Trust Publisher, Nsukka-Enugu
- Obioma, G. (ND). *Continuous Assessment Practice of Primary and junior Secondary School Teachers in Nigeria*. Nigeria Educational Research and Development Council. (NERDC). Abuja. Nigeria.
- Ong, S. L. & Sireci, S.G. (2008). Using Bilingual Statements to Link and Evaluate Different Language Versions of an Exam. Retrieved March 27, 2009 from <http://www.teacher.org.cn/doc/ucedu20081105.pdf>.

- Onjewu, M. A. (2007). Assuring fairness in the continuous Assessment component of school Based Assessment practice in Nigeria. Paper Presented at the 33rd Annual Conference of the International Association for Educational Assessment. Baku, Azerbaijan.
- Onuka, A. U. (2003). *The Relevant of continuous Assessment in Instruction and learning in the school system*. Paper presented at the 31st Annual conference of the International Association for Educational Assessment Abuja, Nigeria.
- Onunkwo, G.I.N. (2002). *Fundamentals of educational measurement and evaluation*. Cape Publishers International: Owerri, Imo State.
- Owolabi, H.O. (2003). Antecedents of current procedures of evaluating learning outcomes in the Nigerian educational system. Retrieved September 16, 2011, From http://www.unilorin.edu.ng/publications/owolabiho/ANTECEDENTS_OF_CURRENT_PROCEDURES_OF_EVALUATION
- Petersen, S.N. (2008). A Dessension of Population Invariance of Equating. *Applied psychological measurement*, 32(1): 98-101 Retrieved April 3, 2009 from <http://apm.sagepub.com/cgi/content/abstract/32/1/98>.
- Peterson, N. S., Cook, L.L. & Stocking (1988). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics* 8(2), 137-156.
- Peterson, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling Norming and Equating. In RL Linn (Ed.) *Educational Measurement (3rd ed.)*, 221-262. New York: Macmillan.
- Peterson, N. S., Marco, G. L. & Stewart, E.E. (1982). *A Test of Adequacy of Linear score equating models*. In P.W. Holland & D.B. Rubin (Eds.). Test Equating 147-177. New York: Academic Press.
- Pinsonneault, T. B. (1996). Equivalency of computer-assisted paper-and-pencil administered version of the Minnesota Multiphasic Personality Inventory-2. *Computers in Human Behavior*, 12, 291-300.
- Pollit, A. B. (1993). *Items Banking in Primary Mathematics*. Edinburgh: Godfrey. Thomson Unit. University of Edinburgh.
- Ponocny, I. (2002). The Applicability of some IRT Models for Repeated Measurement Designs; Conditions, consequences and Goodness of fit test. *Methods of*

- psychological Research online* 7, 1-12. Retrieved June 9, 2008, from <http://www.nicd.edu.na>:
- Raju, N. S., Laffitte, L. J., & Byrne, B.M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517-529.
- Rapp, J. & Allallouf, A., (2002). *Evaluating cross lingual equating*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Schumacker, R. E. (2005). *Test Equating*. Retrieved March 21st 2009 from <http://www.appliedmeasurementassociates.com/white%20papers/TEST%EQUATING.Pdf>.
- Shoemaker, D. M. (1980). A Note on Allocating items to sub-test in multiple matrix sampling and estimates with the jackknife. *Journal of Educational Measurements* 3(2): 211-220.
- Silvestre-tipay, J. L. (2009) Item Response Theory and Classical Test Theory: An Empirical Comparison of Item/Person Statistics in a Biological Science Test. *International Journal of Educational and Psychological Assessment* 1 19-31
- Skaggs, G. (2005). Accuracy of Random Groups Equating with very small Samples. *Journal of Educational Measurement* 42(4): 309-330.
- Stage, C. (2003) *Classical Test Theory or item response theory*: The Swedish experience.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Svend, K. & Christensen K. B. (2002) Analysis of local Independence multidimensional in graphical loglinear Rasch Models. *Education and Psychological Measurement* 45 (6), 856-865.
- Svend, K. & Christensen, K.B. (2002). Analysis of local independence and multidimensionality in graphical linear Rasch models. *Educational and Psychological Measurement*, 37, 221-244.
- Tam. H. P., Griffith, W.D. and Li, Y.H. (1997). Equating Multiple Tests via an IRT Linking Design: Utilizing a Single set of Anchor Items with fixed Common Item Parameters during the Calibration Process. Retrieved July 9, 2008. from <http://eric.ed.gov/ERICwebportal/homeportal?nfpb=true ERICeXGSEARCH> Description= Result.

- Thissen, D and Wainer, H (2001) *Test Scoring*, Philadelphia, USA: Lawrence Erlbaum Associates.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* 67-113. Hillsdale NJ: Erlbaum.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of Item Response theory in the study of group differences in trace lines. In H. Wainer and H. I. Braun (Eds.), *Test Validity*. 147 ï 169. Hillsdale NJ: Erlbaum.
- Tianyou, W., Won-Chan, L., Brennan, R. L. & Kolen, M. J. (2006). A Comparison of Frequency Estimation and Chained Equipercentile Methods under the Common Item Non-Equivalent Groups Design. Retrieved March 21st, 2009 from <http://www.education.ulowa.edu/casmal/documents/17compparechfreg.rpt.pdf>
- Tianyou, W. & Brennan, R. L. (2006). A Modified Frequency Estimation Equating Method for the Common Item Non-Equivalent Groups Design. Retrieved March 21, 2009. from <http://www.education.ucowa.edu/casma/documents/19modifreg.pdf>.
- Tim, M. (2008). Using Kernel Equating Method of Test Equating for Estimating the Standard Errors of Population Invariance Measures. *Journal of Educational and Behavioural Statistics*. 22(2): 137-157.
- Ugwuanyi, C. L. & Ugodulunwa, C.A. (1999). *Understanding Educational Evaluation* Jos: Anieh Nig. Ltd.
- Usman, K.O. (2002). A Review of Studies on Process Error by students in solving mathematical problems. *Journals of Nigeria Education Research Association* 15(1), 76-82.
- Van der Linden W. J. (2006). Equating scores from Adaptive to Linear Tests. *Applied Psychological Measurement*, 30(6), 493-508.
- Van der Linden, W. J. & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item Response Theory*. New York: Springer-Verleg.
- Van der, Lindern, W. J. (2006). Equating error in observed score equating. *Applied Psychological Measurement*, 30(5), 355-378.
- Von Davier, A. A. & Ricker, K. L. (2006). The role of anchor test in the non-equivalent groups design. Unpublished research report.

- Von Davier, A. A. & Kong, N. (2005). A unified Approach to Linear Equating for the Nonequivalent groups Designs. *Journal of Educational and Behavioural Statistics*, 30: 313-342.
- Von Davier, A. A., Holland, P.W. & Thayer, D.T. (2004). The chain and post-stratification method for observed-score equating: Their Relationship to population-Invariance. *Journal of Educational Measurement*, 41(1), 15-32.
- Von Davier, A. A., Holland, P.W. & Thayer, D.T. (2004). *The Kernel Method of Test Equating*. New York. Springer-Verlag.
- Von Davier, A.A., Holland, P.W., Livingston, S.A., Casabianca, J., Grant, M.C. & Martin, K. (2006). An Evaluation of the kernel equating method in a non-equivalent groups design with an external anchor. A special study with pseudo-test from real test data AERA/NCME paper presented in 2006.
- Wang, T., Won-Clan, L, Brennan, R.L. & Kolen, M.J. (2006). A Comparison of the Frequency Estimation and Chained Equapercentile Method under the Common-Item Non-Equivalent Groups Design: Retrieved March 21, 2009 from <http://www.education.ulowa.edu/casma/documents/17comparechfreq.rpt.pdf>.
- Warm, T.A. (1998). *A primer of item Response Theory*. Technical Report, No 941078. Oklahoma City. USA Coast Guard Institute.
- Wendy, Y. (2002). *Scaling and Equating*. A paper presented at the New York Technical Conference, October 12, 2002. ETS. Retrieved May 18, 2008 from <http://www.p12.nysed.gov/osa/assesspubs/pubsearch/scalingandequating.pdf>
- Wen-Ling, Y., & Rui, G.F. (2008). Invariant of Score Linking Gcross gender Groups for Forms of a Testlet-Based College Level Examination Program. *Examination Measurement*, 32(1): 45-81.
- Wilberg, M. (2004). *Classical Test Theory VS. Item Response Theory – An Evaluation of the Theory test in Swedish Driving i License Test* (Em, No.50), Sweden: UMED University.
- Wilcox, R. R. (1987). Some Empirical and Theoretical Results on an Answer-Until-Correct Scoring Procedure. *British Journal of Mathematics and Statistics Psychology* 35(2): 57-70.
- Woldback, T. (1998). Basic concepts in modern methods of Test Equating.

- Wright, B. & Stone, M. (1999). *Measurement Essentials* (2nd Ed.). Wilmington, Delaware, WIDE RANGE INC.
- Wright, B. D. & Bell, F. (1984). *Best Design: A Handbook for Rasch Measurement*. Chicago: Scientific Press.
- Wright, B. D. & Panchapakesan, N. (1996). A Procedure for Sample-Free Item analysis. *Journal of Educational and Psychological Measurement*, 29(3): 23-48.
- Yang, W. (1997). The Effect of Content Mix and Equating Method on the Accuracy of Test Equating using Anchor. Item Design. Retrieved July 9, 2008 from <http://eric.ed.gov/ERICwebportal/homeportal?nfpb=true ERICeXGSEARCH>
Description= Result
- Yang, W. L., & Gao, R. (2008). Invariance of Score Linkings Across Gender Groups for Forms of a Testlet-Based College-Level Examination Program Examination. *Applied Psychological Measurement*, 32, 45-61.
- Ye, T. & Kolen, M. J. (2005). Assessing Equating Result on Different Equating Criteria *Applied Psychological Measurement*. 29(6): 418-432. Retrieved April 2, 2009 from <http://apm.sagepub.com/cgi/content/abstract/29/6/418>.
- Yen, W. M. (1993). Scaling performance Assessment: Strategies for Managing LID. *Journal of Educational Measurement*, 30(3), 187-213.
- Yen, W. M. & Ferrera, S. (1997). The Maryland School Assessment Program: Performance Assessment with Psychometric quality suitable for high-stake usage. *Educational and Psychological Measurement*, 57, 60-84.
- Zeng, L. (1991). *Standard Error of Linear Equation for the Single-group design*. ACT Research Report series 91-4
- Zwick, R., Thayer, D. T., and Mazzeo, J. (1997). Descriptive and Inferential Procedures for assessing differential item functioning in polytomous items. *Applied Measurement in education*, 10(4), 321 ï 344.
- Zwick, R. (1991). Effect of item order and context on estimation of NAEP Reading Proficiency. *Educational Measurement, Issues and Practice*, 16(3), 10-16.

APPENDIX A

Department of Science Education
Faculty of Education
University of Nigeria,
Nsukka.
12th Dec. 2009

Dear Sir/Madam,

REQUEST FOR VALIDATION OF INSTRUMENT

I am a Ph.D student of the Department of Science Education of the University of Nigeria, Nsukka. I am carrying out a study on the "Relative efficiency of test scores equating methods in the comparison of student's continuous assessment measures". In pursuance of this research, I have developed two parallel tests. Each test is made up of 40 multiple choice test items, in which 20 of the items are common to both tests.

I solicit for your assistance to help me validate the instrument. You are expected to:

- judge the adequacy of the items of the instruments
- undertake a systematic comparison of the test content with subject content to check if the test content adequately sampled the subject content
- say if the items on the test can be used to realize the objectives of this study
- say whether the instruments leave out any important behaviour or lay too much emphasis on one particular aspect of the content of the syllabus
- assess the brevity and unambiguity of the items
- solve the items of the test and indicate the correct options.
- Give your comment about the instrument.

Attached to this letter are: the instrument, purpose of the study, research questions and hypotheses.

Thanks for your anticipated cooperation.

Yours faithfully,

Agah, John Joseph
PG/Ph.D/06/40715

APPENDIX B

Department of Science Education
Faculty of Education
University of Nigeria,
Nsukka.
2nd Jan. 2010.

Dear Sir/Madam,

LETTER OF INTRODUCTION

I am a Ph.D student of the Department of Science Education of the University of Nigeria, Nsukka. I am carrying out a study on "Relative efficiency of test scores equating methods in the comparison of student's continuous assessment measures". In pursuance of this research goal, your maximum cooperation is solicited to make this study worthwhile. You are expected to supply me with the cumulative continuous assessment score for all SSIII students in your school, and also get them ready for a 2 hours test on mathematics.

The information supplied by your school will be use by the researcher only, and hence the result will not have any negative effect on your school.

Thanks for your anticipated cooperation.

Yours faithfully,

Agah, John Joseph
PG/Ph.D/06/40715

APPENDIX C

MATHEMATICS ACHIEVEMENT TESTS (MAT) PAPER A¹

INSTRUCTION

- i. Time Allowed: 1 hr 30 min.
- ii. Use ball-pen only
- iii. Ensure that you fill / tick the blank spaces / box provided
- iv. Read each question carefully before answering it.
- v. Answer all questions, each question carry equal mark
- vi. Circle the correct answer from among the four options A ï D provided.
- vii. You are free to change your choice of option by just canceling your previous choice and circling the new one
- viii. Please work independently and do not discuss anything with anybody
- ix. You are free to seek for clarification from the supervisor /invigilator

Section A

Personal Information of Examinees

Instruction: Kindly complete the space below and tick where necessary

Name of School:é ..

Class:é ..

Date:é é é é é é é é é é é é . Sex: Male ☐ Female ☐

Candidate Number:é ..

State:é ..

Continuous Assessment Score in Maths from School:é é é é é é é é ..

Section B
PAPER A¹

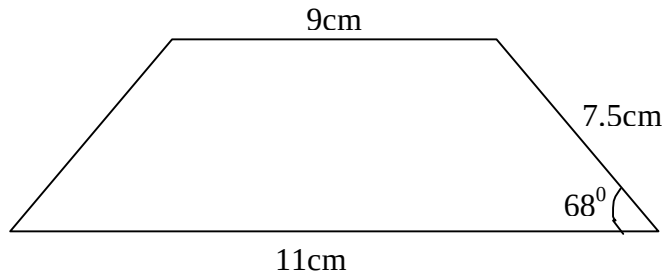
1. The intersection of two sets is a set which:
(A) contains all the element being discussed (B) contains all the elements common to two sets under consideration (C) contains countable elements (D) contains uncountable elements.
2. Arrange $11/13$, $5/15$, $6/8$ and $1/3$ in ascending order of magnitude
(A) $1/3$, $6/8$, $8/15$, $11/13$ (B) $1/3$, $8/15$, $6/8$, $11/13$
(C) $6/8$, $1/3$, $8/15$, $11/13$ (D) $6/8$, $1/3$, $11/13$, $8/15$
3. Simplify $4\sqrt{98} - 3\sqrt{72} + 5\sqrt{18}$
(A) $25\sqrt{2}$ (B) $15\sqrt{2}$ (C) $18\sqrt{2}$ (D) $28\sqrt{2}$
4. Change 7_{10} to base 8
(A) 701_8 (B) 801_8 (C) 107_8 (D) 108_8
5. In a class of 42 students, each student offers at least one of mathematics and physics. If 22 students offer physics and 28 offer mathematics. How many students offer mathematics only
(A) 28 (B) 14 (C) 21 (D) 20
6. A primary school teacher earned ₦8,400 in 2007 and earned 18% more in 2008. If the teacher paid a tax of $12\frac{1}{2}\%$, Find the tax paid in 2008.
(A) ₦826 (B) ₦912 (C) ₦1239 (D) ₦1512
7. If $4^{6n-2} = 1$, Find the value of n
(A) $1/2$ (B) $1/3$ (C) $1/4$ (D) $1/6$
8. The 3rd term of an AP is 8 and the 16th term is 47. What is the sum of the first term and common difference
(A) 2 (B) 3 (C) 4 (D) 5
9. The coefficient of P in the expression $2p \div 3q + 4p$ is
(A) 6 (B) 4 (C) 3 (D) 2
10. Given that $a \div c = d$, express b in terms of a, c and d
(A) $(a \div c)/d$ (B) $(a + c)/d$ (C) $(a + c)/d$ (D) $(d + c)/a$
11. A mother is four times as old as her daughter. In four years time she will be three times as old. What are their ages now
(A) 12 and 32 (B) 8 and 32 (C) 7 and 36 (D) 12 and 36
12. A lorry carrying cements weight 10,000kg, when loaded. The cements weights seven times as much as the lorry. What is the weight of the lorry?
(A) 1250kg (B) 2000kg (C) 2250kg (D) 2500kg
13. Solve for m in the equation $4m^2 - 9m + 5 = 0$
(A) 4,5 (B) $4/5$, 5 (C) $5/4$, 1 (D) $5/4$, 9
14. What is the quadratic equation whose root is given as -3 and +6
(A) $x^2 - 6x - 18 = 0$ (B) $x^2 + 3x - 18 = 0$ (C) $x^2 - 9x - 18 = 0$ (D) $x^2 - 3x - 18 = 0$
15. The length of a rectangle are $4x + 3$ and $x + 6y$ while its width are $4x - y$ and $3x - 1$, what is the value x and y
(A) 4,3 (B) 3,2 (C) 2, 4 (D) 5,4
16. If a = 1 and b = 3 solve for x in the equation $\frac{a}{a-x} = \frac{b}{x-b}$

- (A) $\frac{3}{2}$ (B) $\frac{3}{4}$ (C) $\frac{2}{3}$ (D) $\frac{4}{3}$

17. The area of a triangle is given as: $= \sqrt{s(s-a)(s-b)(s-c)}$ if $a = 15$, $b = 4$ and $c = 13$, what is the value of s

- (A) 32 (B) 19 (C) 17 (D) 16

18. Find the area of the figure below



- (A) 27.5cm^2 (B) 69.5cm^2 (C) 67.5cm^2 (D) 82.5cm^2

19. Calculate the length of an arc of a circle of radius 7cm which subtends an angle of 84° at the centre of the circle.

- (A) 10.3cm (B) 14.4cm (C) 17.3cm (D) 22.4cm

20. A cuboid has the length 13cm, width 9 cm and height 8cm. Calculate the length of its diagonal.

- (A) 93.6cm (B) 68.6cm (C) 40.0cm (D) 17.7 cm

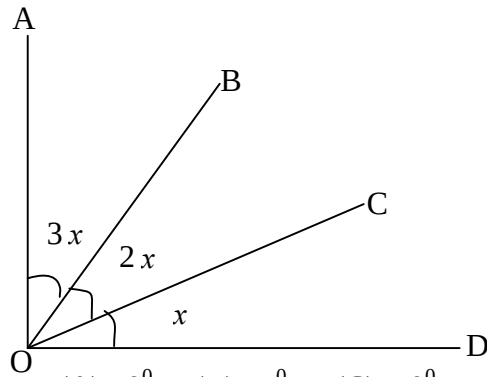
21. What is the total surface area of a cone if its base radius is 6cm and height 9cm

- (A) 204.0cm^2 (B) 214.0cm^2 (C) 317.2cm^2 (D) 339.4cm^2

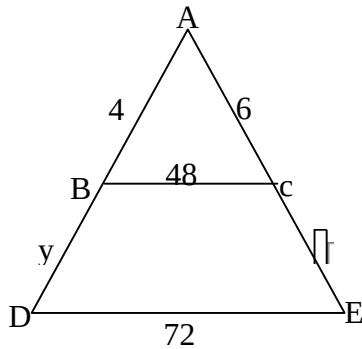
22. Which of the following expression is true if θ is an obtuse angle?

- (A) $180^\circ < \theta < 360^\circ$ (B) $90^\circ < \theta < 180^\circ$ (C) $\theta = 180^\circ$ (D) $\theta = 90^\circ$

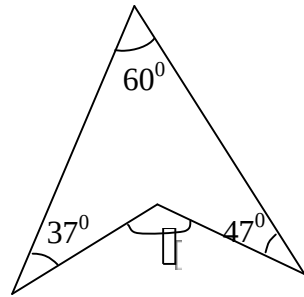
23. If $\overline{AO}$ is perpendicular to $\overline{OD}$, find the value of x in the figure below



- (A) 10° (B) 15° (C) 20° (D) 25°
 24. Each angle of a regular polygon is 165° . How many sides has the polygon?
 (A) 15 (B) 24 (C) 30 (D) 36.
 25. What is the value of Y in the diagram below?



- (A) 2 (B) 6 (C) 9 (D) 12
 26. Find the value of $\angle$ in the diagram below



- (aA) 84° (B) 103° (C) 113° (D) 150°
 27. The length of the chord of a circle is 16cm. if the chord is 6cm from the centre of the circle, calculate the radius of the circle.
 (A) 14cm (B) 10 cm (C) 8cm (D) 6cm
 28. Given that $\sin y = 5/13$, without using table, what is the value of $\cos Y$? (a) $12/13$
 (b) $11/13$ (c) $7/13$ (d) $5/12$
 29. $\sin 45^\circ$ is the same as;
 (A) $1/2$ (B) $\sqrt{2}/2$ (C) $\sqrt{3}/2$ (D) $-1/\sqrt{2}$

30. A hunter decides to climb a tree of height 10cm to await an approaching animal. At what angle to the tree will he aim at the animal in order to kill it at a distance of 14m from the tree?
 (A) 54.46° (B) 37.53° (C) 35.37° (D) 27.53°
31. Solve the equation $\sin \check{U} = \cos 2 \check{U}$
 (A) 30° (B) 45° (C) 60° (D) 75°
32. The difference between the highest and the lowest value in a distribution is called
 (A) Range (B) Class boundary (C) Class limit (D) Class mark.
33. What is angle B of a triangle with the following dimensions $a=12\text{cm}$, $b=11\text{cm}$ and $c=9\text{cm}$
 (A) $72^{\circ} 58^1$ (B) $68^{\circ} 49^1$ (C) $64^{\circ} 22^1$ (D) $61^{\circ} 13^1$
34. Find the number of values in a distribution whose mean is 34 and sum of all values is 408
 (A) 20 (B) 12 (C) 13 (D) 8
35. Which of the following statements clearly define a histogram?
 (i) Rectangular bars with equal width
 (ii) Rectangular bars with no gap between the bars
 (iii) The area of each rectangular bar of a histogram is proportional to the corresponding frequency
 (A) I and II (B) I and III (C) II and III (D) I, II and III
36. What is the variance of the following set of scores: 3,4,5,6,8 and 10
 (A) 2.38 (B) 3.00 (C) 5.67 (D) 6.00

Use the table below to answer question 37 i 38

Score	4	7	8	11	13	18
Frequency	3	5	2	7	2	1

37. The square of the mode of the above distribution is
 (A) 25 (B) 36 (C) 49 (D) 64
38. What is the probability of having the score of 11 and above?
 (A) $7/20$ (B) $3/20$ (C) $1/5$ (D) $1/2$
39. Two hunters aimed at a target, the probability that the first hit it is $2/5$ while the probability that the second hit is $2/7$. What is the probability that one of them will hit the target?
 (A) $6/35$ (B) $17/35$ (C) $12/35$ (D) $9/35$
40. Two dice are tossed once, what is the probability of obtaining a sum of 7
 (A) $5/36$ (B) $7/36$ (C) $1/6$ (D) $5/6$

APPENDIX D
MATHEMATICS ACHIEVEMENT TESTS (MAT) PAPER B¹

INSTRUCTION

- i. Time Allowed: 1 hr 30 min.
- ii. Use ball-pen only
- iii. Ensure that you fill / tick the blank spaces / box provided
- iv. Read each question carefully before answering it.
- v. Answer all questions, each question carry equal mark
- vi. Circle the correct answer from among the four options A ï D provided.
- vii. You are free to change your choice of option by just canceling your previous choice and circling the new one
- viii. Please work independently and do not discuss anything with anybody
- ix. You are free to seek for clarification from the supervisor /invigilator

Section A

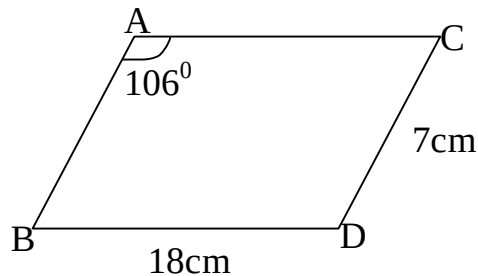
Name of School:é ..
Class:é ..
Date:é é é é é é é é é Sex: Male ☐ Female ☐
Candidate Number:é ..
State:é ..
Continuous Assessment Score in Maths from School:é é é é é é é é é ..

Section B

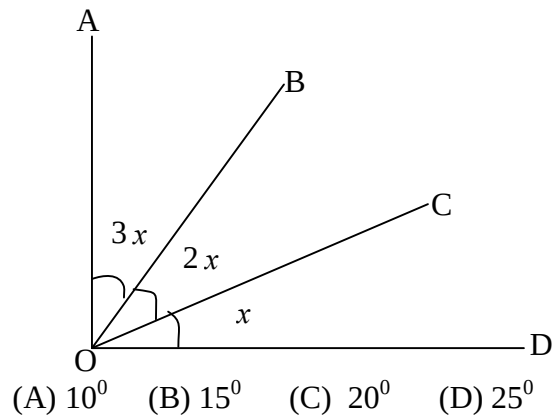
PAPER B¹

- (1) The intersection of two sets is a set which:
 (A) contains all the element being discussed (C) contains all the elements common to two sets under consideration (C) contains countable elements (D) contains uncountable elements .
- (2) Arrange $\frac{15}{19}$, $\frac{7}{8}$, $\frac{10}{11}$, and $\frac{3}{4}$ in descending order
 (A) $\frac{15}{19}$, $\frac{3}{4}$, $\frac{7}{8}$ and $\frac{10}{11}$
 (B) $\frac{3}{4}$, $\frac{15}{19}$, $\frac{7}{8}$ and $\frac{10}{11}$
 (C) $\frac{7}{8}$, $\frac{10}{11}$, $\frac{15}{19}$ and $\frac{3}{8}$
 (D) $\frac{3}{4}$, $\frac{7}{8}$, $\frac{15}{19}$ and $\frac{10}{11}$
- (3) Simplify $4\sqrt{98} - 3\sqrt{72} + 5\sqrt{18}$
 (A) $25\sqrt{2}$ (B) $15\sqrt{2}$ (C) $18\sqrt{2}$ (D) $28\sqrt{2}$
- (4) Convert 213_5 to base 10
 (A) 38 (B) 48 (C) 58 (D) 69
- (5) In a class of 42 students, each student offers at least one of mathematics and physics. If 22 students offer physics and 28 offer mathematics, how many students offer mathematics only
 (A) 28 (B) 14 (C) 21 (D) 20
- (6) A primary school teacher spent 24% of his salary on food, 25% of the rest on rent and the balance left is N3420. How much was spent on food?
 (A) N821 (B) N1,285 (C) N1,676 (D) N1,440
- (7) If $4^{6n-2} = 1$, Find the value of n
 (A) $\frac{1}{2}$ (B) $\frac{1}{3}$ (C) $\frac{1}{4}$ (D) $\frac{1}{6}$
- (8) The 28th term of an AP is -5. Find its common difference if the first term is 31.
 (A) $-5\frac{3}{5}$ (B) $-6\frac{1}{5}$ (C) $-3\frac{1}{3}$ (D) $-1\frac{1}{3}$
- (9) The coefficient of P in the expression $2p^2 - 3q + 4p$ is
 (A) 6 (B) 4 (C) 3 (D) 2
- (10) Given that $V = \frac{1}{2}Lbh$, express b in terms of V, L and h
 (A) $\frac{Lh}{V}$ (B) $\frac{Lh}{2V}$ (C) $\frac{V}{Lh}$ (D) $\frac{2V}{Lh}$

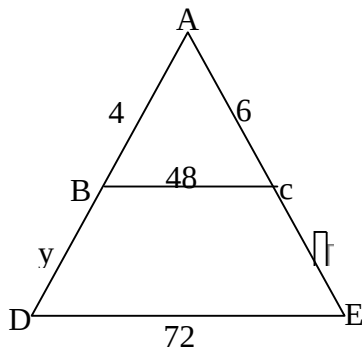
- (11) A mother is four times as old as her daughter. In four years time she will be three times as old. What are their ages now
 (A) 12 and 32 (B) 8 and 32 (C) 7 and 36 (D) 12 and 36
- (12) A lorry carrying timber weights 8220 kg when loaded. The timbers weight four times as much as the lorry. What is the weight of the lorry?
 (A) 1644kg (B) 2055kg (C) 2740kg (D) 4110kg
- (13) Solve for m in the equation $4m^2 - 9m + 5 = 0$
 (A) 4,5 (B) $\frac{4}{5}$, 5 (C) $\frac{5}{4}$, 1 (D) $\frac{5}{4}$, 9
- (14) What is the quadratic equation whose root are given as +4 and -5
 (A) $x^2 - 9x + 20$ (B) $x^2 + 9x - 20$ (C) $x^2 = x - 20$ (D) $x^2 + x - 20$
- (15) The length of a rectangle are $4x + 3$ and $x + 6$ while its width are $4x - y$ and $3x - 1$, what is the value x and y
 (A) 4,3 (B) 3,2 (C) 2, 4 (D) 5,4
- (16) IF $a = 2$, $p = 8$ and $q = 4$, find the value of x in the equation $\frac{1+ax}{1-ax} = \frac{p}{q}$
 (A) $\frac{1}{4}$ (B) $\frac{1}{6}$ (C) $\frac{1}{8}$ (D) $\frac{1}{10}$
- (17) The are of a triangle is given as: $= \sqrt{s(s-a)(s-b)(s-c)}$ if $a = 15$, $b = 4$ and $c = 13$, what is the value of s
 (A) 32 (B) 19 (C) 17 (D) 16
- (18) Find the area of the figure below



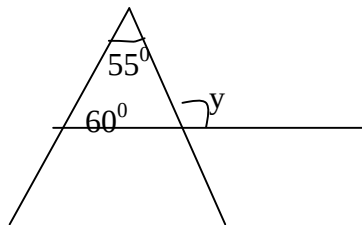
- (A) 126.9 cm^2 (B) 121.1 cm^2 (C) 116.4 cm^2 (D) 108.9 cm^2
- (19) Calculate the length of an arc of a circle of radius 7cm which subtends an angle of 84° at the centre of the circle.
 (A) 10.3cm (B) 14.4cm (C) 17.3cm (D) 22.4cm
- (20) A cuboids has the length 12cm, width 7cm and height 18cm. Calculate its volume.
 (A) 392 cm^3 (B) 492 cm^3 (C) 572 cm^3 (D) 672 cm^3
- (21) What is the total surface are of a cone if its base radius is 6cm and height 9cm
 (A) 204.0 cm^2 (B) 214.0 cm^2 (C) 317.2 cm^2 (D) 339.4 cm^2
- (22) Which of the following pairs of angle that will give you an obtuse angle when added
 (A) 120 ; 5800 (B) 460 ; 580 (C) 134 ; 580 (D) 220° ; 58°
- (23) If $\overline{AO}$ is perpendicular to $\overline{OD}$, find the value of x in the figure below



- (24) What is the sum of interior angles in a hexagon
 (A) 540° (B) 720° (C) 900° (D) 1080°
- (25) What is the value of Y in the diagram below?



- (A) 2 (B) 6 (C) 9 (D) 12
- (26) Find the value of the exterior angle mark y in the figure below



- (A) 65° (B) 55° (C) 120° (D) 115°
27. The length of the chord of a circle is 16cm. if the chord is 6cm from the centre of the circle, calculate the radius of the circle.
 (A) 14cm (B) 10 cm (C) 8cm (D) 6cm
- (28) Given that $\tan\theta = \frac{8}{15}$, without using table, what is the value of $\cos\theta$
 (A) $\frac{23}{17}$ (B) $\frac{7}{17}$ (C) $\frac{12}{15}$ (D) $\frac{15}{17}$

- (29) $\sin 45^\circ$ is the same as; (a) $1/2$ (b) $\sqrt{2}/2$ (c) $\sqrt{3}/2$ (d) $-1/\sqrt{2}$
- (30) A pole 12 cm long leans against a vertical wall with its foot 3m from the wall. Calculate the angle which the pole makes with the ground.
(A) 25.52° (B) 36.52° (C) 75.52° (D) 86.52° .
- (31) Solve the equation $\sin \tilde{U} = \cos 2 \tilde{U}$
(A) 30° (B) 45° (C) 60° (D) 75°
- (32) The sum of all values the data set has, divided by the number of values contained in the data set is called
(A) Mean (B) Median (C) Mode (D) Frequency
- (33) What is angle B of a triangle with the following dimensions a=12cm, b=11cm and c=9cm
(A) $72^\circ 58'$ (B) $68^\circ 49'$ (C) $64^\circ 22'$ (D) $61^\circ 13'$
- (34) What is the mean of the distribution of 42, 56, 59, 38, 41, 86 and 56.
(A) 52 (B) 54 (C) 56 (D) 86
- (35) Which of the following statements clearly define a histogram?
(i) Rectangular bars with equal width
(ii) Rectangular bars with no gap between the bars
(iv) The area of each rectangular bar of a histogram is proportional to the corresponding frequency
(A) I and II (B) I and III (C) II and III (D) I, II and III
- (36) What is the standard deviation of the following set of scores: 10, 8, 6, 5, 4 and 3
(A) 6.00 (B) 5.67 (C) 2.38 (D) 2.00

Use the table below to answer question 37 to 38

Score	4	7	8	11	13	18
Frequency	3	5	2	7	2	1

- (37) The square of the mode of the above distribution is
(A) 25 (B) 36 (C) 49 (D) 64
- (38) What is the probability of having the score 7 and below
(A) $\frac{7}{20}$ (B) $\frac{3}{20}$ (C) $\frac{3}{5}$ (D) $\frac{2}{5}$
- (39) Two hunters aimed at a target, the probability that the first hit it is $2/5$ while the probability that the second hit is $2/7$. What is the probability that one of them will hit the target?
(A) $6/35$ (B) $17/35$ (C) $12/35$ (D) $9/35$
- (40) Two fair dice are tossed once, what is the probability of obtaining a pair of Odd number
(A) $\frac{1}{6}$ (B) $\frac{1}{5}$ (C) $\frac{1}{4}$ (D) $\frac{1}{2}$

APPENDIX E
Key to Test A¹ and B¹

Key to test A ¹				Key to Test B ¹			
1	B	21	C	1	B	21	C
2	B	22	B	2	B	22	D
3.	A	23	B	3.	A	23	B
4	C	24	B	4	C	24	B
5.	D	25	D	5.	D	25	D
6	D	26	D	6	D	26	D
7	C	27	B	7	C	27	B
8	D	28	D	8	D	28	D
9	A	29	B	9	A	29	B
10	D	30	C	10	D	30	C
11	B	31	A	11	B	31	A
12	A	32	A	12	A	32	A
13	C	33	D	13	C	33	D
14	D	34	B	14	D	34	B
15	B	35	D	15	B	35	D
16	B	36	C	16	B	36	C
17	D	37	C	17	D	37	C
18	B	38	D	18	B	38	D
19	A	39	D	19	A	39	D
20	D	40	C	20	D	40	C

APPENDIX F
VALIDATOR'S COMMENT:

NAME OF VALIDATOR:.....
SIGNATURE:.....
DATE:.....

APPENDIX G

Table 8 Table of Specification

Cognitive level	Knowledge 8%	Comprehension 12%	Application 23%	Analysis 24%	Synthesis 18%	Evaluation 15%	Total 100%
Content							
Number and Numeration 20%	1 (1)* (1)**	1 (2)* (2)**	2 (3,4)* (3,4)**	2 (5,6)* (5,6)**	1 (7)* (7)**	1 (8)* (8)**	8
Algebraic process 18%	1 (9)* (9)**	1 (10)* (10)**	2 (11,12)* (11,12)**	2(13,14)* (13,14)**	1 (15)* (15)**	1 (16)* (16)**	8
Mensuration 14%	-	1 (17)* (17)**	1 (18)* (18)**	1 (19)* (19)**	1 (20)* (20)**	1 (21)* (21)**	5
Geometry 16%	-	1 (22)* (22)**	1 (23)* (23)**	2 (24,25)* (24,25)**	1 (26)* (26)**	1(27)* (27)**	6
Trigonometry 14%	-	1(29)* (29)**	1 (28)* (28)**	1 (30)* (30)**	1 (31)* (31)**	1 (33)* (33)**	5
Statistics and Probability 18%	1 (32)* (32)**	1(35)* (35)**	2 (34,36)* (34,36)**	2 (37,38)* (37,38)**	1 (39)* (39)**	1 (40)* (40)**	8
Total 100%	3	6	9	10	6	6	40

* Paper A Questions, ** Paper B Questions

APPENDIX H

Table 7: Item Facility Index and Discrimination Index for Paper A¹

Items	P	D	Item	P	D
1	0.80	0.40	21	0.55	0.50
2	0.75	0.10	22	0.65	0.10
3	0.50	0.40	23	0.55	0.30
4	0.50	0.20	24	0.60	0.00
5	0.35	0.30	25	0.35	0.50
6	0.45	0.10	26	0.50	0.20
7	0.55	0.30	27	0.50	0.40
8	0.60	0.60	28	0.50	0.40
9	0.70	0.00	29	0.40	0.60
10	0.60	0.20	30	0.40	0.40
11	0.50	0.00	31	0.40	0.20
12	0.55	0.30	32	0.40	0.60
13	0.50	0.60	33	0.40	0.60
14	0.60	0.40	34	0.40	0.40
15	0.45	0.10	35	0.50	0.60
16	0.65	0.40	36	0.50	0.20
17	0.70	0.40	37	0.35	0.50
18	0.30	0.20	38	0.20	0.40
19	0.55	0.50	39	0.40	0.80
20	0.50	0.20	40	0.25	0.50

P = Item Facility Index

D = Item discrimination index

APPENDIX I

Table 8: Item Facility Index and Discrimination Index for Paper B¹

Items	P	D	Item	P	D
1	0.80	0.30	21	0.55	0.50
2	0.75	0.30	22	0.65	0.30
3	0.65	0.50	23	0.65	0.30
4	0.65	0.50	24	0.60	0.60
5	0.65	0.50	25	0.45	0.50
6	0.60	0.40	26	0.45	0.30
7	0.50	0.40	27	0.55	0.50
8	0.60	0.60	28	0.55	0.50
9	0.70	0.60	29	0.45	0.70
10	0.55	0.30	30	0.45	0.30
11	0.45	0.30	31	0.45	0.50
12	0.50	0.20	32	0.40	0.60
13	0.45	0.40	33	0.55	0.70
14	0.45	0.80	34	0.50	0.60
15	0.55	0.30	35	0.55	0.50
16	0.65	0.30	36	0.55	0.50
17	0.55	0.50	37	0.40	0.60
18	0.45	0.50	38	0.40	0.60
19	0.55	0.50	39	0.30	0.60
20	0.50	0.40	40	0.35	0.70

P = Item Facility Index

D = Item discrimination index

APPENDIX J
Cross River State Population Distribution for Twenty Three Schools

S/N	Name of School	Male	Female	Total
1	Army Day Secondary School Ikot Ansa	83	41	124
2	Govt. Girls Secondary School Big Qua Town	-	139	139
3	Govt. Secondary School Atu	145	130	275
4	Govt. Secondary School Uwanse	56	72	128
5	Akwa High Secondary School Ifiang Nsung	29	35	64
6	Community Secondary School Itiate	75	65	140
7	Dan Archibong Memorial Sec. School	41	42	83
8	Govt. Day Secondary School Akamkpa	58	49	107
9	Community Secondary School Adim	49	17	66
10	Biase Secondary School Ehom	31	42	73
11	Govt. Secondary School Ikom	50	50	100
12	Community Secondary School Akparabong	35	40	75
13	Community Secondary School Etomi	46	38	84
14	Boki Community Sec. School Okundi	48	38	84
15	Agbo Comprehensive Sec. Sch. Egboronyi	39	57	96
16	Girls Secondary School Ugep	-	63	63
17	Secondary School Idomi	41	55	96
18	Mbembe Commercial Secondary School	50	69	119
19	Bedia Secondary School Obudu	41	33	74
20	Government Secondary School Obudu	42	36	78
21	Basang Comprehensive Secondary School	28	33	61
22	Army Day Secondary School Ogoja	21	32	53
23	Yala Sec. Commercial School Okpoma	30	26	56
	TOTAL	1038	1128	2266

APPENDIX K
Rivers State Population Distribution for Twenty Two Schools

S/N	Name of School	Male	Female	Total
1	Community Secondary School Abuloma	85	82	167
2	Government Secondary School Elekahia	80	74	154
3	Community Secondary School Alesa	71	82	153
4	Government Secondary School Eneka	72	69	141
5	Community Secondary School Ubima	22	28	50
6	Community Secondary School Obele	39	43	82
7	Government Secondary School Emohua	23	23	46
8	Community Secondary School Ulakwo	39	52	91
9	Government Secondary Tec. College Okehi	13	16	29
10	Community Secondary School Okoroagu	38	37	75
11	Community High School Nkoro	23	36	59
12	Community Secondary School Oyigbo	67	89	156
13	Government Secondary School Ngo	27	29	56
14	Community Secondary School Unyeachi	29	17	46
15	Government Secondary School Nkpor	42	36	78
16	Community Secondary School Elelenwo	84	-	84
17	Government Secondary School Kpite	21	33	54
18	Community Secondary School Bunu-Tai	12	20	32
19	Community Secondary School Bori	32	21	53
20	Government Secondary School Kaa	16	30	46
21	Girl Community Secondary School Ndashi	-	96	96
22	Community Secondary School Omadame	18	26	44
	TOTAL	853	939	1792

APPENDIX L
Sample Distribution of the Examinees in twenty three Schools from
Cross River State

S/N	Name of School	Male	Female	Total
1	Army Day Secondary School Ikof Ansa	56	27	83
2	Govt. Girls Secondary School Big Qua Fovon	-	93	93
3	Govt. Secondary School Atu	97	87	184
4	Govt. Secondary School Uwamse	38	48	86
5	Akwa High Secondary School Ifiang Nsung	19	24	43
6	Community Secondary School Itiate	50	44	94
7	Dan Archibong Memorial Sec. School	27	28	55
8	Govt. Day Secondary School Akamkpa	38	33	71
9	Community Secondary School Adim	33	11	44
10	Biase Secondary School Ehom	21	28	49
11	Govt. Secondary School Ikom	33	34	67
12	Community Secondary School Akparabong	23	27	50
13	Community Secondary School Etomi	31	25	56
14	Boki Community Secondary School Okundi	32	43	75
15	Agbo Comprehensive Sec. School. Egboronyi	26	38	64
16	Girls Secondary School Ugep	-	42	42
17	Secondary School Idomi	27	37	64
18	Mbembe Commercial Secondary School	34	46	80
19	Bedia Secondary School Obudu	27	22	49
20	Government Secondary School Obudu	28	24	52
21	Basang Comprehensive Secondary School	19	22	41
22	Army Day Secondary School Ogoja	14	21	35
23	Yala Sec. Commercial School Okpoma	20	17	37
	TOTAL	1693	821	1514

APPENDIX M
Sample Distribution of Examinees in Twenty Two Schools from Rivers State

S/N	Name of School	Male	Female	Total
1	Community Secondary School Abuloma	66	64	130
2	Government Secondary School Elekalia	62	58	120
3	Community Secondary School Alesa	55	64	119
4	Government Secondary School Eneka	56	53	109
5	Community Secondary School Ubima	17	22	39
6	Community Secondary School Obele	30	34	64
7	Government Secondary School Emohua	18	18	36
8	Community Secondary School Ulakwo	30	41	71
9	Government Technical College Okehi	10	12	22
10	Community Secondary School Okoroagu	29	29	58
11	Community High School Nkoro	18	28	46
12	Community Secondary School Oyigbo	52	69	121
13	Government Secondary School Ngo	21	22	43
14	Community Secondary School Unyeada	23	13	36
15	Government Secondary School Kpor	32	28	60
16	Community Secondary School Elelenwo	65	-	65
17	Government Secondary School Kpite	16	26	42
18	Community Secondary School Bunu-Tai	9	16	25
19	Community Secondary School Bori	25	16	41
20	Government Secondary School Kaa	13	23	36
21	Girls Community Secondary School Ndashi	-	74	74
22	Community Secondary School Omadame	14	20	34
	TOTAL	661	730	1391

APPENDIX N
Computation of the Reliability of Test A¹

S/N	R	W	p	q	pq
1	23	7	0.77	0.23	0.18
2	20	10	0.67	0.33	0.22
3	15	15	0.50	0.50	0.25
4	17	13	0.57	0.43	0.25
5	16	14	0.53	0.47	0.25
6	9	21	0.30	0.70	0.21
7	17	13	0.57	0.43	0.25
8	15	15	0.50	0.50	0.25
9	18	12	0.60	0.40	0.24
10	19	11	0.63	0.37	0.23
11	14	16	0.47	0.53	0.25
12	16	14	0.53	0.47	0.25
13	11	19	0.37	0.63	0.23
14	19	11	0.63	0.37	0.23
15	14	16	0.47	0.53	0.25
16	17	13	0.57	0.43	0.25
17	20	10	0.67	0.33	0.22
18	9	21	0.30	0.70	0.21
19	19	11	0.63	0.37	0.23
20	14	16	0.47	0.53	0.25
21	16	14	0.53	0.47	0.25
22	17	13	0.57	0.43	0.25
23	15	15	0.50	0.50	0.25
24	14	16	0.47	0.53	0.25
25	11	19	0.37	0.63	0.23
26	16	14	0.53	0.43	0.25
27	14	16	0.47	0.53	0.25
28	15	15	0.50	0.50	0.25
29	13	17	0.43	0.57	0.25
30	12	18	0.40	0.60	0.24
31	15	15	0.50	0.50	0.25
32	13	17	0.43	0.57	0.25
33	11	19	0.37	0.63	0.23
34	11	19	0.37	0.63	0.23
35	14	16	0.47	0.53	0.25
36	12	18	0.40	0.60	0.24
37	9	21	0.30	0.70	0.21
38	6	24	0.20	0.80	0.16
39	11	19	0.37	0.63	0.23
40	7	23	0.23	0.77	0.18

R	=	Number of Examinees that choose correct option
W	=	Number of examinees that choose wrong options
p	=	Proportion of examinees that choose correct option
q	=	Proportion of examinees that choose wrong options
pq	=	Product of proportion of those that choose correct option and those that choose wrong options

$$\begin{aligned}
 r_A &= \frac{n}{n-1} \left[1 - \frac{\Sigma pq}{S_{a^1}^2} \right] \\
 &= \frac{30}{30-1} \left\{ 1 - \frac{9.15}{44.93} \right\} \\
 &= \frac{30}{29} (1 - 0.20) \\
 &= 1.03 (0.8) \\
 &= 0.83
 \end{aligned}$$

APPENDIX O

Computation of the Reliability of Test B¹

S/N	R	W	p	q	pq
1	24	6	0.80	0.20	0.16
2	23	7	0.77	0.23	0.18
3	21	9	0.70	0.30	0.21
4	17	13	0.57	0.43	0.25
5	21	9	0.70	0.30	0.21
6	17	13	0.57	0.43	0.25
7	13	17	0.43	0.57	0.25
8	19	11	0.63	0.37	0.23
9	24	6	0.80	0.20	0.16
10	15	15	0.50	0.50	0.25
11	15	15	0.50	0.50	0.25
12	16	14	0.53	0.47	0.25
13	13	17	0.43	0.57	0.25
14	13	17	0.43	0.57	0.25
15	17	13	0.57	0.43	0.25
16	16	14	0.53	0.47	0.25
17	19	11	0.63	0.37	0.23
18	12	18	0.40	0.60	0.24
19	14	16	0.47	0.53	0.25
20	15	15	0.50	0.50	0.25
21	15	15	0.50	0.50	0.25
22	19	11	0.63	0.37	0.23
23	16	14	0.53	0.47	0.25
24	15	15	0.50	0.50	0.25
25	14	16	0.47	0.53	0.25
26	13	17	0.43	0.57	0.25
27	16	14	0.53	0.47	0.25
28	15	15	0.50	0.50	0.25
29	12	18	0.40	0.60	0.24
30	17	13	0.57	0.43	0.25
31	13	17	0.43	0.57	0.25
32	11	19	0.37	0.63	0.23
33	12	18	0.40	0.60	0.24
34	17	13	0.57	0.43	0.25
35	17	13	0.57	0.43	0.25
36	14	16	0.47	0.53	0.25
37	14	16	0.47	0.53	0.25
38	12	18	0.40	0.60	0.24
39	8	22	0.27	0.73	0.20
40	8	22	0.27	0.73	0.20

R	=	Number of Examinees that choose correct option
W	=	Number of examinees that choose wrong options
p	=	Proportion of examinees that choose correct option
q	=	Proportion of examinees that choose wrong options
pq	=	Product of proportion of those that choose correct option and those that choose wrong options

$$\begin{aligned}
 r_A &= \frac{n}{n-1} \left[1 - \frac{\Sigma pq}{S_{b^1}^2} \right] \\
 &= \frac{30}{30-1} \left\{ 1 - \frac{10.19}{70.42} \right\} \\
 &= \frac{30}{29} (1 - 0.14) \\
 &= 1.03 (0.86) \\
 &= 0.89
 \end{aligned}$$

APPENDIX P

BILOG-MG V3.0
REV 19990329.1300

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** BILOG-MG ITEM MAINTENANCE PROGRAM ***

*** PHASE 2 ***

CALIBRATION OF STATE A DATA

—

>CALIB ACCEl = 1.0000;

CALIBRATION PARAMETERS
=====

MAXIMUM NUMBER OF EM CYCLES: 20

MAXIMUM NUMBER OF NEWTON CYCLES: 2

CONVERGENCE CRITERION: 0.0100

ACCELERATION CONSTANT: 1.0000

LATENT DISTRIBUTION: NORMAL PRIOR FOR EACH GROUP

PLOT EMPIRICAL VS. FITTED ICC'S: NO

DATA HANDLING: DATA ON SCRATCH FILE

CONSTRAINT DISTRIBUTION ON ASYMPTOTES: YES

CONSTRAINT DISTRIBUTION ON SLOPES: YES

CONSTRAINT DISTRIBUTION ON THRESHOLDS: NO

SOURCE OF ITEM CONSTRAINT DISTRIBUTION

MEANS AND STANDARD DEVIATIONS: PROGRAM DEFAULTS

1

CALIBRATION OF MAINTEST

TEST0001

STATE A

SUBTEST TEST0001; ITEM PARAMETERS AFTER CYCLE 15

ITEM DF	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)
ITEM0001 8.0	0.077	0.498	-0.155	0.446	0.403	16.5
	0.174*	0.076*	0.371*	0.068*	0.080*	(0.0350)
ITEM0002 8.0	-0.625	0.433	1.443	0.397	0.282	15.3
	0.195*	0.093*	0.232*	0.085*	0.053*	(0.0531)
ITEM0003 8.0	-0.336	0.605	0.556	0.518	0.207	8.0
	0.125*	0.080*	0.150*	0.068*	0.047*	(0.5328)
ITEM0004 8.0	-0.570	0.596	0.957	0.512	0.238	9.3
	0.155*	0.095*	0.144*	0.082*	0.044*	(0.6270)
ITEM0005 8.0	-0.624	0.570	1.096	0.495	0.209	13.2
	0.150*	0.090*	0.136*	0.078*	0.041*	(0.8171)
ITEM0006 8.0	-0.857	0.593	1.444	0.510	0.241	11.9
	0.186*	0.112*	0.131*	0.097*	0.037*	(0.2211)
ITEM0007 8.0	-0.932	0.632	1.474	0.534	0.324	5.1
	0.212*	0.127*	0.138*	0.107*	0.034*	(0.1023)
ITEM0008 8.0	-0.527	0.600	0.878	0.514	0.153	10.8
	0.113*	0.073*	0.113*	0.063*	0.036*	(0.2147)
ITEM0009 8.0	-0.738	0.564	1.309	0.491	0.215	7.0
	0.164*	0.096*	0.135*	0.084*	0.040*	(0.2001)
ITEM0010 8.0	-0.718	0.576	1.247	0.499	0.258	14.6
	0.174*	0.102*	0.143*	0.088*	0.041*	(0.1023)
ITEM0011 8.0	-0.753	0.613	1.228	0.523	0.246	9.3
	0.168*	0.102*	0.127*	0.087*	0.038*	(0.1967)
ITEM0012 8.0	-0.927	0.754	1.230	0.602	0.183	8.2
	0.155*	0.110*	0.085*	0.088*	0.028*	(0.1132)
ITEM0013 8.0	-0.969	0.751	1.289	0.601	0.211	9.5
	0.167*	0.115*	0.091*	0.092*	0.028*	(0.4214)

ITEM0014	-0.637	0.654	0.975	0.547	0.157	11.4
8.0	0.124*	0.086*	0.100*	0.072*	0.034*	(0.1779)
ITEM0015	-1.104	0.804	1.373	0.627	0.255	14.3
8.0	0.197*	0.138*	0.093*	0.108*	0.026*	(0.1115)
ITEM0016	-0.669	0.620	1.080	0.527	0.160	15.6
8.0	0.128*	0.083*	0.107*	0.071*	0.034*	(0.1498)
ITEM0017	-0.784	0.701	1.118	0.574	0.151	6.4
8.0	0.129*	0.091*	0.087*	0.074*	0.029*	(0.0917)
ITEM0018	-0.859	0.632	1.359	0.534	0.281	7.1
8.0	0.194*	0.120*	0.128*	0.101*	0.036*	(0.5280)
ITEM0019	-0.497	0.643	0.774	0.541	0.139	13.2
8.0	0.104*	0.073*	0.101*	0.061*	0.034*	(0.1059)
ITEM0020	-0.618	0.593	1.044	0.510	0.104	5.8
8.0	0.093*	0.063*	0.093*	0.055*	0.028*	(0.0684)
ITEM0021	-0.678	0.666	1.018	0.554	0.089	12.4
8.0	0.087*	0.065*	0.076*	0.054*	0.023*	(0.1561)
ITEM0022	-0.948	0.919	1.032	0.677	0.160	32.8
8.0	0.140*	0.114*	0.064*	0.084*	0.023*	(0.0001)
ITEM0023	-0.965	0.676	1.427	0.560	0.231	9.0
8.0	0.181*	0.117*	0.109*	0.097*	0.031*	(0.0719)
ITEM0024	-0.925	0.698	1.325	0.573	0.150	12.0
8.0	0.140*	0.096*	0.088*	0.079*	0.027*	(0.0692)
ITEM0025	-1.033	0.791	1.306	0.620	0.145	8.7
8.0	0.148*	0.109*	0.076*	0.085*	0.024*	(0.0701)
ITEM0026	-1.251	1.015	1.232	0.712	0.206	8.0
8.0	0.177*	0.138*	0.064*	0.097*	0.020*	(0.1426)
ITEM0027	-1.167	0.871	1.340	0.657	0.262	66.1
8.0	0.198*	0.142*	0.085*	0.107*	0.024*	(0.0000)

ITEM0028	-1.030	0.870	1.183	0.656	0.087	9.2
8.0						
7	0.113*	0.092*	0.059*	0.069*	0.018*	(0.2022)
ITEM0029	-1.050	0.898	1.170	0.668	0.121	13.5
8.0						
	0.129*	0.101*	0.061*	0.075*	0.020*	(0.1502)
ITEM0030	-1.020	0.816	1.250	0.632	0.106	10.0
8.0						
	0.122*	0.093*	0.066*	0.072*	0.020*	(0.0803)
ITEM0031	-1.438	1.122	1.282	0.746	0.154	8.2
8.0						
	0.170*	0.135*	0.056*	0.090*	0.016*	(0.1132)
ITEM0032	-1.076	0.781	1.378	0.615	0.161	14.0
8.0						
	0.157*	0.112*	0.082*	0.088*	0.024*	(0.0684)
ITEM0033	-1.119	0.892	1.255	0.666	0.080	29.4
8.0						
	0.116*	0.093*	0.058*	0.069*	0.016*	(0.0592)
ITEM0034	-0.960	0.699	1.374	0.573	0.159	82.8
8.0						
	0.146*	0.098*	0.090*	0.080*	0.026*	(0.0041)
ITEM0035	-1.209	1.060	1.140	0.728	0.080	26.2
8.0						
	0.124*	0.109*	0.048*	0.075*	0.015*	(0.0002)
ITEM0036	-1.030	0.813	1.267	0.631	0.123	6.4
8.0						
	0.136*	0.103*	0.070*	0.080*	0.022*	(0.1720)
ITEM0037	-1.183	1.078	1.098	0.733	0.134	51.5
8.0						
	0.145*	0.124*	0.052*	0.084*	0.018*	(0.0000)
ITEM0038	-1.149	0.941	1.222	0.685	0.094	15.0
8.0						
	0.120*	0.095*	0.055*	0.069*	0.016*	(0.3122)
ITEM0039	-0.954	1.048	0.911	0.723	0.082	10.6
8.0						
	0.101*	0.093*	0.045*	0.065*	0.016*	(0.0825)
ITEM0040	-0.802	0.902	0.890	0.670	0.116	14.9
8.0						
	0.106*	0.086*	0.060*	0.064*	0.021*	(0.0640)

* STANDARD ERROR

LARGEST CHANGE = 0.006135

1106.4 320.0
(0.0000)

PARAMETER	MEAN	STN DEV
ASYMPTOTE	0.179	0.074
SLOPE	1.750	1.042
LOG(SLOPE)	0.313	0.228
THRESHOLD	1.145	0.291

QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.4121E+01	-0.3530E+01	-0.2939E+01	-0.2349E+01	-0.1758E+01
POSTERIOR	0.7838E-04	0.7046E-03	0.4595E-02	0.1950E-01	0.5788E-01

	6	7	8	9	10
POINT	-0.1167E+01	-0.5762E+00	0.1462E-01	0.6054E+00	0.1196E+01
POSTERIOR	0.1187E+00	0.1747E+00	0.2064E+00	0.2328E+00	0.1322E+00

	11	12	13	14	15
POINT	0.1787E+01	0.2378E+01	0.2969E+01	0.3560E+01	0.4150E+01
POSTERIOR	0.4249E-01	0.8234E-02	0.1480E-02	0.2321E-03	0.2174E-04

MEAN	0.00000
S.D.	1.00000

62784 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE-
 2

4748 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-
 2

06/25/2011 01:13:29

APPENDIX Q

BILOG-MG V3.0
REV 19990329.1300

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** BILOG-MG ITEM MAINTENANCE PROGRAM ***

*** PHASE 2 ***

CALIBRATION OF STATE B DATA

>CALIB ACCEl = 1.0000;

CALIBRATION PARAMETERS
=====

MAXIMUM NUMBER OF EM CYCLES:	20
MAXIMUM NUMBER OF NEWTON CYCLES:	2
CONVERGENCE CRITERION:	0.0100
ACCELERATION CONSTANT:	1.0000
LATENT DISTRIBUTION:	NORMAL PRIOR FOR EACH GROUP
PLOT EMPIRICAL VS. FITTED ICC'S:	NO
DATA HANDLING:	DATA ON SCRATCH FILE
CONSTRAINT DISTRIBUTION ON ASYMPTOTES:	YES
CONSTRAINT DISTRIBUTION ON SLOPES:	YES
CONSTRAINT DISTRIBUTION ON THRESHOLDS:	NO
SOURCE OF ITEM CONSTRAINT DISTRIBUTION	
MEANS AND STANDARD DEVIATIONS:	PROGRAM DEFAULTS

1

CALIBRATION OF MAINTEST

TEST0001

State B

SUBTEST TEST0001; ITEM PARAMETERS AFTER CYCLE 22

ITEM DF	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)

ITEM0001 7.0	-0.165	0.748	0.221	0.599	0.382	47.1
	0.227*	0.133*	0.270*	0.107*	0.073*	(0.0201)
ITEM0002 7.0	-0.542	0.883	0.613	0.662	0.191	11.7
	0.170*	0.120*	0.127*	0.090*	0.043*	(0.0802)
ITEM0003 7.0	-0.628	0.946	0.664	0.687	0.199	12.7
	0.185*	0.133*	0.119*	0.097*	0.041*	(0.0582)
ITEM0004 7.0	-0.838	1.340	0.626	0.801	0.183	7.7
	0.197*	0.168*	0.083*	0.101*	0.030*	(0.1002)
ITEM0005 7.0	-0.894	1.024	0.873	0.716	0.167	6.3
	0.190*	0.138*	0.092*	0.097*	0.031*	(0.0577)
ITEM0006 7.0	-0.807	0.717	1.126	0.583	0.147	7.8
	0.172*	0.112*	0.117*	0.091*	0.036*	(0.2341)
ITEM0007 7.0	-0.935	0.695	1.345	0.571	0.256	10.0
	0.266*	0.159*	0.155*	0.131*	0.044*	(0.0951)
ITEM0008 7.0	-0.914	0.910	1.005	0.673	0.140	8.9
	0.183*	0.132*	0.091*	0.098*	0.031*	(0.2810)
ITEM0009 7.0	-0.939	0.846	1.110	0.646	0.163	12.3
	0.201*	0.137*	0.104*	0.105*	0.034*	(0.2667)
ITEM0010 7.0	-0.802	0.742	1.080	0.596	0.238	9.2
	0.233*	0.145*	0.147*	0.117*	0.045*	(0.5001)
ITEM0011 7.0	-1.282	0.934	1.372	0.683	0.241	12.6
	0.300*	0.200*	0.104*	0.146*	0.030*	(0.2043)
ITEM0012 7.0	-1.019	0.805	1.265	0.627	0.136	6.6
	0.197*	0.134*	0.100*	0.104*	0.031*	(0.1014)
ITEM0013 7.0	-1.180	0.814	1.450	0.631	0.163	12.9

		0.245*	0.163*	0.110*	0.126*	0.031*	(0.1740)
ITEM0014	-0.986	0.903	1.092	0.670	0.134	8.2	
7.0							
		0.189*	0.136*	0.089*	0.101*	0.029*	(0.2503)
ITEM0015	-1.501	1.059	1.418	0.727	0.216	6.0	
7.0							
		0.317*	0.219*	0.091*	0.151*	0.025*	(0.1046)
ITEM0016	-0.955	0.857	1.113	0.651	0.144	6.4	
7.0							
		0.189*	0.131*	0.096*	0.099*	0.031*	(0.1024)
ITEM0017	-1.103	0.875	1.260	0.659	0.134	5.5	
7.0							
		0.204*	0.142*	0.091*	0.107*	0.028*	(0.3006)
ITEM0018	-1.113	0.810	1.374	0.629	0.277	11.2	
7.0							
		0.297*	0.187*	0.131*	0.145*	0.037*	(0.1163)
ITEM0019	-0.790	0.784	1.008	0.617	0.148	8.8	
7.0							
		0.172*	0.117*	0.110*	0.092*	0.035*	(0.2103)
ITEM0020	-1.109	0.888	1.248	0.664	0.116	3.3	
7.0							
		0.192*	0.135*	0.086*	0.101*	0.026*	(0.3801)
ITEM0021	-1.416	1.064	1.331	0.729	0.108	40.6	
7.0							
		0.230*	0.165*	0.073*	0.113*	0.021*	(0.0071)
ITEM0022	-2.342	1.542	1.519	0.839	0.184	7.0	
7.0							
		0.423*	0.289*	0.070*	0.157*	0.016*	(0.4296)
ITEM0023	-1.662	0.832	1.998	0.639	0.296	4.6	
7.0							
		0.426*	0.254*	0.211*	0.195*	0.027*	(0.0815)
ITEM0024	-2.391	1.574	1.519	0.844	0.212	11.6	
7.0							
		0.468*	0.321*	0.072*	0.172*	0.017*	(0.1149)
ITEM0025	-2.242	1.153	1.943	0.756	0.213	9.0	
7.0							
		0.520*	0.326*	0.162*	0.214*	0.018*	(0.2118)
ITEM0026	-4.073	2.377	1.714	0.922	0.287	26.6	
7.0							
		1.204*	0.721*	0.070*	0.280*	0.015*	(0.0004)
ITEM0027	-5.417	3.296	1.644	0.957	0.356	59.9	
7.0							
		2.144*	1.259*	0.052*	0.366*	0.015*	(0.0102)

ITEM0028	-2.326	1.554	1.497	0.841	0.131	11.4
7.0	0.385*	0.269*	0.063*	0.145*	0.014*	(0.1032)
ITEM0029	-3.736	2.382	1.568	0.922	0.195	6.0
7.0	0.693*	0.440*	0.054*	0.170*	0.013*	(0.0755)
ITEM0030	-2.865	1.712	1.673	0.864	0.178	10.2
7.0	0.572*	0.370*	0.078*	0.186*	0.014*	(0.1785)
ITEM0031	-5.771	3.344	1.726	0.958	0.211	6.1
7.0	2.607*	1.524*	0.049*	0.437*	0.013*	(0.0905)
ITEM0032	-4.171	2.670	1.562	0.936	0.222	9.7
7.0	0.898*	0.564*	0.052*	0.198*	0.014*	(0.1062)
ITEM0033	-4.461	2.576	1.732	0.932	0.136	3.6
7.0	1.122*	0.668*	0.055*	0.242*	0.011*	(0.5013)
ITEM0034	-4.165	2.574	1.618	0.932	0.241	20.5
7.0	1.026*	0.632*	0.056*	0.229*	0.014*	(0.0001)
ITEM0035	-5.744	3.439	1.670	0.960	0.140	4.7
7.0	1.857*	1.096*	0.041*	0.306*	0.011*	(0.1962)
ITEM0036	-4.940	2.899	1.704	0.945	0.228	9.7
7.0	1.752*	1.039*	0.055*	0.339*	0.014*	(0.4701)
ITEM0037	-4.933	3.131	1.575	0.953	0.208	23.4
7.0	1.169*	0.714*	0.048*	0.217*	0.013*	(0.0000)
ITEM0038	-6.587	4.019	1.639	0.970	0.160	4.7
7.0	2.505*	1.468*	0.042*	0.354*	0.011*	(0.3000)
ITEM0039	-4.405	2.593	1.699	0.933	0.184	10.7
7.0	1.187*	0.707*	0.056*	0.254*	0.013*	(0.2401)
ITEM0040	-6.379	3.847	1.658	0.968	0.217	9.0
7.0	2.680*	1.565*	0.044*	0.394*	0.013*	(0.1622)

* STANDARD ERROR

LARGEST CHANGE = 0.371591

1149.0 280.0

(0.0000)

PARAMETER	MEAN	STN DEV
ASYMPTOTE	0.197	0.062
SLOPE	1.654	1.028
LOG(SLOPE)	0.330	0.581
THRESHOLD	1.356	0.386

QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.4032E+01	-0.3457E+01	-0.2883E+01	-0.2308E+01	-0.1733E+01
POSTERIOR	0.6472E-04	0.6015E-03	0.3842E-02	0.1742E-01	0.5650E-01

	6	7	8	9	10
POINT	-0.1159E+01	-0.5841E+00	-0.9472E-02	0.5652E+00	0.1140E+01
POSTERIOR	0.1249E+00	0.1921E+00	0.2003E+00	0.1646E+00	0.1650E+00

	11	12	13	14	15
POINT	0.1714E+01	0.2289E+01	0.2864E+01	0.3438E+01	0.4013E+01
POSTERIOR	0.7239E-01	0.2241E-02	0.2137E-05	0.4277E-09	0.3862E-13

MEAN	0.00000
S.D.	1.00000

62784 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE-
2

4748 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-
2

06/25/2011 07:46:11

APPENDIX R

BILOG-MG V3.0
REV 19990329.1300

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** BILOG-MG ITEM MAINTENANCE PROGRAM ***

*** PHASE 2 ***

DIF FOR STATE OF MATHEMATICS TEST ITEMS
*

>CALIB NQPt = 15;

CALIBRATION PARAMETERS
=====

MAXIMUM NUMBER OF EM CYCLES: 20

MAXIMUM NUMBER OF NEWTON CYCLES: 2

CONVERGENCE CRITERION: 0.0100

ACCELERATION CONSTANT: 0.5000

LATENT DISTRIBUTION: EMPIRICAL PRIOR FOR EACH GROUP
ESTIMATED CONCURRENTLY
WITH ITEM PARAMETERS
REFERENCE GROUP: 1

PLOT EMPIRICAL VS. FITTED ICC'S: NO

DATA HANDLING: DATA ON SCRATCH FILE

CONSTRAINT DISTRIBUTION ON SLOPES: NO

CONSTRAINT DISTRIBUTION ON THRESHOLDS: NO

1

CALIBRATION OF MAINTEST

STATDIF

METHOD OF SOLUTION:

EM CYCLES (MAXIMUM OF 20)
 FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF 2)

QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP 1 STATE A

	1	2	3	4	5
POINT	-0.4000E+01	-0.3429E+01	-0.2857E+01	-0.2286E+01	-0.1714E+01
WEIGHT	0.7648E-04	0.6387E-03	0.3848E-02	0.1673E-01	0.5245E-01

	6	7	8	9	10
POINT	-0.1143E+01	-0.5714E+00	-0.8882E-15	0.5714E+00	0.1143E+01
WEIGHT	0.1186E+00	0.1936E+00	0.2280E+00	0.1936E+00	0.1186E+00

	11	12	13	14	15
POINT	0.1714E+01	0.2286E+01	0.2857E+01	0.3429E+01	0.4000E+01
WEIGHT	0.5245E-01	0.1673E-01	0.3848E-02	0.6387E-03	0.7648E-04

QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP 2 STATE B

	1	2	3	4	5
POINT	-0.4000E+01	-0.3429E+01	-0.2857E+01	-0.2286E+01	-0.1714E+01
WEIGHT	0.7648E-04	0.6387E-03	0.3848E-02	0.1673E-01	0.5245E-01

	6	7	8	9	10
POINT	-0.1143E+01	-0.5714E+00	-0.8882E-15	0.5714E+00	0.1143E+01
WEIGHT	0.1186E+00	0.1936E+00	0.2280E+00	0.1936E+00	0.1186E+00

	11	12	13	14	15
POINT	0.1714E+01	0.2286E+01	0.2857E+01	0.3429E+01	0.4000E+01
WEIGHT	0.5245E-01	0.1673E-01	0.3848E-02	0.6387E-03	0.7648E-04

[E-M CYCLES]

-2 LOG LIKELIHOOD = 143457.119

CYCLE 1; LARGEST CHANGE= 0.22883

-2 LOG LIKELIHOOD = 142974.926

CYCLE 2; LARGEST CHANGE= 0.00250

[NEWTON CYCLES]

-2 LOG LIKELIHOOD: 142949.1124

CYCLE 3; LARGEST CHANGE= 0.00147

INTERVAL COUNTS FOR COMPUTATION OF ITEM CHI-SQUARES

```
-----
-      120.   198.   190.   246.   400.   148.   95.   38.   79.
-----
-
```

INTERVAL AVERAGE THETAS

```
-----
-      -1.652 -1.116 -0.633 -0.150  0.262  0.703  1.180  1.605  2.465
-----
```

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
GROUP THRESHOLD DIFFERENCES

ITEM	GROUP B - A	ITEM	GROUP B - A	ITEM	GROUP B - A
ITEM0001	0.064 0.130*	ITEM0015	0.013 0.121*	ITEM0028	0.192 0.137*
ITEM0002	-0.295 0.116*	ITEM0016	-0.144 0.122*	ITEM0029	0.099 0.131*
ITEM0003	-0.196 0.120*	ITEM0017	-0.002 0.125*	ITEM0030	0.120 0.133*
ITEM0004	-0.446 0.120*	ITEM0018	-0.453 0.117*	ITEM0031	0.073 0.131*
ITEM0005	-0.239 0.120*	ITEM0019	-0.144 0.122*	ITEM0032	-0.072 0.127*
ITEM0006	-0.165 0.119*	ITEM0020	-0.145 0.125*	ITEM0033	0.246 0.142*
ITEM0007	-0.224 0.116*	ITEM0021	0.450 0.130*	ITEM0034	-0.079 0.125*
ITEM0008	0.083 0.122*	ITEM0022	0.238 0.128*	ITEM0035	0.416 0.143*

ITEM0009	-0.115	ITEM0023	-0.388	ITEM0036	0.117
	0.120*		0.120*		0.129*
ITEM0010	-0.316	ITEM0024	-0.157	ITEM0037	0.097
	0.118*		0.126*		0.131*
ITEM0011	0.009	ITEM0025	-0.016	ITEM0038	0.254
	0.119*		0.128*		0.138*
ITEM0012	-0.146	ITEM0026	-0.330	ITEM0039	0.326
	0.124*		0.124*		0.135*
ITEM0013	0.122	ITEM0027	-0.512	ITEM0040	0.227
	0.122*		0.119*		0.129*
ITEM0014	0.083	ITEM0028	0.192		
	0.123*		0.137*		

*STANDARD ERROR

GROUP: 1 STATE A QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.4129E+01	-0.3554E+01	-0.2979E+01	-0.2404E+01	-0.1830E+01
POSTERIOR	0.0000E+00	0.5155E-05	0.3975E-03	0.7720E-02	0.4723E-01

	6	7	8	9	10
POINT	-0.1255E+01	-0.6797E+00	-0.1048E+00	0.4701E+00	0.1045E+01
POSTERIOR	0.1179E+00	0.1604E+00	0.2499E+00	0.2360E+00	0.9248E-01

	11	12	13	14	15
POINT	0.1620E+01	0.2195E+01	0.2770E+01	0.3345E+01	0.3920E+01
POSTERIOR	0.4639E-01	0.2424E-01	0.1022E-01	0.5023E-02	0.2115E-02

MEAN	0.00000
S.E.	0.00000

S.D.	1.00000
S.E.	0.00000

GROUP: 2 STATE B QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.5121E+01	-0.4546E+01	-0.3971E+01	-0.3396E+01	-0.2821E+01
POSTERIOR	0.0000E+00	0.4928E-05	0.1889E-03	0.3085E-02	0.2351E-01

	6	7	8	9	10
POINT	-0.2247E+01	-0.1672E+01	-0.1097E+01	-0.5218E+00	0.5312E-01

POSTERIOR 0.9344E-01 0.2125E+00 0.2489E+00 0.1521E+00 0.1235E+00

	11	12	13	14	15
POINT	0.6280E+00	0.1203E+01	0.1778E+01	0.2353E+01	0.2928E+01
POSTERIOR	0.7581E-01	0.3703E-01	0.2225E-01	0.7195E-02	0.5577E-03

MEAN -0.83813

S.E. 0.03145

S.D. 1.04909

S.E. 0.02608

2 105076 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE-

2 2816 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-

01/27/2012 12:07:30

APPENDIX S

BILOG-MG V3.0
REV 19990329.1300

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** BILOG-MG ITEM MAINTENANCE PROGRAM ***

*** PHASE 2 ***

DIF FOR GENDER OF MATHEMATICS TEST ITEMS
*

>CALIB NQPt = 15;

CALIBRATION PARAMETERS
=====

MAXIMUM NUMBER OF EM CYCLES: 20

MAXIMUM NUMBER OF NEWTON CYCLES: 2

CONVERGENCE CRITERION: 0.0100

ACCELERATION CONSTANT: 0.5000

LATENT DISTRIBUTION: EMPIRICAL PRIOR FOR EACH GROUP
ESTIMATED CONCURRENTLY
WITH ITEM PARAMETERS
REFERENCE GROUP: 1

PLOT EMPIRICAL VS. FITTED ICC'S: NO

DATA HANDLING: DATA ON SCRATCH FILE

CONSTRAINT DISTRIBUTION ON SLOPES: NO

CONSTRAINT DISTRIBUTION ON THRESHOLDS: NO

1

CALIBRATION OF MAINTEST

SEXDIF

METHOD OF SOLUTION:

EM CYCLES (MAXIMUM OF 20)
 FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF 2)

QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP 1 MALE M

	1	2	3	4	5
POINT	-0.4000E+01	-0.3429E+01	-0.2857E+01	-0.2286E+01	-0.1714E+01
WEIGHT	0.7648E-04	0.6387E-03	0.3848E-02	0.1673E-01	0.5245E-01

	6	7	8	9	10
POINT	-0.1143E+01	-0.5714E+00	-0.8882E-15	0.5714E+00	0.1143E+01
WEIGHT	0.1186E+00	0.1936E+00	0.2280E+00	0.1936E+00	0.1186E+00

	11	12	13	14	15
POINT	0.1714E+01	0.2286E+01	0.2857E+01	0.3429E+01	0.4000E+01
WEIGHT	0.5245E-01	0.1673E-01	0.3848E-02	0.6387E-03	0.7648E-04

QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP 2 FEMALE F

	1	2	3	4	5
POINT	-0.4000E+01	-0.3429E+01	-0.2857E+01	-0.2286E+01	-0.1714E+01
WEIGHT	0.7648E-04	0.6387E-03	0.3848E-02	0.1673E-01	0.5245E-01

	6	7	8	9	10
POINT	-0.1143E+01	-0.5714E+00	-0.8882E-15	0.5714E+00	0.1143E+01
WEIGHT	0.1186E+00	0.1936E+00	0.2280E+00	0.1936E+00	0.1186E+00

	11	12	13	14	15
POINT	0.1714E+01	0.2286E+01	0.2857E+01	0.3429E+01	0.4000E+01
WEIGHT	0.5245E-01	0.1673E-01	0.3848E-02	0.6387E-03	0.7648E-04

[E-M CYCLES]

-2 LOG LIKELIHOOD = 141974.254

CYCLE 1; LARGEST CHANGE= 0.44800

-2 LOG LIKELIHOOD = 140082.434

CYCLE 2; LARGEST CHANGE= 0.03554

-2 LOG LIKELIHOOD = 139912.030

CYCLE 3; LARGEST CHANGE= 0.01659

-2 LOG LIKELIHOOD = 139842.144

CYCLE 4; LARGEST CHANGE= 0.01529

-2 LOG LIKELIHOOD = 139801.570

CYCLE 5; LARGEST CHANGE= 0.00759

[NEWTON CYCLES]

-2 LOG LIKELIHOOD: 139773.8378

CYCLE 6; LARGEST CHANGE= 0.00595

INTERVAL COUNTS FOR COMPUTATION OF ITEM CHI-SQUARES

```
-----
-      0.      0.      7.      22.     111.     211.     290.     677.     133.
-----
```

INTERVAL AVERAGE THETAS

```
-----
- ***** -3.518 -2.898 -1.958 -0.996 -0.221  0.621  1.198
-----
```

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
GROUP THRESHOLD DIFFERENCES

ITEM	GROUP F - M	ITEM	GROUP F - M	ITEM	GROUP F - M
ITEM0001	0.106	ITEM0015	0.346	ITEM0028	-0.335
	0.149*		0.117*		0.117*
ITEM0002	0.092	ITEM0016	0.180	ITEM0029	-0.242
	0.104*		0.132*		0.125*
ITEM0003	-0.245	ITEM0017	-0.223	ITEM0030	-0.397

	0.315*		0.331*		0.260*
ITEM0004	-0.158	ITEM0018	-0.109	ITEM0031	0.017
	0.110*		0.009*		0.354*
ITEM0005	0.264	ITEM0019	-0.254	ITEM0032	0.205
	0.137*		0.124*		0.138*
ITEM0006	0.204	ITEM0020	0.136	ITEM0033	-0.211
	0.112*		0.036*		0.387*
ITEM0007	0.209	ITEM0021	-0.202	ITEM0034	0.376
	0.115*		0.147*		0.232*
ITEM0008	-0.326	ITEM0022	-0.201	ITEM0035	-0.324
	0.120*		0.136*		0.197*
ITEM0009	0.075	ITEM0023	0.383	ITEM0036	-0.267
	0.214*		0.217*		0.152*
ITEM0010	0.100	ITEM0024	0.160	ITEM0037	-0.255
	0.009*		0.036*		0.156*
ITEM0011	0.412	ITEM0025	0.089	ITEM0038	0.167
	0.312*		0.343*		0.079*
ITEM0012	-0.191	ITEM0026	0.414	ITEM0039	-0.251
	0.128*		0.130*		0.173*
ITEM0013	0.259	ITEM0027	0.306	ITEM0040	-0.338
	0.323*		0.215*		0.144*
ITEM0014	-0.340	ITEM0028	-0.335		
	0.124*		0.117*		

*STANDARD ERROR

GROUP: 1 MALE QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.8011E+01	-0.6998E+01	-0.5985E+01	-0.4972E+01	-0.3959E+01
POSTERIOR	0.0000E+00	0.0000E+00	0.1172E-04	0.3412E-03	0.3732E-02
	6	7	8	9	10
POINT	-0.2946E+01	-0.1933E+01	-0.9205E+00	0.9248E-01	0.1105E+01
POSTERIOR	0.2042E-01	0.6929E-01	0.1913E+00	0.4127E+00	0.2879E+00
	11	12	13	14	15
POINT	0.2118E+01	0.3131E+01	0.4144E+01	0.5157E+01	0.6170E+01
POSTERIOR	0.1433E-01	0.1046E-04	0.0000E+00	0.0000E+00	0.0000E+00

MEAN 0.00000
S.E. 0.00000

S.D. 1.00000
S.E. 0.00000

GROUP: 2 FEMALE QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.1641E+01	-0.6281E+00	0.3848E+00	0.1398E+01	0.2411E+01
POSTERIOR	0.0000E+00	0.0000E+00	0.0000E+00	0.0000E+00	0.4550E-03

	6	7	8	9	10
POINT	0.3424E+01	0.4437E+01	0.5450E+01	0.6463E+01	0.7476E+01
POSTERIOR	0.9578E-01	0.5007E+00	0.2081E+00	0.5927E-01	0.4369E-01

	11	12	13	14	15
POINT	0.8489E+01	0.9502E+01	0.1051E+02	0.1153E+02	0.1254E+02
POSTERIOR	0.4026E-01	0.2554E-01	0.1380E-01	0.8011E-02	0.4407E-02

MEAN 5.27139
S.E. 0.05134

S.D. 1.64210
S.E. 0.06736

105076 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE-
2

2816 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-
2

01/27/2012 11:16:28

APPENDIX T

BILOG-MG V3.0
REV 19990329.1300

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** BILOG-MG ITEM MAINTENANCE PROGRAM ***

*** PHASE 2 ***

DIF FOR ABILITY OF MATHEMATICS TEST ITEMS
*

>CALIB NQPt = 15;

CALIBRATION PARAMETERS
=====

MAXIMUM NUMBER OF EM CYCLES: 20

MAXIMUM NUMBER OF NEWTON CYCLES: 2

CONVERGENCE CRITERION: 0.0100

ACCELERATION CONSTANT: 0.5000

LATENT DISTRIBUTION: EMPIRICAL PRIOR FOR EACH GROUP
ESTIMATED CONCURRENTLY
WITH ITEM PARAMETERS
REFERENCE GROUP: 1

PLOT EMPIRICAL VS. FITTED ICC'S: NO

DATA HANDLING: DATA ON SCRATCH FILE

CONSTRAINT DISTRIBUTION ON SLOPES: NO

CONSTRAINT DISTRIBUTION ON THRESHOLDS: NO

1

CALIBRATION OF MAINTEST

ABILDIF

METHOD OF SOLUTION:

EM CYCLES (MAXIMUM OF 20)
 FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF 2)

QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP 1 LOW ABILITY

	1	2	3	4	5
POINT	-0.4000E+01	-0.3429E+01	-0.2857E+01	-0.2286E+01	-0.1714E+01
WEIGHT	0.7648E-04	0.6387E-03	0.3848E-02	0.1673E-01	0.5245E-01

	6	7	8	9	10
POINT	-0.1143E+01	-0.5714E+00	-0.8882E-15	0.5714E+00	0.1143E+01
WEIGHT	0.1186E+00	0.1936E+00	0.2280E+00	0.1936E+00	0.1186E+00

	11	12	13	14	15
POINT	0.1714E+01	0.2286E+01	0.2857E+01	0.3429E+01	0.4000E+01
WEIGHT	0.5245E-01	0.1673E-01	0.3848E-02	0.6387E-03	0.7648E-04

QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP 2 HIGH ABILITY

	1	2	3	4	5
POINT	-0.4000E+01	-0.3429E+01	-0.2857E+01	-0.2286E+01	-0.1714E+01
WEIGHT	0.7648E-04	0.6387E-03	0.3848E-02	0.1673E-01	0.5245E-01

	6	7	8	9	10
POINT	-0.1143E+01	-0.5714E+00	-0.8882E-15	0.5714E+00	0.1143E+01
WEIGHT	0.1186E+00	0.1936E+00	0.2280E+00	0.1936E+00	0.1186E+00

	11	12	13	14	15
POINT	0.1714E+01	0.2286E+01	0.2857E+01	0.3429E+01	0.4000E+01
WEIGHT	0.5245E-01	0.1673E-01	0.3848E-02	0.6387E-03	0.7648E-04

[E-M CYCLES]

-2 LOG LIKELIHOOD = 143999.079

CYCLE 1; LARGEST CHANGE= 0.18259

-2 LOG LIKELIHOOD = 143581.331

CYCLE 2; LARGEST CHANGE= 0.00249

[NEWTON CYCLES]

-2 LOG LIKELIHOOD: 143561.4252

CYCLE 3; LARGEST CHANGE= 0.00099

INTERVAL COUNTS FOR COMPUTATION OF ITEM CHI-SQUARES

```
-----
-      2.    48.   147.   220.   189.   179.   301.   115.   152.
-----
-
```

INTERVAL AVERAGE THETAS

```
-----
-      -2.041 -1.693 -1.289 -0.871 -0.424  0.092  0.494  0.911  1.819
-----
-
```

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
GROUP THRESHOLD DIFFERENCES

ITEM	GROUP L - H	ITEM	GROUP L - H	ITEM	GROUP L - H
ITEM0001	-0.054 0.118*	ITEM0015	-0.019 0.109*	ITEM0028	-0.086 0.122*
ITEM0002	-0.060 0.105*	ITEM0016	0.009 0.111*	ITEM0029	0.078 0.118*
ITEM0003	0.148 0.109*	ITEM0017	-0.095 0.113*	ITEM0030	0.006 0.120*
ITEM0004	-0.189 0.108*	ITEM0018	-0.066 0.107*	ITEM0031	0.156 0.119*
ITEM0005	0.117 0.109*	ITEM0019	-0.115 0.111*	ITEM0032	-0.077 0.115*
ITEM0006	0.056 0.108*	ITEM0020	-0.034 0.113*	ITEM0033	-0.036 0.124*
ITEM0007	-0.258 0.106*	ITEM0021	-0.262 0.116*	ITEM0034	-0.134 0.113*
ITEM0008	-0.044 0.110*	ITEM0022	-0.085 0.115*	ITEM0035	0.168 0.126*

ITEM0009	-0.129	ITEM0023	0.166	ITEM0036	-0.027
	0.108*		0.109*		0.118*
ITEM0010	-0.033	ITEM0024	-0.054	ITEM0037	0.227
	0.107*		0.114*		0.119*
ITEM0011	-0.007	ITEM0025	-0.153	ITEM0038	-0.132
	0.108*		0.116*		0.123*
ITEM0012	-0.181	ITEM0026	0.056	ITEM0039	0.150
	0.113*		0.113*		0.121*
ITEM0013	0.018	ITEM0027	0.053	ITEM0040	-0.122
	0.111*		0.109*		0.115*
ITEM0014	-0.099	ITEM0028	-0.086		
	0.111*		0.122*		

*STANDARD ERROR

GROUP: 1 LOW QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.3907E+01	-0.3368E+01	-0.2830E+01	-0.2292E+01	-0.1754E+01
POSTERIOR	0.2731E-06	0.2538E-04	0.6919E-03	0.8128E-02	0.4610E-01

	6	7	8	9	10
POINT	-0.1215E+01	-0.6770E+00	-0.1387E+00	0.3996E+00	0.9380E+00
POSTERIOR	0.1325E+00	0.1775E+00	0.1721E+00	0.2293E+00	0.1369E+00

	11	12	13	14	15
POINT	0.1476E+01	0.2015E+01	0.2553E+01	0.3091E+01	0.3629E+01
POSTERIOR	0.5388E-01	0.2084E-01	0.1056E-01	0.7278E-02	0.4263E-02

MEAN	0.00000
S.E.	0.00000

S.D.	1.00000
S.E.	0.00000

GROUP: 2 HIGH QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.4147E+01	-0.3609E+01	-0.3071E+01	-0.2533E+01	-0.1994E+01
POSTERIOR	0.1022E-06	0.1419E-04	0.4336E-03	0.5918E-02	0.3715E-01

	6	7	8	9	10
POINT	-0.1456E+01	-0.9176E+00	-0.3793E+00	0.1590E+00	0.6973E+00

POSTERIOR 0.1173E+00 0.2036E+00 0.1912E+00 0.1867E+00 0.1407E+00

	11	12	13	14	15
POINT	0.1236E+01	0.1774E+01	0.2312E+01	0.2851E+01	0.3389E+01
POSTERIOR	0.5600E-01	0.3919E-01	0.1900E-01	0.2458E-02	0.1861E-03

MEAN -0.20262

S.E. 0.02783

S.D. 0.98903

S.E. 0.02157

2 105076 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE-

2 2816 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-

01/27/2012 12:19:33

APPENDIX U

BILOG-MG V3.0

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** LOGISTIC MODEL ITEM ANALYSER ***

*** PHASE 3 ***

CALIBRATION OF STATE A ABILITY DATA

>SCORE ;

PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION

METHOD OF SCORING SUBJECTS:	EXPECTATION A POSTERIORI (EAP; BAYES ESTIMATION)
TYPE OF PRIOR:	NORMAL
SCORES WRITTEN TO FILE	FIRSTSTATE.PH3
TYPE OF RESCALING:	NONE REQUESTED
ITEM AND TEST INFORMATION:	NONE REQUESTED
DOMAIN SCORE ESTIMATION:	NONE REQUESTED

TEST	NAME	QUAD POINTS
1	MATHTEST	10

1

SCORING

PRIOR DISTRIBUTION(S)
=====

EAP SUBJECT ESTIMATION, TEST: MATHTEST

QUADRATURE POINTS AND PRIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.1190E-03	0.2805E-02	0.3002E-01	0.1458E+00	0.3213E+00

	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.3213E+00	0.1458E+00	0.3002E-01	0.2805E-02	0.1190E-03

MEAN 0.0000
S.D. 1.0000

1

SUMMARY STATISTICS FOR SCORE ESTIMATES

=====

CORRELATIONS AMONG TEST SCORES

	MATHTEST
MATHTEST	1.0000

MEANS, STANDARD DEVIATIONS, AND VARIANCES OF SCORE ESTIMATES

TEST: MATHTEST
MEAN: 1.9461
S.D.: 0.8929
VARIANCE: 0.7973

ROOT-MEAN-SQUARE POSTERIORI STANDARD DEVIATIONS

TEST: MATHTEST
RMS: 0.4543
VARIANCE: 0.2064

EMPIRICAL
RELIABILITY: 0.7944

MARGINAL LATENT DISTRIBUTION(S)

=====

MARGINAL LATENT DISTRIBUTION FOR TEST MATHTEST

MEAN = 0.000
S.D. = 0.975

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.2155E-03	0.4813E-02	0.4381E-01	0.1448E+00	0.2401E+00

	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.4039E+00	0.1437E+00	0.1755E-01	0.1041E-02	0.2601E-04

3 44 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE-

3 2968 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-

APPENDIX V

BILOG-MG V3.0

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** LOGISTIC MODEL ITEM ANALYSER ***

*** PHASE 3 ***

CALIBRATION OF STATE B ABILITY DATA

>SCORE ;

PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION

METHOD OF SCORING SUBJECTS:	EXPECTATION A POSTERIORI (EAP; BAYES ESTIMATION)
TYPE OF PRIOR:	NORMAL
SCORES WRITTEN TO FILE	RIVERS.PH3
TYPE OF RESCALING:	NONE REQUESTED
ITEM AND TEST INFORMATION:	NONE REQUESTED
DOMAIN SCORE ESTIMATION:	NONE REQUESTED

TEST	NAME	QUAD POINTS
1	TEST0001	10

1

SCORING

PRIOR DISTRIBUTION(S)
=====

EAP SUBJECT ESTIMATION, TEST: TEST0001

QUADRATURE POINTS AND PRIOR WEIGHTS, MEAN AND S.D.:

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.1190E-03	0.2805E-02	0.3002E-01	0.1458E+00	0.3213E+00

	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.3213E+00	0.1458E+00	0.3002E-01	0.2805E-02	0.1190E-03

MEAN 0.0000
S.D. 1.0000

1

SUMMARY STATISTICS FOR SCORE ESTIMATES

=====

CORRELATIONS AMONG TEST SCORES

	TEST0001
TEST0001	1.0000

MEANS, STANDARD DEVIATIONS, AND VARIANCES OF SCORE ESTIMATES

TEST:	TEST0001
MEAN:	1.8930
S.D.:	0.8643
VARIANCE:	0.7469

ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS

TEST:	TEST0001
RMS:	0.5448
VARIANCE:	0.2968

EMPIRICAL
RELIABILITY: 0.7156

MARGINAL LATENT DISTRIBUTION(S)

=====

MARGINAL LATENT DISTRIBUTION FOR TEST TEST0001

MEAN	=	0.000
S.D.	=	0.995

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.1346E-03	0.3165E-02	0.3354E-01	0.1576E+00	0.3113E+00

	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.2714E+00	0.2058E+00	0.1704E-01	0.5195E-07	0.3764E-13

44 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE- 3

2968 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-3

APPENDIX W

BILOG-MG V3.0

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

*** LOGISTIC MODEL ITEM ANALYSER ***

*** PHASE 3 ***

CONCURRENT CALIBRATION OF STATE A,B ABILITY DATA

>SCORE ;

PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION

METHOD OF SCORING SUBJECTS:	EXPECTATION A POSTERIORI (EAP; BAYES ESTIMATION)
TYPE OF PRIOR:	NORMAL
SCORES WRITTEN TO FILE	JOSEPHABILITY.PH3
TYPE OF RESCALING:	NONE REQUESTED
ITEM AND TEST INFORMATION:	NONE REQUESTED
DOMAIN SCORE ESTIMATION:	NONE REQUESTED

TEST	NAME	QUAD GROUP	POINTS
1	ABILDIF	1	10
1	ABILDIF	2	10

SCORING

PRIOR DISTRIBUTION(S)

=====

EAP SUBJECT ESTIMATION, TEST: ABILDIF

GROUP 1 LOWABLE

QUADRATURE POINTS AND PRIOR WEIGHTS, MEAN AND S.D:

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.1190E-03	0.2805E-02	0.3002E-01	0.1458E+00	0.3213E+00

	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.3213E+00	0.1458E+00	0.3002E-01	0.2805E-02	0.1190E-03

MEAN 0.0000

S.D. 1.0000

EAP SUBJECT ESTIMATION, TEST: ABILDIF

GROUP 2 HIGHABLE

QUADRATURE POINTS AND PRIOR WEIGHTS, MEAN AND S.D:

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.1190E-03	0.2805E-02	0.3002E-01	0.1458E+00	0.3213E+00

	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.3213E+00	0.1458E+00	0.3002E-01	0.2805E-02	0.1190E-03

MEAN 0.0000

S.D. 1.0000

SUMMARY STATISTICS FOR SCORE ESTIMATES BY STATE

=====

STATISTICS FOR STATE: A

CORRELATIONS AMONG TEST SCORES

	ABILDIF
ABILDIF	1.000

MEANS, STANDARD DEVIATIONS, AND VARIANCES OF SCORE ESTIMATES

TEST:	ABILDIF
MEAN:	1.9225
S.D.:	0.5220
VARIANCE:	0.2725

ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS

TEST:	ABILDIF
RMS:	0.8101
VARIANCE:	0.6563

EMPIRICAL	
RELIABILITY:	0.7934

STATISTICS FOR STATE: B

CORRELATIONS AMONG TEST SCORES

	ABILDIF
ABILDIF	1.000

MEANS, STANDARD DEVIATIONS, AND VARIANCES OF SCORE ESTIMATES

TEST:	ABILDIF
MEAN:	1.8902
S.D.:	0.7123
VARIANCE:	0.5074

ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS

TEST:	ABILDIF
RMS:	0.7636
VARIANCE:	0.5831

EMPIRICAL	
RELIABILITY:	0.7653

MARGINAL LATENT DISTRIBUTION(S) OF COMBINED GROUPS

=====

MARGINAL LATENT DISTRIBUTION FOR TEST ABILDIF

MEAN = 0.935
S.D. = 1.368

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.6259E-04	0.1458E-02	0.1505E-01	0.7213E-01	0.1731E+00
	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.2373E+00	0.2278E+00	0.1596E+00	0.8060E-01	0.3289E-01

MARGINAL LATENT DISTRIBUTIONS BY TEST AND GROUP

=====

MARGINAL LATENT DISTRIBUTION FOR TEST ABILDIF GROUP 1 LOWABLE

MEAN = 0.723
S.D. = 0.936

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.1253E-03	0.2920E-02	0.3009E-01	0.1430E+00	0.3237E+00
	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.3362E+00	0.1420E+00	0.2103E-01	0.9677E-03	0.1276E-04

MARGINAL LATENT DISTRIBUTION FOR TEST ABILDIF GROUP 2 HIGHABLE

MEAN = 1.890
S.D. = 1.018

	1	2	3	4	5
POINT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00
WEIGHT	0.6608E-09	0.2280E-06	0.2924E-04	0.1365E-02	0.2291E-01
	6	7	8	9	10
POINT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01
WEIGHT	0.1387E+00	0.3134E+00	0.2979E+00	0.1601E+00	0.6570E-01

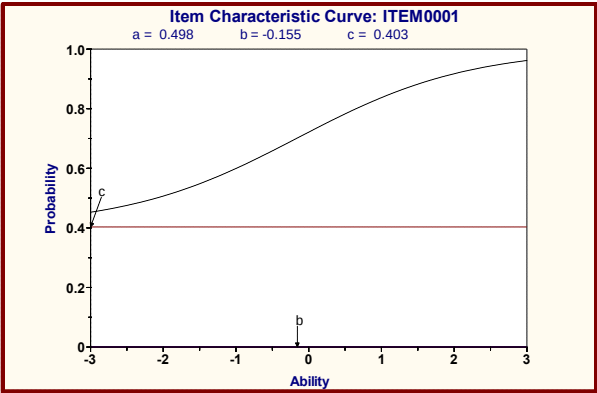
44 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE-3

2976 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE-3

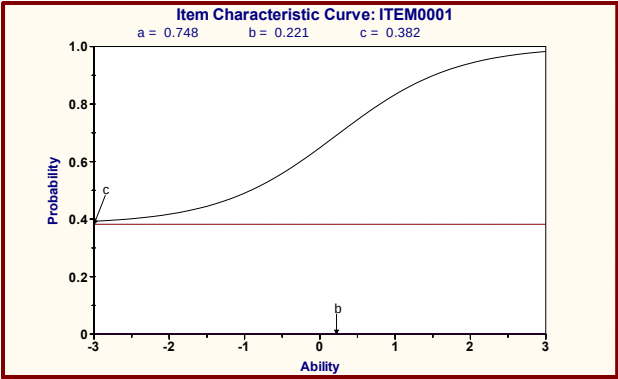
APPENDIX X
ITEM CHARACTERISTIC CURVE

ICCSTATE A

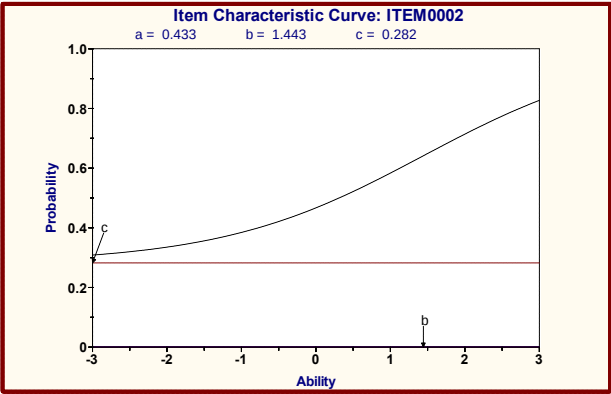
ICCSTATE B



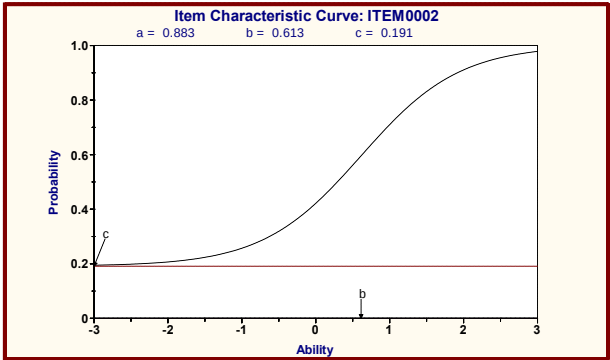
3-Parameter Model, Normal Metric Item: 1
Subtest: TEST0001



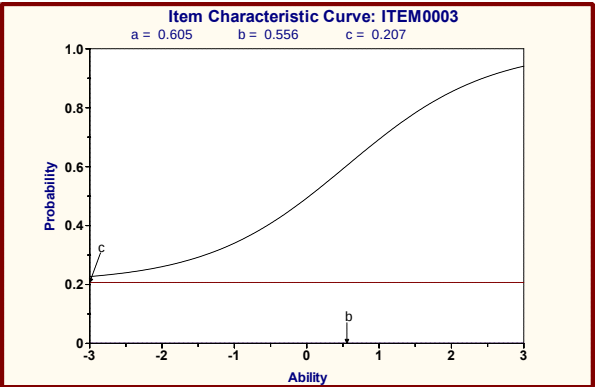
3-Parameter Model, Normal Metric Item: 1
Subtest: TEST0001



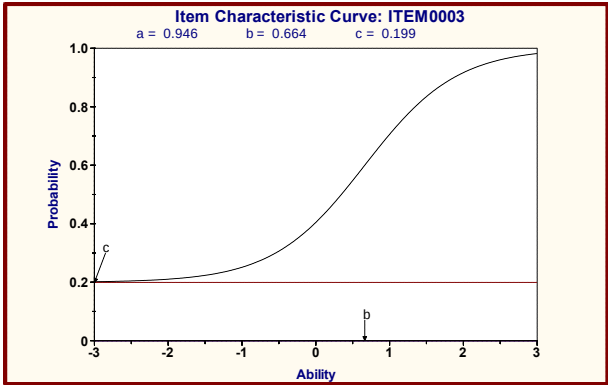
3-Parameter Model, Normal Metric Item: 2
Subtest: TEST0001



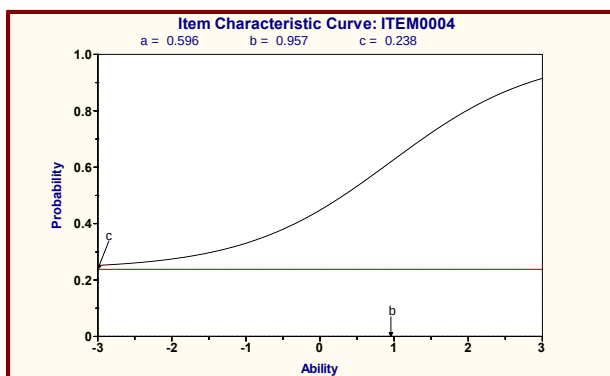
3-Parameter Model, Normal Metric Item: 2
Subtest: TEST0001



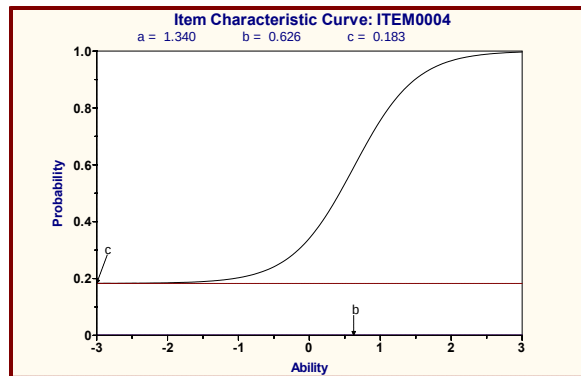
3-Parameter Model, Normal Metric Item: 3
Subtest: TEST0001



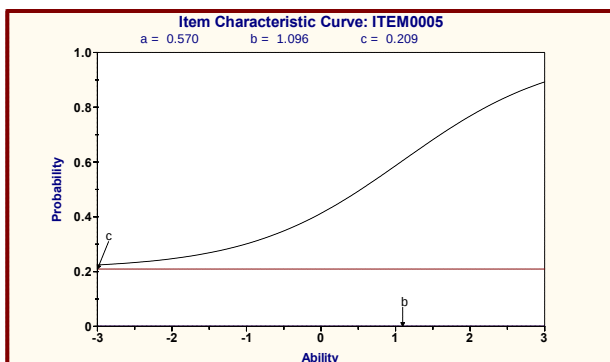
3-Parameter Model, Normal Metric Item: 3
Subtest: TEST0001



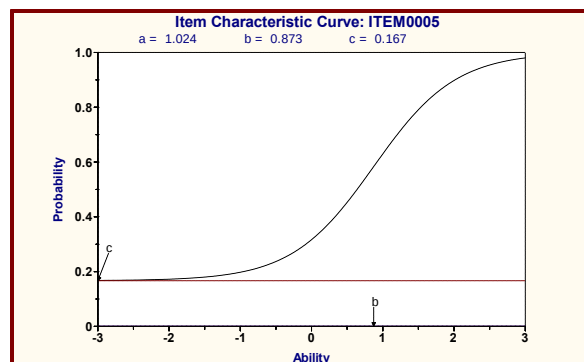
3-Parameter Model, Normal Metric Item: 4
 Subtest: TEST0001



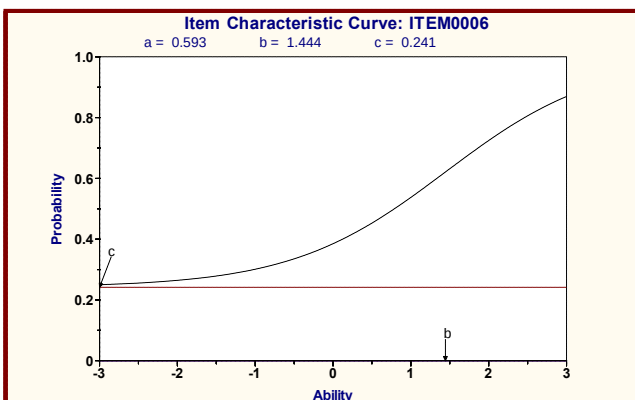
3-Parameter Model, Normal Metric Item: 4
 Subtest: TEST0001



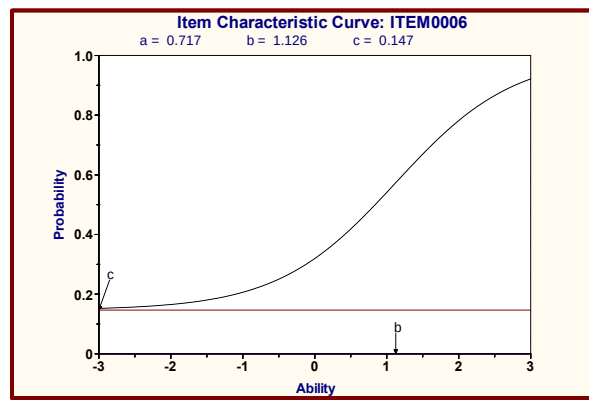
3-Parameter Model, Normal Metric Item: 5
 Subtest: TEST0001



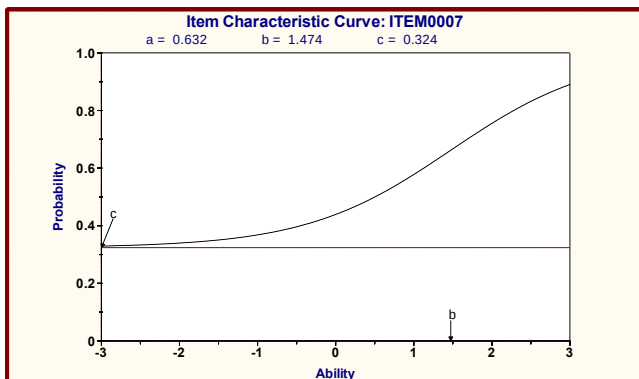
3-Parameter Model, Normal Metric Item: 5
 Subtest: TEST0001



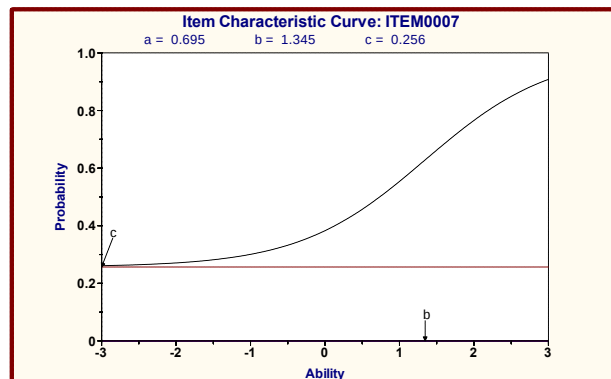
3-Parameter Model, Normal Metric Item: 6
 Subtest: TEST0001



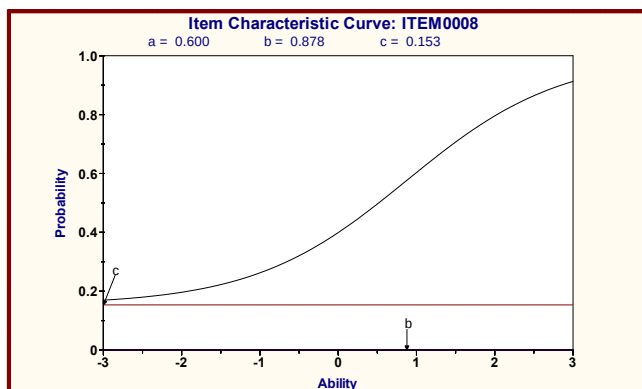
3-Parameter Model, Normal Metric Item: 6
 Subtest: TEST0001



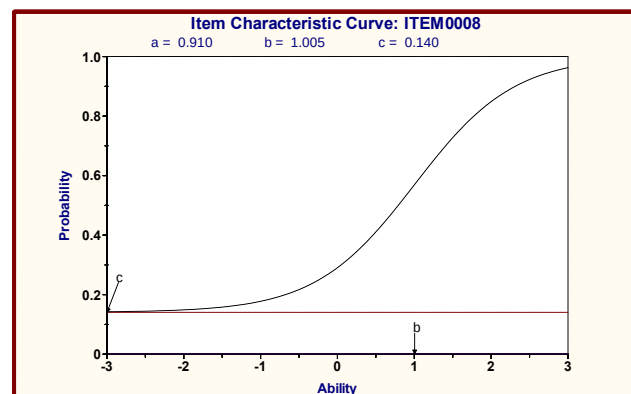
3-Parameter Model, Normal Metric Item: 7
 Subtest: TEST0001



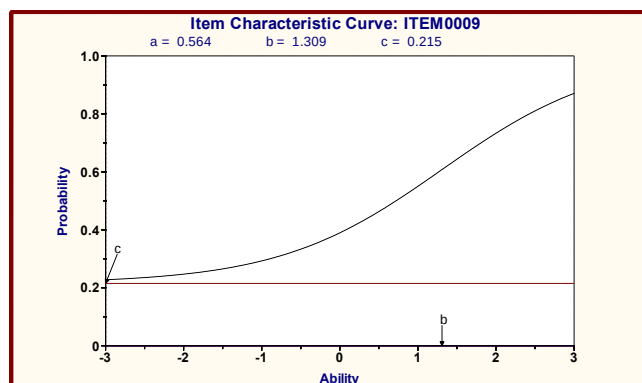
3-Parameter Model, Normal Metric Item: 7
 Subtest: TEST0001



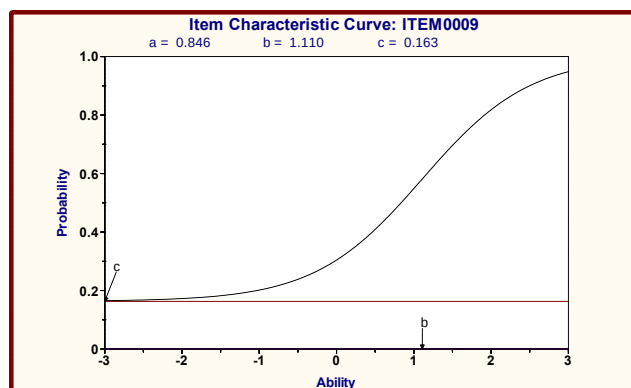
3-Parameter Model, Normal Metric Item: 8
 Subtest: TEST0001



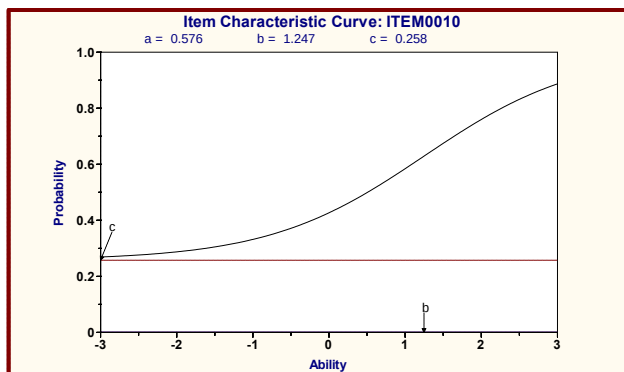
3-Parameter Model, Normal Metric Item: 8
 Subtest: TEST0001



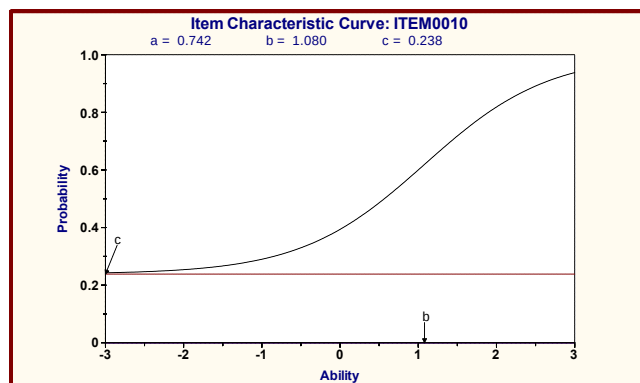
3-Parameter Model, Normal Metric Item: 9
 Subtest: TEST0001



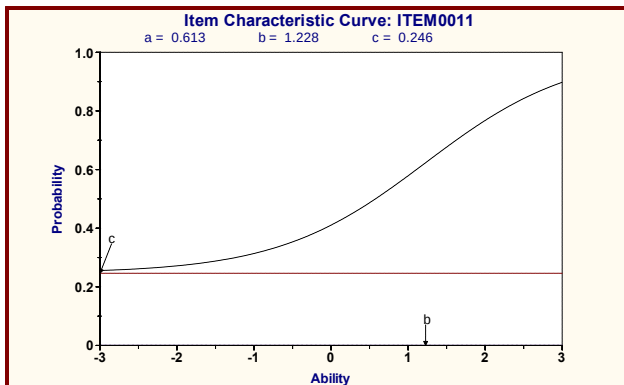
3-Parameter Model, Normal Metric Item: 9
 Subtest: TEST0001



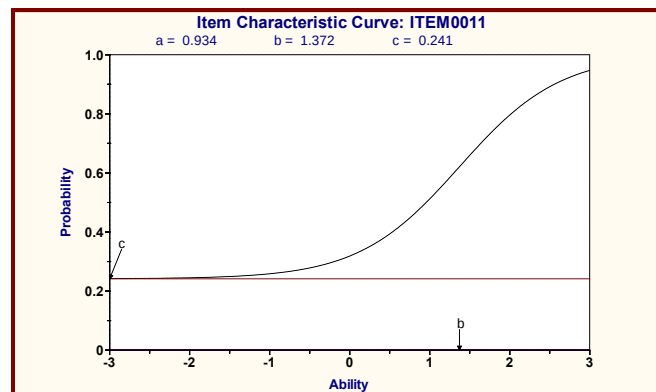
3-Parameter Model, Normal Metric Item: 10
Subtest: TEST0001



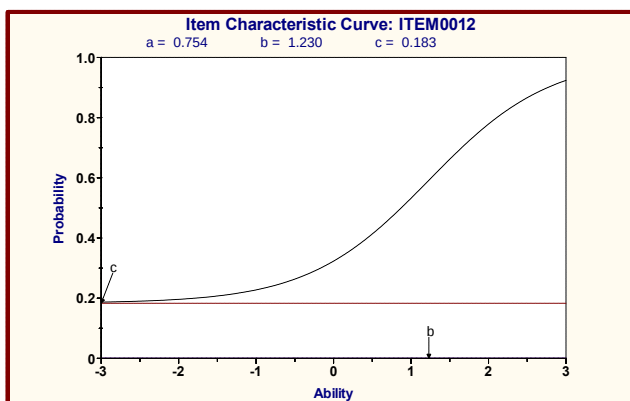
3-Parameter Model, Normal Metric Item: 10
Subtest: TEST0001



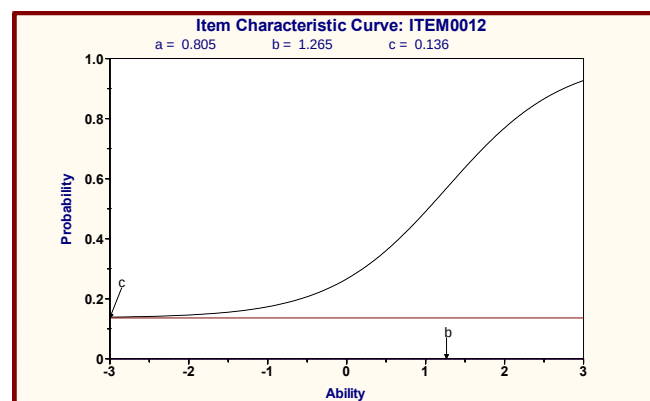
3-Parameter Model, Normal Metric Item: 11
Subtest: TEST0001



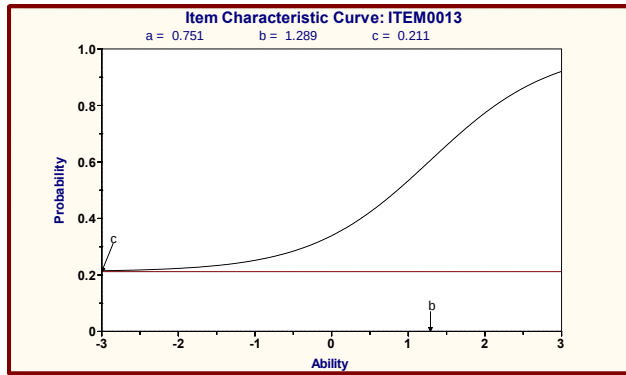
3-Parameter Model, Normal Metric Item: 11
Subtest: TEST0001



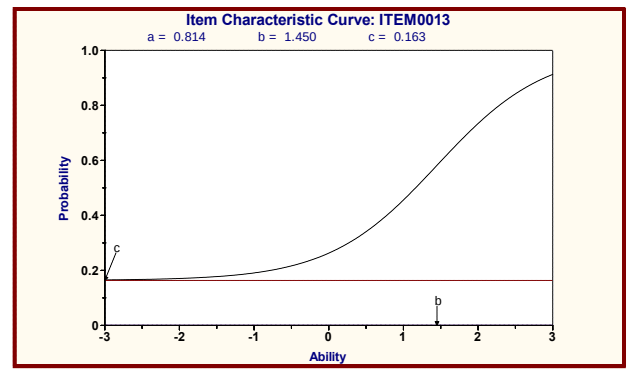
3-Parameter Model, Normal Metric Item: 12
Subtest: TEST0001



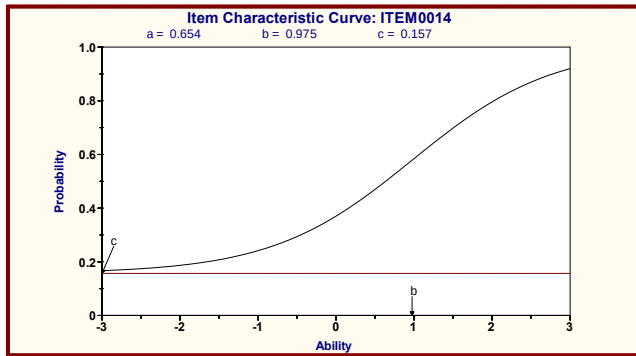
3-Parameter Model, Normal Metric Item: 12
Subtest: TEST0001



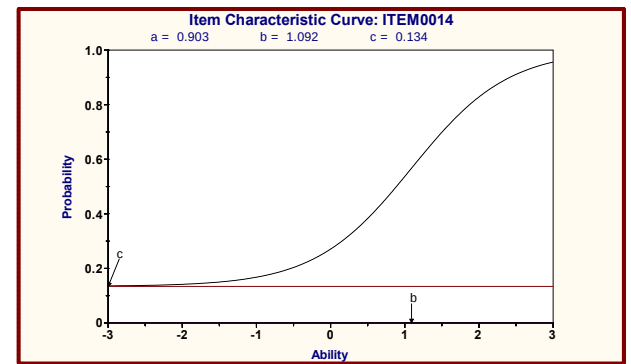
3-Parameter Model, Normal Metric Item: 13
Subtest: TEST0001



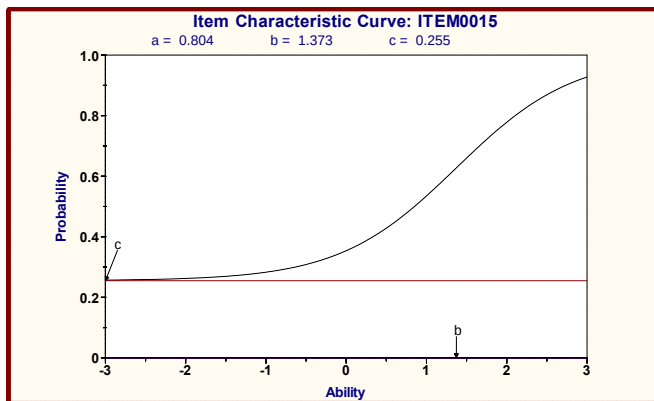
3-Parameter Model, Normal Metric Item: 13
Subtest: TEST0001



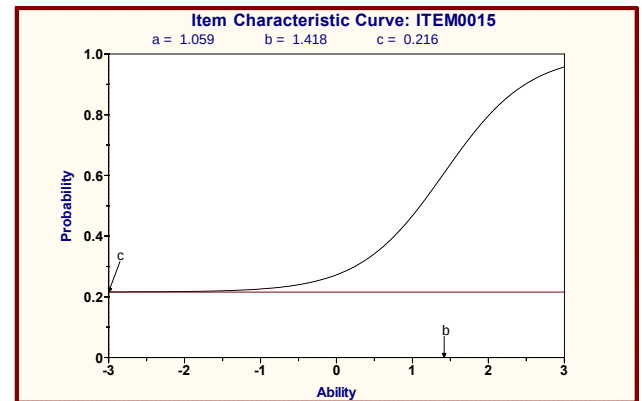
3-Parameter Model, Normal Metric Item: 14
Subtest: TEST0001



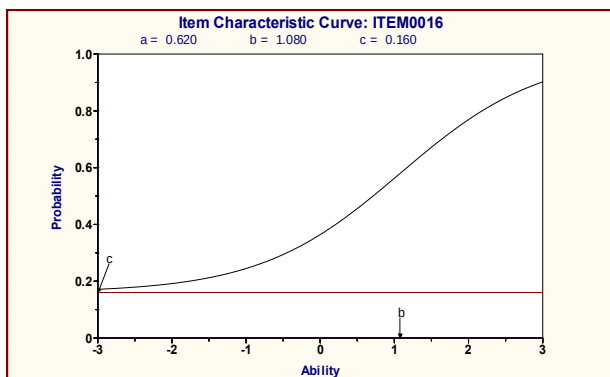
3-Parameter Model, Normal Metric Item: 14
Subtest: TEST0001



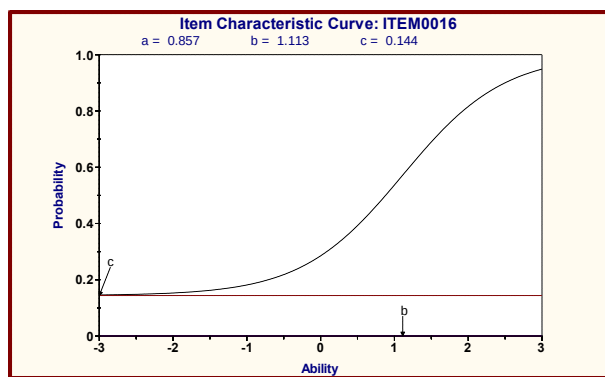
3-Parameter Model, Normal Metric Item: 15
Subtest: TEST0001



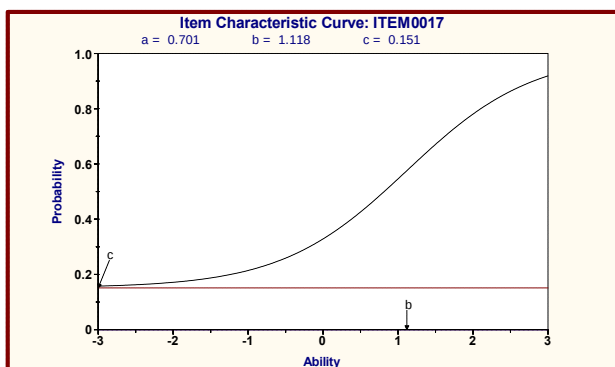
3-Parameter Model, Normal Metric Item: 15
Subtest: TEST0001



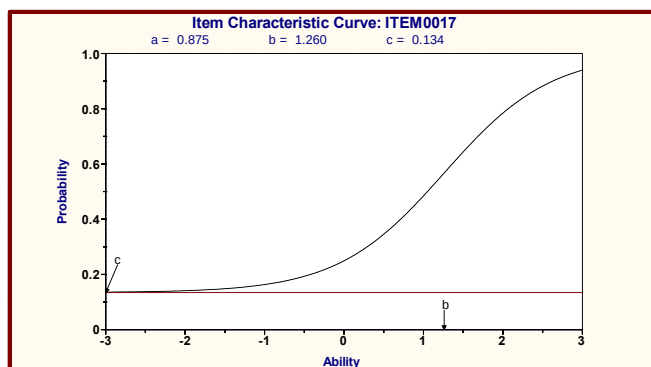
3-Parameter Model, Normal Metric **Item: 16**
Subtest: TEST0001



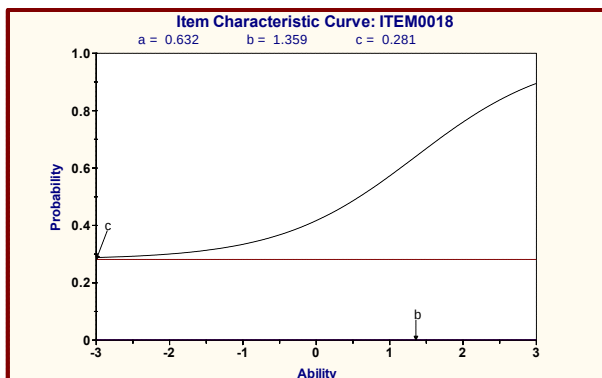
3-Parameter Model, Normal Metric **Item: 16**
Subtest: TEST0001



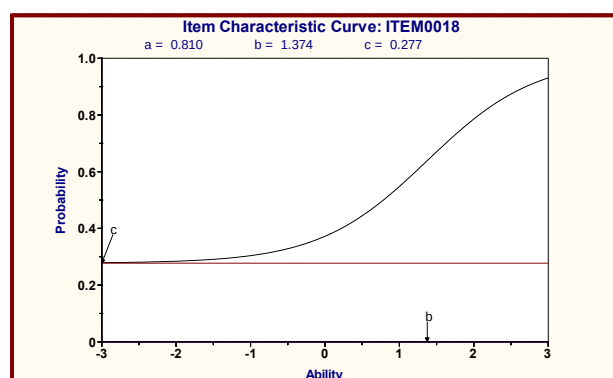
3-Parameter Model, Normal Metric **Item: 17**
Subtest: TEST0001



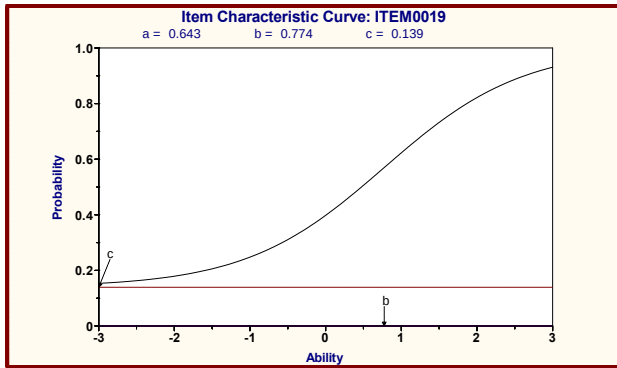
3-Parameter Model, Normal Metric **Item: 17**
Subtest: TEST0001



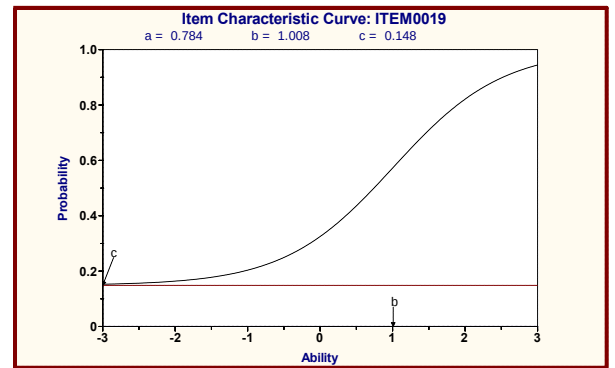
3-Parameter Model, Normal Metric **Item: 18**
Subtest: TEST0001



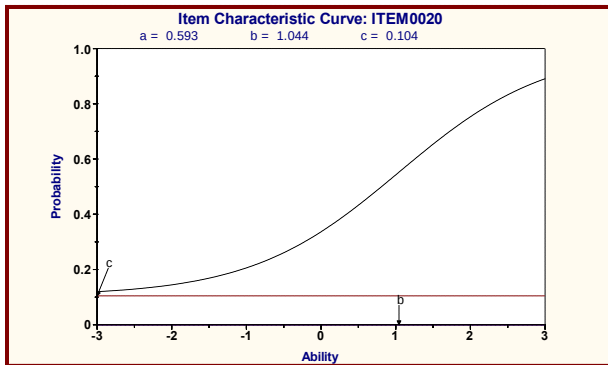
3-Parameter Model, Normal Metric **Item: 18**
Subtest: TEST0001



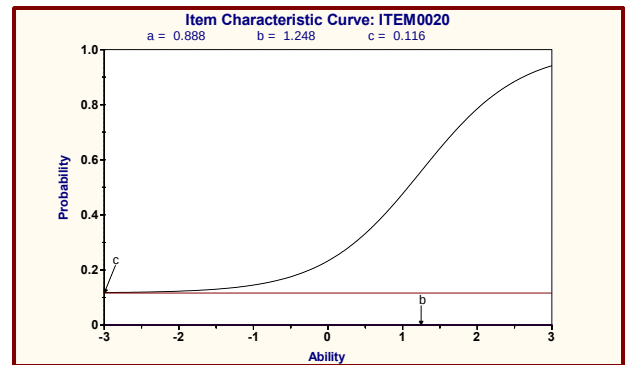
3-Parameter Model, Normal Metric Item: 19
Subtest: TEST0001



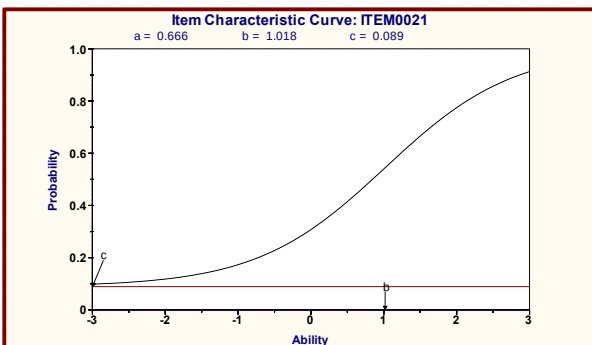
3-Parameter Model, Normal Metric Item: 19
Subtest: TEST0001



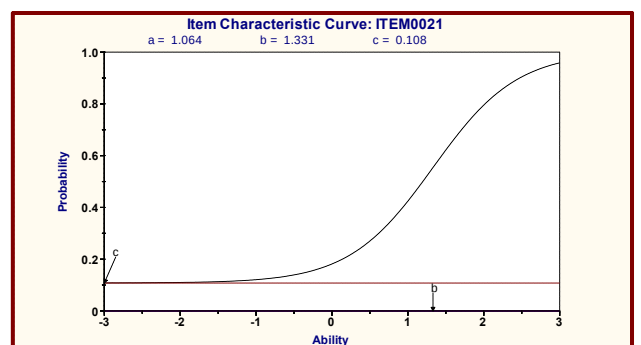
3-Parameter Model, Normal Metric Item: 20
Subtest: TEST0001



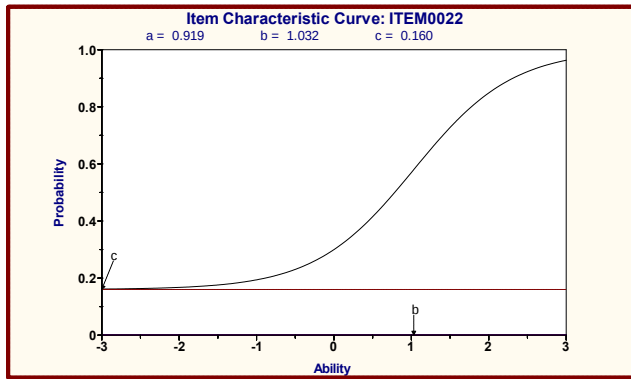
3-Parameter Model, Normal Metric Item: 20
Subtest: TEST0001



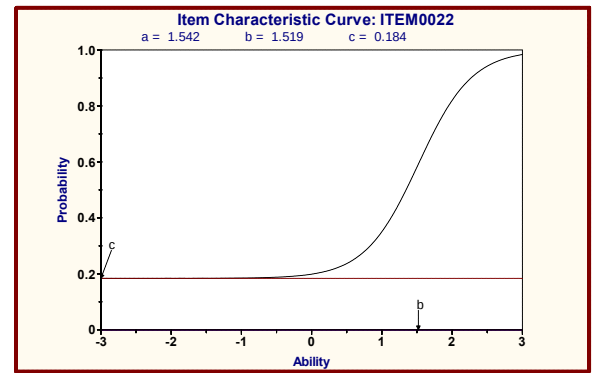
3-Parameter Model, Normal Metric Item: 21
Subtest: TEST0001



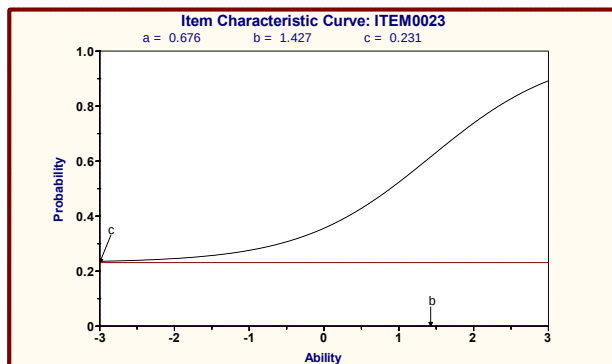
3-Parameter Model, Normal Metric Item: 21
Subtest: TEST0001



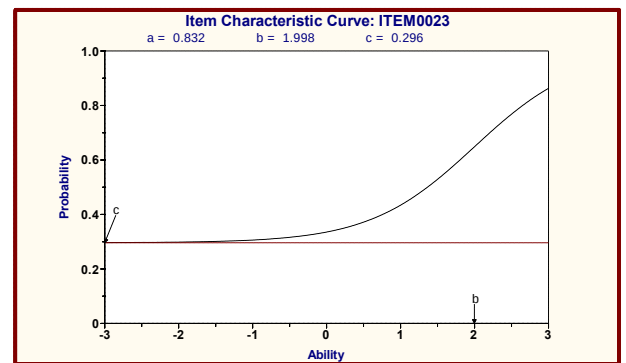
3-Parameter Model, Normal Metric Item: 22
 Subtest: TEST0001



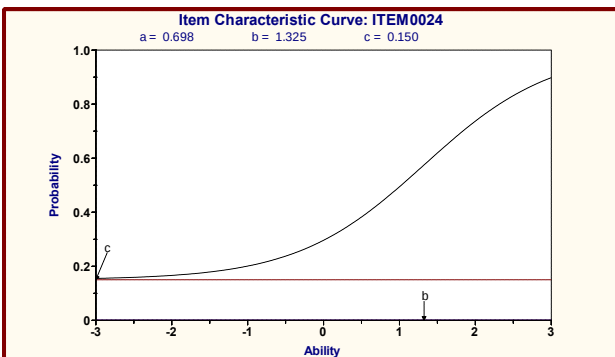
3-Parameter Model, Normal Metric Item: 22
 Subtest: TEST0001



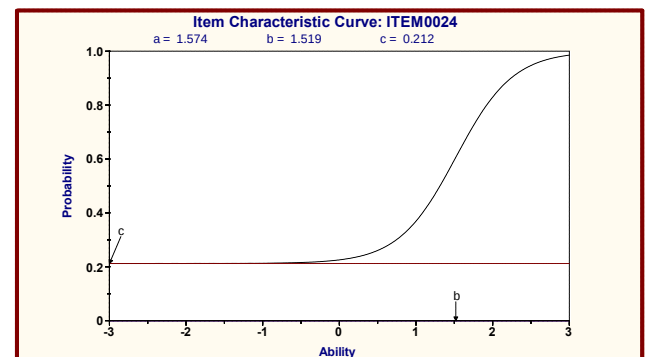
3-Parameter Model, Normal Metric Item: 23
 Subtest: TEST0001



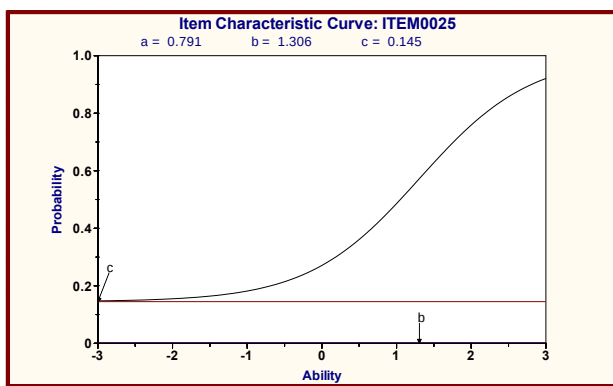
3-Parameter Model, Normal Metric Item: 23
 Subtest: TEST0001



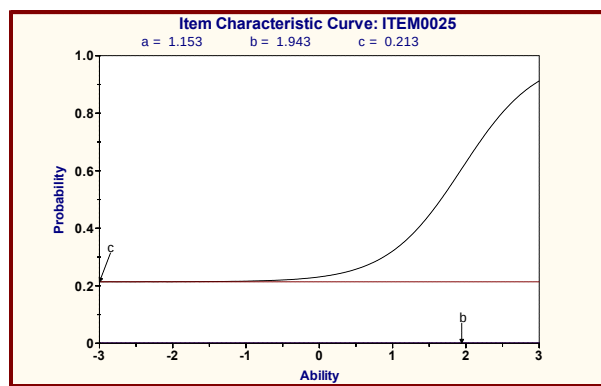
3-Parameter Model, Normal Metric Item: 24
 Subtest: TEST0001



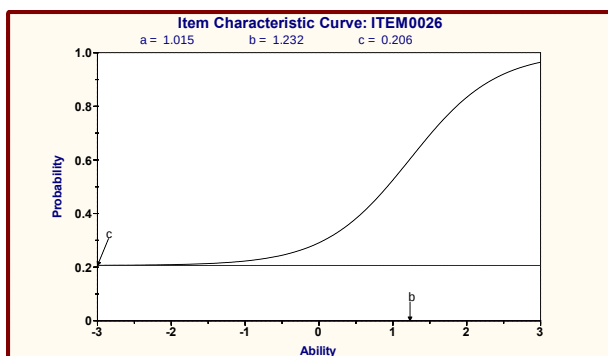
3-Parameter Model, Normal Metric Item: 24
 Subtest: TEST0001



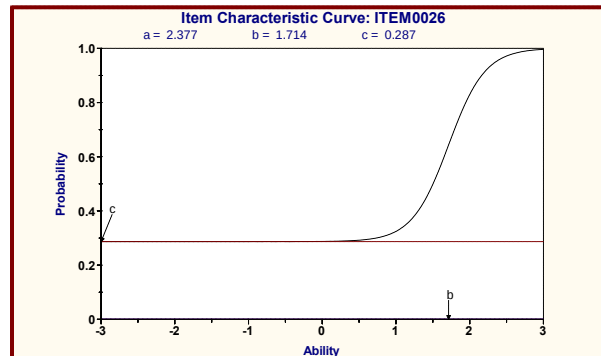
3-Parameter Model, Normal Metric Item: 25
 Subtest: TEST0001



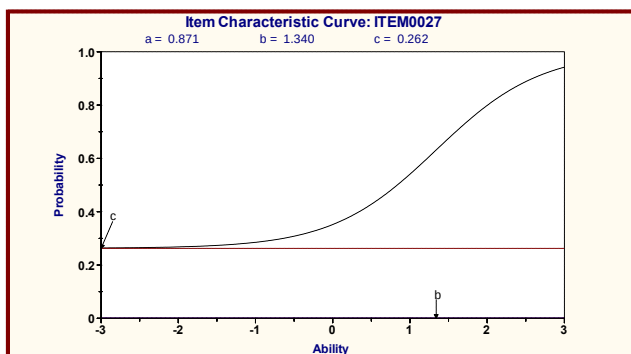
3-Parameter Model, Normal Metric Item: 25
 Subtest: TEST0001



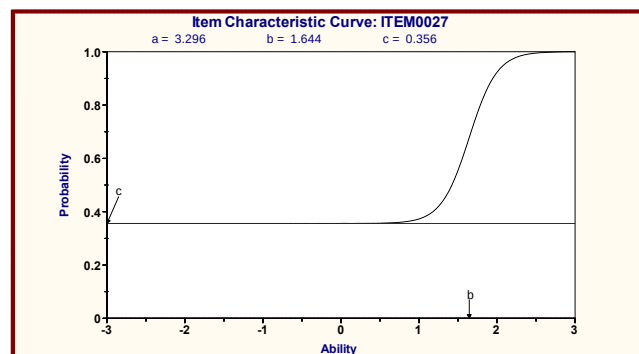
3-Parameter Model, Normal Metric Item: 26
 Subtest: TEST0001



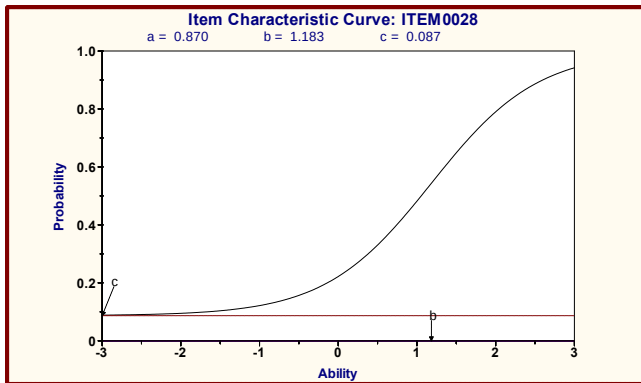
3-Parameter Model, Normal Metric Item: 26
 Subtest: TEST0001



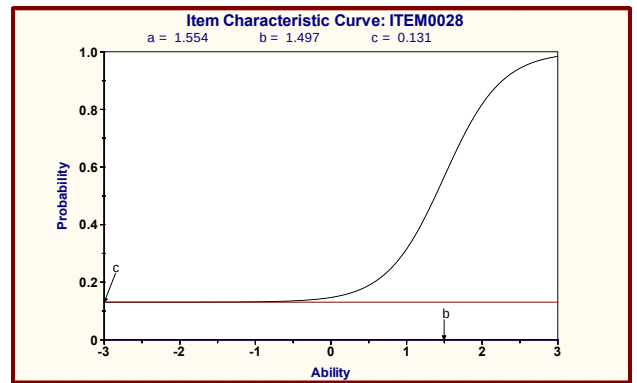
3-Parameter Model, Normal Metric Item: 27
 Subtest: TEST0001



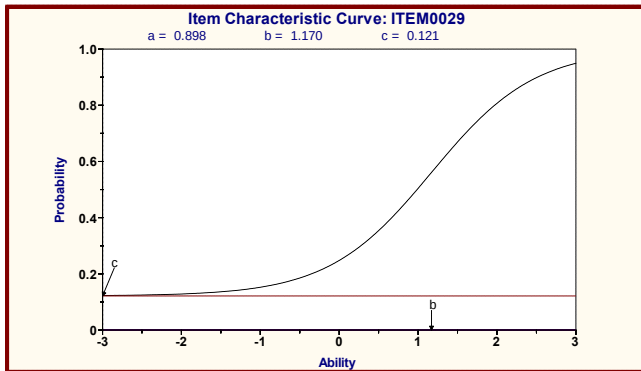
3-Parameter Model, Normal Metric Item: 27
 Subtest: TEST0001



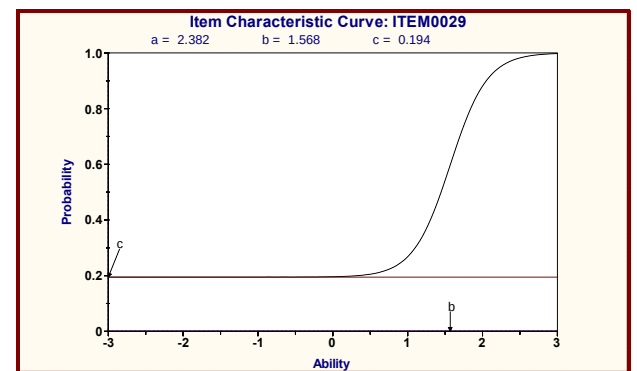
3-Parameter Model, Normal Metric Item: 28
Subtest: TEST0001



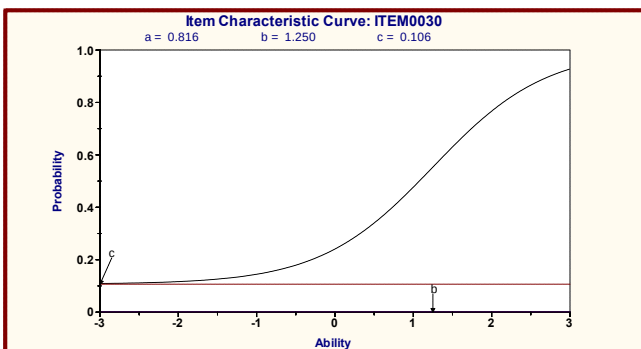
3-Parameter Model, Normal Metric Item: 28
Subtest: TEST0001



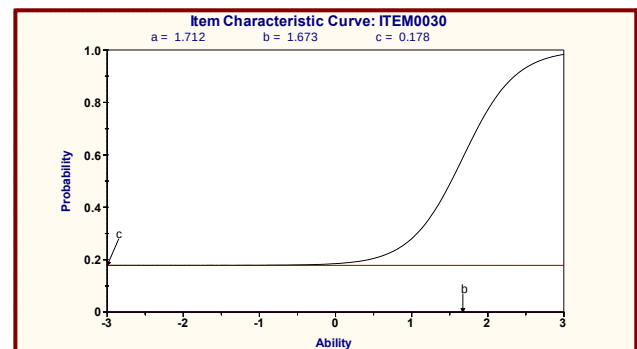
3-Parameter Model, Normal Metric Item: 29
Subtest: TEST0001



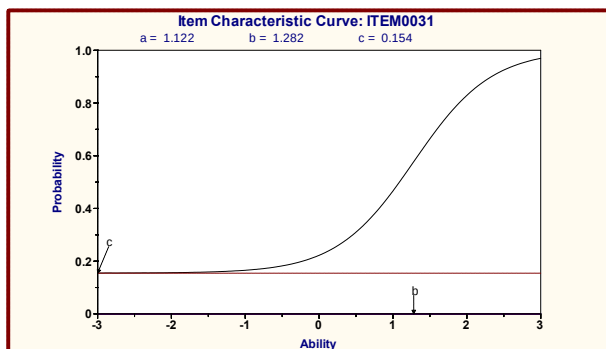
3-Parameter Model, Normal Metric Item: 29
Subtest: TEST0001



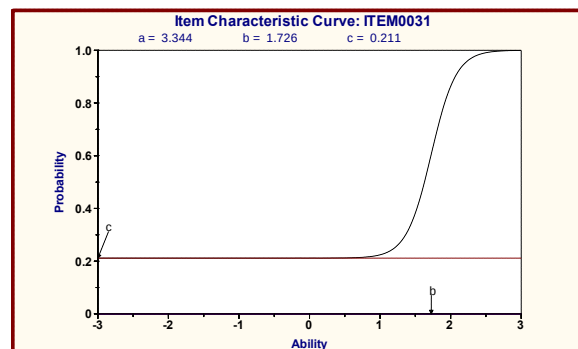
3-Parameter Model, Normal Metric Item: 30
Subtest: TEST0001



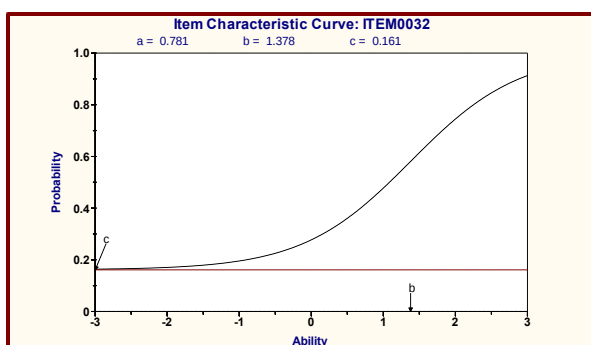
3-Parameter Model, Normal Metric Item: 30
Subtest: TEST0001



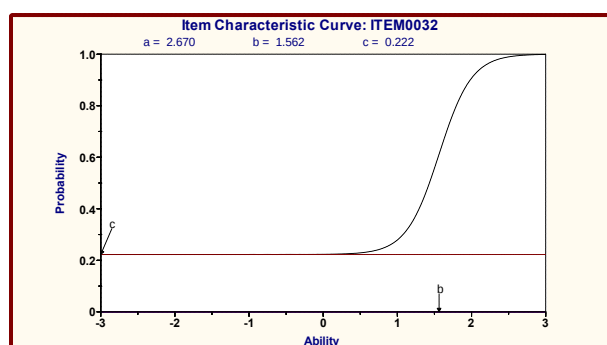
3-Parameter Model, Normal Metric Item: 31
Subtest: TEST0001



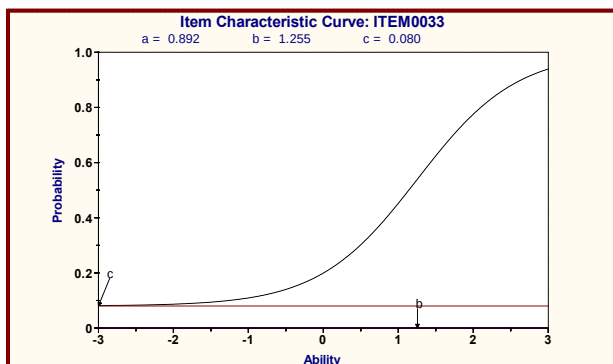
3-Parameter Model, Normal Metric Item: 31
Subtest: TEST0001



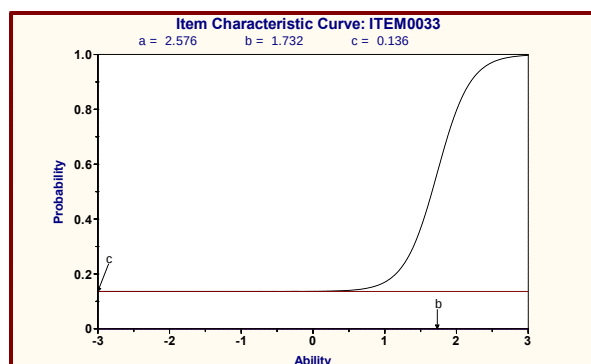
3-Parameter Model, Normal Metric Item: 32
Subtest: TEST0001



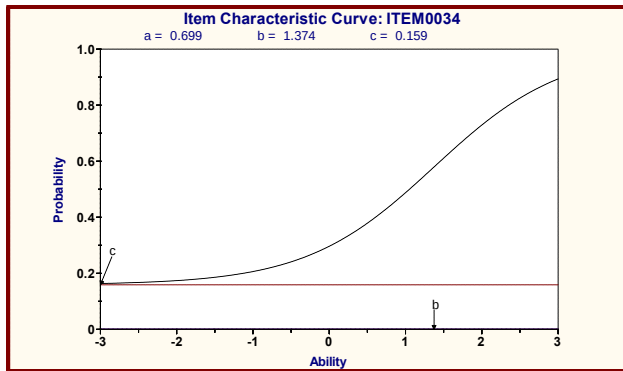
3-Parameter Model, Normal Metric Item: 32
Subtest: TEST0001



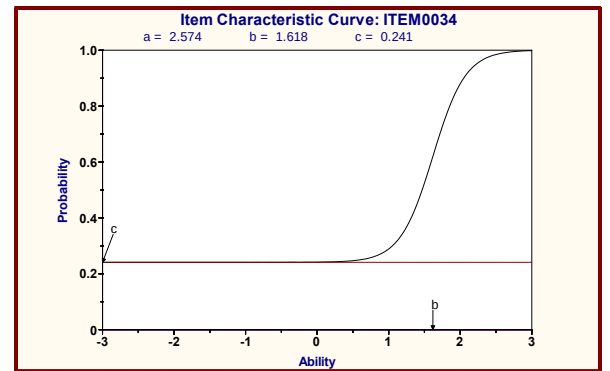
3-Parameter Model, Normal Metric Item: 33
Subtest: TEST0001



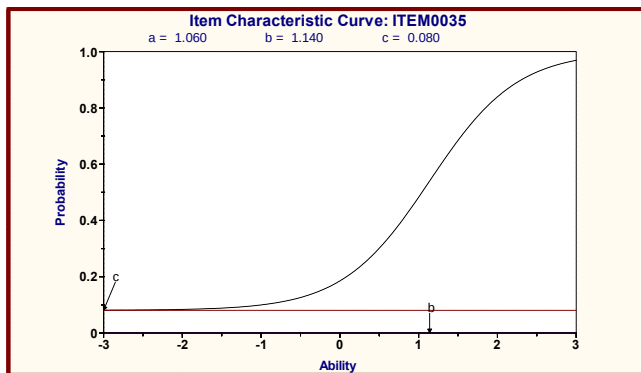
3-Parameter Model, Normal Metric Item: 33
Subtest: TEST0001



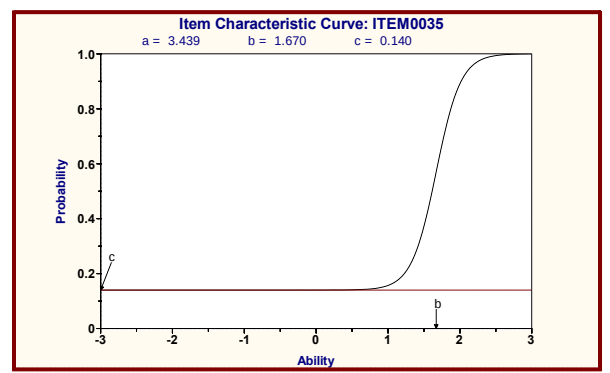
3-Parameter Model, Normal Metric Item: 34
 Subtest: TEST0001



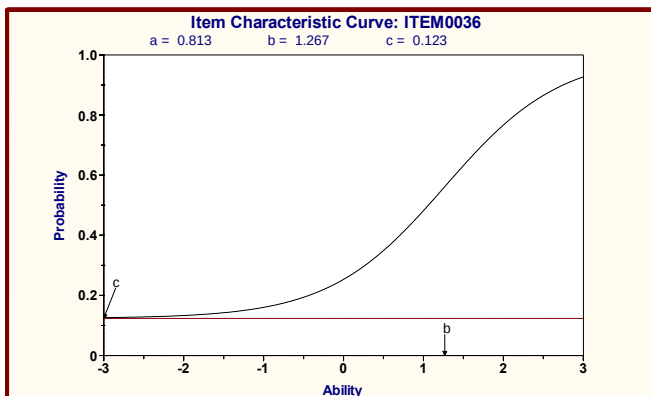
3-Parameter Model, Normal Metric Item: 34
 Subtest: TEST0001



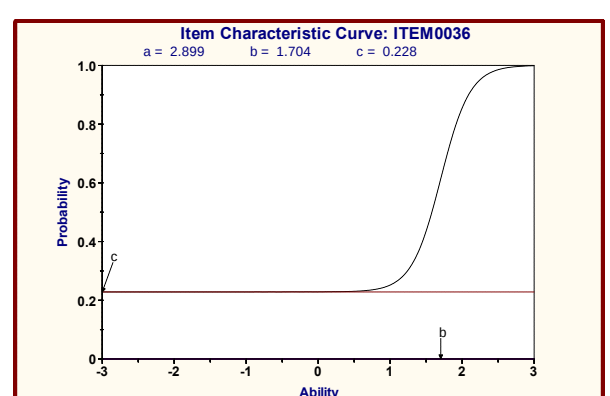
3-Parameter Model, Normal Metric Item: 35
 Subtest: TEST0001



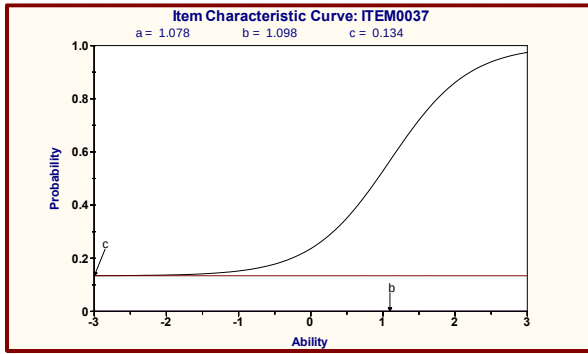
3-Parameter Model, Normal Metric Item: 35
 Subtest: TEST0001



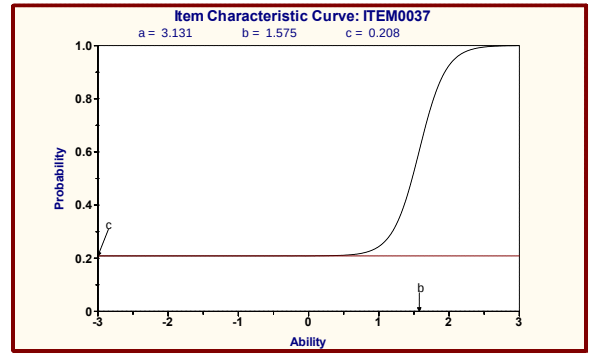
3-Parameter Model, Normal Metric Item: 36
 Subtest: TEST0001



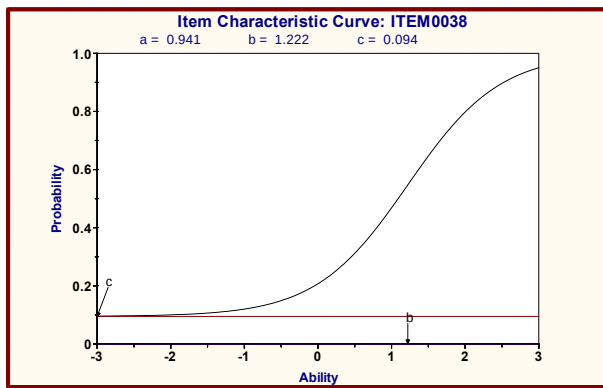
3-Parameter Model, Normal Metric Item: 36
 Subtest: TEST0001



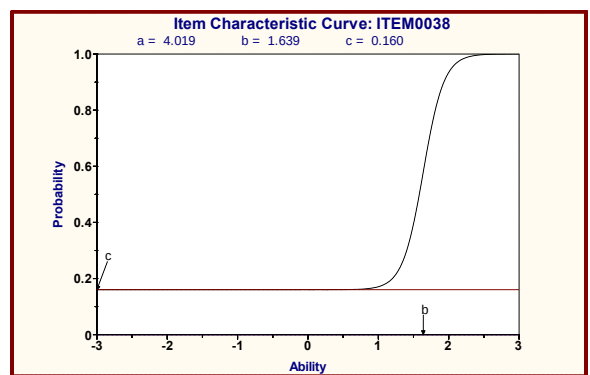
3-Parameter Model, Normal Metric Item: 37
 Subtest: TEST0001



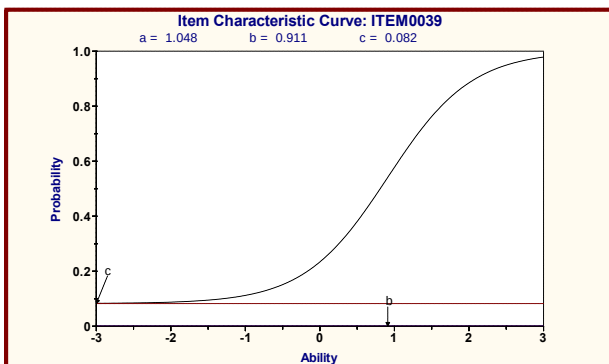
3-Parameter Model, Normal Metric Item: 37
 Subtest: TEST0001



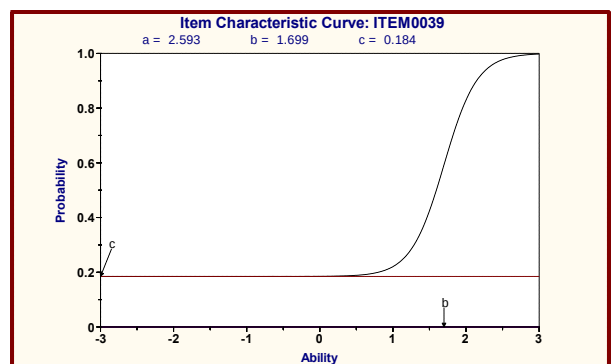
3-Parameter Model, Normal Metric Item: 38
 Subtest: TEST0001



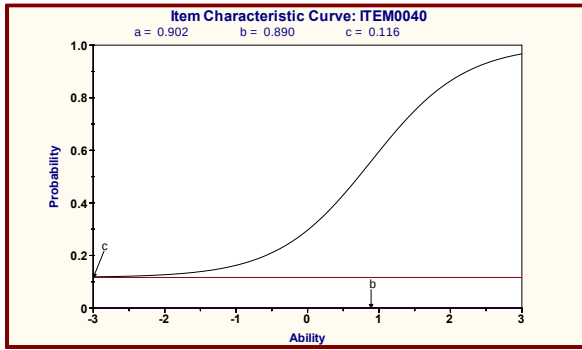
3-Parameter Model, Normal Metric Item: 38
 Subtest: TEST0001



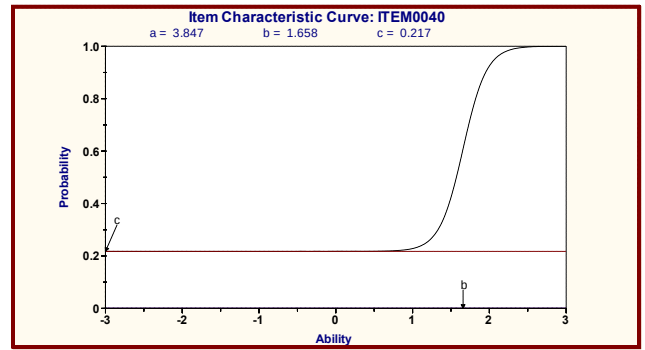
3-Parameter Model, Normal Metric Item: 39
 Subtest: TEST0001



3-Parameter Model, Normal Metric Item: 39
 Subtest: TEST0001



3-Parameter Model, Normal Metric **Item: 40**
Subtest: TEST0001



3-Parameter Model, Normal Metric **Item: 40**
Subtest: TEST0001

APPENDIX Y
LINEAR EQUATING
OUTPUT

STATE A

S/N	CAS A	TS A	LE A										
1	24	26	76	38	22	18	66	81	19	9	54		
2	21	26	76	39	24	18	66	82	18	8	53		
3	20	25	75	40	22	18	66	83	17	7	52		
4	20	24	74	41	20	18	66	84	21	35	88		
5	23	23	72	42	18	18	66	85	20	34	86		
6	22	23	72	43	15	18	66	86	20	32	84		
7	22	22	71	44	19	17	65	87	20	32	84		
8	18	22	71	45	22	17	65	88	24	30	81		
9	21	22	71	46	23	17	65	89	22	30	81		
10	18	21	70	47	20	17	65	90	22	30	81		
11	20	21	70	48	21	17	65	91	18	30	81		
12	24	21	70	49	20	17	65	92	21	30	81		
13	21	21	70	50	18	17	65	93	18	29	80		
14	17	20	68	51	19	17	65	94	20	29	80		
15	16	20	68	52	20	17	65	95	24	27	77		
16	20	20	68	53	18	16	63	96	21	27	77		
17	18	20	68	54	22	16	63	97	17	27	77		
18	20	20	68	55	19	16	63	98	16	27	77		
19	25	20	68	56	18	16	63	99	20	27	77		
20	18	20	68	57	17	16	63	100	18	26	76		
21	19	20	68	58	18	16	63	101	20	26	76		
22	23	20	68	59	19	16	63	102	25	25	75		
23	25	20	68	60	23	16	63	103	18	25	75		
24	20	19	67	61	22	16	63	104	19	25	75		
25	23	19	67	62	21	15	62	105	23	25	75		
26	20	19	67	63	19	15	62	106	25	25	75		
27	19	19	67	64	17	15	62	107	20	24	74		
28	24	19	67	65	16	15	62	108	23	24	74		
29	22	19	67	66	19	15	62	109	20	24	74		
30	20	19	67	67	14	14	61	110	19	24	74		
31	16	19	67	68	14	14	61	111	24	24	74		
32	20	19	67	69	14	14	61	112	22	24	74		
33	19	19	67	70	13	13	59	113	20	23	72		
34	15	19	67	71	13	13	59	114	16	23	72		
35	18	18	66	72	12	12	58	115	20	22	71		
36	18	18	66	73	12	12	58	116	19	22	71		
37	19	18	66	74	12	12	58	117	15	22	71		
				75	11	11	57	118	21	22	71		
				76	11	11	57	119	20	22	71		
				77	20	10	56	120	24	22	71		
				78	20	10	56	121	24	22	71		
				79	20	10	56	122	24	22	71		
				80	19	9	54	123	20	22	71		

124	20	22	71	167	18	18	66	210	18	27	77
125	21	22	71	168	19	17	65	211	16	27	77
126	21	22	71	169	19	17	65	212	18	27	77
127	20	22	71	170	15	15	62	213	21	27	77
128	23	22	71	171	18	13	59	214	22	27	77
129	18	22	71	172	18	12	58	215	21	27	77
130	22	22	71	173	14	12	58	216	23	27	77
131	23	22	71	174	19	11	57	217	27	27	77
132	20	22	71	175	19	9	54	218	26	26	76
133	21	21	70	176	19	9	54	219	22	26	76
134	16	21	70	177	24	36	89	220	20	26	76
135	18	21	70	178	20	36	89	221	25	26	76
136	20	21	70	179	23	35	88	222	20	26	76
137	18	21	70	180	21	33	85	223	24	26	76
138	22	21	70	181	20	32	84	224	22	26	76
139	20	21	70	182	21	32	84	225	20	26	76
140	18	21	70	183	22	32	84	226	23	26	76
141	20	21	70	184	24	32	84	227	21	26	76
142	21	21	70	185	20	32	84	228	25	25	75
143	16	21	70	186	19	31	83	229	21	25	75
144	19	21	70	187	20	31	83	230	22	25	75
145	18	21	70	188	22	31	83	231	27	25	75
146	21	21	70	189	22	30	81	232	21	25	75
147	21	21	70	190	21	30	81	233	22	25	75
148	23	21	70	191	22	30	81	234	24	24	74
149	20	20	68	192	21	30	81	235	21	24	74
150	20	20	68	193	22	30	81	236	20	24	74
151	20	20	68	194	24	30	81	237	26	24	74
152	21	20	68	195	20	30	81	238	27	24	74
153	20	20	68	196	19	29	80	239	21	24	74
154	20	20	68	197	22	29	80	240	23	24	74
155	20	20	68	198	18	29	80	241	21	24	74
156	24	20	68	199	23	29	80	242	20	24	74
157	22	20	68	200	24	29	80	243	19	24	74
158	22	20	68	201	20	28	79	244	22	24	74
159	18	20	68	202	21	28	79	245	18	24	74
160	21	20	68	203	22	28	79	246	17	24	74
161	18	20	68	204	21	28	79	247	19	24	74
162	19	19	67	205	18	28	79	248	20	24	74
163	22	19	67	206	19	28	79	249	15	24	74
164	20	19	67	207	21	28	79	250	18	24	74
165	18	18	66	208	20	28	79	251	17	24	74
166	19	18	66	209	19	27	77	252	18	24	74

253	18	24	74	296	20	19	67	339	21	12	58
254	22	23	72	297	21	19	67	340	18	12	58
255	21	23	72	298	27	19	67	341	18	12	58
256	23	23	72	299	20	19	67	342	17	12	58
257	22	23	72	300	21	19	67	343	21	12	58
258	23	23	72	301	22	18	66	344	21	12	58
259	20	23	72	302	22	18	66	345	22	11	57
260	19	23	72	303	24	18	66	346	23	11	57
261	18	23	72	304	21	18	66	347	21	11	57
262	18	23	72	305	20	18	66	348	24	11	57
263	19	23	72	306	19	18	66	349	22	10	56
264	21	23	72	307	23	18	66	350	21	10	56
265	20	22	71	308	20	18	66	351	21	10	56
266	25	22	71	309	22	17	65	352	24	10	56
267	25	22	71	310	21	17	65	353	18	9	54
268	25	20	68	311	20	17	65	354	16	9	54
269	26	20	68	312	19	17	65	355	15	9	54
270	27	20	68	313	19	17	65	356	16	9	54
271	22	20	68	314	20	17	65	357	17	9	54
272	23	20	68	315	21	17	65	358	20	8	53
273	24	20	68	316	22	17	65	359	18	8	53
274	26	20	68	317	23	17	65	360	19	8	53
275	27	20	68	318	24	16	63	361	20	21	70
276	20	20	68	319	20	16	63	362	21	19	67
277	22	20	68	320	21	15	62	363	18	18	66
278	20	20	68	321	22	15	62	364	22	18	66
279	21	20	68	322	20	15	62	365	22	17	65
280	19	20	68	323	19	15	62	366	19	17	65
281	20	20	68	324	18	14	61	367	21	17	65
282	18	20	68	325	15	14	61	368	22	16	63
283	15	20	68	326	18	14	61	369	20	16	63
284	19	20	68	327	19	14	61	370	20	16	63
285	17	20	68	328	20	14	61	371	18	16	63
286	20	20	68	329	19	14	61	372	18	15	62
287	22	20	68	330	18	14	61	373	20	15	62
288	23	20	68	331	18	14	61	374	20	15	62
289	24	20	68	332	16	14	61	375	19	15	62
290	22	19	67	333	17	14	61	376	22	15	62
291	23	19	67	334	18	13	59	377	21	15	62
292	24	19	67	335	18	13	59	378	21	15	62
293	21	19	67	336	19	13	59	379	20	15	62
294	23	19	67	337	21	13	59	380	17	14	61
295	22	19	67	338	20	13	59	381	22	14	61

382	22	14	61	425	20	9	54	468	21	18	66
383	17	14	61	426	21	9	54	469	22	18	66
384	18	14	61	427	19	9	54	470	19	18	66
385	21	14	61	428	20	9	54	471	20	17	65
386	20	13	59	429	17	9	54	472	21	17	65
387	17	13	59	430	17	9	54	473	22	17	65
388	21	13	59	431	20	8	53	474	21	17	65
389	17	13	59	432	20	8	53	475	18	17	65
390	21	13	59	433	21	8	53	476	19	17	65
391	21	13	59	434	23	8	53	477	21	17	65
392	18	12	58	435	19	8	53	478	19	17	65
393	19	12	58	436	20	8	53	479	17	17	65
394	23	12	58	437	19	8	53	480	18	16	63
395	20	12	58	438	20	7	52	481	20	16	63
396	19	12	58	439	22	7	52	482	21	15	62
397	23	12	58	440	21	7	52	483	18	13	59
398	18	12	58	441	18	7	52	484	16	11	57
399	21	12	58	442	23	6	50	485	17	11	57
400	18	12	58	443	22	6	50	486	19	10	56
401	19	12	58	444	20	6	50	487	17	9	54
402	18	12	58	445	23	5	49	488	18	9	54
403	18	12	58	446	24	5	49	489	19	9	54
404	17	11	57	447	22	24	74	490	25	26	76
405	23	11	57	448	24	23	72	491	23	24	74
406	20	11	57	449	20	22	71	492	24	24	74
407	21	11	57	450	21	22	71	493	21	22	71
408	22	11	57	451	19	22	71	494	20	22	71
409	22	11	57	452	18	21	70	495	21	22	71
410	18	11	57	453	21	21	70	496	20	22	71
411	20	11	57	454	22	21	70	497	19	22	71
412	24	11	57	455	21	21	70	498	20	21	70
413	19	11	57	456	23	20	68	499	22	21	70
414	19	11	57	457	21	20	68	500	23	21	70
415	17	11	57	458	20	20	68	501	21	21	70
416	19	11	57	459	21	19	67	502	20	21	70
417	23	10	56	460	19	19	67	503	18	21	70
418	18	10	56	461	18	19	67	504	19	20	68
419	20	10	56	462	24	19	67	505	20	20	68
420	20	10	56	463	22	18	66	506	21	20	68
421	19	10	56	464	21	18	66	507	20	20	68
422	21	9	54	465	20	18	66	508	18	20	68
423	21	9	54	466	20	18	66	509	17	20	68
424	17	9	54	467	22	18	66	510	17	20	68

511	21	20	68	554	20	17	65	597	23	22	71
512	18	20	68	555	19	17	65	598	21	22	71
513	24	20	68	556	19	17	65	599	18	22	71
514	23	20	68	557	20	17	65	600	19	22	71
515	24	20	68	558	21	17	65	601	21	22	71
516	19	20	68	559	22	16	63	602	17	22	71
517	19	20	68	560	18	16	63	603	19	22	71
518	21	20	68	561	17	15	62	604	18	22	71
519	22	20	68	562	15	15	62	605	20	21	70
520	21	19	67	563	19	15	62	606	21	20	68
521	19	19	67	564	21	15	62	607	21	20	68
522	18	19	67	565	22	13	59	608	23	20	68
523	17	19	67	566	18	13	59	609	18	19	67
524	17	19	67	567	17	13	59	610	22	19	67
525	15	19	67	568	16	13	59	611	20	19	67
526	20	19	67	569	20	12	58	612	19	19	67
527	18	19	67	570	21	12	58	613	18	19	67
528	17	19	67	571	22	12	58	614	19	18	66
529	18	19	67	572	16	11	57	615	20	18	66
530	20	19	67	573	18	11	57	616	21	18	66
531	23	19	67	574	15	11	57	617	22	18	66
532	21	19	67	575	16	11	57	618	18	16	63
533	18	19	67	576	17	10	56	619	21	16	63
534	17	19	67	577	17	10	56	620	17	16	63
535	18	19	67	578	18	9	54	621	17	16	63
536	19	18	66	579	16	8	53	622	19	16	63
537	19	18	66	580	19	8	53	623	18	14	61
538	20	18	66	581	20	7	52	624	22	14	61
539	22	18	66	582	21	5	49	625	21	14	61
540	23	18	66	583	18	5	49	626	20	14	61
541	22	18	66	584	22	28	79	627	18	14	61
542	19	18	66	585	24	27	77	628	19	11	57
543	22	20	68	586	23	24	74	629	19	11	57
544	25	17	65	587	20	25	75	630	21	10	56
545	21	17	65	588	26	25	75	631	21	10	56
546	22	17	65	589	22	25	75	632	22	8	53
547	24	17	65	590	20	23	72	633	17	7	52
548	20	17	65	591	21	23	72	634	18	7	52
549	19	17	65	592	24	23	72	635	19	6	50
550	18	17	65	593	21	23	72	636	17	6	50
551	19	17	65	594	20	23	72	637	20	6	50
552	20	17	65	595	22	23	72	638	21	5	49
553	21	17	65	596	21	22	71	639	26	34	86

640	24	31	83	683	18	21	70	726	18	19	67
641	23	31	83	684	18	21	70	727	19	19	67
642	23	30	81	685	20	20	68	728	13	19	67
643	24	30	81	686	17	20	68	729	14	19	67
644	25	30	81	687	21	20	68	730	15	19	67
645	26	28	79	688	22	19	67	731	18	18	66
646	24	28	79	689	20	19	67	732	19	18	66
647	22	28	79	690	19	19	67	733	18	18	66
648	19	28	79	691	18	19	67	734	21	17	65
649	20	28	79	692	15	19	67	735	22	17	65
650	22	28	79	693	17	18	66	736	18	16	63
651	19	28	79	694	20	18	66	737	18	16	63
652	18	28	79	695	17	18	66	738	12	16	63
653	17	27	77	696	17	18	66	739	13	15	62
654	24	27	77	697	19	17	65	740	18	15	62
655	25	27	77	698	18	17	65	741	18	15	62
656	25	27	77	699	20	17	65	742	16	15	62
657	22	27	77	700	21	17	65	743	17	13	59
658	19	27	77	701	19	17	65	744	18	13	59
659	24	27	77	702	20	17	65	745	18	13	59
660	21	27	77	703	19	15	62	746	19	13	59
661	20	27	77	704	18	15	62	747	19	12	58
662	23	27	77	705	15	11	57	748	20	12	58
663	17	26	76	706	16	7	52	749	21	11	57
664	19	26	76	707	16	7	52	750	20	10	56
665	20	26	76	708	20	6	50	751	15	10	56
666	22	26	76	709	20	6	50	752	16	10	56
667	21	26	76	710	20	27	77	753	15	10	56
668	23	25	75	711	24	23	72	754	24	23	72
669	25	25	75	712	20	23	72	755	20	22	71
670	24	24	74	713	19	21	70	756	21	22	71
671	21	24	74	714	21	21	70	757	19	21	70
672	18	24	74	715	18	21	70	758	22	21	70
673	19	24	74	716	21	21	70	759	21	21	70
674	18	24	74	717	20	20	68	760	23	21	70
675	22	24	74	718	19	20	68	761	24	21	70
676	21	24	74	719	18	20	68	762	20	20	68
677	22	24	74	720	20	20	68	763	20	20	68
678	23	24	74	721	17	20	68	764	21	20	68
679	24	22	71	722	18	20	68	765	18	20	68
680	22	22	71	723	15	20	68	766	19	20	68
681	21	22	71	724	16	20	68	767	20	20	68
682	22	22	71	725	16	19	67	768	22	20	68

769	19	20	68	812	21	21	70	855	21	15	62
770	24	19	67	813	17	20	68	856	22	15	62
771	23	19	67	814	20	20	68	857	19	15	62
772	24	19	67	815	16	20	68	858	18	14	61
773	20	19	67	816	18	19	67	859	20	14	61
774	18	19	67	817	22	19	67	860	21	14	61
775	17	19	67	818	20	19	67	861	18	13	59
776	19	19	67	819	18	19	67	862	23	13	59
777	19	19	67	820	20	19	67	863	20	13	59
778	22	19	67	821	18	19	67	864	20	13	59
779	18	18	66	822	18	19	67	865	23	12	58
780	15	18	66	823	23	19	67	866	19	12	58
781	20	18	66	824	18	18	66	867	16	12	58
782	22	18	66	825	21	18	66	868	17	12	58
783	21	17	65	826	18	18	66	869	20	12	58
784	18	17	65	827	22	18	66	870	24	39	93
785	19	17	65	828	20	18	66	871	21	38	92
786	18	17	65	829	22	18	66	872	20	37	90
787	20	17	65	830	20	18	66	873	22	36	89
788	15	17	65	831	25	18	66	874	20	35	88
789	16	17	65	832	18	17	65	875	24	33	85
790	17	16	63	833	21	17	65	876	22	32	84
791	19	16	63	834	22	17	65	877	23	32	84
792	21	16	63	835	23	17	65	878	20	31	83
793	23	16	63	836	23	17	65	879	20	31	83
794	24	16	63	837	22	17	65	880	21	31	83
795	22	16	63	838	18	17	65	881	22	29	80
796	19	12	58	839	20	17	65	882	19	28	79
797	20	12	58	840	19	17	65	883	20	27	77
798	16	11	57	841	20	17	65	884	18	17	65
799	15	11	57	842	23	17	65	885	17	17	65
800	16	10	56	843	20	16	63	886	23	16	63
801	17	9	54	844	18	16	63	887	24	16	63
802	16	7	52	845	23	16	63	888	18	14	61
803	20	27	77	846	22	16	63	889	20	13	59
804	20	27	77	847	20	16	63	890	24	13	59
805	18	26	76	848	21	16	63	891	22	12	58
806	22	25	75	849	20	15	62	892	18	12	58
807	20	23	72	850	19	15	62	893	19	12	58
808	20	22	71	851	21	15	62	894	22	12	58
809	20	22	71	852	23	15	62	895	16	12	58
810	18	22	71	853	20	15	62	896	14	11	57
811	20	21	70	854	22	15	62	897	15	11	57

898	17	11	57	941	20	20	68	984	21	33	85
899	20	11	57	942	18	20	68	985	19	31	83
900	21	10	56	943	19	20	68	986	19	31	83
901	22	10	56	944	17	20	68	987	20	31	83
902	26	10	56	945	18	19	67	988	18	31	83
903	20	10	56	946	19	19	67	989	23	30	81
904	19	10	56	947	19	19	67	990	21	29	80
905	18	9	54	948	20	19	67	991	22	29	80
906	22	9	54	949	16	19	67	992	19	28	79
907	16	8	53	950	18	18	66	993	18	28	79
908	17	8	53	951	23	18	66	994	19	28	79
909	21	8	53	952	20	18	66	995	22	26	76
910	18	7	52	953	19	18	66	996	24	26	76
911	19	7	52	954	18	17	65	997	21	25	75
912	20	7	52	955	18	17	65	998	19	25	75
913	23	6	50	956	16	17	65	999	17	25	75
914	18	6	50	957	20	17	65	1000	18	24	74
915	17	6	50	958	24	14	61	1001	20	23	72
916	20	6	50	959	20	14	61	1002	18	23	72
917	21	6	50	960	17	12	58	1003	15	23	72
918	16	5	49	961	18	11	57	1004	21	21	70
919	17	4	48	962	16	11	57	1005	22	21	70
920	24	29	80	963	18	11	57	1006	23	21	70
921	24	28	79	964	20	11	57	1007	22	21	70
922	22	28	79	965	21	10	56	1008	21	21	70
923	20	27	77	966	19	10	56	1009	22	21	70
924	24	27	77	967	18	9	54	1010	18	21	70
925	21	27	77	968	19	9	54	1011	19	21	70
926	19	25	75	969	22	9	54	1012	24	21	70
927	20	25	75	970	24	8	53	1013	16	21	70
928	18	25	75	971	16	8	53	1014	17	20	68
929	15	25	75	972	15	7	52	1015	23	20	68
930	20	25	75	973	17	7	52	1016	21	19	67
931	21	25	75	974	16	7	52	1017	20	19	67
932	16	24	74	975	18	6	50	1018	21	19	67
933	21	24	74	976	26	38	92	1019	24	19	67
934	18	23	72	977	23	37	90	1020	21	19	67
935	18	23	72	978	24	37	90	1021	20	19	67
936	22	23	72	979	20	37	90	1022	18	19	67
937	24	22	71	980	23	34	86	1023	19	19	67
938	23	21	70	981	24	34	86	1024	20	18	66
939	21	21	70	982	24	33	85	1025	18	18	66
940	26	20	68	983	22	33	85	1026	22	18	66

1027	19	18	66	1070	21	26	76	1113	18	13	59
1028	17	18	66	1071	23	26	76	1114	20	11	57
1029	15	17	65	1072	21	25	75	1115	21	32	84
1030	18	17	65	1073	19	25	75	1116	20	31	83
1031	19	17	65	1074	18	24	74	1117	21	31	83
1032	18	16	63	1075	18	23	72	1118	19	30	81
1033	20	16	63	1076	19	23	72	1119	22	30	81
1034	16	16	63	1077	18	23	72	1120	19	30	81
1035	18	16	63	1078	19	22	71	1121	17	30	81
1036	20	16	63	1079	20	22	71	1122	18	29	80
1037	17	15	62	1080	21	22	71	1123	16	28	79
1038	17	15	62	1081	19	21	70	1124	18	26	76
1039	24	14	61	1082	22	20	68	1125	19	26	76
1040	22	14	61	1083	18	20	68	1126	20	25	75
1041	19	13	59	1084	19	20	68	1127	22	24	74
1042	20	13	59	1085	19	20	68	1128	18	24	74
1043	18	13	59	1086	18	20	68	1129	17	24	74
1044	16	12	58	1087	22	20	68	1130	15	24	74
1045	18	12	58	1088	24	20	68	1131	18	21	70
1046	18	11	57	1089	26	20	68	1132	19	21	70
1047	17	11	57	1090	24	20	68	1133	17	21	70
1048	18	8	53	1091	20	19	67	1134	17	21	70
1049	16	7	52	1092	21	19	67	1135	18	20	68
1050	16	7	52	1093	18	19	67	1136	19	20	68
1051	25	35	88	1094	20	19	67	1137	18	20	68
1052	25	35	88	1095	17	18	66	1138	19	20	68
1053	20	34	86	1096	21	18	66	1139	20	19	67
1054	21	33	85	1097	17	18	66	1140	19	19	67
1055	22	33	85	1098	16	18	66	1141	20	19	67
1056	24	33	85	1099	19	17	65	1142	22	18	66
1057	20	32	84	1100	18	17	65	1143	21	18	66
1058	21	32	84	1101	19	17	65	1144	15	16	63
1059	20	32	84	1102	23	17	65	1145	16	16	63
1060	19	32	84	1103	23	17	65	1146	15	15	62
1061	20	31	83	1104	22	17	65	1147	17	15	62
1062	22	30	81	1105	26	16	63	1148	18	15	62
1063	21	30	81	1106	24	16	63	1149	17	15	62
1064	18	30	81	1107	20	16	63	1150	15	14	61
1065	19	30	81	1108	19	15	62	1151	17	13	59
1066	22	29	80	1109	14	15	62	1152	16	12	58
1067	23	27	77	1110	17	15	62	1153	14	12	58
1068	24	27	77	1111	23	14	61	1154	15	11	57
1069	28	27	77	1112	21	14	61	1155	16	11	57

1156	16	11	57	1199	21	17	65	1242	22	15	62
1157	23	33	85	1200	19	17	65	1243	22	15	62
1158	20	31	83	1201	17	15	62	1244	20	15	62
1159	22	30	81	1202	23	15	62	1245	20	15	62
1160	24	30	81	1203	21	15	62	1246	22	15	62
1161	19	28	79	1204	20	14	61	1247	23	14	61
1162	17	28	79	1205	19	13	59	1248	15	14	61
1163	18	28	79	1206	18	12	58	1249	20	14	61
1164	20	27	77	1207	20	11	57	1250	16	14	61
1165	19	27	77	1208	22	11	57	1251	22	14	61
1166	15	27	77	1209	21	10	56	1252	23	13	59
1167	20	27	77	1210	24	10	56	1253	21	13	59
1168	22	27	77	1211	21	10	56	1254	18	13	59
1169	21	25	75	1212	18	10	56	1255	21	13	59
1170	22	25	75	1213	19	10	56	1256	18	13	59
1171	23	25	75	1214	19	9	54	1257	20	13	59
1172	21	25	75	1215	21	9	54	1258	21	13	59
1173	20	25	75	1216	22	9	54	1259	16	12	58
1174	19	25	75	1217	15	7	52	1260	23	12	58
1175	21	25	75	1218	17	7	52	1261	23	12	58
1176	18	25	75	1219	18	7	52	1262	23	12	58
1177	22	24	74	1220	19	6	50	1263	23	12	58
1178	17	24	74	1221	21	24	74	1264	22	12	58
1179	16	23	72	1222	24	22	71	1265	14	11	57
1180	23	23	72	1223	21	20	68	1266	15	11	57
1181	21	23	72	1224	22	19	67	1267	22	11	57
1182	20	22	71	1225	14	19	67	1268	18	11	57
1183	20	22	71	1226	20	19	67	1269	17	11	57
1184	19	22	71	1227	20	18	66	1270	20	11	57
1185	18	21	70	1228	19	18	66	1271	20	11	57
1186	16	20	68	1229	25	17	65	1272	23	11	57
1187	17	20	68	1230	20	17	65	1273	23	11	57
1188	19	20	68	1231	24	17	65	1274	18	11	57
1189	20	20	68	1232	22	17	65	1275	23	11	57
1190	23	20	68	1233	18	16	63	1276	18	10	56
1191	19	19	67	1234	20	16	63	1277	21	10	56
1192	20	19	67	1235	20	16	63	1278	20	10	56
1193	18	19	67	1236	21	16	63	1279	20	10	56
1194	18	18	66	1237	20	15	62	1280	20	10	56
1195	17	18	66	1238	16	15	62	1281	19	10	56
1196	16	18	66	1239	22	15	62	1282	15	10	56
1197	18	18	66	1240	19	15	62	1283	16	10	56
1198	20	17	65	1241	20	15	62	1284	20	10	56

1285	21	10	56	1328	18	15	62	1371	19	14	61
1286	20	9	54	1329	17	12	58	1372	17	14	61
1287	16	9	54	1330	18	11	57	1373	15	13	59
1288	20	9	54	1331	19	11	57	1374	16	13	59
1289	15	9	54	1332	16	11	57	1375	14	13	59
1290	19	9	54	1333	16	10	56	1376	19	12	58
1291	27	9	54	1334	16	9	54	1377	14	12	58
1292	18	9	54	1335	16	9	54	1378	18	12	58
1293	19	9	54	1336	15	9	54	1379	15	12	58
1294	22	9	54	1337	15	8	53	1380	19	12	58
1295	18	8	53	1338	15	6	50	1381	20	11	57
1296	22	8	53	1339	15	6	50	1382	22	11	57
1297	20	7	52	1340	13	5	49	1383	16	11	57
1298	19	6	50	1341	14	5	49	1384	15	11	57
1299	15	6	50	1342	14	5	49	1385	16	11	57
1300	20	6	50	1343	18	4	48	1386	19	11	57
1301	21	26	76	1344	19	4	48	1387	23	11	57
1302	20	26	76	1345	15	4	48	1388	20	10	56
1303	21	25	75	1346	16	4	48	1389	18	10	56
1304	20	25	75	1347	17	4	48	1390	19	10	56
1305	20	25	75	1348	18	4	48	1391	20	9	54
1306	18	24	74	1349	14	4	48	1392	17	9	54
1307	18	24	74	1350	25	35	88	1393	18	9	54
1308	20	24	74	1351	23	25	75	1394	21	9	54
1309	19	24	74	1352	26	24	74	1395	22	8	53
1310	17	24	74	1353	24	23	72	1396	21	8	53
1311	22	23	72	1354	22	23	72	1397	20	8	53
1312	16	23	72	1355	21	22	71	1398	19	7	52
1313	15	22	71	1356	23	22	71	1399	16	6	50
1314	16	22	71	1357	20	19	67	1400	18	6	50
1315	20	21	70	1358	19	18	66	1401	15	6	50
1316	21	21	70	1359	21	18	66	1402	23	23	72
1317	19	21	70	1360	18	18	66	1403	20	32	84
1318	21	21	70	1361	16	18	66	1404	21	32	84
1319	20	21	70	1362	20	17	65	1405	19	31	83
1320	17	20	68	1363	23	17	65	1406	20	30	81
1321	18	20	68	1364	17	17	65	1407	22	30	81
1322	19	20	68	1365	19	16	63	1408	18	28	79
1323	16	20	68	1366	21	15	62	1409	20	28	79
1324	17	20	68	1367	24	15	62	1410	19	28	79
1325	20	19	67	1368	19	15	62	1411	20	25	75
1326	19	17	65	1369	23	14	61	1412	21	24	74
1327	19	15	62	1370	20	14	61	1413	20	23	72

1414	18	23	72	1457	22	18	66	1500	23	15	62
1415	19	21	70	1458	18	18	66	1501	18	14	61
1416	21	21	70	1459	19	18	66	1502	21	14	61
1417	19	20	68	1460	19	18	66	1503	20	13	59
1418	21	20	68	1461	20	18	66	1504	18	13	59
1419	20	19	67	1462	17	17	65	1505	19	13	59
1420	23	19	67	1463	20	17	65	1506	20	12	58
1421	18	19	67	1464	22	12	58	1507	18	11	57
1422	19	19	67	1465	24	10	56	1508	17	10	56
1423	20	18	66	1466	23	10	56	1509	16	10	56
1424	18	18	66	1467	20	10	56	1510	19	9	54
1425	22	17	65	1468	19	8	53	1511	17	8	53
1426	20	15	62	1469	20	8	53	1512	20	7	52
1427	19	15	62	1470	21	8	53	1513	19	7	52
1428	19	14	61	1471	19	8	53	1514	17	6	50
1429	16	14	61	1472	20	7	52				
1430	18	12	58	1473	22	6	50				
1431	19	12	58	1474	19	6	50				
1432	17	12	58	1475	27	6	50				
1433	18	11	57	1476	22	6	50				
1434	20	11	57	1477	18	6	50				
1435	21	10	56	1478	24	21	70				
1436	18	9	54	1479	23	20	68				
1437	17	9	54	1480	21	20	68				
1438	15	8	53	1481	19	19	67				
1439	18	8	53	1482	20	19	67				
1440	16	7	52	1483	23	19	67				
1441	19	7	52	1484	21	19	67				
1442	18	7	52	1485	18	19	67				
1443	16	22	71	1486	19	19	67				
1444	22	22	71	1487	22	18	66				
1445	23	21	70	1488	16	18	66				
1446	23	21	70	1489	17	18	66				
1447	19	21	70	1490	20	18	66				
1448	23	20	68	1491	24	18	66				
1449	18	20	68	1492	23	17	65				
1450	24	20	68	1493	22	17	65				
1451	21	20	68	1494	23	17	65				
1452	19	20	68	1495	20	17	65				
1453	18	20	68	1496	19	16	63				
1454	22	20	68	1497	17	16	63				
1455	21	19	67	1498	20	16	63				
1456	24	19	67	1499	22	16	63				

STATE B

S/N	CAS B	TS B	LE B										
1	46	24	76	39	44	10	58	82	36	7	54		
2	41	19	69	40	43	10	58	83	37	7	54		
3	42	15	64	41	43	10	58	84	38	7	54		
4	44	15	64	42	29	10	58	85	42	7	54		
5	35	14	63	43	27	10	58	86	37	7	54		
6	31	14	63	44	28	10	58	87	39	7	54		
7	44	14	63	45	29	10	58	88	36	7	54		
8	41	14	63	46	43	10	58	89	36	7	54		
9	35	14	63	47	37	10	58	90	22	7	54		
10	38	13	62	48	45	10	58	91	28	6	53		
11	41	13	62	49	36	9	57	92	38	6	53		
12	38	13	62	50	40	9	57	93	40	6	53		
13	46	13	62	51	41	9	57	94	43	6	53		
14	31	13	62	52	42	9	57	95	31	6	53		
15	41	12	60	53	42	9	57	96	27	6	53		
16	30	12	60	54	35	9	57	97	26	6	53		
17	39	12	60	55	36	9	57	98	33	6	53		
18	38	12	60	56	41	9	57	99	38	6	53		
19	38	12	60	57	46	9	57	100	41	6	53		
20	29	12	60	58	37	9	57	101	32	5	51		
21	41	12	60	59	36	9	57	102	39	5	51		
22	47	12	60	60	32	9	57	103	32	5	51		
23	41	12	60	61	43	9	57	104	27	5	51		
24	43	11	59	62	42	8	55	105	28	5	51		
25	32	11	59	63	36	8	55	106	21	5	51		
26	33	11	59	64	38	8	55	107	29	5	51		
27	36	11	59	65	42	8	55	108	22	5	51		
28	37	11	59	66	30	8	55	109	20	5	51		
29	40	11	59	67	38	8	55	110	31	5	51		
30	45	11	59	68	46	8	55	111	40	5	51		
31	33	11	59	69	42	8	55	112	34	5	51		
32	44	11	59	70	41	8	55	113	36	5	51		
33	48	11	59	71	38	8	55	114	37	4	50		
34	42	11	59	72	32	8	55	115	38	4	50		
35	37	10	58	73	31	8	55	116	39	4	50		
36	39	10	58	74	38	8	55	117	38	4	50		
37	32	10	58	75	40	7	54	118	37	4	50		
38	41	10	58	76	35	7	54	119	36	4	50		
				77	41	7	54	120	40	4	50		
				78	37	7	54	121	41	4	50		
				79	34	7	54	122	42	4	50		
				80	22	7	54	123	44	3	49		
				81	29	7	54	124	41	3	49		

125	36	3	49	168	32	11	59	211	30	7	54
126	32	3	49	169	41	11	59	212	31	7	54
127	30	3	49	170	38	11	59	213	32	7	54
128	34	3	49	171	40	11	59	214	33	7	54
129	38	2	48	172	41	11	59	215	34	7	54
130	31	2	48	173	31	11	59	216	39	7	54
131	48	30	83	174	29	11	59	217	40	7	54
132	48	25	77	175	33	11	59	218	40	7	54
133	42	24	76	176	21	10	58	219	21	7	54
134	40	24	76	177	31	10	58	220	32	6	53
135	41	24	76	178	40	10	58	221	38	6	53
136	43	24	76	179	42	10	58	222	38	6	53
137	39	22	73	180	38	10	58	223	40	6	53
138	33	21	72	181	29	10	58	224	41	6	53
139	40	21	72	182	41	10	58	225	29	6	53
140	42	21	72	183	30	10	58	226	31	6	53
141	40	21	72	184	40	10	58	227	34	6	53
142	41	21	72	185	41	10	58	228	35	6	53
143	43	20	71	186	41	9	57	229	24	5	51
144	41	20	71	187	44	9	57	230	28	5	51
145	42	19	69	188	28	9	57	231	26	5	51
146	39	18	68	189	21	9	57	232	27	5	51
147	38	18	68	190	22	9	57	233	28	5	51
148	36	18	68	191	21	9	57	234	31	5	51
149	40	18	68	192	29	9	57	235	30	5	51
150	41	18	68	193	31	9	57	236	32	4	50
151	39	17	67	194	34	9	57	237	37	4	50
152	41	16	65	195	29	9	57	238	29	4	50
153	32	16	65	196	31	9	57	239	40	4	50
154	40	16	65	197	28	9	57	240	21	4	50
155	38	16	65	198	31	8	55	241	22	4	50
156	39	15	64	199	45	8	55	242	41	4	50
157	35	15	64	200	41	8	55	243	21	4	50
158	40	15	64	201	42	8	55	244	32	3	49
159	40	15	64	202	31	8	55	245	37	3	49
160	41	14	63	203	30	8	55	246	29	3	49
161	42	13	62	204	31	8	55	247	24	3	49
162	44	13	62	205	38	8	55	248	28	3	49
163	32	12	60	206	31	8	55	249	30	3	49
164	32	12	60	207	22	8	55	250	29	2	48
165	31	12	60	208	20	8	55	251	42	20	71
166	40	11	59	209	40	8	55	252	42	18	68
167	38	11	59	210	21	8	55	253	36	18	68

254	32	16	65	297	21	10	58	340	32	7	54
255	41	16	65	298	24	10	58	341	42	7	54
256	32	15	64	299	33	10	58	342	36	7	54
257	30	15	64	300	37	10	58	343	34	7	54
258	42	14	63	301	32	10	58	344	35	7	54
259	41	14	63	302	29	10	58	345	35	7	54
260	35	14	63	303	28	10	58	346	40	6	53
261	36	14	63	304	33	10	58	347	41	6	53
262	33	14	63	305	41	10	58	348	36	6	53
263	34	14	63	306	42	10	58	349	38	6	53
264	36	13	62	307	41	9	57	350	38	6	53
265	35	13	62	308	40	9	57	351	39	6	53
266	28	13	62	309	40	9	57	352	35	6	53
267	39	13	62	310	45	9	57	353	37	6	53
268	32	13	62	311	38	9	57	354	32	6	53
269	33	13	62	312	32	9	57	355	38	5	51
270	33	13	62	313	40	9	57	356	39	5	51
271	34	13	62	314	28	9	57	357	41	5	51
272	35	13	62	315	28	9	57	358	32	5	51
273	36	13	62	316	30	8	55	359	39	5	51
274	37	13	62	317	31	8	55	360	29	5	51
275	38	13	62	318	34	8	55	361	22	5	51
276	39	13	62	319	37	8	55	362	27	4	50
277	40	12	60	320	36	8	55	363	30	4	50
278	41	12	60	321	36	8	55	364	27	4	50
279	42	12	60	322	36	8	55	365	34	4	50
280	44	12	60	323	37	8	55	366	36	4	50
281	43	12	60	324	33	8	55	367	28	3	49
282	44	12	60	325	34	8	55	368	29	3	49
283	42	12	60	326	32	8	55	369	31	2	48
284	30	12	60	327	28	8	55	370	43	34	88
285	34	11	59	328	27	8	55	371	44	25	77
286	39	11	59	329	27	8	55	372	41	25	77
287	33	11	59	330	30	8	55	373	39	25	77
288	31	11	59	331	36	8	55	374	34	24	76
289	32	11	59	332	41	8	55	375	41	23	74
290	35	11	59	333	32	8	55	376	48	23	74
291	29	11	59	334	36	8	55	377	45	23	74
292	29	11	59	335	31	7	54	378	39	22	73
293	28	11	59	336	33	7	54	379	30	21	72
294	30	10	58	337	39	7	54	380	35	20	71
295	32	10	58	338	43	7	54	381	50	20	71
296	30	10	58	339	35	7	54	382	50	20	71

383	32	19	69	426	34	10	58	469	42	4	50
384	33	19	69	427	31	10	58	470	39	4	50
385	34	19	69	428	28	10	58	471	38	3	49
386	41	17	67	429	29	10	58	472	36	3	49
387	40	17	67	430	40	10	58	473	37	3	49
388	40	17	67	431	38	10	58	474	32	3	49
389	42	16	65	432	32	10	58	475	29	3	49
390	43	15	64	433	38	10	58	476	30	2	48
391	31	15	64	434	43	10	58	477	28	2	48
392	38	15	64	435	27	10	58	478	34	2	48
393	39	14	63	436	34	9	57	479	39	29	82
394	36	14	63	437	37	9	57	480	41	26	78
395	37	14	63	438	38	9	57	481	42	24	76
396	32	14	63	439	36	9	57	482	35	22	73
397	41	14	63	440	40	9	57	483	34	20	71
398	46	14	63	441	49	8	55	484	37	20	71
399	44	14	63	442	41	8	55	485	38	20	71
400	48	14	63	443	33	8	55	486	38	19	69
401	43	14	63	444	35	8	55	487	40	19	69
402	47	13	62	445	38	8	55	488	39	19	69
403	43	13	62	446	38	8	55	489	41	18	68
404	35	12	60	447	41	7	54	490	37	18	68
405	32	12	60	448	33	7	54	491	38	17	67
406	39	12	60	449	38	7	54	492	37	16	65
407	40	12	60	450	43	7	54	493	36	16	65
408	38	12	60	451	44	7	54	494	40	14	63
409	30	12	60	452	29	7	54	495	36	14	63
410	34	12	60	453	37	6	53	496	41	14	63
411	37	12	60	454	34	6	53	497	42	13	62
412	40	12	60	455	33	6	53	498	39	13	62
413	39	12	60	456	33	6	53	499	38	12	60
414	42	12	60	457	42	6	53	500	35	12	60
415	43	12	60	458	46	6	53	501	37	11	59
416	47	11	59	459	44	6	53	502	42	11	59
417	36	11	59	460	41	6	53	503	41	11	59
418	38	11	59	461	32	6	53	504	44	10	58
419	42	11	59	462	38	6	53	505	39	10	58
420	43	11	59	463	39	5	51	506	38	9	57
421	44	11	59	464	37	5	51	507	43	9	57
422	41	11	59	465	30	5	51	508	41	8	55
423	38	11	59	466	44	5	51	509	38	8	55
424	35	11	59	467	31	5	51	510	41	7	54
425	36	11	59	468	33	5	51	511	35	7	54

512	32	6	53	555	38	16	65	598	32	19	69
513	30	6	53	556	35	15	64	599	39	19	69
514	38	6	53	557	36	15	64	600	40	18	68
515	34	5	51	558	30	12	60	601	41	16	65
516	37	5	51	559	32	12	60	602	43	16	65
517	38	5	51	560	41	12	60	603	41	14	63
518	42	34	88	561	42	10	58	604	40	14	63
519	45	32	86	562	44	10	58	605	30	11	59
520	41	32	86	563	38	10	58	606	41	10	58
521	39	31	85	564	39	10	58	607	39	10	58
522	40	31	85	565	38	10	58	608	38	10	58
523	38	30	83	566	38	10	58	609	32	9	57
524	41	29	82	567	38	10	58	610	41	9	57
525	37	29	82	568	42	9	57	611	40	9	57
526	42	27	80	569	30	9	57	612	41	7	54
527	40	27	80	570	43	9	57	613	30	7	54
528	39	25	77	571	40	8	55	614	32	6	53
529	38	25	77	572	29	8	55	615	38	6	53
530	37	23	74	573	31	8	55	616	37	5	51
531	32	23	74	574	38	7	54	617	40	5	51
532	30	22	73	575	34	7	54	618	40	34	88
533	41	22	73	576	33	7	54	619	42	28	81
534	44	21	72	577	38	6	53	620	43	27	80
535	42	21	72	578	36	6	53	621	39	26	78
536	38	21	72	579	34	6	53	622	38	26	78
537	37	21	72	580	33	6	53	623	40	25	77
538	38	20	71	581	34	6	53	624	42	24	76
539	41	20	71	582	42	28	81	625	41	23	74
540	40	20	71	583	40	27	80	626	38	23	74
541	42	20	71	584	41	27	80	627	39	23	74
542	38	20	71	585	43	25	77	628	35	23	74
543	35	19	69	586	42	25	77	629	38	21	72
544	37	19	69	587	38	25	77	630	39	20	71
545	38	19	69	588	39	25	77	631	30	19	69
546	38	19	69	589	38	24	76	632	38	19	69
547	35	19	69	590	40	23	74	633	35	18	68
548	32	19	69	591	41	23	74	634	43	17	67
549	40	19	69	592	42	21	72	635	41	17	67
550	43	19	69	593	43	21	72	636	45	17	67
551	42	18	68	594	44	20	71	637	36	16	65
552	41	18	68	595	41	20	71	638	33	16	65
553	34	17	67	596	39	20	71	639	35	16	65
554	39	17	67	597	38	19	69	640	38	16	65

641	37	16	65	684	37	5	51	727	44	13	62
642	38	16	65	685	38	5	51	728	43	13	62
643	39	15	64	686	37	5	51	729	44	13	62
644	41	15	64	687	38	5	51	730	47	13	62
645	43	14	63	688	35	4	50	731	42	13	62
646	36	14	63	689	38	13	62	732	41	12	60
647	37	14	63	690	39	13	62	733	31	12	60
648	36	14	63	691	37	12	60	734	38	12	60
649	35	14	63	692	38	12	60	735	39	12	60
650	38	14	63	693	37	11	59	736	37	12	60
651	37	13	62	694	35	10	58	737	38	12	60
652	32	13	62	695	32	10	58	738	37	12	60
653	40	13	62	696	38	9	57	739	38	11	59
654	41	13	62	697	36	9	57	740	39	11	59
655	31	13	62	698	34	8	55	741	34	11	59
656	41	13	62	699	34	8	55	742	33	10	58
657	42	13	62	700	38	8	55	743	32	10	58
658	33	13	62	701	38	7	54	744	38	10	58
659	40	13	62	702	36	7	54	745	39	10	58
660	39	13	62	703	41	6	53	746	32	10	58
661	41	12	60	704	32	6	53	747	34	10	58
662	36	12	60	705	30	6	53	748	40	10	58
663	39	12	60	706	38	6	53	749	39	9	57
664	33	12	60	707	37	6	53	750	38	9	57
665	38	12	60	708	34	5	51	751	35	9	57
666	32	12	60	709	35	5	51	752	30	9	57
667	38	12	60	710	39	5	51	753	39	9	57
668	41	12	60	711	48	24	76	754	40	9	57
669	39	12	60	712	43	22	73	755	43	8	55
670	42	11	59	713	44	19	69	756	31	8	55
671	43	11	59	714	42	18	68	757	32	8	55
672	40	10	58	715	41	18	68	758	43	8	55
673	39	10	58	716	35	17	67	759	44	7	54
674	37	10	58	717	39	16	65	760	37	6	53
675	35	10	58	718	38	14	63	761	38	6	53
676	40	9	57	719	34	14	63	762	39	6	53
677	42	8	55	720	37	13	62	763	36	6	53
678	38	8	55	721	32	13	62	764	38	6	53
679	39	8	55	722	37	13	62	765	32	5	51
680	37	8	55	723	35	13	62	766	38	5	51
681	36	8	55	724	41	13	62	767	38	5	51
682	38	6	53	725	42	13	62	768	32	5	51
683	39	6	53	726	43	13	62	769	49	32	86

770	45	31	85	813	29	7	54	856	31	11	59
771	44	29	82	814	27	6	53	857	33	11	59
772	46	29	82	815	28	33	87	858	36	11	59
773	39	28	81	816	44	19	69	859	32	11	59
774	42	24	76	817	38	19	69	860	31	11	59
775	38	24	76	818	35	16	65	861	31	11	59
776	41	21	72	819	36	16	65	862	28	11	59
777	40	20	71	820	26	16	65	863	35	11	59
778	41	20	71	821	34	15	64	864	35	11	59
779	42	20	71	822	34	14	63	865	32	11	59
780	36	19	69	823	36	14	63	866	32	10	58
781	38	19	69	824	42	14	63	867	40	10	58
782	39	19	69	825	36	13	62	868	31	10	58
783	36	18	68	826	38	13	62	869	32	10	58
784	32	16	65	827	36	13	62	870	38	10	58
785	38	16	65	828	35	13	62	871	38	10	58
786	34	14	63	829	43	13	62	872	37	10	58
787	37	14	63	830	41	13	62	873	40	10	58
788	38	13	62	831	37	13	62	874	36	10	58
789	39	13	62	832	37	13	62	875	38	10	58
790	37	13	62	833	37	13	62	876	40	10	58
791	40	13	62	834	35	13	62	877	36	10	58
792	35	13	62	835	32	13	62	878	40	10	58
793	42	13	62	836	34	13	62	879	40	10	58
794	39	12	60	837	32	12	60	880	32	10	58
795	37	12	60	838	35	12	60	881	31	10	58
796	36	12	60	839	40	12	60	882	30	10	58
797	40	12	60	840	35	12	60	883	32	10	58
798	41	12	60	841	47	12	60	884	38	9	57
799	37	12	60	842	40	12	60	885	39	9	57
800	35	11	59	843	22	12	60	886	38	9	57
801	38	11	59	844	36	12	60	887	38	9	57
802	29	11	59	845	40	12	60	888	36	9	57
803	31	11	59	846	37	12	60	889	32	9	57
804	36	10	58	847	38	12	60	890	40	9	57
805	37	10	58	848	33	12	60	891	31	9	57
806	34	10	58	849	40	12	60	892	37	9	57
807	33	9	57	850	32	11	59	893	45	9	57
808	29	9	57	851	34	11	59	894	34	9	57
809	39	8	55	852	41	11	59	895	37	9	57
810	38	8	55	853	36	11	59	896	37	9	57
811	41	8	55	854	36	11	59	897	38	9	57
812	37	7	54	855	37	11	59	898	42	9	57

899	41	8	55	942	48	23	74	985	40	20	71
900	38	8	55	943	38	22	73	986	37	20	71
901	38	8	55	944	39	21	72	987	35	19	69
902	36	8	55	945	34	20	71	988	36	19	69
903	37	8	55	946	38	20	71	989	35	19	69
904	31	8	55	947	40	19	69	990	34	18	68
905	36	8	55	948	31	19	69	991	34	18	68
906	40	8	55	949	42	19	69	992	39	18	68
907	36	8	55	950	35	19	69	993	38	18	68
908	30	7	54	951	36	19	69	994	40	17	67
909	45	7	54	952	37	18	68	995	32	17	67
910	33	7	54	953	39	18	68	996	40	17	67
911	34	7	54	954	38	18	68	997	42	17	67
912	42	7	54	955	40	18	68	998	41	17	67
913	39	7	54	956	42	18	68	999	44	17	67
914	32	7	54	957	35	17	67	1000	39	17	67
915	31	7	54	958	32	17	67	1001	38	16	65
916	40	7	54	959	37	17	67	1002	32	16	65
917	29	7	54	960	38	17	67	1003	39	16	65
918	46	7	54	961	39	17	67	1004	33	14	63
919	41	7	54	962	42	16	65	1005	41	14	63
920	31	7	54	963	41	16	65	1006	37	13	62
921	36	7	54	964	42	14	63	1007	38	12	60
922	36	7	54	965	38	14	63	1008	41	12	60
923	40	6	53	966	39	14	63	1009	40	11	59
924	36	6	53	967	42	14	63	1010	37	10	58
925	41	6	53	968	44	14	63	1011	36	10	58
926	37	6	53	969	33	13	62	1012	35	9	57
927	37	5	51	970	41	13	62	1013	41	9	57
928	36	5	51	971	32	12	60	1014	35	8	55
929	37	5	51	972	31	11	59	1015	46	31	85
930	34	5	51	973	40	11	59	1016	41	29	82
931	36	5	51	974	41	11	59	1017	39	28	81
932	38	5	51	975	38	10	58	1018	41	24	76
933	32	5	51	976	32	10	58	1019	38	24	76
934	38	4	50	977	36	9	57	1020	42	23	74
935	39	3	49	978	37	8	55	1021	40	23	74
936	43	29	82	979	44	27	80	1022	43	22	73
937	39	27	80	980	43	25	77	1023	35	22	73
938	41	27	80	981	43	23	74	1024	41	21	72
939	37	26	78	982	40	23	74	1025	40	20	71
940	37	25	77	983	36	22	73	1026	42	20	71
941	47	23	74	984	38	21	72	1027	36	20	71

1028	37	20	71	1071	36	8	55	1114	38	17	67
1029	43	19	69	1072	39	8	55	1115	34	16	65
1030	41	19	69	1073	40	7	54	1116	32	15	64
1031	38	19	69	1074	34	6	53	1117	34	14	63
1032	41	19	69	1075	41	28	81	1118	34	14	63
1033	42	19	69	1076	40	27	80	1119	40	14	63
1034	44	18	68	1077	42	27	80	1120	36	11	59
1035	41	18	68	1078	48	26	78	1121	37	11	59
1036	39	18	68	1079	41	26	78	1122	42	11	59
1037	36	18	68	1080	39	26	78	1123	37	11	59
1038	37	18	68	1081	44	26	78	1124	34	10	58
1039	38	18	68	1082	40	25	77	1125	32	10	58
1040	39	18	68	1083	38	25	77	1126	34	9	57
1041	42	17	67	1084	39	24	76	1127	37	9	57
1042	29	17	67	1085	40	23	74	1128	38	9	57
1043	36	17	67	1086	32	23	74	1129	40	9	57
1044	27	17	67	1087	39	23	74	1130	29	8	55
1045	30	16	65	1088	37	23	74	1131	30	8	55
1046	33	16	65	1089	36	21	72	1132	31	8	55
1047	35	16	65	1090	43	20	71	1133	34	8	55
1048	40	15	64	1091	41	20	71	1134	33	7	54
1049	34	15	64	1092	38	20	71	1135	37	7	54
1050	41	14	63	1093	42	20	71	1136	38	7	54
1051	36	14	63	1094	39	20	71	1137	33	7	54
1052	33	13	62	1095	37	19	69	1138	40	6	53
1053	32	13	62	1096	38	19	69	1139	32	6	53
1054	39	13	62	1097	38	19	69	1140	44	28	81
1055	34	13	62	1098	36	19	69	1141	41	26	78
1056	38	13	62	1099	30	19	69	1142	38	26	78
1057	38	13	62	1100	41	19	69	1143	41	21	72
1058	40	13	62	1101	39	19	69	1144	39	21	72
1059	36	12	60	1102	38	19	69	1145	37	21	72
1060	32	12	60	1103	35	19	69	1146	38	20	71
1061	37	12	60	1104	40	19	69	1147	40	20	71
1062	38	11	59	1105	33	18	68	1148	36	20	71
1063	38	11	59	1106	34	18	68	1149	37	20	71
1064	40	11	59	1107	35	18	68	1150	35	20	71
1065	41	10	58	1108	41	18	68	1151	35	19	69
1066	39	10	58	1109	39	17	67	1152	39	19	69
1067	42	10	58	1110	40	17	67	1153	38	19	69
1068	39	9	57	1111	42	17	67	1154	37	19	69
1069	43	9	57	1112	36	17	67	1155	37	19	69
1070	38	9	57	1113	34	17	67	1156	39	19	69

1157	42	18	68	1200	32	9	57	1243	32	5	51
1158	35	18	68	1201	35	8	55	1244	30	5	51
1159	37	18	68	1202	36	8	55	1245	39	4	50
1160	38	18	68	1203	40	7	54	1246	38	3	49
1161	40	18	68	1204	35	7	54	1247	40	3	49
1162	42	18	68	1205	32	7	54	1248	48	24	76
1163	36	18	68	1206	34	7	54	1249	46	23	74
1164	34	18	68	1207	48	30	83	1250	46	23	74
1165	41	17	67	1208	43	29	82	1251	46	21	72
1166	36	17	67	1209	44	16	65	1252	42	21	72
1167	32	17	67	1210	40	14	63	1253	39	20	71
1168	33	15	64	1211	41	14	63	1254	40	20	71
1169	36	14	63	1212	39	13	62	1255	42	20	71
1170	32	12	60	1213	33	13	62	1256	39	20	71
1171	37	12	60	1214	42	13	62	1257	38	19	69
1172	38	12	60	1215	40	13	62	1258	35	19	69
1173	38	11	59	1216	43	12	60	1259	30	19	69
1174	40	10	58	1217	43	12	60	1260	37	19	69
1175	39	10	58	1218	36	12	60	1261	38	17	67
1176	42	9	57	1219	39	11	59	1262	36	17	67
1177	35	9	57	1220	37	11	59	1263	37	17	67
1178	38	7	54	1221	32	11	59	1264	38	17	67
1179	34	7	54	1222	42	11	59	1265	39	16	65
1180	32	6	53	1223	37	10	58	1266	38	16	65
1181	32	5	51	1224	34	10	58	1267	30	14	63
1182	25	34	88	1225	39	10	58	1268	32	14	63
1183	30	32	86	1226	42	9	57	1269	38	14	63
1184	28	32	86	1227	36	9	57	1270	29	14	63
1185	31	27	80	1228	38	9	57	1271	32	13	62
1186	40	26	78	1229	34	8	55	1272	37	13	62
1187	37	14	63	1230	37	8	55	1273	40	13	62
1188	36	14	63	1231	36	8	55	1274	42	13	62
1189	37	13	62	1232	40	8	55	1275	39	13	62
1190	30	13	62	1233	34	8	55	1276	43	12	60
1191	37	13	62	1234	37	8	55	1277	33	12	60
1192	36	13	62	1235	34	7	54	1278	39	12	60
1193	30	12	60	1236	33	7	54	1279	38	12	60
1194	38	12	60	1237	37	7	54	1280	40	11	59
1195	32	12	60	1238	42	7	54	1281	41	9	57
1196	28	12	60	1239	40	7	54	1282	38	9	57
1197	29	12	60	1240	33	7	54	1283	40	8	55
1198	39	11	59	1241	34	6	53	1284	44	33	87
1199	28	10	58	1242	29	6	53	1285	43	32	86

1286	44	30	83	1329	41	14	63	1372	36	22	73
1287	48	30	83	1330	37	14	63	1373	32	21	72
1288	43	28	81	1331	34	14	63	1374	35	21	72
1289	29	28	81	1332	34	14	63	1375	37	21	72
1290	38	27	80	1333	40	14	63	1376	36	19	69
1291	34	24	76	1334	41	13	62	1377	38	19	69
1292	42	24	76	1335	29	13	62	1378	34	17	67
1293	44	24	76	1336	31	13	62	1379	37	16	65
1294	47	23	74	1337	33	13	62	1380	40	16	65
1295	48	23	74	1338	35	13	62	1381	39	15	64
1296	39	21	72	1339	40	13	62	1382	42	14	63
1297	36	20	71	1340	37	13	62	1383	36	13	62
1298	37	20	71	1341	39	13	62	1384	35	12	60
1299	36	20	71	1342	42	13	62	1385	35	11	59
1300	34	20	71	1343	43	12	60	1386	40	11	59
1301	35	20	71	1344	36	12	60	1387	38	11	59
1302	33	20	71	1345	30	12	60	1388	36	10	58
1303	34	20	71	1346	32	12	60	1389	37	10	58
1304	38	19	69	1347	30	12	60	1390	33	10	58
1305	35	19	69	1348	41	12	60	1391	34	9	57
1306	30	19	69	1349	36	12	60				
1307	29	19	69	1350	31	11	59				
1308	40	19	69	1351	36	10	58				
1309	35	19	69	1352	37	10	58				
1310	35	19	69	1353	30	10	58				
1311	33	18	68	1354	32	10	58				
1312	40	18	68	1355	34	10	58				
1313	41	18	68	1356	37	9	57				
1314	38	18	68	1357	36	9	57				
1315	39	18	68	1358	43	31	85				
1316	41	18	68	1359	45	30	83				
1317	35	18	68	1360	44	30	83				
1318	43	18	68	1361	40	28	81				
1319	41	18	68	1362	39	27	80				
1320	36	17	67	1363	39	27	80				
1321	42	17	67	1364	40	27	80				
1322	40	17	67	1365	41	27	80				
1323	38	17	67	1366	28	25	77				
1324	35	17	67	1367	39	24	76				
1325	38	17	67	1368	36	24	76				
1326	38	15	64	1369	38	23	74				
1327	40	15	64	1370	37	22	73				
1328	44	14	63	1371	34	22	73				

APPENDIX Z

T-Test**Group Statistics**

STATE		N	Mean	Std. Deviation	Std. Error Mean
LINEAR EQUATING	STATE A	1514	66.0709	8.52874	.21919
	STATE B	1391	61.6494	8.12876	.21795

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2- tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	1.683	.195	14.275	2903	.000	4.42157E0	.30974	3.81424E0	5.02889
Equal variances not assumed			14.304	2.899E3	.000	4.42157E0	.30911	3.81547E0	5.02766