EXPLORING 4-PARAMETER LOGISTIC AND LOGNORMAL RESPONSE-TIME MODELS IN CALIBRATING COMPUTER-BASED MATHEMATICS TEST AMONG SENIOR SECONDARY SCHOOL STUDENTS IN OYO AND LAGOS STATES, NIGERIA

RASIDAT OMONIKE, LAWAL
MATRIC NO: 125511

JULY, 2021

EXPLORING 4-PARAMETER LOGISTIC AND LOGNORMAL RESPONSE-TIME MODELS IN CALIBRATING COMPUTER-BASED MATHEMATICS TEST AMONG SENIOR SECONDARY SCHOOL STUDENTS IN OYO AND LAGOS STATES, NIGERIA

BY

RASIDAT OMONIKE, LAWAL
MATRIC NO: 125511
HND. (Ibadan Poly), PGDE., B.Sc. (Ibadan), M.Ed.(Ago-Iwoye)

A THESIS SUBMITTED TO THE
INTERNATIONAL CENTRE FOR EDUCATIONAL EVALUATION (ICEE)
INSTITUTE OF EDUCATION
UNIVERSITY OF IBADAN

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD
OF
DOCTOR OF PHILOSOPHY
UNIVERSITY OF IBADAN, IBADAN, NIGERIA

JULY, 2021

CERTIFICATION

I certify that this research work was carried out by **Rasidat Omonike LAWAL** under my supervision in the International Centre for Educational Evaluation (ICEE), Institute of Education, University of Ibadan, Nigeria.

……………………………………………………………………

Supervisor
Prof. J. G. Adewale
B. Ed, M. Ed, Ph.D (Ibadan)
Professor of Science and Educational Evaluation
Institute of Education
University of Ibadan

DEDICATION

This research work is dedicated to the lover of my soul, my redeemer, my enabler and the only one in whom my horn is exalted, who in His infinite mercy and abounding grace has painstakingly guided and guarded me through every stage of this academic rigour. To Him alone be all the glory!

ACKNOWLEDGMENTS

Eccl 3[1] "To everything there is a season, a time for every purpose under heaven". I am deeply grateful to God, my Eternal Father, who has allowed me to attain this milestone in my academic pursuit.

My profound gratitude goes to the anchor-man, a supervisor per excellence, my mentor, Professor J. G. Adewale who has never taken impossibility as part of an English word as far as I am concerned. Although he might seem tough at times, he always ensures that the right thing is done timely and appropriately, to achieve a brilliant performance. I remain ever grateful for his not giving up on me even when I deliberately stopped showing up because I thought I would not be able cope with the work. May his seed find a softer ground to tread wherever they go by the mercies of God.

I am immensely grateful to Dr M. A. Metibemu on his valuable contributions to make the research area (Item Response Theory Approach) meaningful and realistic. I also appreciate all my other lecturers: Prof. F.V. Falaye, Prof. Y.W. Yoloye, Prof. P.N. Okpala, Prof. C. O Onocha, Prof. E. Adenike Emeke, Prof. Ifeoma. M. Isiugo-Abanihe, Prof. A.O.U. Onuka, Prof. J.A. Adegbile, Prof. Eugenia A. Okwilagwe and Prof. Monica N. Odinko.They also include Dr. O.E. Ibode, Dr. Modupe M. Oshokoya, Dr. B.A. Adegoke, Dr. J.O Adeleke, Dr. Serifat E. Akorede, Dr. Ikmat O. Junaid, Dr. J.A. Abijo, Dr. E.O. Babatunde, Dr. Olutayo T. Omole, Mr B.K. Babatunde and Mr O.M. Desmennu. All the non-teaching staffsare well appreciated as well, for making the administrative part of the study a smooth sailing trip!

To other contributors, Dr A.O Ariyo at Joint Admission and Matriculation Board Headquarters, he encouraged me more on the research title. He also pioneered my understanding of basic things the research work entailed by making available various textbooks, both soft and hardcopy all the way from Abuja, Nigeria. Mrs Alamu at Oyo State Planning, Research and Statistics Unit, Ministry of Education and Mr O.G Adesina at Lagos Education District I headquarters Alausa Ikeja, for their assistance in assessing schools with functional computers at both Oyo and Lagos states public schools. The support that the individual schools' Principals and their deputies provided as well as the roles Mathematics and Computer teachers played in the different schools are worth appreciating. The effort of a great sister and friend, Mrs Ibironke

Adeleyecannot but be cherishedfor the prominent role she played at locating schools and their principals for easy access. Great is thy reward ma!

My special thanks also go to my colleagues: Dr Nathaniel, Mr Pilla, Mr Inimfon, Mr Godwin, Mr Azeez, Mrs Nkembuchukwu and Mrs Joke (of blessed memory). They were always giving moral and physical supports as at when due. We shall all get to our expected end shinning brighter by God's grace and eat the fruit of our labour thereof in Jesus name. My computer programmer (Mr Chinazom) who worked round the clock to ensure that the e-examination mode (computer-based testing) was successful in spite of the many challenges we faced at the onset.

Also worth mentioning is the support of my able and ever willing research assistants who went out of their way, to ensure that the great task of data collection was adequately and timely accomplished. The likes of Mr Oluwole, Mr Selim, Mrs Kolade, Mr Kolade, Mr Jesutofunmi (the precious gift from my loins), and Miss Folakemi. Their immense contributions are highly appreciated.

And to the brain behind the scene, my encourager and sponsor in all ramifications, who understands that Omonike is a "Bookworm" ever since we started the journey of marriage together (Arch. Omotosho Lawal). He has always been providing the enabling environment to soar high. I am forever grateful for having him as my darling husband and teacher.

My amiable parent of blessed memory is acknowledged. Mr and Mrs Abeeb Alani who set my feet upon the academic ladder. Continue to rest in peace dear parent!

I cannot but appreciate myself as well, for staying put and enduring to the end. At a time, it was like I could no longer continue, but God's grace saw me through. I am once again grateful unto my Maker, the gracious Father.

ABSTRACT

Analysis of students' response and response-time data availed test-developers, pyschometricians and researchers the opportunity toimpartially measureexaminees'learning outcomes with the advent of Item Response Theory (IRT). Records showed that the usage of 1-, 2- and 3-Parameter Logistic (PL) models inthe calibration and estimationof examinees' ability and item parameters had actually enhanced students' accurate estimates of their academic performances. However, the newly invented4-PL and Lognormal Response-time (LNIRT) models with their inherent advantages have not been sufficientlyexplored, whereas, the capacity they have to eliminate biases make them stand out. This study, therefore, was designed to explore the applicability of 4-PL and LNIRTmodels in calibratingComputer-Based Mathematics Achievement Test (CBMAT) among senior secondary school (SSS) students in Oyo and Lagos States, Nigeria.

Instrumentation design wasadopted and the study, hinged on IRT approach, was carried out in two phases. Fourteen senior secondary schools in Oyo State, having functional computers were purposively selected for Phase I, which involvedconstruction, validation and calibration of pooled 114-item CBMAT. Stratified sampling in proportion to number of available computers was used to select 731 SSS II students. Forty-item CBMAT with marginal reliability of 0.89 scaled through validation process. Phase II entailed a purposive selection of Lagos State based on the availability of large numbers of functional computers in Agege, Ifako/Ijaye and Alimosho Local Government Areas. Three schools each in Agege and Ifako/Ijaye and two from Alimosho were selected. In each of the eightschools, CBMAT was administered to 874 examinees (in two batches). Data wereanalysed with Dimtest statistic, Yen Q3 test, IRT Logistic Models, LNIRTmodel and Pearson product moment correlation at $\alpha = 0.05$.

Both pooled and final CBMAT revealed that only mathematics ability trait is dominant in the Dimtest results (unidimensionality) (T=1.028; T=0.06).Two pairs of items were locally dependent (Yen Q_3values = 0.38, 0.31;both were >|0.2|).Model-fit analysis indicated that the test data found a better fit with 4-PL model (-2loglikelihood=97274, AIC=97282 and BIC=97293). Comparisons of parameter estimates among 1-, 2-, 3- and 4PL models were significant (discrimination; T=122.68; difficulty: T=24.45; guessing: T=2.09; ability estimate=16.89)although, estimates of 4-PL model performed better. Also, time-intensity in the LNIRT model estimated an approximation of 60minutes to correctly respond to the final CBMAT. The observed response time showed that 4% of the examinees exhibited aberrance responses. There is a negatively low relationship between examinee parameters ($r_{\theta\zeta} = -0.06$) and moderate and significant correlations existed among item parameter estimates of the LNIRT model(r_{12}=0.39, r_{34}=0.48). LNIRT model produced better examinees' ability estimates ($\bar{x}$=0.0015) when compared to that of the conventional IRT models ($\bar{x}$=0.0003).

Calibrating with 4-parameter logistic in the unidimensional category and lognormal response-time models were effective in estimating examinees' mathematics ability in Oyo and Lagos States. Test-developers are encouraged to use lognormal model for a more objective measurement of examinees' ability.

Keywords: Unidimensionality of mathematics ability trait, Computer-based item calibration,Lognormal response-time model, Parameter estimates in 4-PL model

Word count: 488

TABLE OF CONTENTS

Contents **Pages**

Chapter One: Introduction

Chapter Two: Review of Literature

Chapter Three: Methodology

Chapter Four: Results and Discussions

Chapter Five: Summary,Conclusion and Recommendations

LIST OF TABLES

LIST OF FIGURES

Figures **Pages**

LIST OF APPENDICES

Appendices **Pages**

CHAPTER ONE
INTRODUCTION

1.1 Background to the Problem

The world is undergoing major multidimensional transformations which are helping to make life more meaningful. The global change is all encompassing, as it affects all walks of life including education which is accepted universally as a unique tool for transforming the various aspects of development any society could think of. The quality of education, by implication, dictates the level of development and quality of life of the citizenry. Quality education enables individuals to develop their potentials to the extent that they can contribute maximally to the growth of society. The thirst for quality education has made measuring students' learning outcomes through testing a fundamental issue in education since the quality of learners' performances at the end of every assessment process reflects the quality of the contents they have been exposed to as well as the quality of the process of assessment.

In pursuance of quality education for sustainable development, world leadersat a United Nations Summit on 25th September 2015, adopted the 2030 Sustainable Development Agenda with 17 Sustainable Development Goals (SDGs) and 169 Targets in-between them. Goal 4 specifically talks about quality education that will ensure inclusive and equitable quality education, and promote lifelong learning opportunities for all by year 2030. This goal is to further enhance the realization of a developed society.

To meet up with the 21st century global events for sustainable development and technological advancement through quality education, the Institute of Education, University of Ibadan in her 5th International Conference, came up with a theme, "Quality Issues in Education". Uwadiae (2017) in his keynote address at the conference highlights some key elements of quality education. These include quality learners, quality processes, the quality of the learning environment, quality contentsand quality outcomes. However, the quality of learner's outcome that will be commensurate with the achievable level of development the nation desires is yet to be attained.

Having understood the positive relationship between quality of education and national development, the importance of mathematics education becomesa major focus in Nigeria's quest for technological, social and economic growth. Ukeje in Zakariya and Bamidele (2015) acknowledge the role and significanceof mathematics in the modern culture of science and technology and states that:

> When an order of preference is critically looked at, it seems that mathematics is considered as the foundation on which science, modern technology and modern society are pivoting. An implication of this is that mathematics is considered the originator and the ideal on which science and technology and the essential component of the development of modern society rely (p.1-5).

The ideas of mathematics have facilitated the observed revolution in electronics whichin a way has transformed the present way we think and live. The information technology (IT) of today (hi-technology and internet super highways) has changed the world into a global village with advancements in science and technology through numerous developments in pure mathematics. So, it seems that no nation of the world can make any meaningful accomplishment, particularly in economic development, without technology, whose foundation lies in science and mathematics.

In spite of the importance of mathematics to a nation's development, Ojerinde (1999) rates students' mathematics achievement over the years at both internal and external assessments as unsatisfactory and disturbing. Mamman and Eya (2014) investigates the patterns of students' performance rate in mathematics in the West African School Certificate Examination (WASSCE) for 10 years (2004-2013). Their analysis shows a flunctuating trend in performance patterns with erratic means and variances indicating a stochastic movement over time at both credit (A1- C6) and outright failure (F9) levels. This is worrisome for a nation that is eager to meet up with her counterparts in other developing and developed climes in terms of technological advancement.

Meanwhile, researchers at various times had studied various factors that are attributed to the flunctuating pattern in students'performance in mathematics. These factors range from teachers'(Ekwueme, Meremikwu and Kalu, 2013) tostudents' factors (attitude and commitment, self-esteem, emotional problems and study habits; Aremu and Oluwole, 2001). School factor isalso seen as lack of conducive environment for effective teaching and learning (Asikhia, 2010; Umar-Ud-Din Khan and Mohamood, 2010; Michael, 2011) as well as absence of diagnostic assessment and after-school programmes for low-achieving mathematics students (Ariyo, 2017).Other governmental factor like lack of not enoughworkshopsand seminarssessionsfor teachers is worthy of mention.Omotayo (2017) identifies certain seemly difficult topic like Geometry in mathematics while home factor as parental involvement, parental education, socioeconomic status, language spoken at home and family sizeare considered by Lawal (2009).

In order to bring students' mathematics performance at both school-based and external assessment levels to a steady and improved state, several interventions have been made. These interventions hadbrought the little surge in the performance rate that has been seen so far. However, asatisfactory achievement level where more than 70% of examinees will consistently pass at credit and distinctionlevelsis yet to be attained. Therefore, acontinuous search for a way out is merited.In order to achieve this byway of deviating from usual intervension strategies, this study viewed examinees' performance from an assessment-based perspective where a holistic approach of probing assessment instrument used in examining students in a modern-day test theory framework wasconsidered.

It is worthy of note that the consideration of the psychometric properties, purposes for which assessmentsare given and how well a test-giver is able to technically construct test items might relate with how well students perform in that assessment exercise. The purpose for which assessment is carried out differs as far as different stakeholders in education are concerned. So, for any type of assessment instrument to really measure the desired ability that will elicit the expected performance, high quality items that are devoid of any form of mistakes are needed. This is the reason Ariyo and Lemut (2015) stated that when a strict adherence to due process in test development is followed and items of high quality are framed, qualified candidates will be selected and placed appropriately according to merit.

The development of a good achievement test is not a matter of chance. Okpala, Onocha and Oyedeji (1993) point out that the technique of test construction should involveplanning, item development, item analysis and marking scheme development.These stages are howeversubject to the type of measurement framework employed by the test developer.However, poorly worded test items with vague meanings may be confusing to test takers and, in providing responses to such items, examinees' final marks might not be accurate representation of their actual proficiency. This becomes one of the problems of students' erratic performance rate in school subjects, especially in mathematics.The inability to objectively measure students'ability becomes a fundamental issue in educational testing. Thus, right measurement and assessment aid the design and selection of instruments that will produce minimum error such that valid and reliable assessment results are assured.

Educational researchers, especially psychometricians, are often concerned that the latent trait to be measured to make the purpose of assessment a reality is done with utmost care. This is because of the intrinsic nature of the inherent characteristic. Nenty (2004) is of the view that assessing psychological characteristics like emotions, aptitude, attitudes, interest, behaviour and ability is error prone because of the indirect nature of measurement that is basicallyinferential.So, testing and assessment instruments that are meant to extract the best performance in learners for meaningful decisions must be handled with all seriousness and professionalism (Nenty, 2004). This is to ensure that student's true ability is depicted in hisperformance on the assessment instrument.

Therefore, a theory of measurement that will provide guidelines and directions in an attempt to measure and estimate the given ability level possessed by a respondent is essential in item construction. It means that constructing a reliable and valid instrument is technically tasking and requires a lot of know-howthat will be guided by operationalisable test theories or models (Nenty, 1998). Test theories are carefully worked-out notions that help to validate and explain definite problems on how to develop and utilize tests as well as offer procedures for answering the problems (McDonald, 1999).

Different theories and models with their assumptions are assumed to handle measurement errors differently. For instance, if errors are assumed to follow a normal distrution in a model, another model might appear silent about the distributional assumptions of the error in such model. Alordiah (2015) gives an analogy that an examinee with zero or hundred percent mark in a subject cannot be said to have no or all understanding of that subject in the process of measurement. Scores interpretation is therefore hinged on the theory or model adopted when measuring. This study considered how assessment theory could alleviate the stochastic nature of academic performance either at school-based or public assessment level.

However, two major theories with their accompanying models in educational measurement are the classical test theory (CTT) and item response theory (IRT). These theories are used for developing and analysing psychometric properties of instruments, as well as examinees' abilities and performances.The traditional approach (CTT), views observed score (X_o) (score that resulted from measurement) as the combination of true score (X_T) (expected score) and error component (e) (some unobservable measurement errors that are random and normally distributed) (Crocker and Algina, 2008). CTT model is represented as: $\boldsymbol{X_o = X_T \pm e}$**…….eqn. 1.1**

The prevalent usage of the basics of CTT, its popularity for over nine decades in standardized testing and measurement technologywith its common practice in psychometric analysis were seen by the studies of Ojerinde (2013); Wallace and Bailey (2010) and Morizot, Ainsworth and Reise (2007). Ojerinde (2013) affirms that CTT conceptual details, assumptions as well as its fundamental proofs have permitted improvement of certain exceptional psychometrically stable instruments in testing as far as the Africa continent is concerned. This was made possible due to the ease of interpreting students' learning outcomes in test settings (Hambleton, 1989). Although evidences abound in supporting CTT's wide usage in assessment, its application has however been linked with some deficiencies. Nenty(1998) aguesthat although,CTT usage has sustained educational measurement for long, its measurement is likened to a kind of material that gives an unstable result when exposed to known/unknown extraneous factors.

The other type of approach that is gradually trending and has penetrated both measurement and assessment world in educational setting is IRT. This approachdemands a more objective way of measurement and it overcomesthe many prominent limitations of the CTT approach. This study is anchored on the modern method to measurementbecause of the many advantages it possesses over the traditional method. Objectivity in measurement demands that every effort at measuring the same ability in individuals should give the same result no matter the specific instrument used, the person doing the measurement or the persons with whom the individual is measured.

By the nature of measurement, Wright and Stone in Nenty (2004) stress that if any alteration is observed in what is being measured, the result of such measurement will be termed subjective, and cannot give a valid value of what is being measured. Troy-

Gerard (2004) andAdedoyin (2010) assert that educational measurement is undergoing various reforms with the aim of meeting up with the increasing requests for effective explanation of respondents'assessment score.

IRT approach states that the possibility of responding to a question accurately or of reaching a specific response level is exhibited as a function of a person's profeciency as well as item characteristics (Hambleton and Swaminathan, 1985).Alordiah (2015) posits that an individual's observed score (X_i) in IRT is mathematically given as: $\mathbf{X_i = \theta_i + \lambda_i + \mathcal{E}_i}$**…….eqn.1.2**

Where θ_i is the true examinee ability, λ_i is the systematic error variable and $\mathcal{E}_i$ is the unsystematic error. The acknowledgment of systematic error in IRT is a keydevelopment over CTT. IRT method rests on the idea that an examinee's achievement in a specific item is dependent on two factors: his proficiency and the features of the item (Fulcher and Davidson, 2007).

However, certain assumptions must be met in IRT approach for its effective usage and appropriate interpretations so as to have precise and useful results. These are trait dimensionality, item local independence and monotonicity of response assumptions. The theory originally accepts that a distinct dominant ability is adequately enough to describe examinee's performance (unidimensional models) but research has shown new development in multidimensional traits for multidimensional models (Olonade, Metibemu and Adewale, 2017).

The assumption of local independence states that an item will strictly attract the probability of a right answer from the respondent, based on studentability level on that item and not on the performance on any other item of the test. Item characteristic curve (ICC) that is likened to monotonicity of response function is another fundamental mechanism upon which IRT methods is pivoted (Henard, 2000 and Baker, 2001). ICC is a graphical conceptthat relates the probability of accurate response on a question with the measured ability of the examinee. It takes a normalcurve shape and two technical properties (item difficulty and discrimination) are employed to describe it.

However, in applying IRT functions to analyze assessment questions, the assessment itself as well as the score patterns can only be valid if any of its models holds (Cees and Rob, 2003). IRT models are collections of different mathematical models that permit prediction of examinees' test performance from a particular person's trait and the characteristics of the items that make up a test (Hambleton and Swaminathan, 1985; Ercikan and Koh, 2005). Other assumption according to Hambleton and Swaminathan (1985) that is inherent to every IRT model is that the assessment to which any of the models will fit must not be given in any speeded condition. This implies that sufficient time should be provided for the examinees in responding to the items of the instrument. This provision is such that failure on assessment instrument will not be accrued to insufficient time but only to incapacitated ability.

Various types of IRT models are available in analysing response data. They range from unidimensional (UIRT) to multidimensional IRT (MIRT) models of either dichotomous or polytomous response-type formats. Although IRT models were initiallyestablished for items that are dichotomously scored with unidimensional models, itideas and approaches have been stretched to a wide range of MIRT models. MIRT models are expansion of UIRT which were created to model more precisely estimates of items and examinees in circumstances where items measure more than a dominant attribute (Peterson, 2014).

Meanwhile, unidimensional models with response format that is dichotomous (true/false or correct/incorrect)are the focus models for this study. There are 1-, 2-, 3- and 4-parameter logistic models. One-parameter logistic (1PL) model is termed the most common and simplest of the IRT models. Examinee's correct response is assessed by the possessedability level and the difficulty of the item (b_i) for 1PL model. Item response function of 1PL model is mathematically defined as:

$$\boldsymbol{Pi(\theta_s) = \mathrm{Pr}(X_{is} = 1|\theta_s, b_i) = \frac{1}{1+e^{a(\theta_s - b_i)}}} \text{........}\textbf{eqn 1.3}$$

In spite of the simplicity and popularity of the usage of 1PL model, it is limited due to the fact that it cannot be used with large numbers of examinees. Items are said to only differ in how tough they are before examinees can provide answers to them and not how good they are in assessing the different high and low-ability examinees.Two-parameter logistic (2PL) model was a further formulation of 1PL where an additional estimate was added(discrimination parametera_i) to make it have a better fit. However, 2PL model is strictly applicable to items where guessing is very unlikely. Item response function (IRF) of a 2PL model is given as:

$$\boldsymbol{P_i(\theta_s) = \mathrm{Pr}(X_{is} = 1|\theta_s, a_i, b_i) = \frac{1}{1+e^{a_i(\theta_s - b_i)}}} \qquad \text{..............}\textbf{eqn 1.4}$$

Three-parameter logistic (3PL) model on the other hand, was developed for test items where several alternatives are likely as options. An additional pseudoguessing parameter (c_i) known as lower asymptote was inculcated and the likelihood of guessing becomes a factor. This helps to improve upon 2PL model. Ojerinde, Popoola, Ojo and Ariyo (2014) reiterate that there are some possibilities that some examinees willanswersome items aright by lucky guessing. Then, the influence of chance selection to the probability of a right response brought a lost of some mathematical properties of the logistic function (the additive property). The Item Response Function of a 3PL model is;

$$\boldsymbol{P_i(\theta_s) = \mathbf{Pr}(X_{is} = 1|\theta_s, a_i, b_i\,, c_i) = c_i + (1 - c_i)\frac{1}{1+e^{a_i(\theta s - b_i)}}} \quad \textbf{.....eqn 1.5}$$

Some of the studies where the familiar and commonly used 3, 2 and 1PL models were evident in fitting response data are the works of Steinberg and Thissen (1995); Lanza, Foster, Taylor and Burns (2005); Amarnani (2009); Ojerinde *et. al.* (2012);Ojerinde (2013); Adegoke (2013; 2014), Enu (2015); Metibemu (2016); Olonade (2017) and Fakayode (2018).

In order to overcome some estimation errors of the 3PL model, 4PL model was formulated, where an upper asymptote known as carelessness parameter was added to the model. This was made possible such that a high-ability respondent who as a result of carelessness that might result from mistake, stress, tiredness, inattention, anxiety,lack of familiarity with computer techniques, distraction by poor testing conditions and misreading of questions; responded to an easy item incorrectly. 4PL model mathematical formulation is given as:

$$\boldsymbol{P\mathrm{i}(\theta_s) = \mathbf{Pr}\left(X_{is} = 1\middle|\theta_s, a_i, b_i, c_{i,}d_{i,}\right) = c_i + \left(d_{i,} - c_i\right)\frac{1}{1+e^{a_i(\theta s - b_i)}}}\textbf{...........eqn1.6}$$

This study laid emphasis on theusage of 4PL model in exploring its applicability to test data vis-á-viz other types of uni-dimensional (3-, 2- and 1-PL) models. Meanwhile, the utility of 4PL model did not get significantattention as others models when it was earlier suggested by Barton and Lord (1981). Some of its setbacks are that (a) its application and utility were seen as isolated where no clear agreement on its need was reached (Barton and Lord, 1981; Hambleton and Swaminathan, 1985) (b) the use of maximum-likelihood (ML) method in fitting its model and estimating itsparameters was traditionally seen as a challenging task (Waller and Reise, 2009) (c) 3PL model usage across literature was seen as dominating, and the lack of agreement on the usefulness of 4PL model was a strong argument against its effective usage as pointed out by Linacre (2004) and Loken and Rulison (2010).

This however generated a concern that the estimate of the 4th parameter (d_{j})would be inconsistent since estimating guessing parameter (c_j)with maximum likelihood estimation approachdo pose a problem (Embretson and Reise, 2000; Rupp, 2003; Baker and Kim, 2004).However, 4PL model was reassessed with an improvement in its computational power and resources such that more sophisticated statistical modeling software such as the use of Bayesian method was developed. This constituted a substantial innovation towards a wider deliberation and usage of 4PL model for definite purposes.

Although few studies have been carried out in support of the usage of 4PL model (Chang and Yin, 2008;Rulison and Loken, 2009; Loken and Rulison, 2010; Liao, Ho, Yen and Cheng, 2012),there is however a more pressing demand for its effective usage in calibrating item parameters and estimating examinees' true ability. This is done to enhance the model's utility and popularity for more objective measurement. This study therefore explored 4PL model's usage to further utilise the additional benefits it poses, in term of minimising random error for a more accurate and unbiased measurement.

In the course of this study, computer-based testing (CBT) as a test administration mode was evident. This was to enable one of the variables under consideration to be captured. Almost all assessment bodies, private or public, academic institutions or professional bodies in Nigeria are working towards engaging e-examinations for conducting either assessment or online-registration for their candidates. Assessment bodies as National Examinations Council (NECO), West African Examinations Council (WAEC) and Joint Admission and Matriculation Board to mention but a few.

The computerisation of assessments has acted as a stimulating factor in response time modelling andhas created opportunities for examiners/assessment bodies to examine activities relating to the time respondents spend on individual items of a given assessment instrument. This was previously impossible as aggregate testing time and responses were only accessible in a paper-pencil test. Thus, a shift to this mode of testing is gradually becoming common-place in and outside the classrooms, even in high and low-stakes assessments. This is generating an innovative development in psychometrics where more sophisticated approaches to measuring some variables that were initially difficult to measure in the pre-CBT era are now readily measured in test settings. One of such variables is item response time (RT) which could automatically be recorded in CBT mode.

Schnipke and Scrams (1999) state that response time analysis allows researchers to study the interactions in respondent ability, item parameters as well as examinee response speed. Time taken by respondents is said to relate with examinee true ability (θ) with some measured errors which IRT models are capable of estimating. Some measurement errors incurred in testing could be lessened if more research is concentrated on the area of timing in assessment.Van der Linden (2009) advocates that test theorists are facinatedby the relationship that existed between test item responses and time spent by examinee to give correct response. Therefore, the need to utilise this so-called collateral information (response time) was one of the considerations that informed the study.Suh (2016) affirms that various types of information (speededness, schemes used in pacing and time boundary) in assessment settings could also be evaluated from the response time data gotten from examinees.

Meanwhile, some studies had attempted to proffer solutions to the incessant rising and falling trend in students' performances over the years, such attempts have not yielded satisfactory result for stakeholders. The study however attempteda different approach from assessment-based perspective of the modern-day test theory to analyse the researcher's self-developed computer-based mathematics test instrument. The researcher was of the opinion that the nation's present and future challenges of technological advancement could be alleviated by adopting the right approach of model parameter estimation in assessing examinee's true ability that will surport the valid judement of whom individual student has been made to be.

The use of 1, 2 and 3PL models in addressing measurement problems in assessment is gradually and positively aidingestimating students' true ability and performance, a connotation of an objective measurement of students' latent construct. However, more effort should be intensified until consistent and excellent performances either in school-based or external examinationis attained. If this is achieved, an individual with the right ability would be able to tackle any challenge that is due tohis or her ability level. This will thereby enable the breed of the right set of learners that wouldappropriately handle technological advancement.

However, the usage of the newest and most recent 4PL model whose features take care of other measurement errors so that better estimates of model parameters are provided is worth exploring and applying. Surprisingly, adopting the usage of 4PL as well as response-time models here in Africa is unpopular up till the time the research work is carried out.

1.2 Statement of the Problem

In the class of unidimensional IRT models, the dominant usage of 1-, 2- and 3PL models across the globe in attendance to how model parameters should be objectively estimated have received much attention. This approach reflected impartial measurement of examinees' true ability via their performances. A new 4PL model was later formulated and suggested but its usage became unpopular because of some disparities and technical difficulties resulting from calibrating its parameters. This is in spite of the many inherent advantages it possesses over the previously used models. However, few recent studies revealed the development of more sophisticated statistical software that could easily estimate the seemingly complex and heavily parameterized 4PL model parameters. Therefore, suggestions on its potential applications in educational assessment to impartially estimate students' ability as a reflection their performances in mathematic is empirically required.

Other area where it seems research has not really focused in this clime is in the item response time in testing situation. Response time data seems to allow the understanding of examinees' response behavioural patterns from data-based perspectives. Such patterns help to further investigate the type of relationship that ensues between examinees' ability (performance) and their response speed. If it is acceptable that accuracy of responses to test items is connected with student true ability, then response time should, as well, be given appropriate attention for its effects on examinees' performance in assessment setting. It is worthy of note that no clear reasons were stated across reviewed local literature on why CBT, that affords automatic record of response time data had not been explored in spite of the few assessment bodies noted to have commenced the usage of CBT in Nigeria.

The sufficiency of time allotted to students in supplying correct response to items posed before them was another problem this study explored. An implicit assumption that is common to all IRT models is that giving adequate time is vital to correctly responding to items of a scale. This was to prevent adducing failure to insufficient timing, and not lack of ability. Hence, an empirical documentation is necessary to affirm timing effect.This study was however informed by the need to further inquire the effect of item response time and the use of 4PL model. 4PL and LNIRT models in the calibration of computer-based mathematics achievement test among students in senior secondary schools in Lagos and Oyo States, Nigeria was explored.

1.3 Research Questions

The following research questions were answered based on the problems stated above:

1. Which of the four IRT models for dichotomous test best fits the pooled Computer-Based Mathematics Achievement Test (CBMAT) response data?
2. What is the quality of the pooled CBMAT items under other dichotomous IRT models and the model that best fits the test data?
3. Is there any significant mean difference in the item parameter estimates of the other IRT models and the model that best fits the pooled CBMAT response-data at the developmental stage?
4. How consistent is the model used in calibrating the pooled CBMAT response data at the development stage to the model used in calibrating the final CBMAT response data?
5. Is there any significant mean difference in the examinee's parameter estimates of the other dichotomous IRT models and the model that best fits the final CBMAT response data?
6. Is there any significant mean difference in the item parameter estimates of the other dichotomous IRT models and the model that best fits the final CBMAT response data?
7. What are the estimates of item and examinee's parameters of the Lognormal Response Time IRT (LNIRT) model when the final CBMAT response and response time data are used?
8. Is there any significant relationship between item and examinee's parameters of the LNIRT response time model?
9. What are the patterns of the person-fit statistics for detection of aberrant response behaviour in the CBMAT response time data?
10. How comparable are the item and examinees parameter estimates of the traditional IRT model to the LNIRT response time model?

1.4 Scope of the study

This research work focused on and was limited to the application of the unidimensional IRT models for dichotomously scored data with emphasis on 4PL model. This was as a result of the various suggestions from literature that more studies to further establish the utility of 4PL model are needed. Item response time that served as collateral information to the response given in the CBMAT instrument was another emphasis of the study with the adoption of Lognormal response time IRT model in the investigation of examinees response time viz-a-viz their responses.

The instrument used was limited to the Computer-Based Mathematics Achievement Test (CBMAT) in a multiple-choice response format type developed from the revised New General Mathematics for SSI (2011 edition) and the 2008 Mathematics curriculum of Nigerian Educational Research and Development Council (NERDC). The coverage for the study was restricted to all government-owned secondary schools II (SS1) students in Lagos and Oyo States, Nigeria with functional computer laboratories. Markov Chain Monte Carlo (MCMC) diagnostics was used in the study. This is a Bayesian statistical approach designed to calibrate complex and heavily parameterized models such as the 4PL model and LNIRT response time model.

1.5 Significance of the study

The development of different IRT models (1PL, 2PL, 3PL and 4PL) has always been an attempt towards improving the quality of assessment items such that objectivity and minimal measurement error are fostered in educational assessment. However, the choice of a right model has always been a herculean task, because for quality to be assured, the right model must be applied. Exploring the applicability of 4PL model in developing, validating, scoring and calibrating test items has provided a landmark insight in estimating person and item parameters better. This is as a result of the shift to a more objective measurement test theory (IRT). The study is therefore significant to psychometricians and test developers to producing high quality items that will elicit intended construct in the examinees.

The 4PL IRT model was also developed to handle guessing problems and other careless mistakes examinees might incur in the cause of measurement. No other model among the more popular 1, 2 and 3PL models has been able to cater for the careless mistakes highly-abled examinees commit in the assessment process. This study will enable stakeholder such aspublic examining bodies like WAEC and NECO in the usage of the latest uni-dimension 4PL model.It is significant in the provision of theoretical knowledge for potential researchers as well.

The record of response time in the CBMAT instrument aided the improvement of ability estimation in the LNIRT model used for this study. When estimates are improved, better and true ability as it relates to examinees performances (scores) are obtained. This will in-turn give great confidence to the users of test scores as students, parents, teachers, higher institutions of learning, policy and decision makers. Response time analysis allows researchers to devise appropriate models, secure test validity and further examine otherforms of human behaviour that could influence students' performance positively. Examinees with normal and those with aberrant responses were detected in this study. The study thereby provides researchers with information on respondent's behavioural patterns in the cause of responding to items of a scale.

A further benefit is the reinforcement ofcomputer usageawareness in testing among senior secondary school students. This enables efficient, unbiased test administration and scoring approach that could aid quality and improved standard of education in Nigeria.

1.6 Definition of Terms

1.6.1 Conceptual definition of Terms

Items: Thisis a series of units carefully constructed to facilitate responses from the respondents that are meant for assessment of contents taught within a period of time in an educational testing.

Item Response Theory: This is a modern educational and psychological measurement paradigm for designing, analysing and scoring of tests and similar instruments that are meant to assess students' learning outcomes. The approach uses sophisticated statistics to evaluate the three domains in learning.

Item Parameters (*a. b, c* and *d*):These are the difficulty (location), discrimination (slope), pseudo-guessing (lower asymptote) as well as the carelessness (upper asymptote) parameter estimates that were calibrated inthe models in other to estimate the likelihood of correct response by the examinees.

Item Calibration: This has to do with estimating the parameters of the test and determining the amount of trait level an examinee must possess in order to get an item correctly.

Response Time (RT): This is the amount of time an examinee employs on individual item of a CBT test. It is a variable that is automatically recorded in a CBT to see its effect on examinee ability or performance and pattern of responses.

Lognormal Response Time Model (LNIRT): A joint model that incorporated response and response time model which include time-discrimination and time intensity parameters with the other three parameters that are available on the traditional IRT model to yield unbiased item/person parameter estimates.

Response type Dichotomy: These are two response formats for individual items that can either be right or wrong, true or false, correct or incorrect.

Item Characteristic Curve: This is an assumption that establishes a relation between examinee correct response to an item and his ability trait that is assumed to be influencing the correct response. It can be assessed by plotting with any IRT logistic model using any IRT software package.

1.6.2 Operational Definition of terms

Item Response Theory Models: These are mathematical logistic functions that attempt to typify the connection between an unobserved construct known asa person's latent trait, and the likelihood of correct response to a specific item of an assessment instrument. These models refer to 1-, 2-, 3- and 4PL IRT dichotomous models that could be estimated using different IRT softwares.

Person Parameters (θ and ζ_i): These are examinees' ability/proficiency and the working speed parameters in an academic area (mathematics) that could be estimated using either traditional IRT or Lognormal response time (LNIRT) model.

A fixed-Form Computer-Based Test: A CBMAT, where all test takers answer the same questions. It is a kind of paper-pencil test that is administered on the computer and adopted as the instrument for the study.

Collateral Information: This is additional information known as response time that could give clue to the cognitive processes an examinee exhibits in an attempt to appropriately responde to a question in a test. It is said to have significantly contributed to how examinees mental capability can be assessed.It is estimated with the LNIRT model.

Time intensity: It is one of the lognormal model'sparameters that measures the expected time an item (on a logarithmic scale) should take before a correct response is made. It is estimated with LNIRT model.

1.7 List of Acronyms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CAT	Computer-Adaptive Test
CBMAT	Computer-Based Mathematics Achievement Test
CBT	Computer-Based Test
DIC	Deviance Information Criterion
EAP	Expected á Posteriori
EM	Expectation Maximization algorithm
ICC	Item Characteristics Curve
MAP	Maximum á Posteriori
MCMC	Marcov Chain Monte Carlo Method
MLE	Maximum Likelihood Estimation
NCME	National Council on Measurement in Education
PL	Parameter Logistic
PPMC	Posterior Predictive Model Checking
PPT	Paper and Pencil Test
RT	Response Time

CHAPTER TWO
LITERATURE REVIEW

In exploring the justification and usage of four-parameter logistic model within item response theory context, a theory that is largely anchored on the premise of the invariance property of item and person parameters with a consideration of time of response is essential. Studies have been carried out on the usage of different IRT models in evaluating cognitive processes. For this study, relevant literature and empirical studies that provided both theoretical and conceptual background were examined. Literature reviewed was appraised and certain gaps the study permeated were also evident.

2.1 Theoretical Background

There are paradigms for establishing relations between observations made and constructs to be measured. Such paradigms attempt to explain all the facts with which such construct can be approached. The application of theories in the analysis of items of assessment instrument tends to bring fair and reliable judgement to the outcome of any measurement made. This study therefore was anchored on one of the major paradigms used in the educational assessment realm known as item response theory (IRT).

2.1.1 Theories and Models Relating to Test

The historical development of test theory has been linked to one of the disciplines in science known as psychology. The efforts of great psychologists both in Europe and United Statesbrought about the evolvement of test theory as they studied a variety of psychological and educational problems. Test theory is seen as part of the science that consists of statements that are well spelt out, containing acceptable guidelines and techniques of inquiry. It is a principle that guides action or assists understanding or decision making (Ariyo, 2015).

The justification of test theory is the provision it makes for understanding examinees' abilities and their true observed scores. Test theories are used to solve measurement problems especially in the validation and improvement of the quality of test items or assessment instruments. Test models, on the other hand, help to understandbetter the relationship that exists between the observed and unobserved scores in a testing situation (De Champlain, 2010). Therefore, as assessment is progressively becoming universal, educational measurement is being appropriately positioned to gain from theories.

Theories, functions, models or principles are needed to provide a guide and basis for estimating what the researcher is looking for from that which is being observed. Scores that emanate from the effort of measurement, to some extent, are not error-free in representing the trait levels of individuals being measured. Therefore, there is need to measure aright the scores in other to estimate the trait levels of examinees more validly. However, different theories are being accompanied with their own models.

Although test theories and their accompany models are in most times used exchangeablly, theoryis fundamentally the assemblage of mathematical conceptions which validate and explain certain investigations about the construction and usage of tests, thereby providing methods for answering them (McDonald, 1999). They also provide an overall structure for relating observable variables such as item and test scores; to the latent variables,like the true and proficiency scores.

Theories which present conceptual scores as accurate, observed and error scores can not be adjudged suitable or inadequate unless they are wholly represented by specific models (Nenty, 2004). Test models could be expressed as specified by the concepts of that theory. Such concepts are details that show relationships in a set of theoretical concepts that accompany some fixed assumptions on how to make use of them.The suitability of test models can be assessed by subjecting them to specific sets of empirical data which can be achieved by carrying out research or model fit analysis.

2.1.2 Significance of Theories and Models in Testing

Theories and their associated models appear significant in the measurement practises of education and psychology as a result of different frameworks provided in considering issues, addressing technical problems and handling of measurement errors. Test theories make experts to be mindful of the mathematical models and other rationalitythat enables standard practices in test development and use. Nenty (1998) posits that a functional theory or model provides the help needed in measuring what is to be measured aright so that what is expected can be predicted.

Wiberg (2004) and Hambleton and Jones (1993) stress on the fact that any test theory that is beneficial will helps professionals to know how importanterrors of measurement are in (i) an attempt to estimatean individual’s capability and how such error could be lessened (ii) correlating variables, and (iii) assessing correct proficiency scores with their related confidence bounds.

Various models as well as the theories that uphold them deal with inaccuracies in measurement in a different way. For instance, if errors are expected to have a distribution that is normal in a model, non -normality assumption might be applicable in another model. Additionally, the extent of estimation errors can be consistent in a test-score scale in one model while in another model, error sizes could be ascribed to the respondent's true score. Therefore, how error is specified in a model will tell the extent at which such error scores are assessed and conveyed.

Acceptable test modelshould be able to state precisely the relationships between items of a test and proficiency scores such that test design work can be cautiouslydone to give the anticipated assessment score distributions andtolerable error estimate. An instance is observed in computer adaptive testing, where models that connect ability estimates to item characteristics are required toinform the processes of item selection. Test theory therefore plays asignificant role in the generalprocess of research methodology by presentinga broad method of measuring variables of interest as well as testing the sensitivity and accuracy of the procedural methods needed in measurement (Croaker and Algina, 2008).

Thus, two major popular test theories exist that are classified into traditional and modern-day test theories for addressing assessment issues and solving assessment problems in educational settings. Such problems arise in test construction, test-score equating and biased test items identification. The theories are classical (CTT) and item response (IRT) theories.

2.1.3 A Summary of the Traditional Approach (CTT) and its Model

The traditional test theory also known as the classical test theory methods to test development and scoringwas predominantly used in the early 1900 within intelligence testing context (Ojerinde, Popoola, Onyeneho and Akintunde, 2013). Its methodology begins with a presumption that organised effects noted in the reactions of examinees are expected within a variety of abilities. All other probable bases of difference that occur in the testing situations can either be due to the external or internal states of the respondents. An essential model of CTT states that observed test scores (X_0) consisted of trait score (X), which is a reflection of the exact value of the respondent's ability and an error score ($\boldsymbol{X_e}$), a reflection of the consequence of extraneous effects of the measurement procedure when measurement is done (error of measurement). The accurate and error scores do not depend on each other. This concept was established by Spearman (1904) and Novick (1966) and illustrated as:

$$X_0 = X + X_e \qquad \text{……………………..eqn. 2.10}$$

Where the true score is$X= X_0$ - X_e and its variance is given as,

$$\sigma_X^2 = \sigma_{X_0}^2 - \sigma_{X_e}^2 \qquad \text{……..…………eqn. 2.11}$$

Some underlying assumptions that made it to be relatively simple to interpret are: the error and true scores from the same test must have a correlation of zero.It is assumed that expected mean of zero is attached to the error terms and errors from parallel measurements do not relate (Lord, 1980). It is assumed that errors that result from a measurement are uncorrelated with those obtained from a different measurement. These assumptions and the CTT model form the basis of the psychometric concept of reliability and validity coefficient. The relationship between the observed scores on an instrument and the corresponding trait scores (true scores) is the index reliability for the instrument (Croaker and Algina, 2008). Therefore, using the variances of the true scores (σ_X^2) and observed scores $(\sigma_{X_0}^2)$, the population reliability of a test scores in CTT is obtained as:

$$\rho_{xx'} = \frac{\sigma_X^2}{\sigma_{X_0}^2} \qquad \text{…...................eqn 2.12}$$

And an assessment of the variance of the errors of measurement in any set of observed scores is $\sigma_{X_e}^2 = \sigma_{X_0}^2\,(1 - \rho_{xx'})$ whose square root, σ_{X_e} is termed the standard error of measurement. It is also known as the standard deviation of the errors of measurement that are related to the observed scores for a particular set of respondents.

CTT method is associated with how observed score, standard error of measurement and the true score are determined and compared with some preset criteria so as to interpret the score (Amarnani, 2009). It basically makes individual test generally discrete and an atomic whole in which specific items irrespective of their difficulty or predictive power, contribute equally to the raw score an examinee gets in the test, and to the scaled scores and eventually to the assessment itself. Some item statistics have been developed as regard test development in CTT, and the frequently used ones are difficulty and discrimination of an item. Item difficulty is presented as the proportion of individuals in a sample that answer an item correctly. So, value of difficulty index that is near zero (0) depicts a hard item while those with values near one (1) connote an easy item. Item discrimination on the other hand is usually projected by correlating the item scores with the total test scores. This is also known as item-total correlation. It is used to describe the differences in item difficulty for high and low performing respondents. Values of the item-total correlation at or below zero (0) indicate a poorly functioning item while strong negative correlations are signs of mis-keyed item.

CTT also takes a safer route by essentially compiling the psychometric power and the standard error of each item to produce a robust test that can withstand the subversions of other individual differences that may otherwise adversely affect the eventual score in a test (Sijtsima and Junker, 2006). This theory uses the standard deviation of errors as its main proportion of error called the standard error of measurement. Hypothetically, the amount of measurement error is gauged by the standard deviation of the distribution of random errors for each individual. It is typically accepted that the conveyance of random errors will be equal for all examinees.

Wallace and Bailey (2010) and Morizot, Ainsworth and Reise (2007) confirm the prevalent usage of the fundamentals of traditional test theory and its popularity for over nine decades in standardized testing and measurement technology. They observed that a common practice in psychometric analysis is the usage of CTT which compares the difference in the observed versus true participant scores.Its reliability solely depends on parameters that are strongly reliant on the sample. McDonald (1999), Zickar and Broadfoot (2009) likewise proposed that CTT approach is as perfectly healthy and worth utilizing for some applications.

Hambleton and Jones (1993) list some of the benefits obtainable through the application of CTT to measurement problems: (a) Smaller sample sizes are required for analyses (b) Simpler mathematical analyses compared to item response theory (c) Model parameter estimation is conceptually straightforward (d) Analyses do not require strict goodness-of-fit studies to ensure a good fit of model to the test data because of a simple assumption that the model cannot be disproved by any set of data since a respondent’s true score is latent and unknown and then the associated error with the observed score is unknown.Ariyo (2015) observed however that increasing the number of items on an assessment instrument could aid precision in measurement with CTT framework but such advice seems questionable.

Yet CTT approach is fraught with many shortcomings that made researchers' attention to shift unto the modern-day theory approach to enable more objectivity in measurement. Nenty (1998) also corroborate that though CTT usage has sustained educational measurement for almost a century,it is a measurement with an elastic ruler that shrinks and stretches under pressure from known and unknown extraneous forces. Hence, it produces results that are at best meaningful only in extremely limited circumstances.Ariyo (2015) noted that complex strategies have been proposed to overcome some of the restrictions CTT approach is confronted with.

2.1.4 Issues with the continual usage of Classical Test Theory Approach

Ojerinde (2013) affirms that the fundamental principles of the traditional approach have been the foundation of measurement theories for almost a century and that the theory has made the advancement of some great psychometrically stable instruments to be realizable in educational measurement in Africa. It was feasible because of the simple way of analysis that is conveniently associated with interpreting examinees learning outcomes in assessment (Hambleton, 1989). The prevalent usage of the fundamentals of CTT in standardized testing and measurement technology and its common practice in psychometric analysis were also noted by Wallace and Bailey (2010) and Morizot; Ainsworth and Reise (2007).

In spite of its prevalent usage, CTT assumptions are termed weak; its item and person statistics are group and test dependent. This is supported by Hambleton, Swaminathan and Rogers (1991), Nenty (1998), Hambleton and Jones (1993), Adedoyin (2010) and Ojerinde (2013). They saw the usefulness of the results produced by CTT approach as meaningful in extremely limited circumstances. The following shortcomings were observed with a continual usage of CTT approach:

a) Item statistics, such as item difficulty and item discrimination, depend on the particular examinee samples based on what they obtained, i.e. they are group and test dependent. The average level of ability and the range of ability scores among a sample of examinees influence, often substantially, the values of the item statistics.
b) Examinees' scores on their ability are solely test dependent. The examinees ability changes depending on different occasions they take the test which results in poor consistency of the test (Ariyo, 2015).

c) Reliability is established through the concept of parallel tests, and this is difficult to achieve in practice. This is because individuals are never exactly the same on a second administration of a test because people tend to forget.They develop new skills or their motivational or anxiety levels might change.

d) CTT approach reflects on test level information to the exclusion of item level information. It therefore deals with individual's total score and not their ability at the individual item level. Test level information is an additive process.It is the sum of the information across items, and item level information is only the information for a particular item. Nenty (2004) admits that researchers in educational processes in Africa have unquestionably accepted this, and have applied the results from measurement based on the CTT without much caution.

e) Another problem is that CTT assumes that all items are equal; differences in difficulty, discrimination and vulnerability to guessing do not play direct role in generating raw scores. The ability level of an individual examinee is determined only by the quantity and not the quality of items (Lord, 1980).

f) CTT approach lacks the ability to determine what a particular examinee might do when confronted with a test item. It cannot predict how a test taker will perform with respect to a hypothetical item, unless the item has been previously administered to a similar individual.

g) Mode-data fit analysis cannot be carried out with the simple assumption of CTT model because of the unknown nature of the true scores and the associated error of the observed scores (de Ayala, 2009).

With the limitations stated above, a more objective measurement theory was sought. It can be seen that CTT deals with the individual's total score, and not their ability at the individual item level (Hambleton, Swaminathan and Rogers, 1991). Therefore an alternative to CTT approach was needed that brought about the existence of item response theory (IRT) approach.

2.1.5 An Overview of Item Response Theory (IRT) and its Models

The contributions of Lawley (1943), Birnbaum (1957, 1958), Rasch (1960) as well as Lord and Novick (1968) to psychometric test theory evolved into IRT methods (Hambleton and Swaminathan, 1985). This theory examines the relationship between the possibility of a student correctly answering a test item and the trait of the respondent on the scale (Hambleton and Swaminathan, 1985; Lord, 1980). The essence of IRT is that the possibility of responding to a question aright can be modelled as a function of examinee's ability (θ) and the characteristics of the item. This individual ability or latent trait can be measured on a transformable scale with a midpoint of zero, a unit measurement of one and ranges from negative infinity (-∞) to positive infinity (+∞) but ranges in practice from negative three (-3) to positive three (+3).

Under IRT, $P(\theta)$ is used to represent this probability. Hedeker, Mermelstein and Flay (2006) are of the view that IRT approach supplies a better statistical method for analysing response data that relates to educational and psychological scale. Ojerinde, Popoola and Ariyo (2015) affirm that IRT approach is the most significant development in psychometrics that assumes an existence of a fairly common feature or characteristic which is used to determine individual capability to attain a specific task. This theory incorporates measurement assumptions about examinee, item and test performance and how the performance relates to knowledge as measured by individual items on a test.

Before the advent of item response theory, the traditional theory was the only theory employed in estimating student's score on a scale. As soon as IRT strategy was developed, several restrictions that CTT approach was fraught with evaluating test data were overcome. In CTT approach, the test is taken as the unit of analysis while IRT method emphasizes on the test item. The focus of IRT method on items as the unit of analysis efficiently solved the many problems confronting CTT (Hedeker, Mermelstein and Flay, 2006).

Hambleton and Jones (1993) stress that for researchers that value invariant items and person statistics, the way out lies in the concepts, models and methods associated with IRT. This was the point made by Lord in his doctoral thesis as a psychometric monograph in 1952 and in an article in 1953. Nering and Ostini (2010) see IRT method as a paradigm for the design, analysis and scoring of tests, questionnaire and similar instruments measuring abilities, attitudes, or other variables. While Osterind and Wang (2012) viewed it as an approach to modern educational and psychological measurement that posits a particular notion about cognition and sets forth sophisticated statistics to appraise cognitive processes. Embretson (1996) states the view of some devoted IRT test designers, that imbibing classical techniques is a relic which will soon be discarded with better training in measurement. An attention to the deficiencies of CTT and the potential advantages offered by IRT led some measurement practitioners to opt to work within the item response theory framework.

The reason for this change by the psychometric and measurement community was as a consequence of the benefits obtained through the application of item response models to measurement problems.

a) The search for objectivity in measurement demands that the calibration of assessment instruments must be independent of those objects and that the measurement of the object's characteristics must be independent of the instrument that is used for measuring (Wright and Stone in Ojerinde, 2013). Invariance is the bedrock of objectivity and the most desirable scientific property of any measurement.Lack of invariance raises a lot of questions about the scientific nature of psychological measurement (Nenty, 2004; Adedoyin, 2010).

b) An optional provision to the CTT method as a basis for analysis. IRT approach presents a powerful psychometric paradigm for developing, delivering, analysing and scoring assessments. In order to utilise IRT with the aim of obtaining accurate results, assessment data must be calibrated with sophisticated software designed for that purpose. All these were made possible as a result of the development of software like Logist, Parscale, Pascal, Xcalibre, Bilog, Bilog-Mg (Zimowski, Muraki, Mislevy, and Bock, 1996), Multilog (Thissen, 1991), WinBUGS, Irtdif, Facets, Winstep, Stata, Mplus, SAS and R-Packages.

c) IRT typically provides a more flexibe and sophisticated information, for tasks that could be accomplished through the CTT approach, given opportunity to the researchers to improve the reliability of assessment (Ojerinde, Popoola, Onyeneho and Akintunde 2013). The concept of reliability and measurement error that is handled by item information function is calculated for each item (Lord, 1980) which gives a sound premise for picking items in test development. Item information function considers all item parameters and shows how effective a measured item is at different ability levels.

d) To the practitioners, the possibility of item banking and adaptive testing become beneficial because of thenoticeable advantage of IRT techniques (Cella, Dymond, Cooper and Turnbull 2007). The opportunity of knowing the type of item examinee will respond to, in terms of the particular attribute such is having. Saving tester and testee's time becomes an important benefit of adaptive testing.

e) Reise and Waller (2009) advocate that only with proper IRT scoring can the large differences between individuals who differ only slightly on total score be detected. This is the case for extremely high or low trait scores that are not in the normal range. It is an argument in favour of all IRT models and merely implies that items should be "difficult" enough for the levels of the trait of interest.

f) Item difficulty in IRT is established independent of participant abilities, and so item and scale information eliminate the dependence on statistical reliability. The approach is also capable of clarifying the extent of discrimination between two participant groups, that is, to differentiate between individuals at different trait levels (Morizot, Ainsworth and Reise, 2007)

Even though most of the early uses of IRT approaches happened to be in the field of education, it applications currently have been stretched to different domains like the social sciences. These include areas as psychopathology (Reise and Waller, 2003; Waller and Reise, 2009), attention deficit and hyperactivity syndrome (Lanza, Foster, Taylor and Burns, 2005) and criminal behaviour (Osgood, McMorris, and Potenza, 2002). Other obvious areas are in attachment (Fraley, Waller and Brennan, 2000) and personality (Ferrando, 1994; Gray-Little, Williams and Hancock, 1997; Rouse, Finger and Butcher, 1999; Steinberg and Thissen, 1995). IRT is obviously valuable for scale development and acquiring true characteristic estimates in different areas of request.

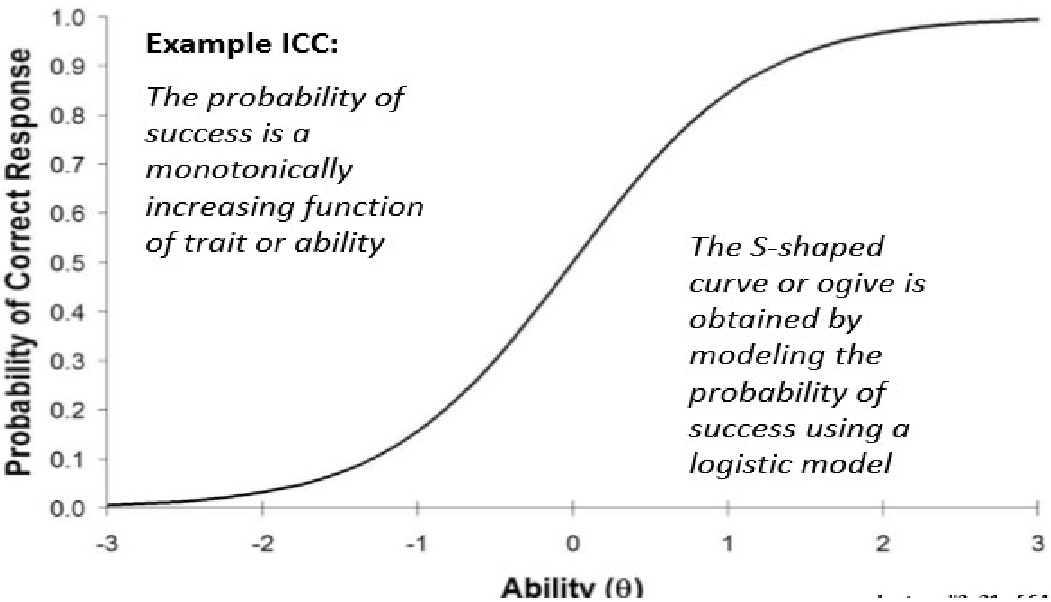
However, most applications of IRT approach assume unidimensionality, and all IRT models assume local independence and monotonicity of item characteristics curve (ICC) assumptions. Unidimensionality means that only one construct is measured by the items in a scale. This means that responses made to questions are directed by one dominant underlying trait (de Ayala, 2009). Local independence trails logically from the unidimensionality assumption (Lord, 1980) and it means that the items are uncorrelated with each other when the latent trait or traits have been controlled for (McDonald, 1981).In the words of Reeve (2000), this assumption asserts that responses to a question do notdepent on the responses to another question in as much as there is a control on the basic variable measured by the instrument. Items are statistically autonomous; every one displays how valuable it is as well as the exhibition of respondent's sufficient capability, which contributes to how functional a specific item is (Yen, 1993).

In other words, local independence is obtained when the complete latent trait space is specified in the model. If the assumption of unidimensionality holds, then only a single latent trait is influencing item responses and local independence is obtained (Hays, Morales, MPH and Reise, 2000). Assumption of local independence is most times debased when items are crowded within reading passages while multidimensionality appears in situations where mathematical word problems that necessitate reading and application of mathematical skills by examinees to give a high probability of supplying right answer. Unidimensionality assumption is also faulted more commonly in circumstances where mixed or innovative item formats are seen (Peterson, 2014).

The relation between the anticipated responses to an item and the latent trait is known as the item-characteristic curve (ICC).With dichotomous items, there tends to be*s*-shaped relationship between increasing respondent trait level and increasing probability of answering an item. The ICC displays the nonlinear regression of the probability of a particular response (*y* axis) as a function of trait level (*x* axis). Items that produce a non-monotonic association between trait level and response probability are unusual, but nonparametric IRT models have been developed to take care of such (Santor, Ramsay and Zuroff, 1994).

Two technical properties are used to describe the curve. These are given as:(a) Item difficulty- this property describes where the item will function along the ability scale. This means that an easy item will function among the low-ability examinees while a seemly difficult item will function among the high-ability examinees. It therefore serves as a location index. (b) Item discrimination- this other property tells how effective a question will distinguish examinees that are having abilities below the item location and those students having abilities above the item location on the continuum. Item discrimination helps to reflects how steep the ICC will look in its middle part. Item tends to discriminate better, when the curve appears steeperwhile the curve will look flatter whenit seems that the item discriminating power is poor since the probability of correct response at low ability levels is nearly the same as it is, at high ability levels. Figure 2.1 is showing a typical ICC where different ability levels span the horizontal axis and the probabilities of correct response are indicated on the vertical axis of the graph.

The middle of the ICC is steeper in slope, implying large changes in probability of answering with small changes in trait level. Item discrimination corresponds to the slope of the ICC. The ICC for items with a higher probability of correct response (easier items) are located farther to the left on the trait scale, and those with a lower probability of correct response (harder items) are located further to the right.



Example ICC:

The probability of success is a monotonically increasing function of trait or ability

The S-shaped curve or ogive is obtained by modeling the probability of success using a logistic model

Fig 2.1: A typical Item Characteristic Curve. Source: Templin (2011)

In Figure 2.1, x-axis is the proficiency levels that ranges from -3 to +3. In actual fact, there may not be examinees that can reach a proficiency level of +3 or fail so miserably as to be in the -3 group. To study the performance of an item given a person whose θ is +3, the probability of giving the right answer is close to 1. The ICC indicates that when θ is zero, which is average, the probability of answering the item correctly is almost 0.5. When θ is -3, the probability is almost zero while the probability increases to almost 1when θ is +3.

Although IRT models were originally developed for dichotomous items and assumed that an individual score on the test is a unidimensional latent trait, extensions for polytomous items and multidimensional latent traits are evident (Thiessen and Steinberg,1986; Embreston,1991; 2000; Mellenbergh, 1995; van der Linden and Hambleton, 1997; Embreston and Reise, 2000; De Ayala, 2009). The progression experienced in the usage of IRT framework comes with the desire to carefully reflect the various parametric forms for IRT models as well as having meaningful explanation and values for inference.

Items in a given task in IRT approach are characterized by (i) difficulty parameter, *b* (ii) discrimination parameter, *a* (iii) vulnerability to guessing parameter, *c* and (iv)carelessness parameter, *d.* The *b* or threshold parameter tells how easy or difficult an item is. It is used in the one-parameter 1PL IRT model. The *a*-parameter tells how effectively item can discriminate between highly proficient students and less proficient students. The 2PL IRT model uses both *a* and*b*-parameters. There is a tendency that some ICCs will cross over each other because of the discrimination parameter in 2PL where we have several ICCs on a graph, unlike the 1PL model where no two ICCs can cross each other.

IRT models are mathematical equations describing the association between a respondent's underlying level on a latent trait and the probability of a particular item response using a nonlinear monotonic function (Reise, Widaman and Pugh, 1993). A variety of IRT models have been developed for dichotomous and polytomous data within the context of IRT. These models are used for two basic purposes: (a) to obtain scaled estimates of θ as well as to calibrate items and examine their properties (Lord, 1980). These measurement models are often applied to cognitive data like achievement test data or attitudinal, behavioural, personality or other non-cognitive data.

2.1.6 Dichotomous and Polytomous Item Response Theory Models

Different kinds of IRT models are distinguished by the functional form specified for the relationship between underlying ability and item response probability (the ICC). IRT models available in the single-trait (unidimensional) ability framework with dichotomous response format (true/false, yes/no or correct/incorrect) are four. They are 1-, 2-, 3 and 4-parameter logistic IRT models. These models are collection of different mathematical models that permit prediction of examinees' test performance from an individual's standing on a trait and the characteristics of the items that make up a test (Hambleton and Swaminathan, 1985 and Ercikan and Koh, 2005). In dichotomous item response models, the only type of response data is binary (0 or 1).

On the other-hand, items that are meant for polytomous models indicate the ones with more than two options of response and their use is quite common in the behavioural sciences where respondents typically express their opinions. These models are not named after numbers like their dichotomous counterpart; instead they are called by different names. For example, a questionnaire on attitude, using Likert-scale response-type items, may result in five categorical responses (strongly disagree, disagree, neutral, agree, and strongly agree). IRT polytomous models for all types of psychological variables that are measured by rating scales of various kinds (vander Linden and Hambleton, 1997) and a variety of multidimensional IRT models. Examples of which include the Partial Credit Model, the Generalized Partial Credit Model (Masters,1982; Muraki, 1992) andthe Multilog and Parscale Graded Response Models(Embretson and Reise, 2000).

Templin (2011) stated that three major kinds of polytomous models exist. These are the Ordered Category Models (Graded Response or Modified Graded Response Model), the Partially Ordered Category Models (Partial Credit or Generalized Partial Credit Model) and the Unordered Category Models (Nominal Response Model). Although, he affirmed that there are many more polytomous models, the aforementioned are the major categories.

The normal ogive form of the Samejima's model for graded responses is given as:

$$\pi_{ijk} = P(Y_{ij} = k|\theta_i, \kappa_{jk}, \kappa_{j,k+1}) = \frac{1}{\sqrt{2\pi}} \int_{\alpha_j\theta_i - \kappa_{j,k+1}}^{\alpha_j\theta_i - \kappa_{jk}} e^{-t^2/2}\, dt$$

....eqn 2.14

For the Graded Response Model, the probability that the i^{th} examinee's response will fall in the k^{th} category on item j is written as:

$$\pi_{ijk} = P(Y_{ij} = k|\theta_i) = P^*_{ik} - P^*_{i,k+1} \qquad \textbf{............eqn2.15}$$

The Partial Credit Model (PCM), an extension of the dichotomous 1PLM (Martelli, 2014) is expressed as;

$$P\left(Y_{ij} = k|\theta_i\right) = \frac{exp\{\sum_{u=1}^{k}(\theta_i - k_{ju})\}}{\sum_{v=1}^{k_j} \exp\{\sum_{u=1}^{k} D\alpha_j(\theta_i - k_{ju})\}} \textbf{..................eqn2.16}$$

While the Generalized Partial Credit Model is a modification of the PCM with item discrimination parameter added to the model. Martelli (2014) expressed the model mathematically as:

$$P\left(Y_{ij} = k|\theta_i\right) = \frac{exp\{\sum_{u=1}^{k} D\alpha_j(\theta_i - \beta_j + K_{ju})\}}{\sum_{v=1}^{k_j} \exp\{\sum_{u=1}^{k} D\alpha_j(\theta_i - \beta_j + K_{ju})\}} \textbf{..........eqn2.17}$$

Where **D** = 1.702 is the scaling constant and assumption of constant discrimination parameter of test items is relaxed, in fact α_j parameters may vary across items. Reckase (2009) provides an exhaustive illustration of such models.

Usually, in IRT models it is assumed that there is one (dominant) latent variable θ that explains test performance. However, it may be a priori clear that multiple latent variables are involved or the dimensionality of the latent variable structure might not even be clear at all. In such case, multidimensional IRT (MIRT) models can serve confirmatory and explorative purposes. Meanwhile, models in the MIRT have some comparisons with the ones used in the UIRT strategy where the chance of a student responding aright to a certain item is stated as a function of item and person characteristics and the relationship is demonstrated in *m*- dimensions. Therefore, response probabilities are restricted on numerous latent abilities denoted by a vector of θ_s. Multidimensional item response model in its general form is given as:

$(U_i|\boldsymbol{\theta}) = (\boldsymbol{\theta},)$ **......................eqn2.18**

whereU_i indicates response to an item for a specific person; θ stands for the vector of person's ability parameters while and γ gives the vector of item parameters (Reckase, 2009). MIRT models are categorized into compensatory and non-compensatory subject if the ***m*** latent traits are indicated to follow either a multiplicative or additive relationship as presented by the item response function (Peterson, 2014). Non-compensatory models show a kind of relationship that is multiplicative in nature in a way that the levels of ***m*** latent traits influence the response probability. For compensatory models, if a trait is operating at a high level, it can reimburse the insufficiencies on other trait because of the additive nature of the relationship. A multidimensional two parameter logistic (M2PL) model is presented as:

$$P\ (U_{ij} = 1 |\ \theta_j,\ \boldsymbol{a_i}, d_i) = \frac{e^{a_i\,\theta'_j + \mathrm{d}_i}}{1+e^{a_i\,\theta'_j + \mathrm{d}_i}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\mathbf{eqn2.19}$$

Where $a_i\ \theta'_j = a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \ldots\ldots + a_{im}\theta_{jm} + d_i = \sum_{i+1}^{m} a_{il}\ \theta jl\ + di$

andθ is a 1 x m vector of person coordinates in *m*-dimensional space. *a* is a 1 x m vector of discrimination parameters. The parameter *d* is a scalar which represents the items intercept.

However, the different IRT models that are available for estimating person and item parameters in the uni-dimensional and dichotomous framework are the one, two, three, four-parameter and even the five-parameter (a model used in the clinical sciences with immunoassay data that are asymmetric, Gottschalk and Dunn; 2005) logistic IRT models.For the purpose of this research work, which is an exploratory study on the 4PL IRT model, emphasis will be on dichotomous/uni-dimensional models with much concern on the 4PL model which is considered as latest model in the calibration of test-item characteristics and the estimation of examinees' abilities (Rulison and Loken, 2009; Loken and Rulison, 2010; Chang and Yin, 2008 and Liao, Ho, Yen and Cheng, 2012).

One-Parameter Logistic* (1PL) */Rasch Model:This is the most common and simplest IRT model. It was published by the Danish mathematician, George Rasch in the 1960s (van der Linden and Hambleton, 1997). According to the Rasch model, an individual's response to abinary item is determined by histrait level and the difficulty of the item. Ojerinde, Popoola, Ojo and Oyeneho (2012) specify that the probability that an individual with a particular trait level of difficultywill correctly answer an item. Under this model, the discrimination parameter ***a*****,** is set to***1*****,** with the assumption that there is equal discrimination across items, and guessing parameter ***c*****,** is constrained to ***0*****,** assuming little or no impact of guessing. This results in one-parameter models having the property of specific objectivity.This means that the rank of the item difficulty is the same for all respondents independent of ability, and that the rank of the person ability is the same for items independent of difficulty. Thus, one-parameter models are sampled independent, a property that does not hold for two, three or four-parameter models. 1PL Item Response Function (IRF) is mathematically defined as;

$$\boldsymbol{P_{ix}\,(\theta_s) = Pr(X_{is} = 1|\theta_s, b_i) = \frac{1}{1+e^{a(\theta_s-b_i)}}}\text{............}\textbf{eqn 2.20}$$

X_{is} =1 is the correct response (X) made by subject *s*to item i, $\theta_s$is the trait level of subject s, b_i is the difficulty level of item i, a is item discrimination =1, exp is the base of the natural logarithm (e = 2.718) and Pr (X_{is}=1|θ_s, b_i) is the probability that subject *s*will respond to item *i*correctly given the subject's trait level θ and the item's difficulty b.

The 1PL model is the most commonly used as a result of its simplicity because the test score is a sufficient statistics for estimating θ, the number of examinees that correctly respond to an item is a sufficient statistic for estimating difficulty index and the model fits right with the number right scoring. However, 1PL model is limited in a way that it cannot be used with large number of examinees and does not give what it ought to be known about the factor structure of items i.e. items only differ in how hard they are to answer, but did not differ in how well they assess the latenttrait. Alordiah (2015) applied the 1PL model on mathematics achievement.

Two-Parameter Logistic (2PL) Model:This model was named after Birnbaum (1968) who proposed an item characteristics curve (ICC) that additionally estimates item discrimination which 1PL model considered to be same for all items across the test. The addition of the discrimination parameter$a$, not only makes the model to be a better fit to the data but also leads to non-parallel trace lines. a-parameteris proportional to the slope of the tangent line at the point on the θ scale equal to the $\boldsymbol{b}$-parameter. The higher the value, the more an item contributes to a student's ability estimate. Thus, an item can be both harder at low levels of ability and easier at high levels of ability.

Items differ not only in how hard they are to be answered, but also in how well they assess the latent trait. This made item discrimination parametera possibility, and this has the effect of magnifying the importance of the variation in the subject's proficiency as well as how hard the item is. 2PL model proposed that the probability of the performance of any individual confronted with an item involves the power of how hard and differentiating such item is in addition to the latent ability. IRF of a 2PL model is mathematically defined as;

$\boldsymbol{P_i(\theta_s) = \Pr(X_{is} = 1|\theta_s, a_i, b_i) = \frac{1}{1+e^{a_i(\theta_s-b_i)}}}$ **...........eqn 2.21**

All other identities remain the same as in 1PL model except for item discrimination **a**-parameter with the subscript i that will take on different values. Yet 2PL model is limited in the sense that no consideration is given to the pseudo-guessing parameter in case there are multiple choice test items. The 2PL is equivalent to the 3PL model with $c_i = 0$. This model is appropriate for testing items where guessing the correct answer is highly unlikely, such as fill-in-the-blank items or where the concept of guessing does not apply as such like the personality, attitude or interest items (agree/disagree).

The Three-Parameter Logistic* (3PL) *Model*:** The original work on IRT was developed for items where there was no guessing. But, when taking a multiple choice ability test with n alternatives, it is possible to get some items correct by guessing. Even, if one knows nothing about the content, correct random guessing of at least 1/n% is tenable. Therefore, the3PL model was developed by Lord (1980),who included an extra parameter known asguessing that estimates the ICC lower asymptote (c***)which happens to be the curve's low point on the horizontal axis as it moves to negative infinity. This explains why people with low latent ability can sometimes rightly endorse an item. In the earlier models (1- and 2-PL), this probability approaches 0 for a weak respondent answering challenging questionsaright and 1 for a brilliant examinee answering not too difficult item. This assumption sometimes might not work foritem that involves ***n*** alternatives as a result of the chance factor that might aidright choice of the key by chance. Yen, Ho,Liao, Chen and Kuo (2012) state thatif partial knowledge of the subject matter is possessed by the student, the likelihood of correct response would even bestronger.

A belief that is taken as part of the realities of life in assessment is the fact that examinees can probably guess some items aright. This indicates that the chance of correct response contains some minor component that could be attributed to guessing (Ojerinde *et al*; 2014). The two previous models did not take the guessing phenomenon into consideration which made the 3PL model to cater for the involvement of guessing the correct response.

Another purpose for including parameter **c** in the model is to attempt to account for the misfit item characteristics at the low end of the ability continuum, where guessing is a factor in test performance. Theoretically, it range is $0 < \boldsymbol{c} \leq 1.0$, but in practice, it is typically 0 <c≤ 0.3 for a four-option multiple choice item (Ojerinde et al; 2012). When guessing is likely to be a factor that will affect examinee's responses in a test, the 3PL model would be an appropriate solution to estimate an examinee's ability (Amarnani, 2009).

The shortcomings of 3PL model are that some measured quality of the logistic function were missing (the additive property) because of the contribution the involvement of guessing to the likelihood of correct response and the definition of the difficulty parameter which was altered. Under the previous two models, ***b*** was the point on the ability scale at which the probability of correct response was 0.5. But now, the lower limit of the ICC is the value of **c** rather than 0.The result is that the item difficulty parameter is the point on the ability scale with 1+c/2. The discrimination parameter *a* is still interpreted as being proportional to the slope of the ICC at the point θ =*b*,but in 3PL model, it is actually 1-c/4. The deviations in the explanations of difficulty (*b*) and discrimination (*a*) appear trivial but stand significant when inferring the result of test analyses.

In spite of the added advantage 3PL model came up with by taking care of the guessing error, it seems to be at an advantaged to low-ability examinees who could guess correctly a hard item anda disadvantaged to brilliant high-ability students who could have responded correctly to an easy item but slightly misses the item as a result of some other factors. At such instance, there is a likelihood of bias in estimating its parameters. Item response function (IRF) of a 3PL model is mathematically defined as:

$$\mathbf{Pr}(Xis = 1|\theta s, a_i, b_i\,, c_i) = c_i + (1 - c_i)\frac{1}{1+e^{a_i(\theta s - b_i)}}\text{..........eqn 2.23}$$

b_i = item difficulty or threshold or location, a_i = item discrimination or slope (higher values means more discriminating and better item), c_i= item lower asymptote or guessing parameter.

Unfortunately, the addition of the guessing parameter increases the likelihood that the trace lines will intersect and thus increases the non-additivity of the item function.

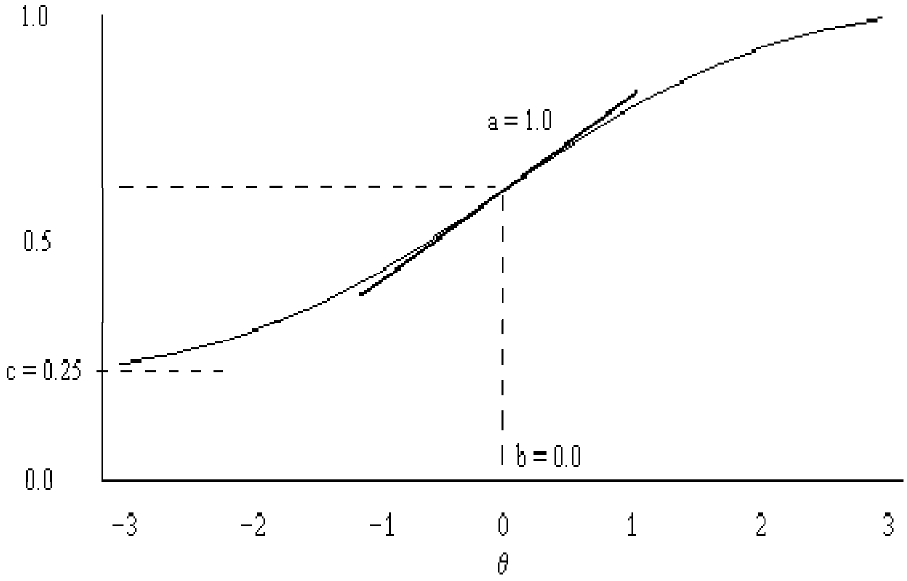


Figure 2.2: Example of 3PL IRF, with dotted lines overlaid to demonstrate parameters

***The Four-Parameter Logistic* (4PL) *Model*:** This is the target model for this study and it was first empirically investigated by Barton and Lord (1981), who wanted to know whether including an upper asymptote that is under 1 will improve ability estimation in standardized tests. The researchers felt that the 3PL model might be unduly punitive to highly able examinees who easily responded to an easy item wrongly (Mislevy and Bock, 1982).

In other words, if the guessing parameter had been taken care of, in the 3PL model for a low ability respondent who correctly guesses a difficult item, there should be provision for a high ability respondent who as a result of carelessness or mistake, incorrectly answers an easy item. This was the purpose for which 4PL IRT model was established.

In view of the underlying characteristics of the traditional IRT model, an examinee's ability would be considerably underestimated if he/she missed early items due to carelessness (Rulison and Loken, 2009). To cope with the underestimation problem, Barton and Lord (1981) introduced the fourth parameter (upper asymptote/carelessness/mistake) parameter into the 3PL model allowing a high-ability student who misses an easy item to have his/her ability not to be drastically lowered. The item response function (IRF) of Baton and Lord's 4PL model was given as:

$\boldsymbol{Pi(\theta s) = \Pr\left(Xis = 1\middle|\theta s, a_i, b_i, c_{i,}d\right) = c_i + (d - c_i)\frac{1}{1+e^{a_i(\theta s - b_i)}}}$**.....eqn2.24**where ***d***,the additional parameter, upper asymptote, was permanent at 1 (yielding the 3PM) and the other three parameters (*a*, *b*, and *c*)were fixed at values earlier estimated utilising the 3PM.

Barton and Lord re-evaluated a huge number of test scores for students taking three American scholarly tests: the Scholastic Aptitude Test, the Graduate Record Examination, and the Advanced Placement Examination. It was presumed that the deviations in ability estimates with ***d***set below 1 were too little to be considered as significant, along the lines proposed by the model of equation 2.24. The model observed by Barton and Lord (1981) utilised a universal *d*to denote a determinate probability of carelessness through all items by all respondents.

For Barton and Lord, their demonstrating approach was not the broadest implementation of the 4PL model as they did not evaluate the fourth parameter but rather fitted models with fixed qualities for ***d*** (Tavares, de Andrade and Pereira, 2004;

Osgood, McMorris and Potenza, 2002; Reise and Waller, 2003).Then, the need for a 4PL model became more evident with the *d*parameter carrying a subscript ***i*** (Waller and Reise, 2009; Loken and Rulison, 2010), to nullify certain discrepancy as to the particular form the model ought to take. It was said that the ***d*** parameter should depict a characteristics of the item and never a propensity of the students.

Therefore, 4PL model yet added another parameter d_i to reflect the tendency to never respond to an item for a high ability respondent. The more general formulation of the 4PL model according to Waller and Reise (2009), Tavares *et al.* (2004), Linacre (2004) and Rupp (2003) suggests$d_i$to be an item-specific upper asymptote that should be less than 1. An upper asymptote, denoted by $\boldsymbol{d_i}$is used where **1-** $\boldsymbol{c_i}$in the 3PL is replaced by $\boldsymbol{d_i}$ **-** $\boldsymbol{c_i}$ .The model is then given as;

$$\boldsymbol{Pi(\theta s) = \Pr\left(Xis = 1 \middle| \theta s, a_i, b_i, c_{i,} di\right) = c_i + (di - c_i)\frac{1}{1+e^{a_i(\theta s - b_i)}}} \text{....eqn2.25}$$

In adding to the standard 3PL model, the upper asymptote permits that profoundly capable respondents may anyway answer a question inaccurately because of stress, tiredness, inattention or carelessness. In its broad structure, 4PL model permits an alternate upper asymptote for each question. This was contrary to the initial recommendation of Barton and Lord (1981) that a typical upper asymptote was needed for all items (d_i= *d*,for all *i*).

However, 4PL model’smain assetis its ability to make a highly-ablestudent have an answering probability of non-zerotoeasy question he incorrectly responded to. This is because such examineeresponding abberantly might result in error of estimation.Rulison and Loken (2009) in a computerized adaptive testing explored this asset and revealed that early mistakes effect onbrilliant respondents couldefficientlybe lessened by applying 4PL model.

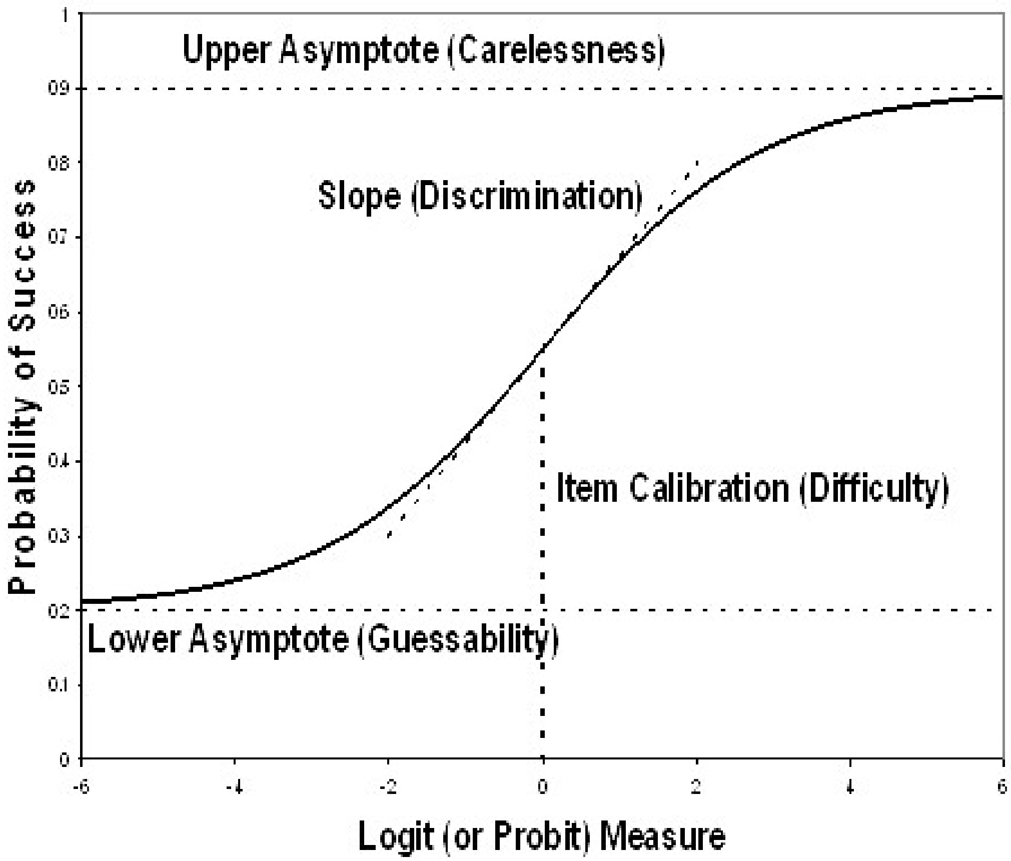


Figure 2.3: ICC showing each of the item parameter in a 4PL Model(Ojerinde, 2013)

2.1.7 The Basis and Justification for the usage of 4PL IRT Model

The need to emphasise the exploration of the 4PL model, which is the basis for this research work, is because of the relatively little attention the model has received since it was proposed by Barton and Lord (Loken and Rulison, 2010). Its usage in calibrating model's parameters is rare, unlike the other more familiar 1-, 2- and 3-parameter logistic models that are commonly used to fit a data set. Linacre (2004), Rupp (2003), Hambleton and Swaminathan (1985) and Barton and Lord (1981) observed that 4PL model was rarely mentioned in literature because it was rarely used in practice as it was believed to have only a little benefit and it is challenging to estimate.

Some of the reasons the 4PL model earlier suggested by Barton and Lord (1981) could not generate enough awareness and was not widely utilised until recently are;

a) There was no strong agreement on the usefulness of the model (Hambleton and Swaminathan, 1985; Barton and Lord, 1981).

b) The model was usually shown to be problematic in evaluating with Maximum-likelihood estimation (MLE) approach (Waller and Reise, 2009),this made it fitting to be considered difficult. This created a sort of worry that its estimates (upper asymptote (d_j))might not be consistent given the difficulties repeatedly met when estimating the lower asymptote (c_j)with maximum likelihood approach (Baker and Kim, 2004; Embretson and Reise, 2000).

c) 3PL modelusage prevalence and disaggrement on 4PL model's effectiveness are strong arguments against its usage (Loken and Rulison, 2010).

However, researchers like Osgood, McMorris and Potenza (2002), Reise and Waller (2003), Tavares, de Andrade and Pereira (2004), Rouse, Finger and Butcher (1999) and Waller and Reise (2009) suggest the need for a 4PL model in their studies and that the 4th parameter should capture the item feature and not the characteristic of the examinee (i.e making the upper asymptote to be item-specific). The conceptual drawbacks 4PL model experienced and the suggestions of various researchers madethe model to be reconsidered.This brought about the nullification of the use of the initial model in equation 2.24that is,

$\boldsymbol{P_i(\theta_s) = \mathrm{Pr}\left(X_{is} = 1 \middle| \theta_s, a_i, b_i, c_{i,} d\right) = c_i + (d - c_i)\frac{1}{1+e^{a_i(\theta s - b_i)}}}$, and made equation 2.25,

$\boldsymbol{P_i(\theta_s) = \mathrm{Pr}\left(X_{is} = 1 \middle| \theta_s, a_i, b_i, c_{i,} di\right) = c_i + (d_i - c_i)\frac{1}{1+e^{a_i(\theta s - b_i)}}}$ageneral formulation for the model such that the upper asymptote was made to be less than one.It was no longer a fixed value in a way that the *d* parameter was meant to be item specific as against capturing the tendency of the examinees. Although 4PL model was an extension of the usual 1-, 2-, and 3PL models, better results can be established when parameters of the 4PL model are appropriately calibrated. Magis (2013) also emphasised that allowing the upper asymptotes to be lesser than 1, as well as making brilliant students to fail some initial items with more likelihood, made the underestimation problem of the 4PL model to be lessened at the first step. This also made the respondent to improve from initial mistakes hastily.

After reconsidering 4PL model, its recent popular usage became possible because of its mathematical power and resources that were enhanced and thedifferent precise software that were formulated. The studies of Linacre (2004) and Rupp (2003)made use of the 4PL model, where the upper asymptote parameter was correctly estimated. Loken and Rulison (2010) developeda Bayesian framework with the help of a Markov chain Monte Carlo (MCMC) method and the WinBUGs software (Lunn, Thomas, Best and Spiegelhalter, 2000). However,WinBUGs software was not initially available when 4PL model was first proposed by Barton and Lord (1981), its development created a landmarkinnovation for a widerattentionof 4PL usage for practical purposes.

Magis (2013) statesthat the core benefit of 4PL model is its allowing a non-zero likelihood of responding to item incorrectly for brilliant examinees. Rulison and Loken (2009) exploited this asset in their study in a setting where computerized adaptive testing (CAT) is used. It was seen that the effect of early mistakes made by brilliant students could be highly reduced by the application of a 4PL model. Later studies on the utility of 4PL model were seen through the works of Green (2011), Liao, Ho, Yen, and Cheng (2012) and that of Yen, Ho, Liao, Chen, and Kuo (2012). Magis (2013) introduced this model as a baseline model for the generation of CAT in the R package catR. In the study of Tendeiro and Meijer (2012), 4PL model was suggested to become a model that will be used toidentify person fit and detect any form of inattention patterns in the nearest future. Therefore, exploring 4PL IRT dichotomous model to encourage its usage was necessary in this study (Magis, 2013).

2.2 Conceptual Background

The process of validly and reliably assessing what students have learnt in order to be sure of their understanding of the acquired knowledge cannot be taken with levity as far as practitioners in the field of education, especially psychometricians, are concerned. This is because any result presented after systematic measurement and assessment procedures are done, will be regarded as the true construct (ability, skill, intelligence) possessed by the subject(s). This result will in turn aid adequate and valid decisions about the respondent or group of respondents. Therefore, the concepts of measurement, assessment, assessment instrument (specifically achievement test), students' response patterns, response time, computer-based testing, mathematics education and students' performances are fundamental to establishing the basis in this research study.

2.2.1 Mathematics and its Importance to National Development

The essence of mathematics education in schools is to make a child think mathematically.This involves bringing a child to the clarity of thought and being able to pursue assumptions to logical conclusions (NCERT, 2006).Several methods of thinking are evident and the type that is applicable to the learning of mathematics is the capability to handle intellections, use abstractions in posing problems, perceive relationships and structures,and solve meaningful problems.

Wheeler (1983) expresses that it is better to be familiar the usage of mathematical method than knowing so much of mathematics. Knowing how to mathematise is of a 'higher aim' that helps in increasing the inward knowledge of a budding child. This can be done by molding the right attitude to problem solving and approach difficulties systematically. However, knowing a lot of mathematics that is considered the 'good and narrow aim' is by making employable grown-ups who will be able to contribute to the social and economic growth of the nation.

Mathematics, as observed by Abiodun (1997), is a main tool for expressing theories in the sciences and other fields. This subject is used to clarify observations from experiments in other fields of inquiry and, as such, the possession of robust mathematical knowledge remains the gateway to virtually all occupations. The necessity for mathematics is said to be as old as mankind as man's desire to count and keep record of things around him keeps increasing. This is why Ambali (2014) asserts that mathematics as a key subject of enquiry contains some ciphering skills that are relevant to an extensive range of logical, hi-tech, scientific, safety, governmental and financial aspects of life.

Mathematical sciences have helped improve the ability to predict weather and measure the effects of environmental hazards. Ciphering skills that are needed to prepare a solid foundation for other educational and professional challenges in life. It is well known that the level of social and economic development of any country is intimately connected with the level of development of that country in science and technology.The National Council on Education (NCE, 2013) came up with far-reaching decisions which included the approval of a 12-point policy framework for educational development in Nigeria. One of such decisions was that federal and state ministry of education (FME), including the Federal Capital Territory, must offer scholarship awards to students studying Physics, Chemistry and Mathematics education in tertiary institutions. The council emphasised the need to brimg the National Mathematics and Science Education National Training Centre to international standard.

Development at both social and economic levels of any country is assumed to be intimately connected to the level of growth of that country in science and technology. Since mathematics is known to be foundational to science and technology, any form of growth, experienced in the social and economic development is closely related to the level of development in the mathematical sciences (Kuku, 2012). Mathematics teaching objectives at primary and secondary school levels make it such a significant and rudimentary subject for attaining success in further academic pursuit and manpower development. Iji (2007) maintains that a nation that desires growth in the sciences, industries and technology must pay attention to mathematics. Knowledge of mathematics is generally believed to be important in understanding other disciplines in education.

Development is widely conceived as a participatory process of social change intended to bring both social and material advancement for the majority of people through their gaining control over the environment (Ambali in Azuka and Kurume, 2015). Some of the elements of development include high standard of living, high agricultural productivity, high technological productivity, adequate exploration and exploitation of the natural and mineral resources of the society, less dependence on imported materials, presence of heavy industries, high literacy and numeracy rate of the citizens, appropriate health care delivery and low unemployment. For these elements to be evident in any nation,mathematics has a central role to play.

No wonderhigher institutions of learning in Nigeria insist on, at least a pass or credit pass in mathematics for admission into any course of study. The Joint Admission and Matriculation Board Brochure (2016/2017) revealed that out of 744 courses available across Nigerian universities, 698 courses, representing 93.8%, require at least a pass or credit pass in mathematics (Ariyo, 2017). This indicates how important mathematics is to national building.

Niss (1996) points out that the fundamental reasons for teaching mathematics in schools include: contributing to the technological, socio-economic, political and culturaldevelopment of the society byempowering individuals to be able to cope with the various spheres of life. This is why Ambali (2014) asserts that mathematical skills are relevant to a wide range of analytical, technological, scientific, security, political and economic applications and that a solid foundation in mathematics prepares onefor other educational and professional challenges.

However, for a country like Nigeria, students' performance rate in mathematics requires an apt attention because of the current pattern of scores in both internal and external assessments. Table 2.1 shows the trend of students' performance in mathematics in West African Senior School Certificate Examinations Council (WASSCE) between 2006 and 2017.

Table 2.1: Statistics of Students' Mathematics Performance in WASSCE between 2006 – 2017

Year	Total No. of Candidates	A_1-C_6	D_7-E_8	F_9
	Who Sat and %	No. and %	No. and %	No. and %
2006	1149277 (98.18)	472674 (41.12)	357325 (31.09)	286826 (24.95)
2007	1249028 (98.33)	584024 (46.75)	333844 (26.72)	30277 (24.24)
2008	1268213 (98.09)	726398 (57.27)	302266 (23.83)	218618 (17.73)
2009	1348528 (98.22)	634382 (47.04)	344635 (25.56)	315738 (23.41)
2010	1306535 (98.13)	548065 (41.95)	363920 (27.85)	355382 (27.20)
2011	1508965 (97.98)	608866 (40.35)	474664 (31.46)	474664 (31.46)
2012	1658357 (97.97)	838879 (50.58)	478519 (28.86)	298742 (18.01)
2013	1656527 (98.19)	897655 (54.18)	462176 (27.90)	245263 (14.80)
2014	1632377 (98.59)	1011608 (61.97)	357555 (21.90)	211941 (12.98)
2015	1581420 (98.69)	901845 (57.02)	425628 (26.91)	219759 (13.89)
2016	1469585 (99.02)	1032175 (70.23)	248676 (19.37)	112328 (7.64)
2017	1550348 (99.05)	1276782 (82.25)	160623 (10.36)	44874 (2.89)

West Africa Examinations Council, Yaba Lagos. (2018)

Table 2.1 shows that from 2006 through 2008 there was an increase in performance rate for higher passes (A_1- C_6) while the lower passes (D_7-E_8) retrogressed appreciably. A decline in performance rate of the higher passes was observed in 2009 through 2011 which in-turn made the performance in the weaker passes to increase. From 2012 through 2015, an upsurge in performance was recorded, which was terminated with a downward trend as indicated in the previous regular patterns. In 2015, the usual 3-year interval of downward trend resurfaced but was surprisingly terminated by an increment of 13.21% in 2016 and 12.02% in 2017.

The increase was indeed a deviation from the norm of irregular pattern in students' performances. If such is the case, it could then be assumed that more research is needed to see to the stochastic patterns and a way out to consistently retain or boost performance of students until a desired achivement level is attained. Therefore continual exploration of research in mathematics as a core subject to increase students' performance and the nation's advancement in science and technology cannot be treated with levity.

2.2.2 The concept of Measurement and Assessment/Testing in Education

At the sound of the word 'measurement', what comes to the mind is the application of one kind of instrument or the other, to gauge the quantity of some possessed attributes that is to beascertained. In educational parlance, Nenty(2004) defines it as the quantity or quality of "something" possessed by the body that is being measured, not the body itself. Measurement has a long fragmented history as a result of irregular documentation process, storage and retrieval of measurement information (Okpala, Onocha and Oyedeji, 1993). An assertion that is credited to Thorndike in Nenty (1998) is that the presence of any trait in a person is in a certain amount and for such to be assessed; a thorough measurement that will give its value and quantity is desired. Measurement should be done with some precision in a way that the knowledge being measured will be conveniently recorded and used.

Then, measurement in education has to do with the use of an instrument such as test or questionnaire to obtain how much students have achieved or understood a learning content to which they have been exposed, or a learner's disposition towards a subject. It is the systematic process of collecting information on entities, objects, students' learning outcome or events, and assigning numerical values to information generated

(Odinko, 2014). It is also a methodical allotment of values to a trait or an attribute in order to determine a person's characteristics with the help some of assessment scales. Osuji, Okonkwo and Nnachi (2006) define it as a logical way of attaining some measured quantity by which an attribute exist in an entity. The main aim of measuring in the educational or psychological sector is to know by what quantity the underlying trait in a particular examinee is present.

Most human traits that are to be measured are concealed and as such it can not be assessed with some certain physical measuring devices applicable to the ones used in the sciences. Underlying constructs like reading, mathematical, scholastic and arithmetic abilities are expressed as behaviours that are observable. To assess these attributes, they have to be provoked in the subjects so as to capture the extent to to which they are present in items that are assoiated with the contents taught (Nenty, 2004). The items must be such that would provoke the observable behaviour under consideration. Therefore, items enable the observer to document the intensity of a provoked latent trait through measurement.

A teacher who wants to know how much his stated objectives have been attained by the students is required to give them a test on the basis of what has been taught. The performance of students in the test is measured by the scores each learner obtains. According to Odinko (2014), the scores are numbers that are quantitative indices of learner's achievement levels in a given task;they do not have values in themselves but merely assist test givers to know how the students have performed. This impies that measurement does not perform the function of passing value judgment on the performance of learners in a given task, but provides only quantitative information on an event, entity, object or individual. There are four basic issues to be considered as far as measurement of latent variables for scaling observations.

a) The first issue involves the consistency of the measures. If after repeated measurements of a latent variable, there is constancy in the dimension of what is measured, the measurement is considered to be highly consistent or reliable which will give a small amount of measurement error and a greater confidence will be imposed on the measurement made. However, if repeated measurements varied wildly from one another, they are said to have low consistency or reliability that could deter the confidence in the measurement. This will, in turn, give a larger amount of error.

b) The validity of the measures is another issue in measurement. It is the extent at which the assessment made is really the manifestation of the trait. However, for measurement to be valid, a high degree of reliability must be attained. It is therefore necessary to be concerned not only with the consistency of measurement, but also with their validity which are of various types. Obtaining validity evidence is part of the measurement process.

c) The scale used in measuring must be independent of the object by which it is being measured. This implies that the instrument must possess the property of invariance. Without invariance, comparison across different objects of measurement would have limited utility.

d) The final issue is the category with which the measurement is scaled. This determines whether what is being measured will follow a ratio, an interval, ordinal or nominal scale such that the assignment of values to characteristics measured is appropriately done for valid interpretation of the different types of information the observation carries.

In education, the use of four distint scales of measurement is evident. These are nominal, ordinal, interval and ratio scales;

Nominal Scale: The simplest and the least of the four scales is the nominal.It entails the allotment of object to groups only and not inferring the extent at which it is allocated. It is used for mere classification, identification or labelling. The interest is in ascertaining if specific objects have their place in one or different classes. Scales that fallin this category do not have some mathematical and statistical tasks such as addition, subtraction, multiplication and division. For example, teachers in the school system can be assigned to two groups. Graduate teachers =1and non-graduate teachers = 0.

Ordinal Scale: In this case, there are magnitude and classification. A real digit may be more, equal or less than another real number. There is an indication of size and as well as ranking. Only few arithmetical operations are attainable at this scale such as classification, counting and ranking to show greater than or less than. Examples of data that could be collected on this scale are: ranking people according to height (tall/taller/tallest), and academic performance (distinction, upper or lower credit and pass). Also, if the grades of four examinees are ordered as 5th, 6th, 7th and 8th positions and their actual scores are 70%, 68%, 50%, and 49% respectively. It is noted that the variance that occurred in-between 5th and 6th positions gives 2%, 18% is observed between 6th and 7th while 7th and 8th positions gives 1%. Unequal intervals were observed on the scale,but equal quantities cannot be depicted.

Interval Scale: This scale is more elaborate than the ordinal and nominal scale. Equal space depicts similar magnitudes. Amount of variations among nearby intervals in the scale is same. For example, difference between 67cm and 69cm is the same as 52cm and 55cm. Tasks like counting, ranking, classification, addition and subtraction are permitted in interval scale and no absolute zero can be seen.

Ratio Scale: Maximum level of measurement occurs here where all the features of other scales exist and an additional feature of absolute zero is present. All arithmetic procedures in terms of addition, subtraction, multiplication and division are contained. An equal interval as well as zero mark is obtainable. Ratio scale is hardly used in the social sciences because if a test is designed to assess student's intelligence, an examinee with a zero score in a test may not necessarily be devoid of any knowledge of the subject matter.Ratio scale is suitably utilized in the physical sciences, whereas interval scale is more suited for education, social science and psychology.

Meanwhile, the continuous search for objective measurement is the basis that informs this study. Since the traitsto be measured are inherent in the body where measurement is to take place, careful measures should be taken. For such concealed trait to be measured objectively, the construction of a precise, reliable and valid assessment instrument that will elicit such trait is needed. Nenty (2004) declares that the development of such metered line is technically a tasking job that calls for a lot of know-how. This is because the measurement of behavioural characteristics is error prone as a result of the indirect nature of measurement.

Measurement and assessment in education are two sides of a coin which cannot be separated. The term assessment connotes a process of organising measurement data into interpretable form to aid decision making. Assessment according to Odinko (2014) is the process of observing, recording as well as documenting what students do and how they do it as a basis for variety of educational decisions that affect learners. Huba and Freed (2000) see assessment as a procedure for assembling and conferring data from diverse sources to get a profound knowledge of what examinees have learnt and can attain with their understanding as regard the educational experiences they have. Though assessment is perceived as a way of knowing the growth of each examinee, it also makes persons, organizations and countries to trail schools quality and educational systems (Braun, Kanjee, Bettinger and Kremer, 2006).Society banks on assessment scores to adjudge the worth of the educational system while policy makers and educational stakeholders use them to decide whether schools are meeting up with the purposes for which they are fashioned.

2.2.3 Test Development in Test Theory and Achievement Test

Measurement began with testing to assess specific variances in adults' skill acquisition. Maskelyne, in 1796, was noted to have sacked his assistant, Kinnebrook, because of the difference he recorded from his own earlier reading on how stars moved in the telescope. Then, between 1820 and 1823, the work of Maskelyne was later enhanced by Tuckman in 1975 where inconsistency in the equations and observations was corrected. It was resorted that changes existed from happening to happening and from person to person which is an implication thata difference could be observed in the time required to respond to a basic boost. Meanwhile, around 2200BC, an informal test adopted by the Chinese to employ individuals into the civil service happened to be the first written test.

Alfred Binet in 1904 examined the contrasts among highly-able and low ability youngsters by building up an assessment test called Binet-Simons understanding test for estimating the intelligence of some kids. Then, Louis Terman and his partners in 1916 at University of Stanford re-examined the Binet-Simon instrument and drew out the Stamford-Binet variant. When the need arise to know the intelligence of officers for appointment into various undertakings and responsibilities during the World War I, a group-test began. Other test developers are David Wechsler who worked on how to arrange individual intelligence instruments from 1939 to 1967, George Fisher, who constructed the first ever achievement test that was a standardized multiple choice test in 1864 and an American, J.M. Rice, who constructed the standard spelling target scale in 1897.

Locally, the first organisation that was established to develop tests in the nation is the West African Examinations Council in 1952 that serves as an examining body for West African nations. This bodyhandles assessments like the Senior School Certificate Examinations, the City and Guilds Examinations and the Royal Society of Arts (RSA). In 1976, the Joint Admission and Matriculation Board (JAMB) was set up with the obligation of carrying out tests for universities, colleges of education and polytechnics.

Meanwhile, a body (National Business and Technical Education Board, NABTEB) was given the duty of arranging tests and certificates for business, technical or vocational education.

Also, another body known as the National Examinations Council (NECO) wasestablished to arrange tests for senior and junior secondary schools as well as entrance tests into unity schools. Likewise, the International Centre for Educational Evaluation (ICEE) is saddled with various assessments and research works in Nigeria's educational system. Aside from bodies that are saddled with national assignments, some states and local education boards have their test units in different departmentswho could create combined assessments efforts for both high and elementary school students/pupils.

However, the two test theories (IRT and CTT) differ in the way items that constitute a test are viewed and analysed. They differ in (i) item analysis (ii) selection of test items, and (iii) reliability assessment. Item analysis usually involves the characterization of test items and the use of statistical information for revising and/or deleting test items (Ojerinde, Popoola, Ojo and Onyeneho, 2012). In CTT, various statistics like p-value, mean, standard deviation, difficulty and discrimination indices, estimates of reliability and validity and T-scores that are utilized are termed sample dependent.

For instance, if the p-value for a group of SS1 students in a mathematics test is 0.25, it means 25% of the students responded to the specific item rightly while 75% missed it. Yet, a p-value of 0.50 for SSII students could be obtained for that same item. Therefore, p-value obtained for an item cannot be taken as a function of the item but that which is used with a particular sample. This is a major shortcoming of the traditional test theory methods as an examinee's score is dependent on the set of items used for analysis.

An assumption of CTT is that items that are termed good differentiate across the wide abilities interval. Larger percentage of examinees having high aggregate scores should be able to respond accurately to the item when likened to examinees with lesser scores. This is as a result of the fact that p-values and point-biserial correlations are taken as average statistics. Test givers cannot say from these estimates alone if an item acts in a way or contrarily.

On the other hand, IRT offers a different way to analyse tests. The attributes of each items of the test are the core of the theory. Students' ability as it relates to the item level performance is what IRT models, instead of the total test performance level. This is due to computing the expected students score from their responses to the individual item. IRT estimated score is sensitive to differences among individual response patterns and thus gives a better estimate of the individual's true level on the ability continuum than CTT assumed scale score (Santor, Ramsay and Zuroff, 1994). As an alternative to assuming that all questions subsidise in the same way to understanding of a person's ability, IRT instead, offers a fine distinction as regards the information that individual items supply about an examinee (Ojerinde, *et al.* 2013). A basic use of IRT is in the assessment and improvement of simple psychometric properties of items and tests. With the information of item properties, assessment designers could pick items that reveal a certain interval of ability levels which possess a strong degree of discriminative ability.

Meanwhile, the purpose for which every test is developed will determine what constitute the assessment instrument. Osuji, Okonkwo and Nnachi (2006) see testingas a human action that comprises of a sort of exchange, which seeksto assess what the students have learnt. While Ojerinde (2013) views it as the core thing in education and that test scores that emanate from it are used to diagnose examinees' academic performances. This is why different assessment bodies satisfy different needs and purposes. There are various types of tests that may be used for gathering data on students. These assessment instruments according to Okpala, Onocha and Oyedeji (1993) include oral test, performance test, achievement test, behavioural assessment test, observational schedule, interview schedule, rating scale questionnaire, sociometric test and logs, dairies and reports. For the purpose of this research, achievement test (Computer-Based Mathematics Achievement Test) which is the main instrument for data collection is discussed.

***Achievement Test*:** An achievement test is a form of cognitive assessment that relates what examinees have learnt to do. It is a test that is used to assess the extent of accomplishment achieved in a definite aspect of learning. Adeleke (2009) specifies that achievement test meaningfully assesses the status of person at all cognitive domains of understanding as suggested by the taxonomy of Bloom's educational objectives. This enables the examiner to measure the complete domain of cognition depending on what has been taught by the instructor or teacher as laid down in the instructional objectives of the curriculum that is used for teaching of instruction. An achievement test can either be teacher-made or standardized essay or objective tests.The objective test could be either supply or select form (Adewale, 2018).

***Essay Test*:** In this type of test, students or examinees possess the liberty of expressing or stating responses to the questions posed in their own words. This is used by teachers to measure achievement or performance from classroom instruction. Essay tests come in two kinds. The free-response (extended) and fixed response kinds. In the extended-response format, items are framed such that responses will demand that the examinee is not restricted in deliberating on the item posed (Osuji, Okonkwo and Nnachi, 2006). In the restricted-response type, examinees are constrained to the nature or organisation of the responses to be given. Directional and desired replies are expected which restrains the respondent's liberty to pick, remember and amalgamate all he knows and supply them as reasonably as he wishes. Essay questions are most suitable in assessing the cognition at the lower level such as remembering, understanding and applying.

***Objective Test*:**Items are framed such that only one right option is made available among the given alternatives. It requires that students identify and choose the most right option. Two types also exist in objective test. The free and fixed- response types. The former which is also known as the supply format is made up of short response and completion items. Fixed response form (selection format) is divided into alternative response, matching and multiple-choice items. Meanwhile, the multiple choice type of the objective formwas the one the researchermade use of in this study. Diagrammatic representation of a teacher-made type of achievement test is presented Figure 2.7.

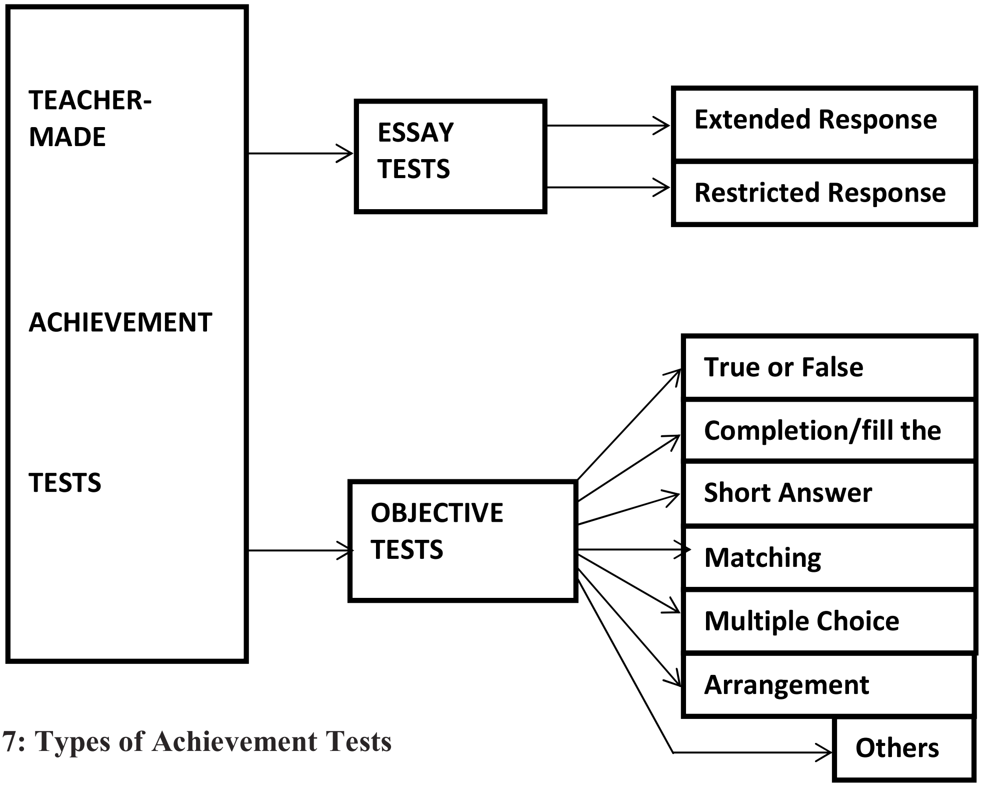


Figure 2.7: Types of Achievement Tests

2.2.4 Ascertaining the Psychometric Properties of Assessment Items and Instrumentthrough IRT Approach

Measuring student's learning outcomes is a central issue in education that should not be treated with levity. This is so because the result obtained from such assessments are used by educators to assess students on how much they have learned and how the information is used to provide feedback for improvement and remediation (Ojerinde, *et.al*, 2015). The most important objective of measurement is to design and select test items/instruments with minimum errors so as to obtain usable and dependable data for adjudging aright the essenceby which assessment process is undergone.

Constructing valid, consistent and operational instruments requires proper arrangement and taking necessary steps using either CTT or IRT framework. This involves having a system which couldgovern test developers in the process of item construction. This become important since assessment instrument is key in learning outcome measurement. The acceptability, dependability and ease of use of such tests rely upon the consideration with which the tests are planned and arranged. Okpala, Onocha and Oyedeji (1993) identify four stages that are involved in item/test construction. These are the planning stage, item development stage, item analysis stage and marking-scheme development stage. The planning stage guarantees that the test covers the specified instructional objectives, topics and planning of the test blueprint during teaching/learning process while giving consideration to the purpose for which that test was developed . Also, decision on item arrangement, the number of items and the time it will take respondents to answer, items assembling, scoring pattern, test administration and analysis of test results should as well be prepared for beforehand.

Item writing stage is so crucial such that poor attempts at this stage can mar the whole test construction exercise. Lindquist in Croaker and Algina (2008) posit that the decision of item-writer must be concrete on the trait to measure and by what method to assess it while writing such items. The thought of how the intended construct is assessed comes to play in item writing and could involve the following activities: i) keeping the test blue-print in mind ii) drafting the test items in unambiguous language iii) preparing extra test items, iv) reviewing items and the choice of items as written in the table of specification and v) having the items examined and criticised by other experts (Osuji, Okonkwo and Nnachi, 2006).

The next stage which is item analysis, also known as the statistical analysis or preliminary items try-out stage is a major concern in test development. It mainly involves ascertaining the psychometric properties of test-items and the whole instrument for a full-fledge field assessment. Ojerinde, Popoola, Ojo and Oyeneho (2012) state that the analysis of items involves characterization of such items with the use of statistical information for revising and/or deleting test items. Before a test developer prints the assessment instrument in its final form for a field test, it is necessary to try out such questions on a representative sample of examinees with similar characteristics to the intended population and record both the responses and the scores. The test developer is expected to use the opportunity to observe examinees' reactions during testing, noting different exhibited behaviours that could indicate confusion about particular items.

After the testing session, a debriefing should take place during which examinees are invited to comment on each item and offer suggestions for possible improvements. An analysis of students' response to the items on a test provides diagnostic information for determining item quality. Item analysis helps in determining those questions with essential characteristics of what it takes to be part of an assessment scale. Osuji, *et al* (2006) states that analysingitems assistin identifyingfunctioning items as intended and those that are not, for retention or deletion purpose.

However, analytical procedures for effective item analysis differ in IRT and CTT. The difference occurs during the stages of i) item analysis ii) selection of test items, and iii) reliability assessment. Croaker and Algina (2008) are of the view that analysing test items and response procedures in classical psychometrics involves determination of mean, standard deviation, difficulty and discriminating indices. Correlational indices of item discrimination in estimating the degree of relationship between item and criterion scores (such as Pearson Product Moment, Point-Biserial, Tetrachoric, Biserial and Phi Correlation Coefficients), effectiveness of distracters and the establishment of test reliability and validity indices and standard error of measurement values are paramount.

Under the CTT approach, assessment questions and their totals are viewed as reliant on the attributes of the examinees answering them. Ojerinde *et. al* (2013) are of the opinion that simple tests are tests in which the vast majority respond to most items effectively and progressively while hard tests are those in which hardly any individual responds to most questions accurately. Various statistics used to portray tests are examinee-dependent based on the achievement of a representative sample of test takers. The theory likewise expect that reliable items differentiate well through a wide interval of abilities to such an extent that more percentage of respondents with high aggregate scores should respond to the question accurately when contrasted with examinees with lower total test scores.

IRT offers a different approach to item analysis item selection and reliability assessment in test development. One of the assumptions in IRT is the fulfilment of monotonically increasing function known as item characteristics curve (ICC). ICC and the item/test information function (IIF/TIF) are major tools used in analysing psychometric properties of both item and assessment instrument in the framework. IRT tries to model students' proficiency with performance at the item level as against performance at the total test level in the CTT approach. Because of the fact that expected examinee score is assessed from the responses to individual items, the estimated examinee's trait level is independent of the item being used. IRT makes available a fine distinction with regard to the information individual items give about an examinee instead of assuming that all items contribute alike to the understanding of a person's abilities.

ICC, a major foundation of the theory (Baker, 2001) has an s-shaped bend that portrays the association concerning the probability of right answers to a question and the ability scale. This curve shows how individual examinees is answer to a test item with the possession of some amount of underlying ability that provides clues to the probabilities that individuals of certain ability level would respond to an item correctly. The curves flow from the left to the right on the horizontal axis to show their difficulty levels. At each ability level, a definite probability *$P(\theta)$,* showing that a respondent possessing that trait will supply a right response to the item is given. Examinees that have low ability will showcase a small amount of this probability unlike those with high ability.

Two technical properties are used in ICC in portraying the features of items in order to know where and how they will function across the ability continuum. These are difficulty and discrimination of item, which seem quite different in meaning to the difficulty and discrimination indices under the classical psychometrics. They are utilized in assessing items and maximizing the quality of the test generally. In the modern-day analysis, the combination of item characteristics as a reflection of the characteristics of the whole test is done. Item difficulty also known as the location index defines wherever such items will function along the ability continuum such that if an item seems to function appropriately as expected, it becomes a yardstick to know if it should be retained or deleted in the whole instrument.

The difficulty of an item implies that any hard item will demand a fairly high ability to respond correctly while a lower level ability will be needed to respond to an easy item correctly. However, a person that possesses an average trait level on an item ($\theta = 0$) is said to have a $50^{0}/_{0}$ likelihood of being able to correctly answer that item; such an item has a difficulty of 0. However, a person with a high ability level ($\theta > 0$) tends to possess a greater probability of responding to the item aright while an examinee having a low ability level ($\theta < 0$) is expected to have a lesser likelihood of responding accurately.

Therefore higher level of difficulty requires higher ability level by the examinee to possess a 50/50 likelihood of replying the item appropriately. For a question with 1.5 difficulty level, a person with a 1.5 trait level will require a probability of 50% to answer the question aright. The same thing goes for a person with a low level of ability. Easy item therefore gets along well among respondents that are categorized as low-ability while a hard question functions in the midst of the high-ability examinees.

The subsequent technical asset of ICC, item discrimination shows just how well a question can distinguish among students with traits below the difficulty level and those with abilities above. Item discrimination in IRT is similar to the item–total correlation in CTT according to Embretson and Reise (2000). The value of item discrimination denotes how relevant an item is to the ability being assessed by the test. Positive valued item discrimination can be said to be consistent with the latent ability being assessed while a large value of discrimination shows a robust consistency between the item and the latent ability. However, an item with discriminating value of zero (0) does not have any correlation with the construct while negatively valued item discrimination is said to be indirectly associated with the latent construct. Therefore, it is largely required that items possess great and positive value in discrimination.

This feature of the ICC basically reveals how steep the curve is at the middle side. If the curve shows so much steepness, it is an indication that the item can really discriminate, which attests to how good the item is to the general test. If the curve is flattened out, the discriminating power of the item will be small since the probability of correct response at low ability levels will be nearly the same as it is at high ability levels. Combining these two features of the ICC, an appropriate item is described by the overall look of the ICC and its technical properties.

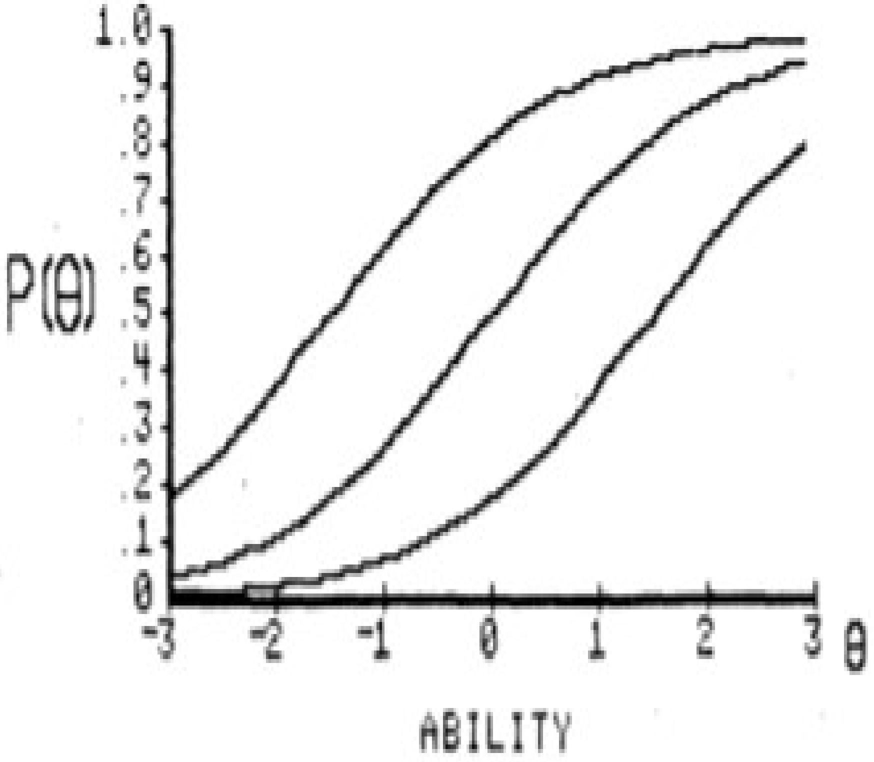


Figure2.5a: Three ICCs having similar discrimination but diverse levels of difficulty

In Figure 2.5a, the idea of item difficulty is presented with three item characteristic curves on a similar graph, having identical levels of discrimination but varying difficulty levels. ICC at the left hand depicts items that are easy since the chance of right answers is much for respondents with low-ability but gets to 1 for highly able persons. The one at the middle signifies an average difficulty item due to the likelihood of right answer that is low at the least ability levels, about 0.5 at the centre of the trait continuum and approaches 1 towards the highest levels of ability. Meanwhile, the right handed curve is showing an uneasy item. The likelihood of right reaction is low for the vast majority of the ability scale and increases just when the higher ability levels are attained. Even at the highest level of ability shown which is +3, the chance of right reaction is only 0.8 for the most difficult item.

Another tool earlier mentioned as one of the measures to ascertain the quality of either test items or instrument apart from using the gauge of item characteristics curve is item/test information functions. One of the traditional indices of the utility of a test from CTT is its standard error of measurement (SE_M). It is assumed that raw scores on tests and test items are a composite of the true score and random error. This SE_M is referred to as the distribution of random errors around the true score (Kline, 2005). Thus, the lower the SE_M value, the more dependable is the test score. A single value for the SE_M is given for the test as a whole. By contrast in IRT, the concept of test and item information is used. Information is said to be indirectly associated with, and it is calculated separately for different ability levels (Thorpe and Favia, 2012). The test information function indicates the appropriate way each ability level is being evaluated by the test (Thorpe, McMillan, Sigmon, Owings, Dawson and Bouman, 2007). This concept is extensively discussed in the next section.

2.2.5The Concept of Reliability in Item Response Theory

Another vital aspect in IRT approach is its item\test information function (IIF\TIF). Reliability in IRT is built upon the concept of IIF which tells where questions function most at a point of assessing the underlying trait. When constructing a test or evaluating items of a test in IRT framework, the best items supply the most valuable information about the examinee’s ability. IRT advances the concept of item and test information to replace test reliability in CTT (Ojerinde *et al*, 2014). Reliability (consistency) happens to be a required asset for any test and the more reliable a test is, the more precise the construct in that test can be measured. So whichever procedure imbibed, either IRT or CTT approach, as reliability increases, the SE_M comes down which makes the observed scores to be closer to the true scores.

In CTT approach, reliability is a one-number summary of test precision and there is corresponding single standard error of measurement that is used for any test score. Reliability is a condition that must be fulfilled to ascertain reproducibility of test scores when such test is re-given in another time to same population of respondents. Therefore, the degree to which individuals' deviation scores or z-scores remain relatively consistent over repeated administration of the same test or alternate test form is reliability (Croaker and Algina, 2008). CTT model links test score to true score as against item score to true score as it is applicable in IRT. Therefore, CTT dependence on test scores limits the utility of person and item statistics in test development. This is as a result of the parameter not being an inherent property of the item but its relativeness to the group on which the item is administered. Assumption about individual items are only made in CTT, they are not really in the measurement model.

Although CTT approach has its measures of item quality, these are only indirectly related to what the reliability of the test will be in terms of item/total correlation. This made the concept of reliability to be the characteristic of a sample and not of a test. Therefore, reliability that is supposed to be the consistency of an individual's score over replication in CTT approach is really not it in the real sense of it (Templin, 2011).

Thus, there are different ways whereby reliability in CTT is estimated. These include (1) computing an estimate of reliability based on the observed correlations or covariances of the items with each other (Cronbach's Alpha, Kuder-Richardson 20 and 21 are all measures of internal consistency coefficients), (2) correlating the results from two alternate forms of the same test or test-retest approach (Pearson Product Moment Correlation Coefficient, used to measure the degree of stability and equivalence), and (3) splitting the same test into two parts and looking at the correlation between the two parts (Spearman-Brown, Rulon and Guttman split-half coefficients are used in ascertaining the extent of equivalence). Few of the mathematical expressions of the methods are given below.

Spearman-Brown formula is given as:

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{xx'}} = \frac{\mathbf{2}\boldsymbol{\rho}_{\boldsymbol{AB}}}{\mathbf{1}+\boldsymbol{\rho}_{\boldsymbol{AB}}} \qquad \text{................}\textbf{...eqn2.25}$$

Where, $\rho_{xx'}$is the reliability projected for the full-length test and ρ_{AB} is the correlation between the half-tests.

Cronbach Alpha formula is:

$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right)$ ……………….**...eqn2.26**

where k is the number of items on the test, σ_i^2 is the variance of item *i* and σ_X^2 is the total test variance.

Rulon Coefficient is given as:

$\boldsymbol{\hat{\rho}_{xx'}} \ \boldsymbol{= 1 -} \frac{\hat{\sigma}_d^2}{\hat{\sigma}_X^2}$, given ***D*** **=** ***A-B***…………….**...eqn2.27**

Where *A* is the examinee’s score on the first half-test and *B* is the score on the second half. The variance of the difference between each student score on both halves is $\hat{\sigma}_d^2$ and $\hat{\sigma}_X^2$is the variance of the total score.

However, the infinite number of measurements in CTT helps to ascertain precision in measurement which in most times is actually not feasible. Some factors that influence reliability of test scores in CTT approach are 1) the length of test, 2) spread of scores (sample of examinees being highly homogenous on the trait been measured), 3) determination of the type of test (speed or power test), 4) objectivity in scoring, and 5) the level of difficulty of a test (Okpala, Onocha and Oyedeji, 1993). The application of different reliability methods probably yields different estimates. This is somewhat worrisome to adjudge how reliable a test is. It is therefore evident that in practice with CTT, precision in measurement can only be enhanced where consecutive samples are fairly representative and do not differ across time. A better approach to achieving precision has been taken care of in IRT method.

Test precision, a term that is similar to the estimate called reliability, is conceptualized as something called ‘the amount of information’ under IRT approach and it is conditional on the trait level being measured. When evaluating reliability in IRT approach, the amount of information an item or a test provides is to be considered. The test as a whole provides better and adequate information about individual items, thus conferring reliability in CTT. Information also known as item quality is a function of ability (*θ*). An item could be very informative for some ability levels and relatively uninformative for others.

Item information being a function of item location as well as discrimination makes it easier graphically to see and explain why some items were chosen. It indicates the usefulness of an item in assessing ability. Item usefulness is measured by how good an item is at distinguishing examinees with low ability levels from those with high ability levels. Items are basically more informative where the slope of the item characteristics curve is steepest. This can only happen when (a) item difficulty (b_j) is relatively close θ_j (b) item discrimination (a_j) is relatively high, and (c) guessing parameter (c_j) is relatively low. If c_j =0, an item provides its maximum information where $\theta_s = b_j$.

Also, if the standard deviation of the ability estimates about the examinee's ability parameter is calculated and squared, it becomes the measure of the precision with which a specified ability level can be estimated. Thus, the amount of information obtained in terms of how large or small, will tell how the examinee ability will be estimated with precision. This indicates that a respondent whose real ability is estimated with precision will have estimates that are reasonably close to the true value.

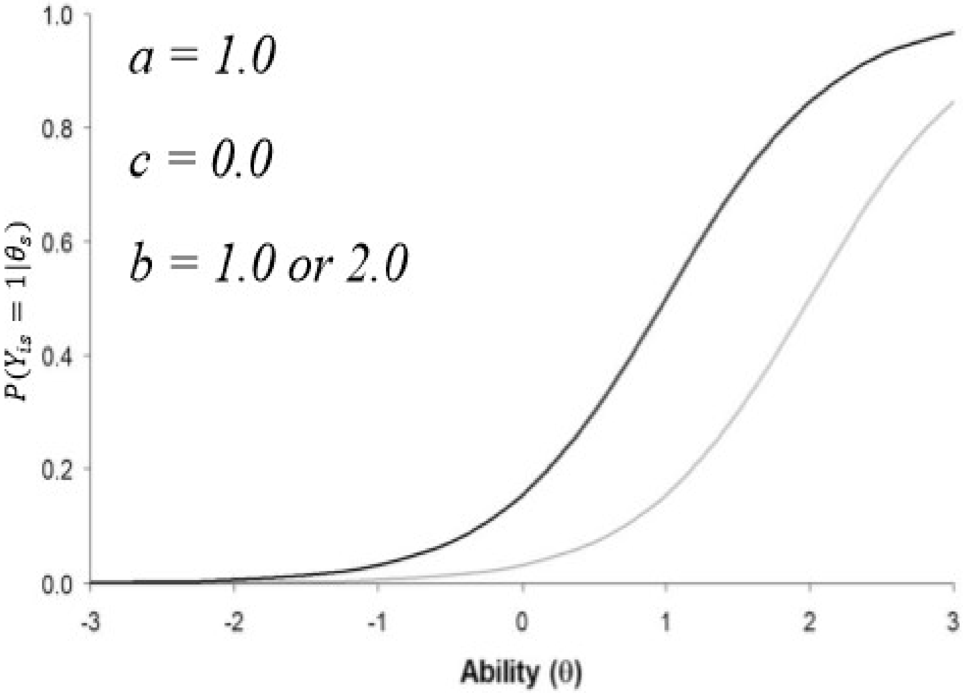


Figure 2.5b: An ICC showing how steep the curve is(Templin, 2011)

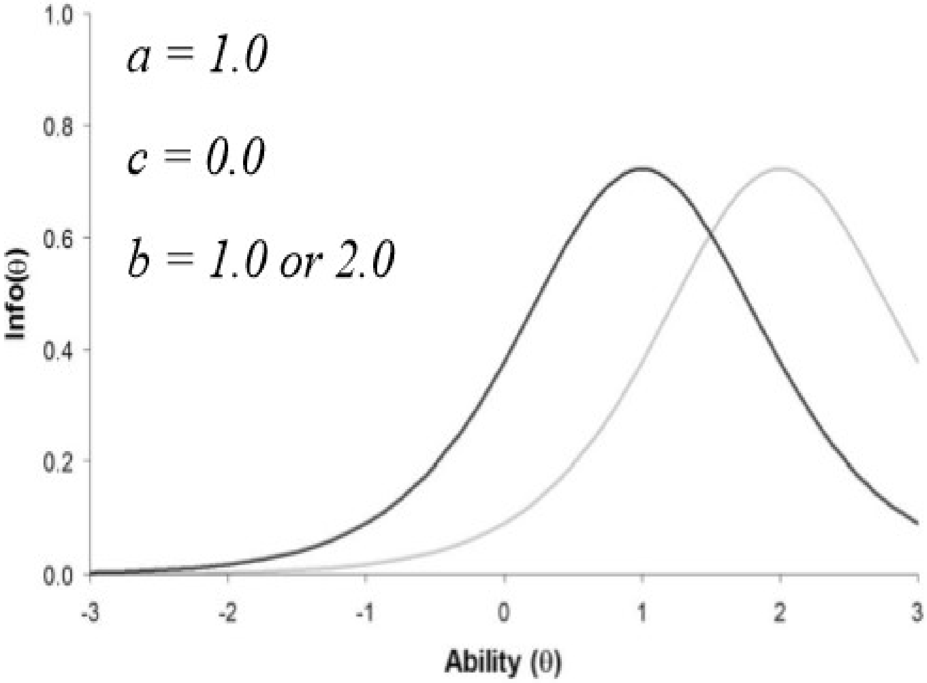


Figure 2.5c: An IIF showing how informative an item is (Templin, 2011)

Figure 2.5c displays that the extent of information is at maximum at an ability level of 1.0 and is about 5 for the ability range of -3<= θ<=1. Within this range ability is estimated with some precision, while outside of it, the amount of information decreases rapidly and the corresponding ability levels are not estimated very well. In a test that is general in use, the best information function would be a horizontal line at some huge value of **I** and all ability levels would be evaluated with the same accuracy. Regrettably, such information function is difficult to attain (Baker, 2001).

Information function (IF) according to Fisher in Baker (2001) is defined as the inverse of the precision by which a parameter could be assessed. Once a parameter is estimated with precision, more information about the estimate of the parameter would be known better than if it is estimated with a smaller amount of precision. Statistically, the variability of an estimate about the value of a parameter is the measure of the precision with which the parameter is assessed. Thus, a measure of precision is the variance of the estimators with which a given ability level can be estimated.This is denoted by $SE(\theta)^2$ (standard error of estimate) while the amount of information $I\ (\theta)$ at that given ability level is the inverse of its variance;

$$\boldsymbol{SE(\theta)} = \frac{\mathbf{1}}{\sqrt{\sum_{j=1}^{N} a_j^2 P_j(\theta)\ Q_j(\theta)}}$$

$\boldsymbol{SE(\theta)^2} = \frac{\mathbf{1}}{I(\theta)}and\boldsymbol{I(\theta)} = \frac{\mathbf{1}}{SE(\theta)^2}$ **……………….eqn2.28**where SE^2 is the variance of the estimator i.e. the square of the standard error of estimation and *I(θ)* is the sum of item information in a test. If information $I(\theta)$ increases with the quality and number of items, the SE conversely decreases.

When estimating ability using IRT, the information for an item is a function of the first derivative of the likelihood function and is maximized at the inflection point of the ICC (McDonald, 1999). The general form of the item information function, given any dichotomousIRT model described by a response probability$P_j(\theta)$ by Lord in Magis (2013) is given as:

$\boldsymbol{I_j(\theta)} = \frac{P_j'(\theta)^2}{P_j(\theta)\ Q_j(\theta)}$ **……………...eqn2.29**

where the $Q_j(\theta)$= 1-$P_j(\theta)$and$P_j'(\theta)^2$is the first derivative of $P_j(\theta)$with respect to θ.

For 1PL IRT model**:** $\boldsymbol{I_j(\theta, b_j) = a^2 P_j(\theta)\ Q_j(\theta)}$ **…….................eqn2.30**

2PL IRT model:$\boldsymbol{I_j(\theta, b_j, a_j) = a_j^2 P_j(\theta) Q_j(\theta)}$ **…................eqn2.31**

3PL IRT model:$\boldsymbol{I_j(\theta, b_j, a_j, c_j)} = \frac{a_j^2 Q_j(\theta)\}\{P_j(\theta)-c_j\}^2}{P_j(\theta)\{1-c_j)^2}$ **……................eqn2.32**

4PLIRT model which is the focus model of this study is given as;

$$\boldsymbol{I_{\mathrm{j}}(\theta, b_j, a_j, c_j, d_j)} = \boldsymbol{a_j^2 \frac{(P_j(\theta)-c_j)^2\,(d_j-P_j(\theta))^2}{(d_j-c_j)^2 P_j(\theta)Q_j(\theta)}} \qquad \textbf{....................eqn2.33}$$

where, $P_j(\theta)$ represents response function of 1, 2, 3 and 4PL models, $Q_j(\theta)$= 1-$P_j(\theta)$,*a*= slope, *b*=threshold, *c*=lower asymptote and *d* = upper asymptote parameter of the particular item.

Loken and Rulison (2010) Information Function for 3 and 4PL models are given as:

$$\boldsymbol{I_J(\theta)} = \frac{\boldsymbol{1.7^2 a_j^2(1-c_j)}}{\boldsymbol{\left[c_j+e^{1.7a_j((\theta-b_j)}\right]\left[1+e^{-1.7a_i((\theta-b_j)^2}\right]^2}} \qquad \textbf{................eqn2.34}$$

$$\boldsymbol{I_j(\theta)} = \frac{\boldsymbol{1.7^2 a_j^2(d_{j\_}c_j)^2}}{\boldsymbol{\left[c_j+d_je^{1.7a_j((\theta-b_j)}\right]\left[1-c_j+(1-d_j)e^{1.7a_j((\theta-b_j)}\right][1+e^{-1.7a_j((\theta-b_j)^2}]}} \textbf{..........eqn2.35}$$

Thus, a large amount of information is an indication that examinee true ability at that level is estimated with precision. This implies that error that will be associated with the ability estimation will be small and such estimations will be sensibly close to the real value. Otherwise, a small quantity of information indicates that the ability will not be evaluated using precision and estimate is bound to be widely scattered about the true ability.

Items that seem highly easy, generally give more information when it comes to the level of students withweak ability while items that are hard and discriminating supply other information about highly-able students at the ability continuum (Ojerinde *et al*, 2014). Therefore, test information is given as the sum of information for all the specific items under conditional independence assumption.

Item information becomes test information when aggregated across items of a test. It is worthy of note that the contribution of item information function $I_i(\theta)$ to test information function $I(\theta)$ does not depend on the particular combination of test items because each item contributes independently to the test (Templin, 2013). Test Information is then defined as the summation of the quantity of item information at a certain ability level. Its function is mathematically written as:

$$\boldsymbol{I(\theta) = \sum_{i=1}^{N} I_i\ (\theta)} \qquad \textbf{......................eqn2.36}$$

Where $I(\theta)$ is the amount of test information at an ability level θ,$I_i(\theta)$ the amount of information for item *i* at an ability level $\theta$and *N* is the number of items in the test. Item information functionis a very big advantage of IRT approach over CTT as reliability can be described conditionally (as information) and it does not depend on the particular set of items. Therefore, test information is defined for a set of items at each point along the ability (θ)scale. Information will continue to increase as test items are added, thus increasing precision.

2.2.6 The Frequentist Estimation Methods used in Calibrating Model's Parameters

Other important aspect of IRT that wasconsidered is the calibration of test items such that the numerical estimates of both item and examinees' parameter estimates in the chosen IRT model are expressed in metric forms. Instrument calibration is one of the primary processes used to maintain instrument accuracy. Calibration is the process of configuring an instrument to provide a result for a sample within an acceptable range. Eliminating or minimizing factors that could cause inaccurate measurements in assessment is fundamental to instrumentation design. Therefore, within the framework of IRT, item calibration involves the estimation of item parameters in the chosen IRT models (Eggen and Verhelst, 2011). Calibration is when a pool of items is developed on the same scale and this could involve estimating item parametersand testing the validity ofthe model.

The main reason for assessing examinee is to know which ability level his or she falls to on the trait continuum under IRT. When an individual ability estimate is known, the examinee can be evaluated in terms of how much underlying ability he possesses in order to get an item correctly. Thus, the need to know the relationship between test item and test performance makes calibration necessary (Ojerinde, Popoola, Ojo and Ariyo, 2015). To utilize IRT approach, a statistical analysis of test data, typically called calibration, is performed with sophisticated software that involves iterative procedures that continue until satisfactory convergence in the estimate is obtained. The complexity of the estimation requires a computer programme. Accuracy of any parameter estimation method used depends on a number of factors that may include, the model chosen, the number of item parameters to be estimated, the dimensionality of the data and the number of items and examinees included in the data set (Ojerinde *et al.*, 2015).

There are various estimation methods in the calibration process. These include the maximum likelihood estimation (MLE) approach that comprises of joint maximum likelihood (JML), conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation methods, all of which are known as the Frequentist estimation method. Other method of estimation is the Bayesian estimation approaches (BEA) which also could be Expected á Posteriori (EAP) and Maximum á Posteriori (MAP) methods (Meng, 2007; González, 2010 and Ojerinde, Popoola, Ojo and Ariyo, 2014). Each approach possesses some different characteristics.

In the maximum likelihood method, two kinds of parameter are involved, the item parameters, which are assumed to be fixed-effect parameters and the persons parameters. Depending on whether persons' parameters are considered as fixed-effect or random-effect parameters, different likelihood-based estimation methods as mentioned above can be considered. Some other known estimation methods which are not meant for dichotomous response-type data and may not be emphasized in this study are the heuristic estimation (HE) and the weighted maximum likelihood (WML) methods.

MLE methods have been greatly used in various studies right from the inception of IRT in diverse computer softwares that are being upgraded constantly. Such software as Apple II, Bical, Bilog-MG, Xcalibre, Logist, Multilog, Noharm and Winsteps have been proven by many researchers in several studies. Amongst them are the works of Wright and Mead, 1976; Mislevy and Bock, 1997; Baker,1985, 2001; Baker and Kim, 2004; Linacre, 2003; Khairani and Nordin, 2011;Imam, Onyeneho, Onoja and Ifewulu,2015; Ojerinde, Onoja and Ifewulu, 2013; Templin, 2011;Ojerinde, *et a,l*2015 and Enu, 2015.

For instance, the Logist software which was popular for a while used joint maximum likelihood estimation (JMLE) method, where θ and item parameters were always simultaneously estimated. In Bilog-MG program, marginal maximum likelihood estimation (MMLE) approach and an Expected-Maximization (EM) algorithm were used such that item and person's parameters were estimated in consecutive steps (Galdin and Laurencelle, 2010).

In general, the likelihood function in MLE approach is maximized jointly with respect to the parameters of item and proficiency to find an estimate of ability (θ) using a standard technique like the Newton-Raphson algorithm. This likelihood function given local independence is;

$\boldsymbol{L = \prod P_i(u_\mathrm{i} \mid \theta)}$ ……………...**eqn2.37**

Where P_i (u_i |θ) produces the probability of answering $\boldsymbol{u_i}$ on item $\boldsymbol{i}$ conditioned on an examinee's true ability score $\boldsymbol{\theta}$, and n depicting items number.

The Joint Maximum Likelihood Estimation (JMLE) approach has the capacity to simultaneously estimate both item and person parameters by maximising the joint likelihood function of persons and items. Estimation is achieved in this approach via Newton Ralphson's method. Both item and person's parameters are considered as fixed effects in this approach. The probability of response in JMLE is given as:

$\boldsymbol{P(\underline{x} \mid \theta, \alpha, \underline{\delta}) = \prod_{j=1}^{L} P_j^{x_j}(1 - Pj)^{1-x_j}}$ ………………..**eqn2.38**

Where $P(\underline{x}|\theta, \alpha, \underline{\delta})$ is the probability of the response vector $\underline{x}$, conditional on the person's location θ, discrimination α and a vector of item location parameters $\underline{\delta}$ (i.e, $\underline{\delta}$ = δ_1, δ_2….. δ_L). The probability for item j, P_j, is calculated according to a particular model.

To obtain the joint likelihood function, L, across persons and items, one multiplies eqn2.31 across N persons:

$\boldsymbol{L = \Pi_{i=1}^{N} \prod_{j=1}^{L} P_j(\theta_i)^{x_{ij}}(1 - P_j(\theta_i))^{1-x_{ij}}}$ ……………**eqn2.39**

The likelihood function is transformed by applying the natural log transformation to eqn2.32, the joint likelihood function is obtained:

$\boldsymbol{lnL = \sum_{i=1}^{N}\sum_{j=1}^{L}[x_{ij} \ln(P_j(\theta_i)) + (1 - x_{ij}) \ln\left(1 - P_j(\theta_i)\right)]}$ ….**eqn2.40**

The values of the θ_s and δ_s that maximize eqn2.33 are taken as the person and item parameter estimates respectively. The basic procedure in JMLE method is relatively simple but its affiliated problem is inconsistency as the statistical properties are rather complex and not very satisfactory (Haberman, 2016). This is because a large number of persons do not ensure that the estimated item parameters converge to the parameter they estimate.

For Conditional Maximum Likelihood Estimation (CMLE) approach, conditioning on sufficient statistics for the person parameters is used (Glas, 2016). If a sufficient statistics$S$(X_i) is constructed for the person parameter θ_i in the presence of the item parameterδ, theprobability of the response pattern can be factored as:

$\boldsymbol{P_{\theta,\delta}(x) = \prod_i P_\delta(x_i \mid s(x_i)).P_{\theta,\delta}(s(x_i))}$ …………..**eqn2.41** where, $P_{\theta,\delta}$ (s (x_i)) is the distribution of the sufficient statistic $S(X_{\text{i}})$, i =1,...,n and the first factor $\Pi_i P_\delta\big(x_i\,|s(x_i\,)\big)$, is the simultaneous conditional probability of the observed responses x, which does not depend on the ability parameters because of the sufficiency of $S(X_{\text{i}})$ for θ_{i}. Estimating the item parameters is done by maximizing this conditional likelihood function with respect to δ:

$$\boldsymbol{L_C(\delta;\big(x\big|s(x)\big) = \Pi_i P_\delta\big(x_i\,|s(x_i\,)\big)} \qquad \textbf{................eqn2.42}$$

In the CML approach, estimating item parameters are considered by random variations of the observations and fixing the values of the conditioning statistics s (x_i). The justification for this depends on whether all random variation that is relevant to the problem (estimating the item parameters, δ) is in the reduced frame of reference (Eggen and Verhelst, 2011). This is easily seen to be heavily dependent on the properties of the neglected part of eqn2.35. If the distribution of the sufficient statistic $s(x_i)$ would be completely independent of the item parameters δ, the justification would be obvious. Eggen (2000) shows that the possible loss of information in CML estimation by neglecting the information on δ in the distribution of $s(x)$, is very small already at short test lengths. A major feature of this method is that it is valid, irrespective of any assumptions on the distribution of the ability of the students taking the test. The individual parameters are only part of the factors in the total likelihood which is neglected.

Both CML and JML estimation methods treat person parameter (θ) as known and a fixed effect. These methods are described as less-used and older in research as a result of some disadvantages pose by their use,such as producing estimators that are biased, inconsistent, with too small standard errors and likelihoods that cannot be used in model comparisons (Templin, 2011).

However, marginal maximum likelihood estimation (MMLE) method that is also known as Gold Standard of Estimation is considered the traditional, most consistent and often used estimation method (Sijtsma and Junker, 2006; Templin, 2011; Burgos, 2010). This method relies on two assumptions of independence: (a) item responses are independent after controlling for θ,i.e. the joint probability (likelihood) of two item responses is just the probability of each multiplied together, (b) persons are independent after controlling for random effects i.e. no clustering or nesting.

Repetitive mathematical methods such as the technique of Newton-Ralphson and Expectation-Maximization (EM) algorithm are used to obtain θ.

EM algorithm is used to find the maximum of a likelihood marginalized over unobserved data. Certain desirable features as asymptotic uniformity and normality that MLE estimates made use of are broadly used in IRT applications. In this method, IRT model is extended by assuming that the ability parameters θ_i are a random sample from a population with probability density function given by $g_\gamma(\theta)$, with γ parameter of the ability distribution. Thus, the response pattern X as well as the ability (θ) trait are considered random variables (Eggen and Verhelst, 2011). $\theta_i$is not as before individual person ability parameters, but realizations of the unobservable random variable θ.

In MML, the marginal distribution of the response pattern X is:

$$\boldsymbol{P_{\beta,\gamma}(x) = \int P_{\beta,\gamma}(x,\theta)d\theta = \prod \int P_{\beta}(x_i|\theta_i)g_{\gamma}(\theta_i)d\theta_i} \text{............... } \textbf{eqn2.43}$$

where$P_{\beta,\gamma}(x,\theta)$ is the simultaneous distribution of the response pattern X and the ability θ.

$P_\beta(x_i|\theta_i) = \prod \int P_{\beta_J}(x_{ij}|\theta_i)$is the IRT model, giving the probability of a response vector i of person, with ability θ_i. Item parameters β are simultaneously estimated with the parameter γ of the ability distribution by maximizing the marginal probability of the observed response pattern x (the marginal likelihood function) with respect to the parameters, that is,

$$\boldsymbol{L_M(\beta,\gamma;x) = \prod \int P_{\beta}(x_i|\theta_i)g_{\gamma}(\theta_i)d\theta_i} \text{.......................} \textbf{...eqn2.44}$$

In spite of the fact that MLE has been widely used and haS gained popularity in producing consistent and reliable estimates of different model parameters, there are still associated problems with its parameters estimation:

a) When the frequentist approach is applied to higher parameterized models like the 4PL model, estimates of both item and respondent parameters are found problematic or difficult to estimate (Loken and Rulison, 2010; Zeng,1997; Balov and Marchenko, 2016; Fox, 2010). In the study of Zeng (1997), it was said that the estimation algorithm of the 3PL model using MLE procedure failed.This resulted in invalid parameter estimate. If, also, a higher number of parameters grow proportionally with the number of observations, it can lead to inconsistent parameter estimate because of the requirement for the inversion of the matrix of second-order derivatives of the likelihood function with respect to all parameters (Glas, 2016).

b) In JMLE approach, simultaneous determination of item and person parameter estimates that maximize the joint likelihood of the observed data has a number of practical implications.

c) For consistent and efficient estimates of item parameters, the need to increase sample size is necessary. Increasing sample size leads to an increase in the number of person parameters that will be estimated. Hence, item parameters estimate would be biased in this approach because additional item information with which to estimate the person parameters will not be provided (de Ayala, 2009).

d) Issue of efficiency of parameter estimates is at stake since item and person parameter estimations are taken together. If one or more items do not exhibit model-data fit, the item will be removed and recalibration of the instrument becomes necessary. But in case of joint estimation, this is not possible.

e) For short instruments like 15 or fewer items, CMLE method produces biased person location estimates that could result in poorly estimated item locations.Having the same responses (response of 0 or1, all through) to one or more items will make item and individual location inestimable. Ability(θ) parameter cannot be estimated if none or all of the items are answered correctly (Templin, 2011).

f) Because ofthe fact that integration (rectangling theta) is required at each step of estimation,the possibility of adoptingMML for IRT models in small sample is not feasible (de Ayala, 2009).

g) MML usually cannot provide absolute fit information, because there are usually not enough people to fill-up all possible response patterns, so there is no valid basis for an absolute fit comparison.

2.2.7 Bayesian Estimation Methods used in Calibrating Model's Parameters

Since large and complex datasets are becoming rampant in educational and psychological research, statistical methods that are crucial for the analysis and interpretation of such data are needed (Gill, Heeringa, van der Linden, Long and Snijders, 2016). Aitkin (2016) observes that as model complexity increases with constant observed data, there were fewer effective observations per parameter and so departures from asymptotic (frequentist) behaviour must be expected. Bayesian methods are becoming increasingly popular because of its adaptability in obliging various models from different fields. So, IRT framework where new models are developed is a good research area through which Bayesian approach is employed in estimating model parameters. This is made possible as a result of the presence of some sophisticated packages as WinBUGS, Mplus and R language that have enabled researchers to explore new possibilities (Burgos, 2010).

Unlike the standard likelihood-based approach where intrigue is placed in the calibration of estimates of parameter that make the best use of likelihood of the observed data in order to yield the parameter estimates of maximum likelihood of concern. Bayesian methodology is a different statistical inference procedure where model parameters are random variables. This methodology has the ability to integrate previous knowledge in the analysis statistically by utilising the distributions of prior on the parameters of interest. Thus, if peradventure such prior information is not available, it will be better as well to utilise non-informative prior on the parameters. Inferences in Bayesian approach depend on samples generated for the distributions of posterior, which could be applied to reduce information about the concerned parameters. It is necessary that MCMC convergence to the target density is monitored and the posterior distribution before utilising the samples to conduct inferences be monitored as well in Bayesian approach.

Van der Linder (2016) state in a preface that when his previous handbook on IRT was produced, Bayesian approaches had already gained some ground but were certainly not common.This gave credence to the computational success of Markov Chain Monte Carlo (MCMC) approach. Although, arguments as confirmed by some studies (Baker, 1998 and Kim, 2001) have expressed that the application of Bayesian approach to estimating simple IRT models' parameters (1PL or 2PL) is typically not superior to that obtained in the frequentist (MMLE) analysis or a Bayes modal procedure. It is however stated that methods of Bayesian that are applied, yielded better estimates where MML or Bayes modal approaches posed statistical problems.

Research in recent years has reconsidered the calibration of parameters in heavily parameterized 4PL model within Bayesian estimation method (Loken and Rulison 2010; Fox 2010; Kim and Bolt 2007). Loken and Rulison (2010) explore the justification and formulation of a 4PL IRT model in a simulation and empirical studies where a Bayesian method was employed to improve effectively parameter estimates for questions and examinees. Data was made available with the aim of calibrating it with a 4PM item response model.Fitting was generally better when 4PM is applied instead of the 3PM or 2PM.

Bayesian methods have also been adapted to estimating IRT models in various ways. Models with multiple raters, multiple item types and missing data as evidenced by Patz and Junker (1997; 1999) and testlet structures by Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow and Du (2000). Other models of latent classes according to Hoijtink and Molenaar (1997), those with a multilevel structure on the ability parameters by Fox and Glas (2001), and the item parameters (Janssen, Tuerlinckx, Meulders, and de Boeck, 2000), as well as multidimensional IRT models (Béguin and Glas, 2001).

However, the requirement for dependency on the complex structures in evaluating with multiple integrals in the process of solving estimation equation in MMLE or Bayes modal framework that has always been difficult motivated the interest in Bayesian inference and MCMC estimation procedures. Patz and Junker (1999) therefore stated that these problems are easily avoided in an MCMC framework.

In Bayesian statistical analyses, all inferences are founded on the posterior distribution of the parameters of concern. Three things convey major contributions into the posterior distribution by which the samples can be taken. First, the likelihood function, that is same as the classical likelihood-based inference. Secondly are the distributions of the parameters of interest (prior), which reflect uncertainty about the true values of the parameters before observing the data and thirdly, the hyper-parameters governing the prior distributions (the parameters that characterize it).

Assuming, it is realised that the parameters of interest follow a probability distribution represented by different parameters before the data is observed. As soon as the data have been observed, the prior knowledge about the parameter of interest is refreshed characterising what is called the posterior distribution. For instance, in Bilog-MG or Multilog, the default estimation procedures are the marginalised Bayesian item parameter estimation (Bayesian Modal Estimation) via an EM algorithm for item parameters and the Bayes Expected á Posteriori (EAP) estimation for θ.

Swaminathan and Gifford (1986) reveal that when comparison of MLE methods are made with Bayesian approach, the latter estimation method enhanced the reliability of the estimates of guessing parameter (c_j) in 3-parameter logistic model. And that Bayesian approach has been evidenced and beneficial in estimating difficult and seriously parameterized models. Therefore, a Bayesian method is seen as a suitable technique to obtain consistent estimates of $\boldsymbol{d}_{\mathrm{j}}$. Balov and Marchenko (2016) also demonstrate how 4PL model was fitted using bayesmh, a stata software. When the Deviance Information Criterion (DIC) of 3PL model was compared to that of 4PL model, the result was that the 4PL model provided a better fit. However, in the cause of applying 4PL model in this study, packages in R software, in the Comprehensive R Archive Network (CRAN)that provide tools for Bayesian inference was employed in estimating the item and person parameters (MIRT). This package was clearly not accessible when 4PL model was initially proposed (Magis, 2013).

The technique that Bayesian approach depends upon is the possibility of refreshing the prior knowledge that is known about the concerned parameter with regards to the information acquired from the current data. The equation is mathematically stated by adopting Bayes Theorem,

$$\mathbf{Pr}\ (\boldsymbol{A}|\boldsymbol{B}) = \frac{\mathbf{Pr}(\boldsymbol{B}|\ \boldsymbol{A})\mathbf{Pr}(\boldsymbol{A})}{\mathbf{Pr}\ (\boldsymbol{B})} \qquad \textbf{......................eqn2.45}$$

If the probability of an occurrence *A* is *Pr(A)* at that point, the function expresses that, once an occurrence *B happens,* the uncertainty about *A* can be expressed as Pr($A|B$) in view of the evidence that *B* provides, according to Bayes equation, The equation is the same for probability density distributions. If f(α) signifies the prior knowledge about a parameter vector α, and the data *Y* has a density function *f*(y|α), at that point equation turn out to be;

$$\boldsymbol{f}(\boldsymbol{\alpha}\ |\mathbf{y}\ ) = \frac{\boldsymbol{f}(\mathbf{y}\mid\boldsymbol{\alpha})\boldsymbol{f}(\boldsymbol{\alpha})}{\boldsymbol{f}(\boldsymbol{y})} \propto \boldsymbol{f}(\mathbf{y}\mid\boldsymbol{\alpha})\boldsymbol{f}(\boldsymbol{\alpha})\textbf{...................eqn2.46}$$

where∝ stands for "proportional to" and it means that the distribution of posterior for that parameter α is proportional to the product of the likelihood function and the prior. It should be noted that all the parameters of the model are contained in vector α.

Weiss and Minden (2012) believe that parameter estimates were generally improved when large sample sizes and long tests were used. Other notable researches where Bayesian estimation methods were applied are the works of Beguin and Glas (2001) to a 3PLM and the study of Klein Entink, Fox and van der Linden (2009) to an IRT model that incorporate response times. Estimation methods that could be found in Bayesian approach include the Maximum a Posteriori (MAP) and Expected a Posteriori (EAP) approaches. Baker and Kim (2004) recommend the use of marginalized Bayesian item parameter estimation (BME). This estimation is quite similar to the MMLE, except that a prior distribution is added on the discrimination parameter (a).

Bayesian approach ensures that the procedure can be completed even in limited cases (when all items have been answered correctly or all incorrectly). Once the items are calibrated, the parameters are obtained by the largely used EAP estimation procedure proposed by Bock and Mislevy (1982).

2.2.8 Comparability of the Frequentist and Bayesian Estimation Approaches

Bayesian method is advocated for, not because its estimates are superior to that of Maximum Likelihood, but because of its usefulness to complex and higher parameter models. The avoidance of the evaluation of multiple integral in solving estimation equations in MML is still another cogent reason. Here are some of its several advantages over MML method.

Table 2.2: Advantages of Bayesian Estimation Approach over Frequentist Approach

S/N	Bayesian Estimation Approach (BEA)	Maximum likelihood Estimation Approach (MLEA)

1.	Estimates of ability score $(\hat{\theta})$ for all response patterns are available, including zero and perfect score patterns.	Ability cannot be estimated $(\hat{\theta})$ for examinees with either zero or perfect score on the test.
2.	Bayes Modal Estimation (EAP) adopts a non-iterative procedure in its estimation method.	MLE uses an iterative procedure in Gaussian quadrature approach to achieve convergence in order to assess model-data fit
3.	The use of MCMC approach in Bayesian method makes the estimation of both simple and complex models to be comparatively easier.	Either JMLE or MMLE techniques requires pre-calculation of derivatives which may not make parameter estimates of complex models to be realistic.
4.	There is no need to derive the theoretical sampling distribution of the statistics in Bayesian method.	Derivation of the theoretical sampling distribution of the statistics is necessary in MMLE, which are sometimes difficult or impossible
5.	Bayesian inference seems to allow many users to achieve reliable results with less effort than the ML approach..	Achieving convergence with the ML approach with low-information data is difficult.
6.	The generality of the procedure in BEA is applicable to all types of IRT models both simple and complicated; this makes it a better estimation technique.	In MMLE, when item estimation is separated from person estimation, Bayesian procedure such as EAP can still be used for estimating person location.
7.	The incorporation of the prior knowledge of the parameter distribution increases the chance of obtaining a better estimate.	Relying on the likelihood alone may not be sufficient.
8.	It is a computationally progressively helpful approach to evaluate IRT models, as models are getting advanced to suit more random effects (both trait and item effects).	Computation becomes infeasible as models become heavily parameterized.
9.	Markov chain convergence to the target density and posterior distribution are strictly monitored to obtain summaries that could give more information than mere point estimate.	Less information about the samples is supplied in likelihood-based approach because of the routines that converge to a point estimate.

2.2.9 Conceptual Issues Relating to Computer-Based Testing

The prevalent assessment mode in education and testing organisations around the nation happens to be paper-pencil type that is as old as the advent of assessment itself. But, as the application of science and technology advances, modern inventions and innovations into every facet of national development are evident. Therefore, the use of ICT-based assessment is one of the rapidly expanding boundaries of educational technology. Ripley (2009) defines computer-based testing (CBT) mode as the utilization of technological know-how to digitize, create a better and proficient modified assessment. He opines that CBT is regarded has been a standardized mechanism for improving all aspects of securing items where both items and students' responses are encrypted in a database.

Currently, a resilient interest in CBT has arisen and several positive advantages of this assessment approach have been identified by advocates. Becker (2006), Salend (2009) and Thompson, Johnstone and Thurlow (2002) highlight some of the advantages of CBT as administering effectively, preference by examinees, immediate result, self-selection results, possibility of shifting attention from assessment to instruction, efficient item development and increased authenticity.

ICT-based assessment includes computer-based test (CBT) and computer-adaptive testing (CAT). Other terms that are used for ICT-testing include Computer-Assisted Assessment or Computer-Aided Assessment (CAA); Computer-Based Assessment (CBA); Online Assessment or Internet-Based Assessment (IBA) and e-Assessment/Testing (Pereira and Scheuermann cited in Ojerinde, Anyaegbu, Onoja and Adelakun, 2013). Ojerinde et al (2013) posited that these terms, although used interchangeably, have distinct meanings in the ways they are applied. For example, e-Testing or CBA is different from the CAT in the sense that the latter is administered to the test-takers according to the level of their abilities.

ICT-based assessment is a system of test administration where candidates' responses are electronically recorded and scored quickly, accurately and securely. Quellmalz and Pellegrino (2009) opine that an increasing indication that advancement in technical know-how enables people to do numerous traditional things in testing globally is evident.Such innovations make assessment better and quicker. Different types of computer-based tests include;

a) A fixed-form CBT: This is a type of CBT in which every examinee have access to similar sets of items. This type is typically a paper-pencil test (PPT) given on the system. An alternative is to randomize the items or present them in different form order for separate examinees.

b) A linear-on-the-fly test (LOFT): This presents a type of test where items are randomly drawn by the system from a pool of items such that individual examinees get a unique test with equivalent content and equivalent statistical characteristics.

c) Computerized-adaptive testing (CAT): A kind of CBT that adapts to examinee's ability level (Tailored Testing). This type of test uses the benefit of the system computational power by helping the computer to score each response as it is given and then pick subsequent items based on the test taker's responses. Normally, more difficult items follow the given of right answers and easier questions follow wrong responses. Wainer (2000) amd Rudner (2012) stated that test takers see few questions that are very difficult or extremely easy for them and have more items of appropriate difficulty.

Like LOFT, CAT is more secured because individual examinees take a different test. It can as well be shorter and more accurate than LOFT because of the tailoring. However, it is more expensive to develop and administer. It is a relatively new and widely accepted method of online psychometric testing. It is used in various areas including employee recruitment. Psychometric tests which are based on CAT include, aptitude tests, reasoning tests, verbal reasoning tests and numerical reasoning tests. Example is the Graduate Management Admission Test (GMAT).

Benefit of Computer-based Test: While both PPT and CBT modes of testing are efficient, experts continue to weigh the greater advantages CBT has over PPT. Some of themare:

a) Increased delivery of calibrated and delineated test items according to their pertinent item characteristics.

b) Efficient administration of examination and scoring of tests.

c) Improved test security resulting from electronic transmission and encryption for total eradication of breaches of examination security.

d) Increased computer awareness by the test-takers.

e) Improving quality and standard by improving the precision of detecting the actual values of the observed variables.

f) Bringing efficiency in collecting and processing information to support decision making and provide rapid feedback for the participants and stakeholders.

g) Enhancing examination discipline and reducing to the barest minimum the problem of examination malpractices.
h) Conforming to the global best practice and international standard offering a computer-based national examination (Ojerinde, 2012).

To further corroborate the advantages of CBT over PPT, the following table shows the statistics of performance in both paper and pencil and CBT examinations.

Table 2.3: 2013 UTME Cumulative Performance Statistics for PPT versus CBT Modes

	Paper-Pencil Test		Computer-Based Test	
Range of Score	**Total No of Candidates**	**%**	**Total No of Candidates**	**%**
200 and Above	151,495	9.67	30,685	39.59
190 and Above	289,495	18.48	43,344	55.93
180 and Above	502,700	32.09	57,743	74.51
170 and Above	756,642	48.30	70,463	90.92
160 and Above	1,012,397	64.63	76,890	99.21
Below 160	554,111	35.37	2,879	3.71

Table 2.3 records a comparative performance statistics of candidates in PPT and CBT of the 2013 Unified Tertiary Matriculation Examination (UTME). CBT records a total number of 30,685 candidate that scored 200 and above representing 39.59% of the total number that sat for the CBT while PPT records 151,495 candidates that scored 200 and above representing 9.67% that sat for PPT.

A global evidence has laid credence to curbing of examination malpractices through e-testing. Most of the international examinations conducted in the world are computer-based. For example, Graduate Management Admission Test (GMAT), Test of English as a Foreign Language (TOEFL), Graduate Record Examination (GRE), Scholastic Aptitude Test (SAT), Cisco Certified Network Associate (CCNA), Oracle and Associate of Chartered Certified Accounts (ACCA). Over time, the international Computer-based Test (CBT) has proven that breaches of examination security can be curbed to a large extent (Ojerinde, 2013).

2.2.10 Relation between e-Assessment (CBT) and Item Response Theory

Joint Admission and Matriculation Board (JAMB) as one of the national public examination bodies that deal with large-scale and high-stake examination was the first ever examination body in Nigeria to conduct computerised testing in 2013. This version was one of the three modes employed in the conduct of the examination which took place over a period of fourteen (14) days, from 18th May to 1st June, 2013 at fifty-six (56) different centres across the country (Ojerinde, 2013). The other two modes were paper- pencil test (PPT) and the dual-based test (DBT) (paper and computer).

This was as a result of the pressures the board had been facing since its inception to administer the popular paper-pencil mode of examination. The problems were high cost of printing and distribution of examination materials such as calculators, rough working sheets, syllabus, brochures, instruction and guideline to invigilators and supervisors, question papers and OMR sheets. Missing scripts, increase rate of examination malpractices, environmental induced factors, the logistics of examination administration in terms of personnel, vehicles, monitors, security cover were evident (Ojerinde, Popoola, Onyeneho and Egberongbe, 2015).

The Nigerian Senate Committee on Education prescribes online tests based on the decision taken by the investigative committee on reasons for poor UTME results. The committee resolved that all written examinations should be online in both secondary and tertiary institutions in Nigeria. This was what prompted JAMB to conduct of CBT as a way of overcoming some of the aforementioned problems that were associated with PPT. Although, the transition was a giant stride, several challenges against the successful implementation of CBT were evident. The board embarked on massive training and retraining of members of staff, development of functional item banking, determination of appropriate software for test authoring, acquisition of software for test delivery and test security, provision of CBT practice tests for candidates, publicity and sensitization of the public.Embracing CBT mode became essential when it was realised that IRT approach provides some new psychometric foundations that were basic to the implementation of CBT (Ojerinde, 2016).

Ojerinde (2013) in the preface of the board compendium titled *"Vital Issues in the Implementation of Computer-Based Testing in Large-Scale Assessment"* states that IRT has provided a modern psychometric basis for implementing CBT. Operators of assessment portfolio in Africa should be conversant with the principle and application of IRT especially with respect to item calibration for trait and parameter estimation. Previously, the board had been adopting traditional test theory (CTT) framework in ascertaining the psychometric properties of her items. But as soon as e-assessment became realistic, the migration into a complete computerization in applying IRT methods in item analyses became possible. The board's transition became advantageous with the help of automated item banking and e-assessment strategies.

Ojerinde further states that because of the many advantages CBT offers, JAMB would not rest on its oars but explore the possibility of venturing into computer-adaptive testing (CAT), a form of CBT that works according to what the respondents can do. Following the relation between e-assessment (CBT) and IRT framework, the following are made possible in modern-day assessment practices:

a. Automatic recording of item response time to individual item
b. Possibility of Computer-Adaptive testing (Famorotimi, 2019)
c. Cloning of test items
d. Enhancement of electronic item banking
e. Creation of Parallel forms of test through IRT approach, etc.

Some of the CBT centres used by JAMB across the nation and one of the PPT centre.

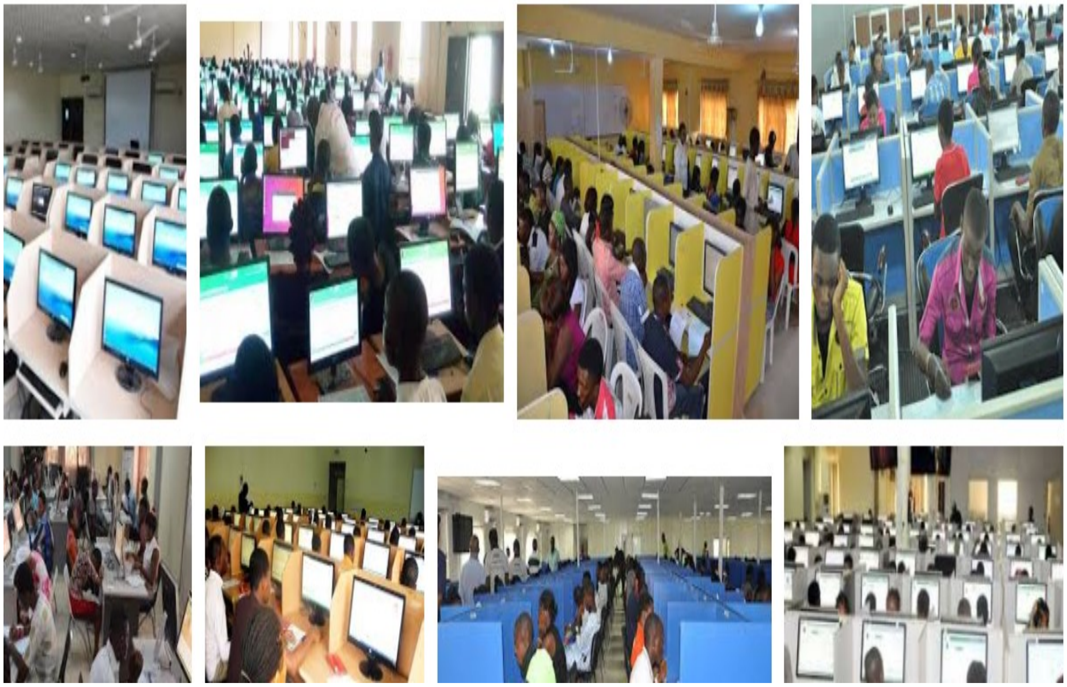


Figure 2.6a: cross-sections of candidates at different e-Testing centres sitting for CBT mode. Retrieved from Google page (2018)

Figure 2.6b: A cross-section of candidates in a JAMB centre sitting for PPT mode. Retrieved from Google page (2018)

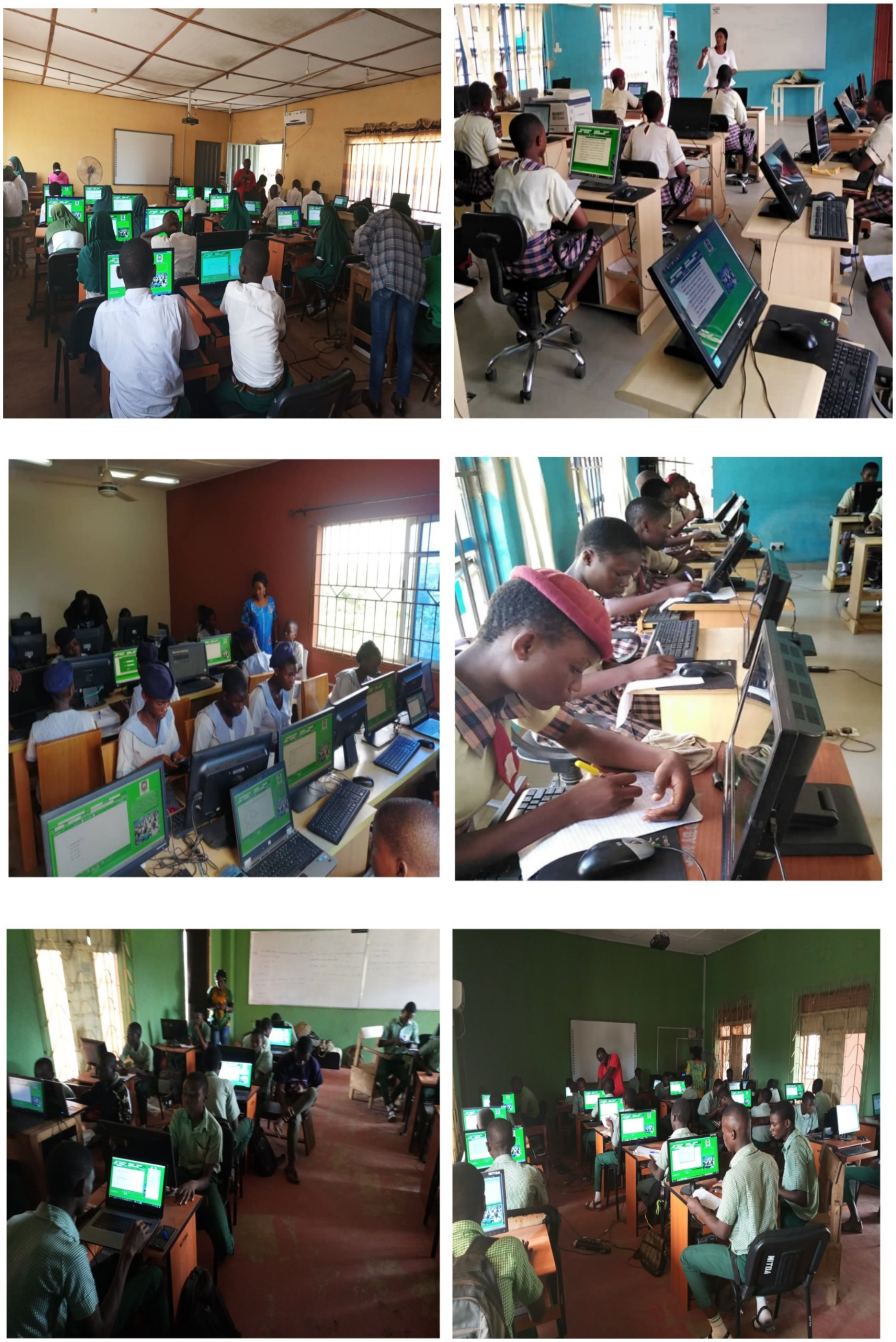
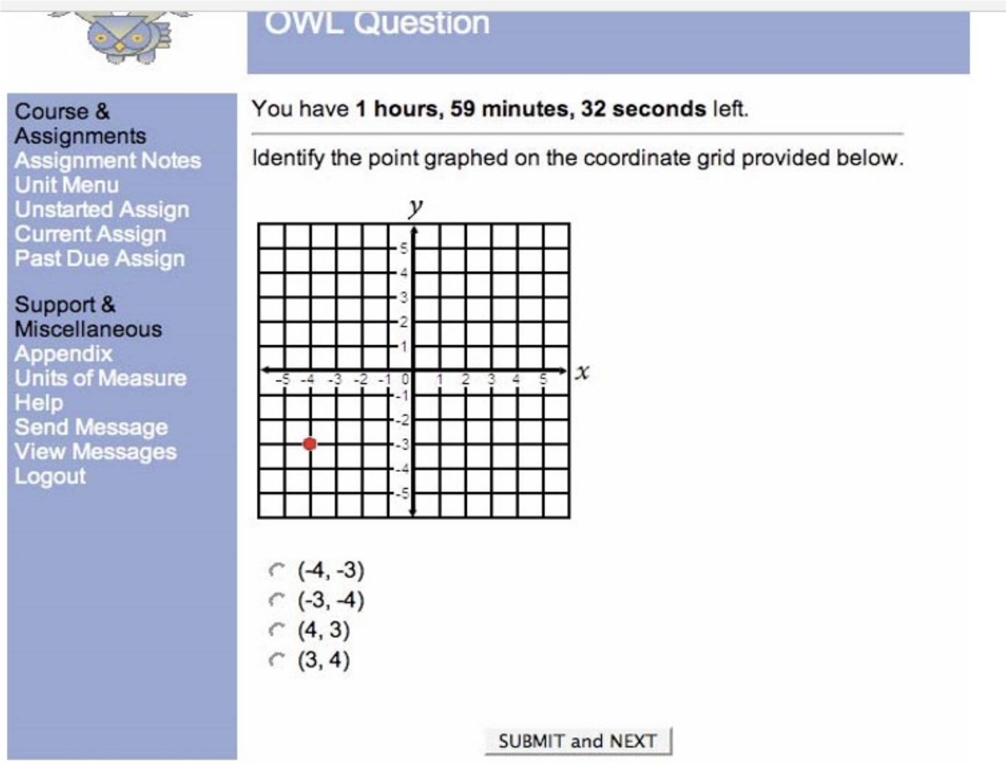


Figure 2.6c: A cross-section of examinees in CBMAT sessions (The Researcher)



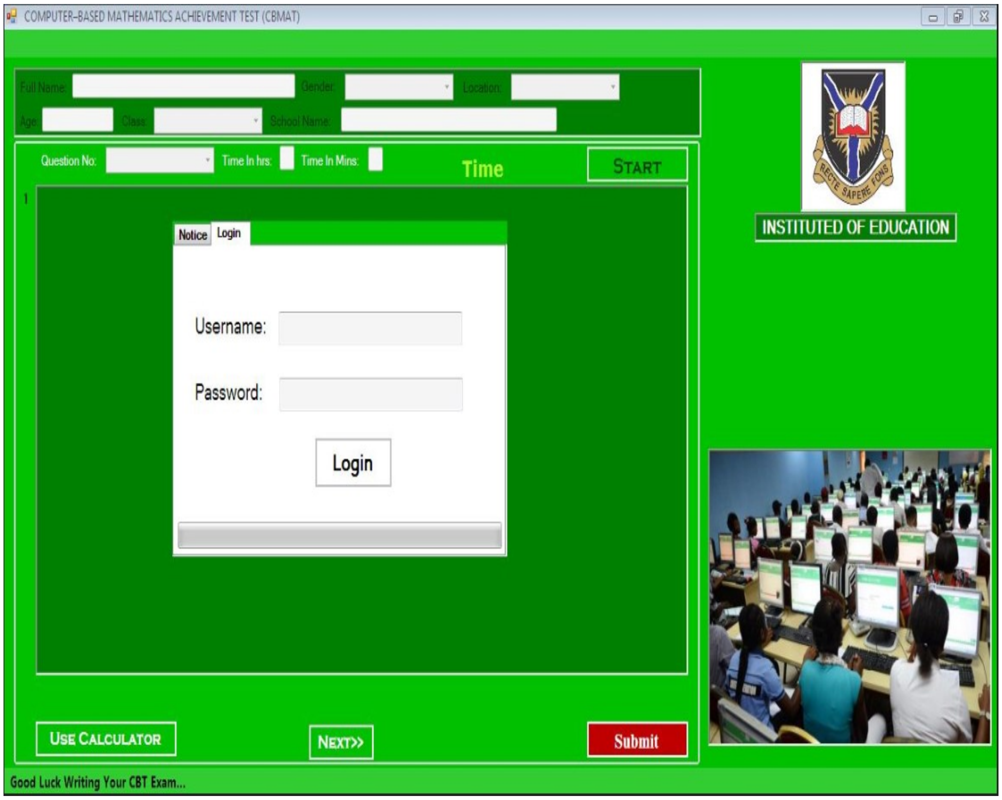
OWL Question

You have **1 hours, 59 minutes, 32 seconds** left.

Identify the point graphed on the coordinate grid provided below.

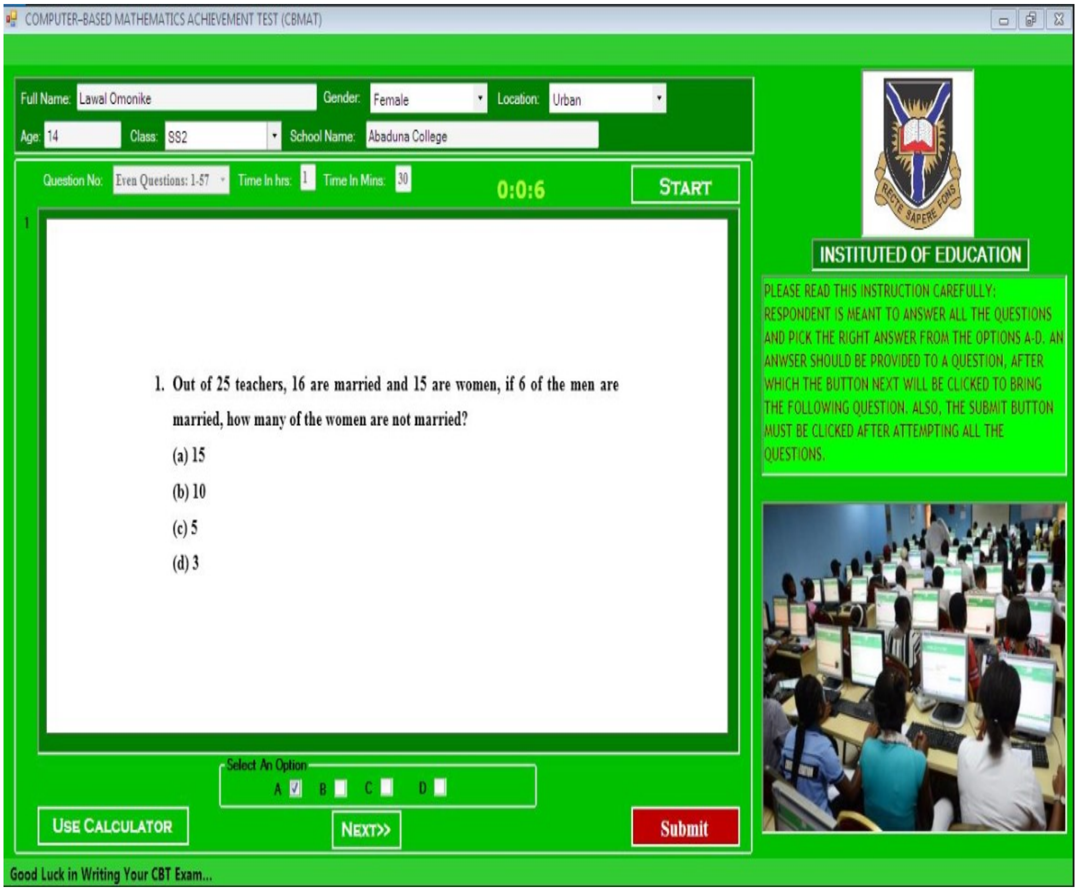
○ (-4, -3)
○ (-3, -4)
○ (4, 3)
○ (3, 4)

Figure 2.6d: Screen capture of a typical CBT interface (Zenisky and Baldwin, 2006)



Good Luck Writing Your CBT Exam...

Figure 2.6e: Screen capture of the login CBMAT login interface (The Researcher)



COMPUTER-BASED MATHEMATICS ACHIEVEMENT TEST (CBMAT)

Full Name: Lawal Omonike

Gender: Female

Location: Urban

Age: 14

Class: SS2

School Name: Abaduna College

Question No: Even Questions: 1-57

Time In hrs: 1

Time In Mins: 30

0:0:6

1

1. Out of 25 teachers, 16 are married and 15 are women, if 6 of the men are married, how many of the women are not married?

(a) 15

(b) 10

(c) 5

(d) 3

Select An Option

A ☑ B □ C □ D □

PLEASE READ THIS INSTRUCTION CAREFULLY: RESPONDENT IS MEANT TO ANSWER ALL THE QUESTIONS AND PICK THE RIGHT ANSWER FROM THE OPTIONS A-D. AN ANWSER SHOULD BE PROVIDED TO A QUESTION, AFTER WHICH THE BUTTON NEXT WILL BE CLICKED TO BRING THE FOLLOWING QUESTION. ALSO, THE SUBMIT BUTTON MUST BE CLICKED AFTER ATTEMPTING ALL THE QUESTIONS.

Good Luck in Writing Your CBT Exam...

Figure 2.6f: Screen capture of the CBMAT questioning interface (The Researcher)

2.2.11 Concept of Response Time (RT)

Response time is seen as a well-known concern in traditional mental measurement that is utilized in examining the connection between human performance and their response speed. As advocated by Suh (2016), the use of data that involve response time was not fully established in educational measurement, but because of the various behavioural patterns students exhibit in test administration, its idea became known. Investigating response time has stimulated the course of exploring and interrogating various mental processes that take place when examinees respond to how items vary. Variations in term of stimulus intensity;how familiarand knowledgeable the respondent is to the question.Studies of Van der Linden, Entink and Fox (2010) as well as Kyllonen and Zu (2016) found that their studies were motivated by the fact that computerized testing has become more popular and brought the formulation of noticeable models when it comes to time spent in responding to items of a scale.

Before IRT era, the differencebetween how fast respondents could tackle questions in a speed test and how hard the enquiries could be solved in a power test had been stressed. Thorndike *et al*. (1926) propose that the envisaged time to finish a test is made up of three components; (a) the time it took an examinee to answer some problems appropriately (b) the time taken to attend to other tasks wrongly, as well as (c) the time used to examine added problems in addition to settling on a choice of not having any desire to answer them.

Some cognitive information-processing treatments that were commonly used to tackle the issue of response time before the advent of IRT are the speed-level distinction, speed-accuracy tradeoff (Heitz, 2014) and the process models application. The speed-level dimensions have normally been assessed with simple speeded test where the sole aim is to know the number of items that are right in a given time among a specific number of items. There is more complicated power teststhat are meant to assess respondent's proficiency level with a focus on assessing number of correct items where no specific time is allotted to complete the items of the scale.

The speed-accuracy tradeoff treatment on the other hand made use of the deadline method (time limit) in estimating response time while process models application provided an elaborate analysis of response time. This is done by recognizing the mental processing taking place between the time an item comes up and the time in which response is made. Process models comprise the ex-Gaussian and the diffusion models (Ratcliff and Tuerlinckx, 2002). Ex-Gaussian distribution is the addition of a Gaussian (normal) and exponential distribution with three parameters,μ and σ of Gaussian as well as β of the exponential function. MATLAB software can be used in estimating the process models (Lacouture and Cousineau, 2008) or R package (R Core Team, 2014).

Schmiedek *et al.* (2007) adoptsthe ex-Gaussian modelin estimating examinee time response distributions of 135 respondents, who answered to eight choice-reaction time (CRT) tasks. A model that is of structural equation was fitted to evaluate the factors of μ, σ, and β. The findings were that a good model fit was found with fairly high factor loadings for μ and β as well as having loadings that are lower for σ. Moderate correlations among the three parameters (0.46, 0.51, 0.75) existed. An absolute correlation with free estimates for reasoning and working-memory capacity occurred for β where $r = -0.72$ to a range of -0.71 as against $r = 0.36$ to 0.56). Also, maximum correlations was seen in σ by a continual speed factor assessed as the quantity of easy questions answered in a time bound written test while the mean had the maximum associations with correctness on the choice-reaction time test.

Ratcliff, Smith, Brown and McKoon (2016) view diffusion model as a common function for signifying reasoning and neural procedures in an easy two-choice inference making. The model is considered as a proof accumulation model that helps in detecting noisy evidence on a stimulus with time till a response condition is gotten. It splits item response time into two. The time required to gather sufficient proof to make a choice on an item (Decision time) and the time necessary to carry out all other processes like stimulus programming and motor response known as non-decision time. Also, the process of evidence amassing in diffusion model is modeled after Brownian motion. Diffusion model parameters are used to settle problems relating to individual and group disparities. An instance that is well known is that people respond more slowly as they advance in age, although the cause of the slow process is indefinite.

Ratcliff, Thapar, Gomez and McKoon (2004) in their study subjected the response data gotten from a lexical-decision task to the diffusion model for more youthful and more established grown-ups. The result was that the lengthier response time by more established grown-ups were because of longer non-decision time and bigger limit detachment, yet this does not explain a general mental handling slow down.

The main problem in response time measurement is the numerous factors that affect it, which naturally make its measurement not to be attributed solely to any specific factor (Kellynon and Zu, 2016). For instance, an examinee that shows slow response on an item might mean that he has either a slow processing speed or that he is exhibiting caution in responding. Another student who answers swiftly and rightly might exhibit a fortunate guess or he might be stimulated to do so. In the same vein, if an examinee does not respond aright, it may be that he does not have the ability or does not spend adequate time to process the information wholly. It could even be that he was confused while responding and decided to leave the item.

Van der Linden (2006) made the following statement in support of a solution to model response time especially at the advent of modern test theory and computer-based testing.

> *"It has long been known that response times on test items are important sources of information on the person's behaviour, but waiting for the advent of computer-based testing to make the recording of response time a routine part of test administration is a dream come true. Now that testing is widely computerized, the question of how to model response time has become urgent".*

Item response theory approach is a method that focuses on response to specific item and every response to the individual item is affiliated to both item characteristics as well as examinee skills. It becomes a useful property for modelling response time. The most important improvement for years back had been the formulation of advanced models relating to psychometrics that combine the measurement of speed and ability parameters. The so called sophisticated models allow that different measurements can be made on response times as regards various item categories with items that are answered appropriately as well as erroneously.

Kyllonen and Zu (2016) observe that assessing and measuring response times often appear worrisome in ability measurement,where measurement is done indirectly,especially in time-boundandthe supposed speeded tests. Although, when issues relating to RT began to gain popularity, the researchers stated that the quality at which ability of respondent is measured has greatly been improved with some other added applications like form assembly and cheating detection. Such added advantages that have been brought to limelight have continued to gain popularity. RT has however been considered as the key outcome variable that is known as reaction time or information-processing test. RT has served a major role in measurement relating to mental capability. A type of RT model known as process model recognizes the mental processing taking place between the time an item comes up and the time at which response is made.

However, the practice of computer assessment mode has brought about important formulation of models of RT as a result of its automatic pool while responding. Computerized testing allows that responses to the test items are scored while time spent to respond are automatically recorded. Such information recorded as response times according to Fox, Etink and van der Linden (2007) helps to improve routine operations in testing, such as item calibration, adaptive item selection, latent ability estimation as well as exploring and measuring factors that influence performance on the test. There is a presumed clue that measuring response time could improve the quality of ability estimation, aid test pacing and form assembly. Item response time is said to help in detecting how students behave in testing. Behaviour such as rapid-guessing (RG) or solution-oriented behaviour (SB) as well as cheating detection could be identified.

Van der Linden (2006) affirms that response times are modeled in IRT model framework because of an assumed interaction between the parameters that govern the distribution of the person's response time and his or her response variables for the items. Luce (1986) also support that RT has been agreed within researchers that it provides information on what transpires in the cognitive processing when response is to be given. RT is however used for constructing test in research and it is referred to as the time a person spends on an item in a test.

Response models are used to estimate students' ability from the responses generated from test. Then, these estimates are measured with error. As a result, if response time as a part is said to be related to respondent true ability, it is assumed that such could lessen measurement error (Wright, 2016). Novick and Jackson (1974) described response time as a collateral information. For example, someone with very low ability may rapidly guess and get a response correct by chance. Incorporating response time into the estimation of the ability estimates could improve estimation.

Lee and Chen (2011) note that analyses of item RTs are not only driven by an interest in RTs (modelling RTs in order to estimate person speed or item time-intensity) or in the relationship between speed component and accuracy component, but also driven by concerns about long-term issues in educational testing. Such an issue is the impact of rapid-guessing behaviour on the estimation of IRT model parameters.Van der Linden, Entink and Fox (2010) and Fox (2018) are of the view that involving response time in either response or response time model could positively aid better estimates of a model's parameters. This supposed variable is serving as an added advantage to the conventional IRT models that have been available in estimating parameters.

The concern about how response time is to be modeled has been viewed from three different perspectives. The first method involves modeling response times with the inclusion of parameters that contain timing to an ordinary IRT model. Examples of this approach are found in Roskam (1997); Theissen (1983); Verhelst, Verstraalen and Jansen (1997). The second approach is characterized by modeling response timeseparately from the responses. Response times are modeled independently of the response variables for the items. Examples include the works of Maris (1993), Scheiblechner (1979), Schnipke and Scrams (1997) and van der Linden, Scrams and Schnipke (1999).

Van der Linden (2006) discussed selecting these models for response times on items of a test. The third methodology presented by Van der Linden (2007) is modeling of RT and reactions in a classified and hierarchical order, models meant to jointly assess response and response time variables. Therefore, some of the essential IRT functions aimed at modeling response and RT are the regression-type models, hierarchical models, cognitive components approach, observations from time-limit tests and diffusion-based IRT models.

Regression-Type Models: These types are earlier formulated IRT models for response and response time. These models either combine response time as predictors in IRT models for responses (Roskam, 1987; 1997; Verhelst, Verhelst and Jansen, 1997; Wang and Hansom, 2005) or integrate parameters of response as predictors for modeling response time (Gaviria, 2005). Time can either be used to forecast examinee's response or their responses are used to predict time in regression-type model.This makes the models to plainly represent a speed–accuracy tradeoff. Example of this type include Thiessen lognormal response time model, whose model was perhaps the leading one in item response theory and associated response times for tests that are considered timed. The standard 2-parameter logistic model was used to model examinee's item response while lognormal model was adopted to model the logarithm of their response time. This was to allow the usual positive skewness of response time distributions. Thiessen model is given as:

$$\boldsymbol{logT_{ij}} = \mu + \boldsymbol{\tau_j} + \boldsymbol{\beta_i} - \boldsymbol{\gamma}[\boldsymbol{a_i}(\boldsymbol{\theta_j} - \boldsymbol{b_i})] + \boldsymbol{\varepsilon_{ij}},\ \boldsymbol{\varepsilon_{ij}} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma^2}) \quad \text{............eqn2.47}$$

Where, μ: Log response time mean; β_i: time intensity; τ_j: person slowness;$a_i(\theta_j - b_i)$: response model; γ: the regression coefficient that revealthe interchangethat exist between response time and accuracy.

Hierarchical Models: These are models that are classified in a two-level category as projected by Van der Linden (2007). These modelsmutually assess reactions and reaction time.The first category is meant for separate students and a set of rigid questions while the subsequent category is meant for all respondents with the questions. The model in the first category accepts that student maintain a steady and consistent speed and ability (a specific area of the person's speed–accuracy tradeoff graph) such that the respondent will not accelerate or retard in speed while writing the test.

Reactions as well as RTs are presented independently with individual and item parameters as speed, ability, difficulty as well as time-intensity. For reactions on item, a 3PL normal-ogive model was used alongside the log odds.The log odd is standing for an individual reacting appropriately to a question which is a component of person ability (θ_j) and item parameters (discrimination, difficulty with guessing). Here, any typical IRT model can be utilized.

Van der linden lognormal model for response time;

$$\boldsymbol{log(T_{ij})} = -\boldsymbol{\tau_j} + \boldsymbol{\beta_i} + \boldsymbol{\varepsilon_{ij}},\ \boldsymbol{\varepsilon_{ij}} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\alpha_i^{-2}}) \quad \text{.....................egn2.48}$$

τj means the speed parameter for individual j (higher estimate shows quicker speed); $\beta_i$is intensity parameter for item *i* while $\alpha_i$gives time discrimination. Estimate α_i implies shows fluctuation of the distribution of log response time on item i through individuals.This implies that item *i* will discriminate examinees with speed that can either be high or low. Also, bigger estimate of β_i depicts that question require additional time in responding. A statistically proper way of making use of response time is through a hierarchical model that involves both traditional item response theory (IRT) and RT parameters.

Application of response time abounds in literature. Kyllonen (2016) focuses on measuring ability and response time for cognitive tests. Other uses of RT in testing include:

a) measuring student motivation levels, especially on assessment with little consequences on students results (Finn, 2015; Lee and Jia, 2014; Wise, Pastor and Kong, 2009).
b) assessing ethnic variations in pacing and time management in taking test (Lee and Haberman, 2015), cheating detection (Van der Linden and Guo, 2008) and assembling parallel forms (Van der Linden, 2005).
c) item selection in adaptive testing (Van der Linden, 2008; Van der Linden, Scrams and Schnipke,1999).
d) guaranteeing that scores are comparable over groups of questions that seem the same in terms of difficulty, then varrying time intensity (Bridgeman and Cline, 2000; 2004). Application of response time model is also evident in personality and attitude assessments (Ranger, 2013; Ranger and Ortner, 2011.

2.2.12 Scoring Models utilising Response Time

Many studies relating to psychometry have focused more on examinees' responses than speed, in spite of the fact that there are numerous experimental studies that have explored response tine psychological research (Schnipke and Scrams, 2002). Studies relating to the time taken in responding in assessments are restricted by some handy reasons which may include records keeping in operational situations and the randomization of ability groups. Suh (2016) corroborates this by stressing that issues relating to response time can not be adequately applied until computerized testing is presented.

More tests are being administered on the computer which makes it a lot simpler to gather time of response information unlike what was obtainable earlier. Right response to questions as well as marks students get in testing related to their proficiency are considered as the usually perceived student behaviour in test administration. Pervious works on time response models that involve the use of data from response time were carefully observed in the old ways of measuring traits that have to do with mental psychology.

Prompt response is taken to be criterion variable when it comes to different models, and assessment of student's capacity to handle the skills they possess. Different models have viewed diverse methods when it comes to specific models that can be used for estimating response time variable (van der Linden, 2006, 2009). They are proper when time items are moderately easy to answer and to process while temporarily, the proficiency is estimated through the processing speed. An example of such includes the usual speed test in intelligence testing.

The utilization of times of response in standardized/ structured tests as they relate with speed and accuracy are seen as dissimilar segments of ability as suggested by Scrams and Schnipke (1997). Models adopted in their study proposed the best approach in utilizing both accuracy in response and response promptness to give distinct estimates for performance. Obviously, models in IRT have been suggested to manage response time issues. Van der Linden (2009) obviously sorted the various models as response time integrating the usual models in IRT and those combining response time.

Thiessen's (1983) model: This researcher suggested a model in response time that combines IRT model for the first time. This is given as follow:

$$\boldsymbol{logT_{ij}} = \mu + \boldsymbol{\tau_j} + \boldsymbol{\beta_i} - \boldsymbol{\gamma}\left[\boldsymbol{a_i}\left(\boldsymbol{\theta_j} - \boldsymbol{b_i}\right)\right] + \boldsymbol{\varepsilon_{ij}},\ \boldsymbol{\varepsilon_{ij}} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}}) \quad \textbf{.........egn2.49}$$

Where, *logT*$_{ij}$is the log response timeof the person *i* who answers to question *j,* , μ depicts overall average, β_j stands for slowness characteristic of question *j*. $\tau_i$gives person *i* slowness characteristic, γ shows the value attached to the log response time as the coefficient of regression for 2PL IRT model, ε_{ij}specifies chance inadequacy.

This model contains examinee and item slowness estimates with the likelihood of precise response the examinee gave when answering the question. The concerned model, consequently indicates dual distinctive trade–offs, the first is between item difficulty and slowness while the second is between examinee ability and slowness. The estimate of the direction of associations between the two trade-offs is interpreted as the regression term (Schnipke and Scrams, 2002). Findings from Thiessen's investigation demonstrated that various types of relationships occur in different tests. The expounded interaction, that occurred in the respondents' reaction speed and accuracy also differ based on the features of the assessment.

Wang and Hanson's (2005) model: Another type of modelling in time response is that of Wang and Hanson that suggested a four-parameter logistic approach that is formulated to estimate the parameters of the model. This approach integrated response time in the estimation process with mathematical function written as:

$$P\left(x_{ij} = 1 \middle| \theta_i, \tau_i, a_j, b_j, c_j, \beta_j, rt_{ij}\right) = c_j + \frac{1-c_j}{1+ e^{-1.7a_j\left[\theta_i-\left(\beta_j\tau_i/rt_{ij}\right)-b_j\right]}} \text{.......egn2.50}$$

rt$_{ij}$is the time of respondent i response item *j*, β_j gives item slowness parameter while τ_i presents person slowness parameter. Both question and respondent slowness characteristics are anchored on the rate at which the probability of a right response increases which is a function of the time of response. Also, the rate at which the right response probability changes along cumulative response timing that is due to specific respondents and a specific item is decided by the product of the two parameters of slowness.

Response accuracy with time of response was later modelled as a joint distribution using one-parameter logistic function of Weibull in advancing the initial one. The reason for this was as a result of the earlier model that came with an independent assumption for item response time and respondent ability parameters. Ingrisome (2008) affirms that such assumption for the earlier model is impracticable in most timed testing circumstances.The latter model was suggested for use because of the independence assumption that was uninvolved. Wang's model was improved upon by involving a two-parameter logistic distribution of Weibull to the response time model with marginal distribution. To estimate models parameter, numerous estimation methods can be used buttechniques like MMLE and MAP showed much improvement on the parameters of the model (Ingrisone, 2008).

Framework with Hierarchical Model: The third category adopted in the formulation of models involving response and response time distribution is that of Van der Linden (2007). His strategy is known as the hierarchical framework that contains reaction time and the usual IRT response model in a two-level framework. Figure 2.6g depicts the pictorial framework representing the model.

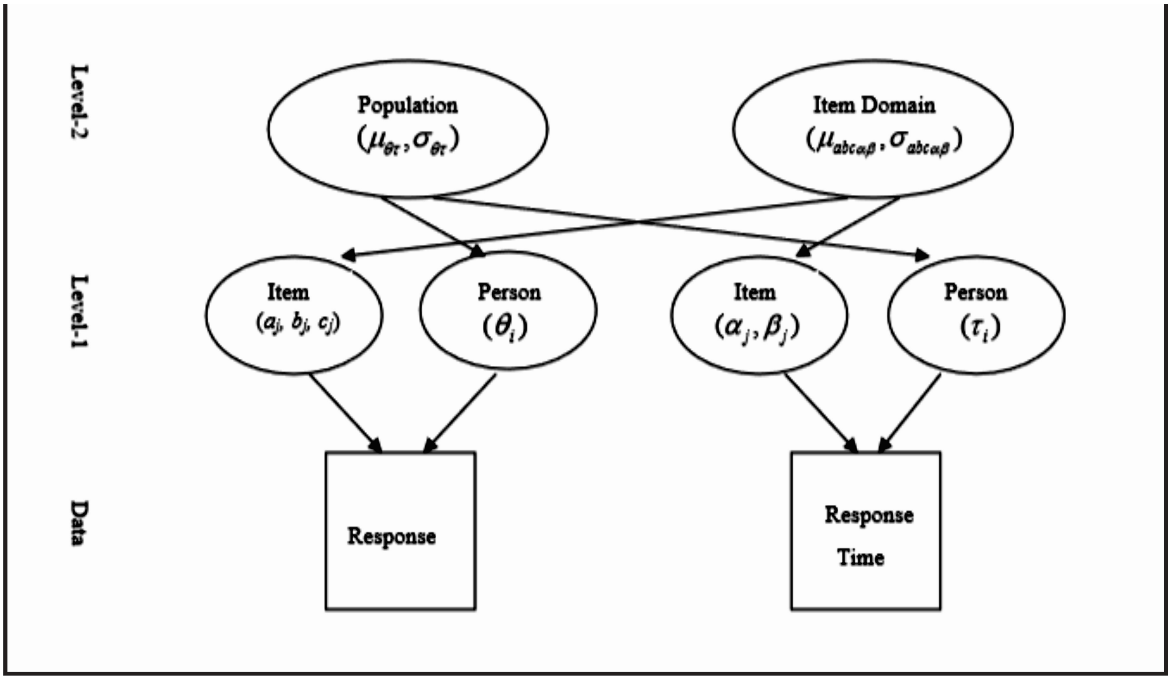


Figure 2.6g: The Joint approach to modelling examinee's parameters on items (van der Linden, 2007)

A usual 3PL IRT model is used in Level-1 for response model, which is given as follow:

$$P(x_{ij}=1|\theta_i,a_j,b_j,c_j)=c_j+\frac{1-c_j}{1+e^{-1.7a_j(\theta_i-b_j)}}.$$

A response time model is a lognormal model as follows:

$$f(t_{ij};\tau_i,\alpha_j,\beta_j)=\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\{-\frac{1}{2}[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))]^2\}, \quad \textbf{...eqn2.51}$$

$t_{ij}$is the time of response of student *i* on item *j*, τ_j represents speed parameter of student *j*, $\alpha_i$gives the discriminating parameter of question *i* with respect to timingwhile β_j depicts the intensity for time for item *j*.

However, a bivariate normal distribution for student's parameter and a multivariate normal distribution for item characteristics for response and reaction time models are contained in the second level of the model.

$$\left.\begin{aligned}&(\theta,\tau)\sim N(\mu_P,\Sigma_P),\\&\text{where}\\&\quad\mu_p=(\mu_\theta,\mu_\tau)\\&\quad\Sigma_P=\begin{pmatrix}\sigma_\theta^2&\rho\\\rho&\sigma_\theta^2\end{pmatrix}\end{aligned}\right\}\text{Distribution of person parameters}$$

And for item parameters,

$$\left.\begin{aligned}&(a_j,b_j,c_j,\alpha_j,\beta_j)\sim N(\mu_I,\Sigma_I),\\&\text{where}\\&\mu_I=(\mu_a,\mu_b,\mu_c,\mu_\alpha,\mu_\beta)\\&\Sigma_I=\begin{pmatrix}\sigma_a^2&\sigma_{ab}&\sigma_{ac}&\sigma_{a\alpha}&\sigma_{a\beta}\\\sigma_{ba}&\sigma_b^2&\sigma_{bc}&\sigma_{b\alpha}&\sigma_{b\beta}\\\sigma_{ca}&\sigma_{cb}&\sigma_c^2&\sigma_{c\alpha}&\sigma_{c\beta}\\\sigma_{\alpha a}&\sigma_{\alpha b}&\sigma_{\alpha c}&\sigma_\alpha^2&\sigma_{\alpha\beta}\\\sigma_{\beta a}&\sigma_{\beta b}&\sigma_{\beta c}&\sigma_{\beta\alpha}&\sigma_\beta^2\end{pmatrix}\end{aligned}\right\}\text{Distribution of item parameters}$$

Every independent time of response and models for response is contained in the first level, but then the structure of covariance of the parameter for first level models is embedded in level 2.

A basic assumption that an examinee must work under constant ability and speed is a criteria, evidence that student's correct proficiency and speeded structures are controlled by the exchange of speed and accuracy. Once the persistent level of the student's speed is known, distribution of response and time will be anchored on the promptness. This will make reaction times to be temporarily independent of speed. Nonetheless, respondents' population, their capability and their reaction speed are dependent.This will make the upper level model's population to indicate how dependent it is (Van der Linden, 2006).

2.3 Conceptual Framework

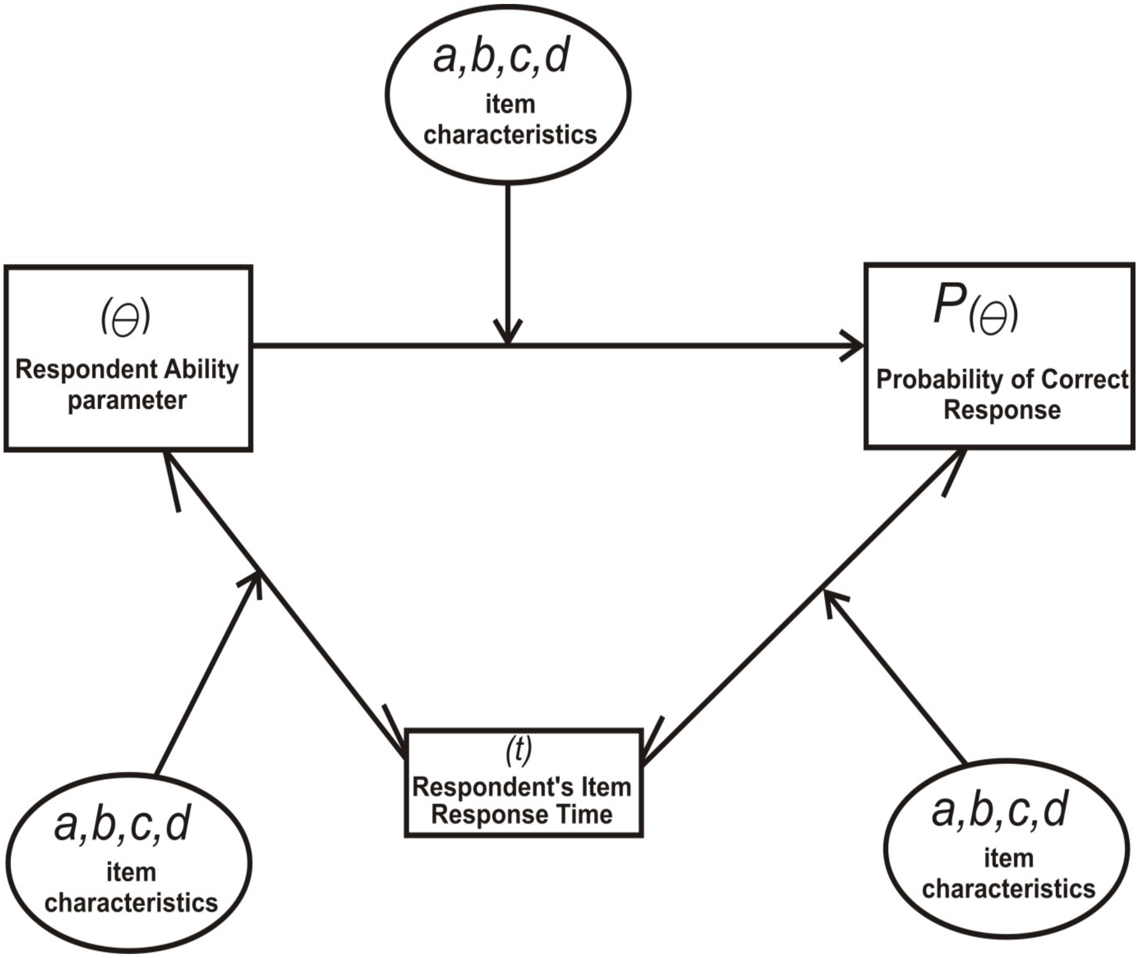
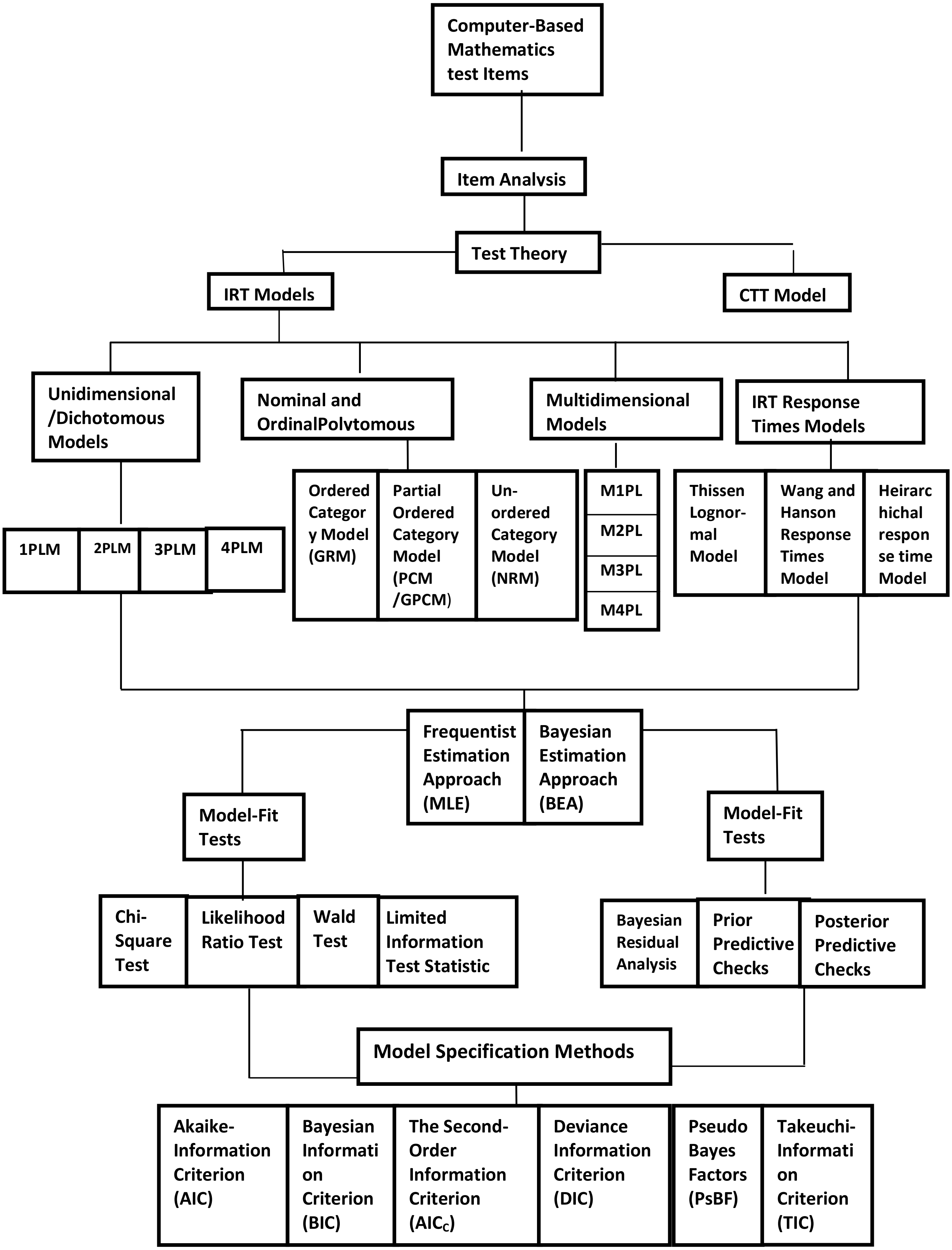


Figure 2.61: Conceptual framework for the study(The Researcher)

Figure 2.61 is showing the interconnectedness of the various components that make up this study. Respondent's ability parameter is depicted to be directly proportional to the likelihood of right answer. An indication that when examinee proficiency increases on the continuum scale, the likelihood of correct response also increases and vice-versa. Aside this, item response time is termed to be inversely proportional to both examinee's proficiency and the likelihood of accurate answer. While lesser response time is tantamount to producing more of the respondent's ability and the likelihood of correct response estimates and vice-versa. Meanwhile, these three components; *θ, P*(*θ*) and *t* mediate *a, b, c* and *d* parameters and this is thereason for direct proportionality among them. Figure 2.62 is showing different calibration methods, model-fits and model selection approaches that are available in IRT measurement framework.



Computer-Based Mathematics test Items
Item Analvsis
Test Theory
IRT Models
CTT Model
Unidimensional /Dichotomous Models
Nominal and OrdinalPolvtomous
Multidimensional Models
IRT Response Times Models
1PLM
2PLM
3PLM
4PLM
Ordered Category Model (GRM)
Partial Ordered Category Model (PCM /GPCM)
Un-ordered Category Model (NRM)
M1PL
M2PL
M3PL
M4PL
Thissen Lognor-mal Model
Wang and Hanson Response Times Model
Heirarchichal response time Model
Frequentist Estimation Approach (MLE)
Bayesian Estimation Approach (BEA)
Model-Fit Tests
Model-Fit Tests
Chi-Square Test
Likelihood Ratio Test
Wald Test
Limited Information Test Statistic
Bayesian Residual Analysis
Prior Predictive Checks
Posterior Predictive Checks
Model Specification Methods
Akaike-Information Criterion (AIC)
Bayesian Information Criterion (BIC)
The Second-Order Information Criterion (AICc)
Deviance Information Criterion (DIC)
Pseudo Bayes Factors (PsBF)
Takeuchi-Information Criterion (TIC)

Figure 2.62: Different Methodological Approaches in IRT framework(The Researcher)

2.4 Review of Related Empirical Studies

Empirical studies of various concepts ranging from one to four parameter logistic IRT models andthe usage of the different response times models such as lognormal models, hierarchical model and Wang and Hanson models. Studies on computer-based testing and those relating to examinees' performance in different learningoutcomes were also considered.

2.4.1 Studies on 1-, 2-, 3- and 4PL Models of Item Response Theory

Alordiah (2015) applies the Rasch/1PL model on mathematics achievement objective test (MAOT) made up of 50 questions. A sample size of 1003 students in public schools in Delta and Edo states were selected using proportionate stratified random sampling approach. The study established how well the items of the MAOT were able to fit the Rasch model and undimensionality of the items was also established. Item difficulty estimates ranged from -1.36 to 1.74 which means that MAOT spreads across wider interval of estimates on the ability continuum of senior secondary school mathematics respondents. The estimated examinees' ability spanned -4.06 to 2.93, and its associated standard errors ranged from 0.29 to 1.01 with mean 0.04.This suggests that 4% of the total variance connected with examinees ability was attributed to error variance whilethe one attributed to true variance is 96%. The result of the study indicated that almost all the items fit the Rasch model apart from 3 items. The study also showed MAOT as an effective and reliable scale that covers a wide range of abilities of the students.The items of the instrument had a good measurement precision. The effectiveness of the model on MAOT was also confirmed.

Ojerinde, *et al*. (2013) examine how students performed in a comparative study of students' performance in English language in the paper-pencil University Matriculation Examination (UME) and that of Post-IRT computer-based Unified Tertiary Matriculation Examination (UTME). 1000 examinees that had scores in the pre and post IRT years were randomly selected and analysis was done with various statistics including Xcalibre to calibrate items and persons parameters for all the candidates' scores. Examination questions in 2012 were analysed with CTT model where discrimination and difficulty indices were determined. Items used in 2013 were also analysed with the 3PL IRT model. The reliability coefficients, means and standard deviations of the items used in 2012 UTME were 0.8202, 52.8201 and 18.1902 while that of 2013 UTME were 0.9473, 52.5983 and 18.4575 respectively. It was however concluded that the items used for 2013 were more reliable than those used in 2012.

A major progress was recorded on examinees that rewrote the test in 2013, compared with the ones who sat for the UTME Use of English in 2012. Analysis indicated that although both measurement frameworks (CTT and IRT) were beneficial to experts in test construction in understanding and assessing mental phenomena and construct,test professionals ought to discover the advantages inherent in the usage of IRT which offers preferable outcomes over CTT system. It was concluded in their study that IRT approach gave a better measurement results by providing more reliable and greater information about the behaviour of items.Model validity was also tested and recommended that IRT application for making usable judgement on properties of items that ought to be used by examiners.

A self-report delinquency instrument that has multinomial responses was analyzed by Osgood *et al*. (2002) adopting the IRT method of 2-parameter graded response function. Then, a good fit of the response data was delivered by generating more reliable information when likened to the direct CTT total score approach in the scale. However, a probability that the most delinquent youth would not exhibit some of the aberrant acts was found. Their submission was that imminent study should review a higher parameter models (4PM model).

Another study by Loken and Rulison (2010) employ Bayesian method in standardizing questions in a 4PL model with MCMC approach The Bayesian approach improved effectively estimates of parameters for item and examinee by the data produced with a 4PL model. The finding also found an overall fit when 4PM was used in place of 3PM or 2PM. Another study where computerized adaptive testing (CAT) environment was used indicates that the influence of prior mistakes committed by brilliant students (as a result stress) could be greatly lessened by applying a four-parameter logistic model (Rulison and Loken, 2009).

Results are likely to be different should a lower parameter model like a 2PM or 3PM be used to analyze 4PM data. For 3PM with WinBUGS software, the slopes were significantly lesser and an approximate value of 0.8 was recorded for 3PM and 0.5 for 2PM with the three types of test when likened to the mean slope of 1.10 for 4PM. Difficulty values moved higher by approximately half a standard deviation for 3PM and shifted backward closer to 0 for 2PM.This reflects the importance of allowing the data at each tail of the ability continuum. Also, in comparing results gotten to the ones for the 4PM, the values for c_j in the 3PM had slightly higher Root Mean Square Error with a bit lower correlation values with the true scores.

The estimates for examinees' trait for 3- and 2-parameter models correlated so greatly with the ones gotten for 4PM and none of the models showed impartially in their parameter estimates. Also, the estimates acquired for the posterior standard errors for 3- and 2PM were lesser than that of 4PM which brought about reduced coverage of the 95% intervals for ability trait, most especially at the tail end of the distribution. Additionally, an unreasonably wide interval existed in the confidence intervals for $\vartheta<-1$ for 3PM and a high coverage too. On the other hand, the ranges are excessively tight at the higher tail ($\vartheta > -1$) which made the interval coverage to be smaller than 95%. In the case of 2PM, there was hindrance at the two tails of the distribution. It was however inferred that despite the separate attribute estimates being practically indistinguishable under the three IRT models, conclusions were vigorously influenced by the utilization of an inappropriate model.

Reise and Waller (2003), in a psychopathology research consider the necessity for the fourth parameter when responses of the Minnesota Multiphasic Personality Inventory (MMPI) were being modeled. MMPI happens to be a scale with dichotomous items in which examinees answer questions that could be right or wrong of them. It was discovered that a lower asymptote that was more than zero was needed by some items as a result of some examinees that appear below average in the latent attribute with a non-zero likelihood of answering the questions. The finding was that many questions necessitated a lower asymptote that is non-zero which signifies that at the initial inputting, an upper asymptote less than 1 should have been modeled with the items. Therefore, the overall fit of the model was not necessarily enhanced by shifting to 3PM, but that when the model was estimated as with 2-, 3- or the reverse code 3-parameter models, the test information function was different. A new model with the fourth parameter, signifying both the careless as well as guessing parameters, is better used in modeling certain medical and personality scales.

Opinions that support 4PM usage for analysis in IRT also surfaced in research that pertains to genetics. Tavares, Andrade de and Pereira(2004) showed that if some genes were made active or incapacitated in persons as a function of certain features of the individual, like anindividual's predisposition to a sickness,items were seen as genes while persons with higher disposition were taken to have better probability of having the genes activated. Although, persons with low disposition could be having the gene being activated, those with better disposition could as well have such genesdeactivated.It is,therefore, necessary that such is tested. 4PL model was proposed by the reseacher.

Ani (2014) develops and validate a 50-item multiple choice economics test for 1005 senior secondary school II students in 46 co-educational schools with the application of IRT framework. Maximum likelihood estimation technique of BILOG-MG computer programming was adopted in analyzing the data generated so as to estimate model's parameters used to fit the data. The results indicated that 49 questions of the economics objective test were good based on the 3PL model used, 31 items of the instrument were difficult and function differentially between male and female examinees in economics. The researcher concluded that since IRT approach provides better information as far as item quality is concerned, examination bodies and teachers are encouragedto adoptthe framework in item development and validation.

Metibemu (2017) carried out a comparative study between CTT and IRT in developing, scoring and equating senior secondary school physics achievement test.Factor analysis was adopted to validate 100 physics multiple-choice items to show that the items measured students' proficiencies in physics. The results discovered that 2PL model deleted items that were too difficult and could not discriminate well among high and low achievers. It was also shown that IRT framework was better in the construction and validation of items of a test.

Kpolovie and Emekene (2016) employ IRT approach in the validation of Raven's Advanced Progressive Matrices (APM) instrument in Nigeria. APM is popularly used in America, Europe and Asia and happened to be a prominentnon-verbal mental ability test globally but has never been validated for use in Nigeria. The scale was meant for identifying persons with strongintellectual skills that can handle demanding study programme and be able to manageboth complex and ambiguous cases in the modern workplace. A sample size of 2100 examinees was randomly drawn and the test yielded a favourable statistics under 3PL IRT model. This recommended the use of the test as it has been localized.

In another study by Anamezie and Nnadi (2018), it was stated that the traditional method of ascertaining statistical quality of test items had been found defective.A cognitive diagnostic modelling (CDM) approach was employed on a 50-item teacher-made physics achievement test. An approach that was an offshoot of IRT models was used. It was meant for estimating four-item parameters as discriminating, difficulty, guessing, slipping/carelessness with the latent skills mastery profile using a Deterministic-Input-Noisy-and-Gate (DINA) model, a sub-model in CDM. DINA is a non-compensatory model in which the examinee's likelihood of answering a question rightly increases if the examinee possesses all the required attributes for a particular item (George and Robitzsch, 2015). Study's outcomerevealed that 18 questions in the test fit the DINA model, 32 items had misfit, and 9 attributes were mastered by the examinees while unmastered ones were 3.

2.4.2 Studies on the usage of the Frequentist and Bayesian Estimation Approachesfor Item Parameters Calibration and Ability Estimation

Loken and Rulison (2010) conduct an imitation study to clarify how a Bayesian method can be applied to assessing 4PL model with three different credible situations that researcher could come across in either education or psychological enquiry. The aim of their research was to exhibit the likelihood of evaluating 4PM. This was carried out using instruments with items of different length; one with 15 items, the other with 30 items while the last instrument had 45 items with a sample of 600 respondents that was produced in a normal population with mean 0 and standard deviation 1. It was concluded that good estimates with a modest sample size of 600 was achieved. However, the extent of steadiness in various settings as well as diverse selections of the distribution of priot for guessing and carelessness parameters has need of systematic investigation.

A study by Swaminathan and Gifford (1986) demonstrates that when Bayesian method was used in estimating the reliability estimates of guessing parameter (c_j) of 3PM against the use of Maximum likelihood method, there was a great improvement in the estimate. Furthermore, the use of Bayesian approaches have demostrated that it gives helpful insight in producing very suitable methods for estimating seemingly complex and profoundly parameterized models that have the chance of non-normal and multimodal characteristics when the upper asymptote (d_i) parameter is seen as item-specific. Consequently, the Bayesian method appears an appropriate approach to attain reliable estimates of d_j.

Cengiz and Ozturk (2013) applied Bayesian methodology in IRT framework in evaluating progress examination of the undergraduate medical students in the 4th academic year at Ondokuz Mayis University. Although IRT method depends on strong assumptions which are beneficial and practicable in various circumstances in tests relating to medical education, some of the assumptions are often time overulled and the statistical inferences arethreatened. As a result of this, the Bayesian approach to estimation in IRT was adopted to analyze such data. 2PL IRT model was used because the items were short-answer scored dichotomously with Winbugs software.

4997 iterations were run with algorithm for the first 1000 iteration as a burn-in. Classic and Bayesian item response models were differently fitted to their progress test data where both Akaike as well as the Bayesian Information Criteria were regarded as the summary assessment of fit. It was discovered,using the two estimated models, that the information criteria values for Bayesian IRT were considerably less than those for classical IRT which made it the one to be preferred (Classical IRT: 314.451, 289.45; Bayesian IRT: 246.853, 214,158). Since the findings found support for the use of Bayesian method, it was therefore concluded that Bayesian method could be adopted for general use in item response analysis.

2.4.3 Studies on Item Response Time Modeling

Kyllonen and Zu (2016) explored response time application in assessing mental ability of the examinees. A speed-level variation, magnitude of speed as well as level of reasoning abilities structures with speed-accuracy tradeoff, especially IRT-based and response time model were considered. Different cognitive psychological models such as the ex-Gaussian and diffusion models together with the employment of other response time models in assessment apart from ability measurement were utilized. Many innovative methods that provided more understanding to speed and level components of mental capacity and speed–accuracy exchange resolutions were discussed in the study. Such methods as item to level time bounds, the feedback approach (cumulative sum; CUSUMs), unambiguous scoring guidelines that combine the information for speed and accuracy (count down timing) in addition to thinking psychology models which are CTT-based were also discussed.

Thiessen model according to Kyllonen and Zu (2016) happened to perhaps be the foundational item response theory model for responses and accompanying reaction times for timed tests. The study suggests joint model of response and response times and adopted the regular 2PL model in modeling the response function with effective ability (θ_j), item difficulty (b_i) and item discrimination (a_i). Thiessen model was used with data from three dissimilar tests that included Verbal Analogies, the Progressive Matrices, as well as, the Clocks.

Result from the analysis of the Progressive Matrices data showed that examinee ability and slowness seemed to be highly positively correlated which indicated that assessment done in limitless period basically estimated slowness (a moderately stress-free test). On the other hand, data on Verbal Analogies showed that ability and slowness were not as associated as in the progressive matrices data.Nevertheless, a positive relationship existed. Whereas data in Clocks scale revealed a very low and inverse relationship which was a pointer to separate spatial ability and speed estimates. It was therefore concluded that assessing proficiency in a joint response time model raised an inquiry that relates ability measurement in the conventional IRT model.

An exploratory research that was carried out by Partchev and De Boeck (2012) explores the impression that different kinds of processing existed qualitatively and so diverse abilities are raised within a presumably single-construct test. A dichotomy of singular person's response times was made from two out of the three tests used by Thiessen study. This was classified into fast and slow group that were centered on either examinee's median or median response time to individual question. The response classifications were based 4-Likert type response categorywhere 1-parameter logistic model was fitted. The finding was that three distinct underlying abilities were needed to assess the data in the speed, slow and fast intelligence. It was as well discovered that though fast and slow intelligence were greatly associated, however, there were distinctions.

Van der Linden (2007) models both a public and powerful two-level hierarchical function strategy of response/reaction times. This model assumed a restrictive independence of the two variables on some unique and dissimilar questions with examinee's ability and speed. A 3-parameter normal-ogive item response model was used with the log odds for examinee answering rightly to a question. The study indicated that when response time is considered, the tendency of reduction in bias rate and error incurred in ability estimates is guaranteed. Then, correlation between proficiency and speediness estimates increased which in turn increased the extent at which the estimates are improved. Glas and van der Linden (2010) also subscribed to the fact that error associated to examinees' ability is lessened.

Thiessen's (1983) lognormal response time model was applied to three different data.This raised the question of how ability measurement made in the joint model (response/response time) was related to the ability measurement made in the stand-alone traditional model (response-only). It was found that the nature of the ability being measured was changed by time-limit.

The application of 3PL model was done by Scrams and Schnipke (1997) to CBTs of verbal, quantitative and thinking aptitudes. The finding revealed reasonable correlations amidst examinees' reaction speed and capability with difficulty of item for three tests. Meanwhile, Swygert (1998) made use of the revised form of Thiessen's (1983) model in evaluating item reaction time on Graduate Record Examination with Computer Adaptive Testing. Reasonable positive association existed between reaction quickness and student ability values for the two test forms. Thiessen's model was also utilized by Ingrisone (2008) where comparison of MMLE method with a maximum a posteriori (MAP) procedure was made. Both strategies were seen as steady and precise when three different simulation studies were carried out to estimate both item and person parameter estimates.

2.4.4 Studies on E-Assessment (Computer-based testing)

Eccles, Haigh, Richard, Mei and Choo (2012) in a study on implementing physics e-assessment, estimate the influence of high-stakes CBT on students in six schools in Singapore. They emphasised the idea of making sure that students' viewpoints were sufficiently reflected in the inculcation of new knowledge into educational practices. During their survey, an appreciable understanding of students' familiarity of e-assessment was gained imbibing triangulated research methods to investigate response from learners and assessors. This approach was added to certain quantitative methods of item-based test statistic.

Three facets of high-stakes CBT were reported: the examinees' know-how of test, how CBT fared as an assessment scale, and the managerial take on running a test. The test was taken in 45 minutes by 144 students who were requested to initially take a 15 minutes practise exam to get acquainted with the assessment condition. The findings indicated that examinees can cope with CBTs assuredly and are ready to be tested in a computer-based setting. Further outcomes revealed that having the privilege of beingassessedin the world of technology could bring about more attractive and reliable testing state that might not be achieved within the constraints of paper-based testing.

Okorie and Mojiboye (2015) made a comparative analysis of the test necessity for the conducting UTME PPT mode in 2011 and UTME CBT mode in 2015. The essence of the research was to view the influence of computer on the items in other to plan the conduct of the different tests in the two years. Their research was a descriptive one that involved the use of data collected on human and material resources and other logistic requirements for the conduct of the examinations in the two years. The results showed that in the 2011 UTME, more personnel were involved than in 2015 UTME. On the number of material requirement and planning arrangement, 2011 UTME required more materials than 2015 UTME.On the number of application and personnel deployment, it was shown that 1,409,462 applicants and 424,315 (98%) personnel were deployed in the PPT mode which was much higher than the number deployed in 2015 CBT mode. It was concluded that the use of computer systems in the modern day test practice is one of the developments that unravelled most problems that accompanied the conventional assessment practice.

2.5 Appraisal of Literature and Gap

The adoption of the modern-day theory of measurement in the analysis of various types of assessment scales/instruments has come to stay. Many researchers, psychometricians, test developers and even test-users have made use of the various approaches embedded in the new theory as a result of its invariance property and themore objective result it gives.

A research study, comparative analysis of the application of the traditional and modern day theories was carried out on the analysis of data collected using English Language as a case study. The findings of the study proved that IRT method was a better measurement theory as far as test development and item analysis are concerned. This was due to the provision of more information it supplied about the behaviour of items and ability estimates of the candidates. Meanwhile, the IRT parameter model used was limited to 3PL model. Model-data fit assessment was not extended to 4PL model as it was considered in this study and response times of the items by the students were not applicable to view their effects on the abilities that were estimated.

In another study, comparability of the traditional as well as the modern approaches to testing in constructing, scoring and test equating of a 100 objective physics achievement test was assessed. The result indicated that IRT fairedwellmore than CTT framework in calibrating and analysing test items.This also agrees to the findings of the previous study.

A study on implementing e-assessment in Singapore highlight the importance of CBT type as it adds to the educational value of the present age. The researcher suggests upcoming research to take account of the examinee perspectives on assessment of CBT. Difficulty that students encountered in interpreting graphs and the use of computer-generated devices (ammeter, stopwatch) with responding to items on CBT were also examined. It was suggested that further studies could recognise which segments of technology could improve the forms of items students respond to so as to inquire intensely into students' knowledge by providing images or appropriate on-screen tools to aid the reading of graphs.

Other worries pinpointed in the work could be addressed by continually making available the needed technological know-how while teaching and learning is taking place. Availing student necessary opportunity to practise test exercises on computers throughout their course of study in school. This study however, improves the form of CBT that was given to the students in terms of provision for soft-devices on the screen while answering to the CBMAT scale used for data collection for this study.

In another development, one of the recent studies on the utility of 4PL model was able to achieve worthy estimates using 600 examinees as sample size in an empirical study. Provision of a new look at a popular delinquency scale was achieved. It was then noted that the extent of reliability with diverse conditions and selections of prior distribution or the lower and upper asymptotes parameters necessitated systematic investigation. A suggestion was made that more observed studies be conducted to justify greater use of the 4PM. However, this study made use of a larger sample size (874).A suggestion that ensued from the previous work was that more could be done to investigate model fit and other standard measures of item and person fit analysis. Meanwhile, the recognition of this fact that much more works are needed to further establish calibration and interpreting the parameter estimates of 4PL model prompted this research work.A model fit assessment was done with the estimates of -2Loglikelihood, AIC and BIC information criteria which gave good parameter estimates.

In recent times, attention has been shifted to the several advantages and benefits IRT framework has brought either in overcoming the many shortcomings CTT approach was fraught with or in complementing its methods to further enhance objective measurement. Although, several models are inherent in IRT approach which are usable in diverse assessment procedures ranging from one to two and three unidimensional models up to the several multidimensional and polytomous models. Many of these models are still limited, in terms of what they could do. This study utilized newer and recent models to verify their underlying advantages as against the many previous available models research has recorded.

From literature reviewed so far, it appears there was no other study that had attempted the usage of the 4PL IRT unidimensional model in this clime. Of course, the major reason has been ascribed to the lack of appropriate statistical software that could estimate the heavily parameterized model. Another laudable gap this research work filled is the fact that the response times that were automatically recorded as students respond to the computer-based mathematics test (CBMAT) items, were used in estimating their abilities in a joint response/response time and in stand-alone response model. A comparison of ability estimates in the stand alone conventional model as well the Lognormal response time model was done.Research in the developed world has produced a more reliable and better ability estimates with 4PL model. This study seems probably the first in Nigeria's research community to have explored the combination of both response time and the conventional IRT models in estimating model parameters.

Also, the deviation from the popularly used frequents approaches (MLE) in estimating parameters of model to the use of Bayesian approach which seems suitable for estimating heavily parameterized model, is another gap this study filled. This approach is considered better with the use of other model-fit approaches involving MCMC diagnostics of Bayesian method as against the popularly used model-fit method, Deviance Information Criteria (DIC). Model-fit assessment result with Full Information Item Factor Analysis (FIFA) of Multidimensional Item Response Theory

(MIRT) package via R-platform software showed that 4PL model fitted the pooled and the final CBMAT response data.

It is on this note that 4PL model was developed with the introduction of the 4th parameter (carelessness) to cater for some other measurement errors the previous models could not address. Such errors were caused by extraneous variables that could hinder objective measurement. But for the lack of consensus and suitable software to estimate its parameters, there were constraints in its utility before now. This study therefore explored the applicability of 4PL model and employed Logmormal response time model in calibrating computer-based mathematics achievement test among senior secondary schools in Lagos and Oyo States, Nigeria.

CHAPTER THREE
METHODOLOGY

This chapter is concerned with the description of the various procedures and methods the research work followed. These include the design of the research work, population of the study, sampling technique and sample, instrumentation, procedures for how data was collected and analysed with the methodological challenges confronted in the study.

3.1 Research Design

Instrumentation design was adopted for the study. The designs enabled the use of systematic statistical procedures in the collection of numerical data to assess, explain and authenticate the research questions posed for the study. It also seemed preferably suited as it helped in the scientific study of educationally significant problems, provided suitable information within sampling error and assisted in solving some of the large disputed issues in education.

3.2 Population of the Study

The population for this study consisted of the entire senior secondary school two (SSSII) mathematics students in all government-owned senior secondary schools that had functional computer laboratories in Lagos and Oyo States.

3.3 Sampling Technique and Sample

Multistage sampling procedure was considered in this study. Sampling was done in two phases such that sample sizes needed for schools' and examinees' representatives at both trial-testing and main study stages were achieved.

The six south-western states are naturally stratified into two distinct strata (the Coastland and the Inland regions). The coastland is made of two states (Lagos and Ondo) while the inland region had four states (Ekiti, Ogun, Osun and Oyo). Purposive sampling was adopted in selecting onestate from each of the two area. Lagos state was selectedfrom the Coastland while Oyo was chosen from the Inland area. The reason behind the choice of Oyo and Lagos was due to their proximity to the researcher and more importantly some laudable 21st computer educational projects that had beendone in most of their secondary schools.Examples of such projects are the Google Africa Code Week in 2018 and 2019, the alumni support (Old Student's Association of schools), foundations and international donors support (World Bank, Unicef, UNESCO and the British Council) as well as the Lagos Eko Project.

3.3.1 Sampling Procedure and Sample for Phase I: Oyo State

The first phase of sampling involved the trial-testing stage of the instrument. This was done for the validation and calibration of the pooled 114 items of the computer-based mathematics achievement test (CBMAT). There were six educational zones in Oyo state and three of them (zones 1, 2 and 4) were purposively selected. This was as a result oftheseemingly available functional computer resource centres and computer systems in the zones.Fifteen schools with available computer systems were also chosen purposefully from the 3 zones. One of the 15 schools was randomly selected to test run the self-developed pooled CBMAT program to ascertain its effective usage in the cause of eliciting right response from the examinees. The remaining 14 schools were used for phase I.

Stratified random sampling was adopted in selecting examinees from each of the sampled schools with respect to the different arms of SSS II (science, commercial and art) classes.Seven hundred and thirty one examinees were selected as the sample size for phase I. This was shown in Table 3.1.

Table 3.1: Sample distribution for phase I

State	Education Zone	LGAs	No of selected schools	Sampled Examinee
Oyo	Zone 1	Ibadan North	385	
		North East	2	83
		North West	1	20
		SouthEast	2	109
		SouthWest	3	174
	Zone 2	Egbeda	1	70
	Zone 4	Atiba	1	56
		Oyo-East	2	134
Total		**15731**		

3.3.2 Sampling Procedure and Sample for Phase II: Lagos State

Lagos state was already stratified into six educational districts with 20 educational zones that cut across all the 20 local government areas (LGAs) of the state. Out of the six education districts, Lagos Education District I that comprised Agege, Alimosho and Ifako/Ijaye LGAs was purposively sampled. The district was selected due to the landmark achievements in the provision of computers. This was made possible due to the district collaboration with some Old Students Associations, elected public office holders, Rotary Club, Parents Forum, MTN Network Provider, Redeemed Christian Church Support, GT Bank and Etisalat in 2015, 2016 and 2017 (Education District 1 website, 2018).

A list of senior secondary schools with functional computer laboratories was collected from the district headquarters, which granted the researcher access to inspect the systems. Thereafter, the researcher purposively selected eight schools across the three education zones. In both Agege and Ifako/Ijaye education zones, three schools were chosen from each of them while from Alimosho zone, two schools were selected.Examinees from SSS II were naturally stratified into three strata (science, commercial and art classes). Randomly sampling was adopted to select students from each of the classes and 874 students were selected.Table 3.2 gives the sampledistribution for phase II.

Table 3.2: Sample distributionfor phase II

Educ. District	Educ. Zones	Name of selected schools	Sampled Examinees
1	Agege	Government Senior College	121
		Girls Senior High School	56
		State Senior High School	127
	Ifako/Ijaye	Sonmori Senior School	75
		Vetland Senior Grammar School	144
		Keke Senior High School	144
	Alimosho	Lagos State Model School	126
		Tomia Community Senior Sec. School	81
Total	**3**	**8**	**874**

3.4 Instrumentation

An initial draft of the scale with 120 items was constructed by the researcher according to the carefully arranged and revised 2008 mathematics curriculum for Senior Secondary School I of the Nigerian Educational Research and Development Council (NERDC). The development of items was aided with the revised New General Mathematics for SSI textbook (2011 edition) that reflected the scheme of work that was logically arranged to cover different topics.

Computer-Based Mathematics Achievement Test (CBMAT) was the only instrument used in this study. The CBMAT instrument was of two types. (i) The pooled CBMAT that consisted of 114 items was validated and IRT item analysis was carried out to delete poor items. This scale was used for the trial-testing phase of the research. (ii) The 40-item final CBMAT scale was used to elicit responses and response time data for the calibration and estimation of item and examinee parameters of the uni-dimension IRT dichotomous response-format models and the adopted joint Log-normal response time model (LNIRT) for the main-study phase.

Multiple-choice type of objective test was adopted with four (4) response options (A-D), three of which served as distracters while the remaining one was the key. Response and response time data for the items were automatically recorded while testing with CBMAT instrument,students' responses were dichotomously scored (1: correct response; 0: incorrect response) and response time data was recorded to certain seconds or minutes' precision on the CBMAT. Time recorded in seconds against an item became the aggregate time spent on that particular question in the cause of responding to that item. The CBMAT instrument was of two sections.

Section A elicited information pertaining the examinees demographic data which includedstudent name, name and location of school, respondent's gender and class type (science, commercial or art). Section B contains items relating to the content of the curriculum which covered the entire three terms (first, second and third) in a session. The CBMAT scale developed was administered to the examinees through computers. A computer program with necessary command codes, algorithms and flowchart to run the CBMAT electronically was developed by a programmer. Each examinee's task was executed as soon as any of the options was selected. Examinees' responses with their reaction time to the questions were automatically documented according to how the programming was set and dichotomously scored for analysis by the system.

3.4.1 Procedures for Construction of Computer-Based Mathematics Achievement Test (CBMAT) Instrument

a) Defining the Purpose of the Test

The purpose for which the CBMAT instrument was developed was to examine respondents' performances with respect to the general knowledge they have acquired in mathematics after teaching and learning had taken place fora whole session. This performance enabled the researcher to estimate respondents' abilities and calibrating item parameters of 4-parameter logistic model and lognormal IRT models. The essence of this was to model students responses correctly so as to depict their true ability for the purpose of which measurement had taken place. SSII students were examined in the first term of a new session (2018/2019) because they had completed the scheme of work for SSI curriculum for the three (3) terms in their previous class.

b) **Outlining the Content**

As itemized in Table 3.3, senior secondary school I mathematics curriculum has four (4) themes and 13 topics. The items of the CBMAT scale were newly constructed by the researcher from the syllabus.The themes and topics are hereby given.

Table 3.3: The Summary of SSSI Curriculum Content

THEMES	TOPICS
1. Numbers and Numeration	1. Number Based System 2. Modular Arithmetic 3. Standard Form 4. Logarithms 5. Sets
2. Algebraic Processes	6. Simple Equations and Variations 7. Quadratic Equations 8. Logical Reasoning
3. Geometry	9. Constructions. 10. Proofs of some Basic Theorems 11. Trigonometric Ratio 12. Mensuration
4. Statistics	13. Data Presentation

Revised NERDC Curriculum for Senior Secondary School (2007)

c) Preparing the Table of Specification

A two-way grid table termed table of specification that link content areas to the behavioural objectives to ensure content validity and comprehensiveness of the test was used. This table is showing the three levels of cognitive domain as remembering, understanding and thinking (the combined higher levels of cognitive domain) (Anderson and Krathwohl, 2001). Percentage weight of items assigned to each cell is decided according to the allotted time of teaching, depth and importance of the topic in the curriculum content. The table of specification for the instrument is given in Table 3.4.

Table 3.4: Table of Specification for CBMAT Instrument

Content Area	**Behavioural Objectives**			**Total**
	Remembering (12%)	**Understanding (20%)**	**Thinking* (68%)**	**(100%)**
Numbers & Numeration (33%)	5items (1,4,6,42,76)	4items (8,9,11,57)	29items (2,12,13,14,17,18,19,20 21,22,23,24,25,26,27, 28,29,30,31,32,33,34, 35,36,37,73,79, 89,92,)	**38**
Algebraic Processes (17%)	3items (44,45,86)	8items (48,49,50,51,52,53, 54,55)	8items (3,7,38,40,41,43,47,56, 93)	**19**
Geometry (35%)	3items (112,113,114)	7items (58,62,80,81,82,90,91)	30items (5,10,15,16,39,46,59,64 68,69,70,71,72,74,75, 77,78,83,84,85,87,88, 94,95,96,97,98,99,100, 101)	**40**
Statistics (15%)	2items (110,111)	5items (102,103,104, 105,106)	10items (60,61,62,63,65,66,67, 107,108, 109,)	**17**
Total (100%)	**13**	**24**	**77**	**114**

Thinking* = Applying + Analysing + Evaluating + Creating

d) Writing, Editing and Assembling of Mathematics items

Item writing was guided accordingly by the table of specification presented in Table 3.4. Items cover the 13 topics to reflect different contents of the curriculum for the three terms in a session. In the course of developing the items, the researcher was mindful of the purpose for which the test was to be administered, the nature of the behaviour being measured and decision on a crude estimate of the level of difficulty of each item.

Test format (Multiple-choice) and content representativeness were taken into consideration in the course of writing the mathematics items. Avoidance of ambiguous statements and irrelevant clues in the item stem and root (correct option and distracters) were bone in mind while writing and editing of items. The quality of an achievement test is the reflection of the extent to which a test constructor has objectively and meticulously operated the procedures of test development to ensure that items conform to purpose (Okpala andOnocha,1995). Therefore, item writing was done in such a way that the expected underlying ability trait, which is to be assessed by the items,was done accordingly (Baker, 2001).

Editing the constituent parts of CBMAT instrument involved experts in test construction and in the subject-area. This process was done with reviewing and critiquing each item with a view to detecting and modifying technical errors. The critiquing was with respect to clarity and conciseness of language of such item. The totality of the draft items that made up the draft copy of CBMAT instrument was 120. This instrument alongside with the table of specification was given to experts at University of Ibadan, Institute of Education and three mathematics teachers in senior secondary schools. This was to establish face and content validity of the pooled CBMAT scale which entails the appropriateness, coverage and clarity in terms of the words used in constructing the items.

Out of the 120 items developed, six items were suggested for deletion by the experts as a result of vague expressions in their stems and ambiguity in some of the options. Eightitems were reframed and readjusted in the remaining items and 114 questions were left in the instrument (Appendix III). These items were separated into two equal halves with odd and even numbered items in distinct parts (Appendices IV and V) for easy administration. This was done to prevent prevent boredom in the course of answering to 114 items at a stretch. The items were organised in a fashion suitable for administration using appropriate computer fonts with the provision of simple and adequate instructions on how to respond to them. Figures 3.1 and 3.2 (pages 132-133) present the program flowchart and the notations of each symbol used in the effective running of the computer-based mathematics achievement test (CBMAT) while administering.

e) Requirements for building the softcopy of the CBMAT program

The CBMAT software was archived on a software development tool called Visual Basic Studio of version 2012. This programming language is normally used for designing a stand-alone software, school management system, supermarket payment software and computer-based testing. Microsoft Excel was as well synchronized in this program to enable effective database utility. The program demanded necessaryhardware and software compatibility.

For proper implementation of the program, the following hardware specifications were used:

- A Pentium 4 processor or higher
- Ram size of at least 512MB
- Enhances keyboard and mouse
- UPS (uninterrupted power supply)

While software requirements include;

- A working Operating System (windows 7, 8 and 10)
- A 64 bits system
- A working Antivirus
- **.**Net Framework version 4.5
- Microsoft Excel (from version 2007 to 2016)

The CBMAT scale was carefully designed for this research work to automatically record respondents’ responses to the items of the scale and their response times which were used for analysis. An offline mode of administration was employed for the program and appropriate measures for a hitch-free test administration process were provided as well. Figures 3.1 and 3.2 depicted the program flowchart and what individual shape implies. Every process of execution is shown as soonexaminees open the CBMAT folders on their different desktops in assessing both the demographic section and the items as appropriate responses were given.

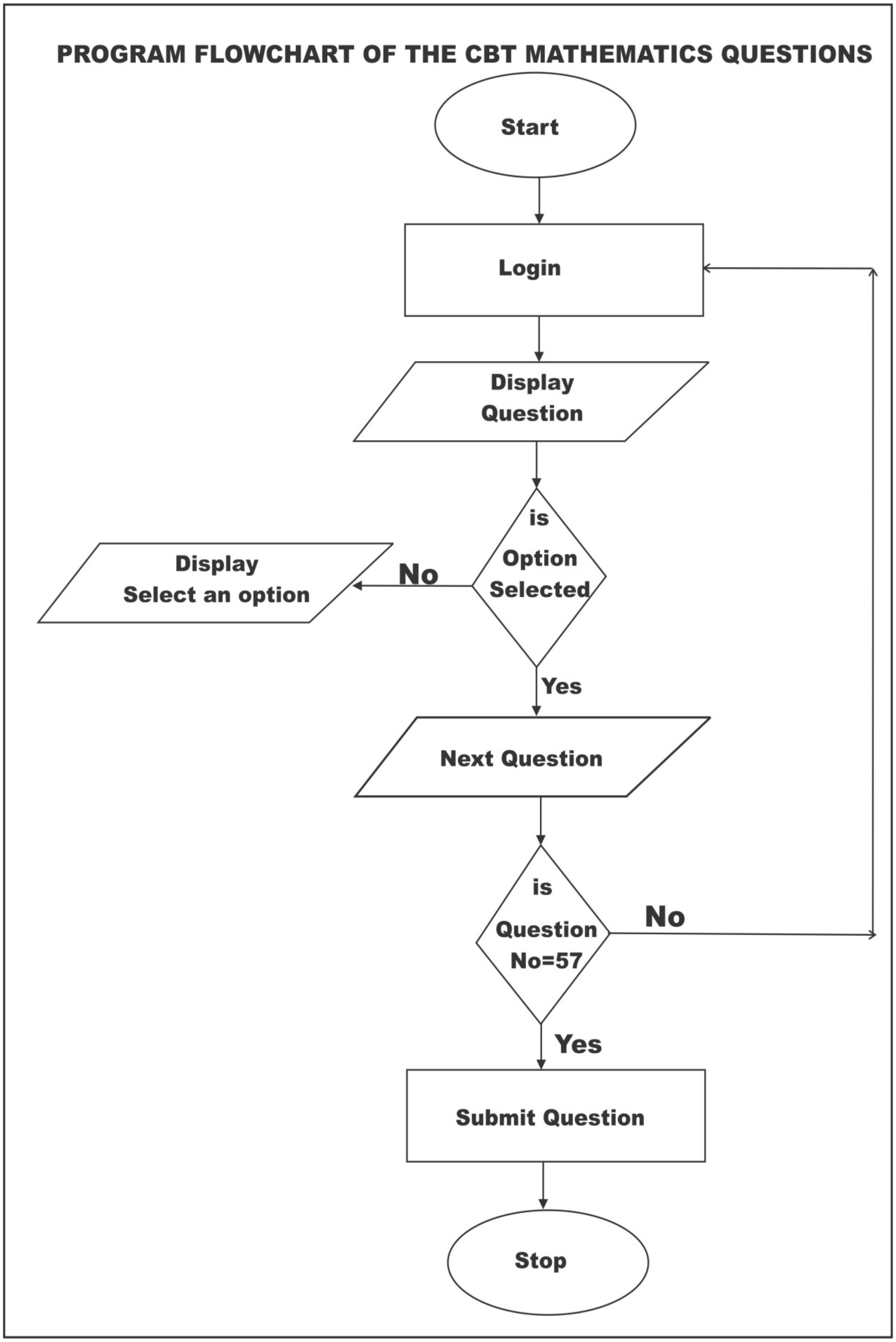


Figure 3.1: The CBMAT Program Flowchart

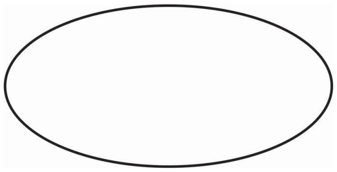
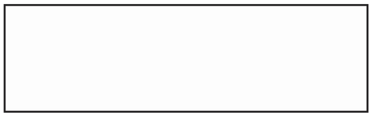
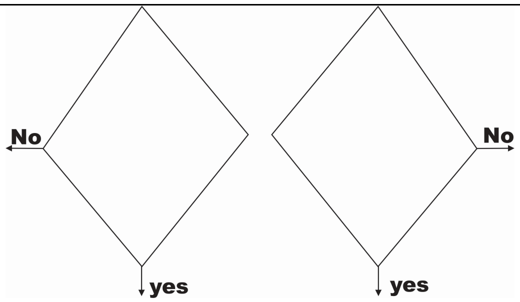
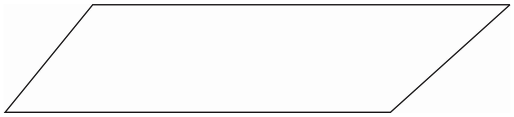
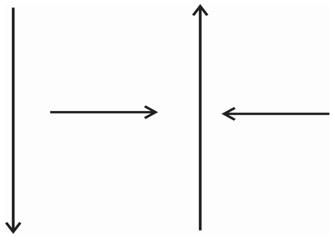
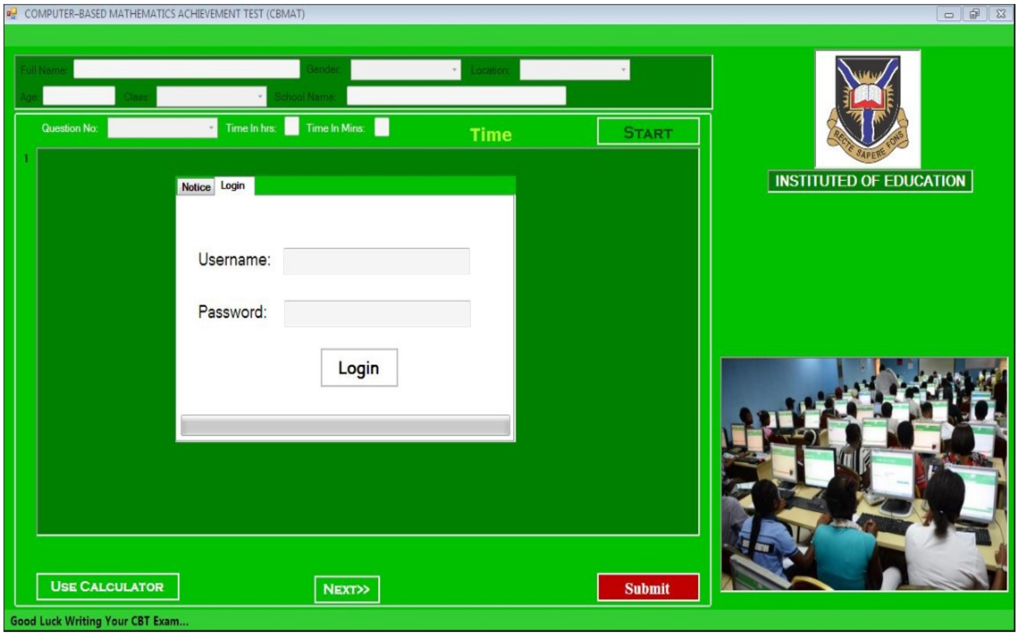
S/N	FLOWCHART SYMBOL	MEANING
1.	Start/Stop 	An oval symbol indicates the start and stop (end) of the flowchart.
2.	Process 	A process symbol indicates the activities that pass through the program.
3.		The decision symbol indicates a control structure/statement. When the condition is true, the Yes direction is followed but when false the No direction is taken.
4.	Input/output 	An input/output symbol is used in displaying output and can as well be used to accept data from the user as an input.
5.	Arrow 	The Arrow symbol represents the logical flow on the program. It shows the direction of the activities or process in the flowchart.

Figure 3.2: Flowchart Symbols and their meanings

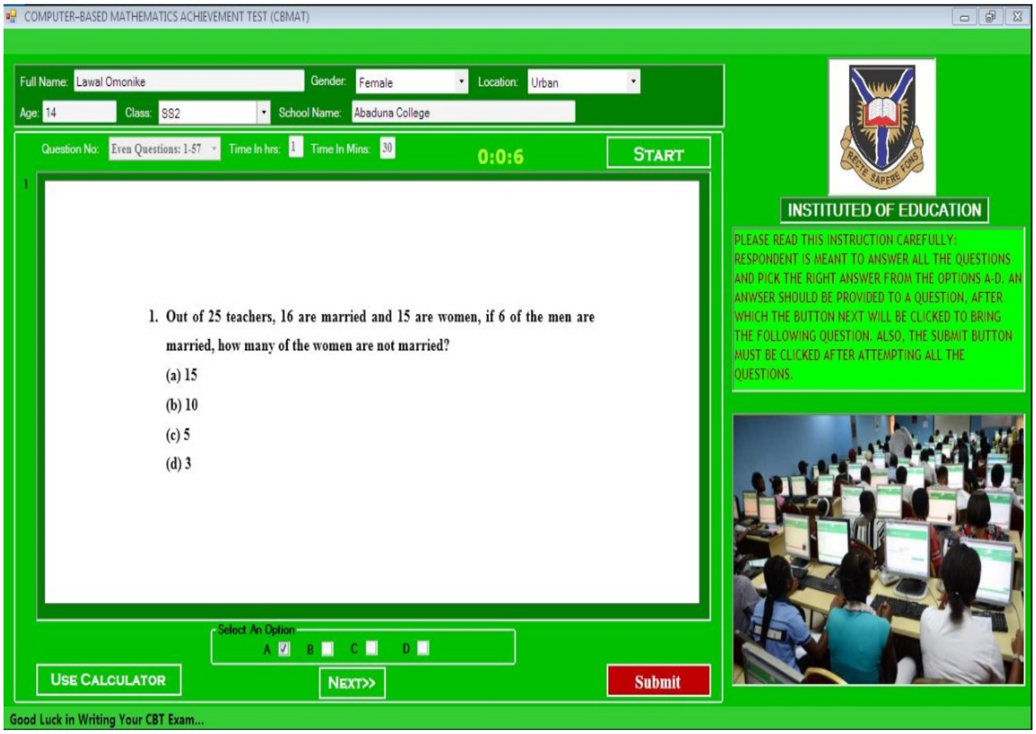
./



Good Luck Writing Your CBT Exam...

Figure 3.3: The CBMAT Log-in screen interface(The Researcher)

Figure 3.3 gives the screen interface of the CBMAT program. An interface that immediatly appeared on the desktop of every respondent's system as soon as the program is launched. Verbal instructions were given on how to kick-start the assessment process. Username and password were provided to them in order to login. Once the respondent clicked on login, the interface automatically keeps loading until 100% loading is reached. It was at this point that the respondents had access to enter their demographic data while timing had not started counting for them.



COMPUTER-BASED MATHEMATICS ACHIEVEMENT TEST (CBMAT)

Full Name: Lawal Omonike Gender: Female Location: Urban

Age: 14 Class: SS2 School Name: Abaduna College

Question No: Even Questions: 1-57 Time In hrs: 1 Time In Mins: 30

0:0:6

1

1. Out of 25 teachers, 16 are married and 15 are women, if 6 of the men are married, how many of the women are not married?
   (a) 15
   (b) 10
   (c) 5
   (d) 3

Select An Option
A ☑ B □ C □ D □

PLEASE READ THIS INSTRUCTION CAREFULLY:
RESPONDENT IS MEANT TO ANSWER ALL THE QUESTIONS AND PICK THE RIGHT ANSWER FROM THE OPTIONS A-D. AN ANWSER SHOULD BE PROVIDED TO A QUESTION, AFTER WHICH THE BUTTON NEXT WILL BE CLICKED TO BRING THE FOLLOWING QUESTION. ALSO, THE SUBMIT BUTTON MUST BE CLICKED AFTER ATTEMPTING ALL THE QUESTIONS.

Good Luck in Writing Your CBT Exam...

Figure 3.4: The CBMAT screen questioning interface (The Researcher)

Figure 3.4 presents the item interface that showed up as soon the respondents were done with supplying their background information and the ‘Start’ menu is clicked. At this point, time automatically starts counting for them. This means that item one (1) for the first time was displayed on the screen. Respondents were instructed to click on any of the option A, B, C or D below the ‘select an option’ menu and click on ‘Next’ menu. The next item displayed automatically as the examinees continues and the same process was followed until the respondent gets to the last item. Then the ‘Submit’ button automatically surfaces. This is clicked and a box that read ‘your exam is successfully completed’ appeared and the box ‘ok’ was clicked to automatically submit the items and immediate score for the examinee is recorded.

f) Validation of the pooled CBMAT instrument

After the face and content validity of the CBMAT scalehas been done and appropriate corrections made, an introduction letter was given by the Institute of Education, University of Ibadan. This letter coupled with a valid University Identification Card of the resercher was shown and submitted to the offices of the Head of Service and the Honourable Commissioner, Oyo State Ministry of Education (Appendix VI and VII; pages 303 and 304). A personal letter was also demanded from the researcher by these offices (Appendix VIII, page 305). All of these were done to gain permission into schools for both students and computer usage. The researcher employed the support of four (4) research assistants that were trained accordingly to help in the data collection process.

However, the pooled CBMAT instrument was validated in two (2) ways. The developed 1.0 version of the CBMAT program was first trial tested to establish its usability and appropriateness in eliciting information from the respondents. This was to enable that the different menus/buttons in the graphical user interface were functioning aright as intended in the test. It was administered on the sampled students from one of the selected schools in Oyo State whose location was off the other schools. Ten (10) respondents were randomly selected from each of the science, commercial and art classes. Thirty examinees were chosen to respond to the instrument at the same time, having taken them through various instructions on how to effectively respond to the test items.

On concluding the test, some anomalies were observed in the course of administering. Few test results (scores) were not submitted into the Excel database as programmed.It was realized that there was no provision for pause menu, should a respondent choose to use the conveniences. It was also discovered that some systems did not have the required number of pixels (resolutions) to display the full page of the questioning interface. However, the initial 1.0 version of the CBMAT program was improved upon and other requirements needed to make it more functional and effective on the schools' computer systems were inculcated. The readjusted 2.0 version of the CBMAT program became the instrument used at both trial-testing and main study phases in data collection.

The second validation process was to ascertain the psychometric properties of the test and items of the scale (construct validity). However, the 14 sampled schools for the trail-testing phase had earlier been visited to establish a good relationship with both mathematics and computer study teachers. The generality of SSII students were also intimated on the modalities and purpose for which the research is being carried out. Computers were inspected to ascertain their functionality and sources of power supply that could disrupt the smooth running of testing were also sorted out with necessary alternatives as the case warranted. The pooled CBMAT items were later administered to 731 examinees.

Examinees' response data from the pooled CBMAT was analysed using IRT approach. This is a modern-day theory that has been recently adopted for test development and item analysis, which is known to yield fine grained information on the suitability of test items for selection purposes (Ariyo and Lemut, 2015; Ojerinde, *et. al*., 2013). The Full Information Item Factor Analysis (FIFA) of Multidimensional Item Response Theory package via R-platform software was used to assess model-data fit result. The test-data was subjected to the four unidimensional IRT models for dichotomously scored response data and model convergence was attained for each of the iteration processes. Thereafter, model-fit result of each of the four models was compared with different information criteria that were available for model fit.The result showed that 4PL model fitted the pooled CBMAT response data because its information criterior showed the smallest value.

IRT empirical reliability of 0.893 was recorded for the instrument. Item and examinee parameters were calibrated and estimated with 4PL model and criteria as benchmarks for all parameter estimates were used to select good items out of the pooled 114 items. At the end of the validation process, 77 items of the pooled CBMAT instrument survived and were retained. This is presented in Table 3.5.

Table 3.5: Table of Specification for the survived items of the pooled CBMAT Instrument

Content Area	**Behavioural Objectives**			**Total**
	Remembering (14%)	**Understanding (18%)**	**Thinking* (68%)**	**(100%)**
Numbers & Numeration (36%)	4items (4,6,42,76)	3items (8,11,57)	20items (12,14,18,19,21,24, 26,28, 29, 31, 32, 33, 34,35, 36, 37, 73, 79, 89,92,)	**27**
Algebraic Processes (16%)	1item (44)	4items (51,52,53,55)	8items (3,7,38,40,41,43,47, 93)	**13**
Geometry (32%)	2items (113,114)	4items (80,82,90,91)	20items (5,10,15,16,39,46,59,6468,69,70,71,74, 77,84, 85,87,99,100,101)	**26**
Statistics (16%)	1item (110)	4items (102,103,105,106)	6items (60,63,66,107,108, 109,)	**11**
Total (100%)	**8**	**15**	**54**	**77**

Thinking* = Applying + Analysing + Evaluating + Creating

g) Selection of items for the final CBMAT instrument for Phase II

Item characteristics curve/item response function (ICC/IRF) that is known as an elementary component of IRT relates respondent's latent trait to the likelihood of responding to an item.It is a strong indicator of how good an item is. This curve, when plotted with the aid of any IRT software, shows how steep individual itemis. It isan indication of how informative a particular item of a scale is to the entire test. Out of the 114 items in the pooled CBMAT instrument, 77 items that survived item analysis were retained. Appendix IX gives the item characteristics curves (ICCs) showing how informative each of the items of the scale is, in terms of the steepness of the ICC graphs. The curves that appeared steepest at the middle indicated (a S-like shape) were the items that constituted the final CBMAT instrument. Appendix X shows the ICCs of the final 40 items used.

This selection constituted the items for the final CBMAT instrument that was used to collect data for the main study. However, consideration of content comprehensiveness was also borne in mind with respect to the table of specification in Table 3.5. The final CBMAT instrument thereafter was made up of the very best 40 items of the scale. Table 3.6 shows the test blueprint of the final items of the CBMAT instrument.

Table 3.6: Table of Specification for the final CBMAT Instrument

Content Area	**Behavioural Objectives**			**Total**
	Remembering (16%)	**Understanding (20%)**	**Thinking* (64%)**	**(100%)**
Numbers & Numeration (34%)	2items (4,6)	2items (11,57)	8items (21,29,31,32 36,37,89,92)	**12**
Algebraic Processes (18%)	1 item (44)	2items (53,55)	4items (3,40,43,47)	**7**
Geometry (30%)	1 item (114)	3items (80,90,91)	10items (10,15,16,39,46, 64, 70,71,74,99)	**14**
Statistics (18%)	1item (110)	2items (102,103)	4 items (60,66,107,108)	**7**
Total (100%)	**5**	**9**	**26**	**40**

Thinking* = Applying + Analysing + Evaluating + Creating

3.5 Procedure for Data Collection

Having gone through validation process, the 40-item final CBMAT instrument was used for data collection for the main study on the sampled schools and students in Lagos State Educational District I. Another letter of introduction was collected and presented to Lagos State Ministry of Education and the office of the Tutor General/Permanent Secretary of Lagos Education District 1, where approval was granted and letters to assess schools were given (Appendices XI, XII, XIII).

As applicable in Oyo State, test administration was also in two batches depending on the number of computers that were available in a school and as many as the researcher were able to add during test administration period. Data collection involved eliciting examinees' responses and response times for individual item and the whole instrument for calibrating item parameters and estimate examinees' ability with the uni-dimensional IRT models especially the model of ultimate interest, four parameter logistic model (4PLM). The software was programmed to record time, once the respondent starts responding to the items.

The researcher and her trained research assistants were involved in instruction, monitoring, supervision and control for some hitches that surfaced in the course of test administration. As soon as testing was done in a particular school, each examinee's result was collated through the school main server. Collation on the other hand was done from individual systems for schools where no single system that connected others was made available as the main server. However, after collection of data, some of the schools requested for appreciation letter for record purpose. A sample of one the letters given by the researcher can be found in Appendix XIV. However, excerpts of some of the pictures taken in different schools during installation, test-running and the real administration of CBMAT program were included in Appendix XXI.

3.6 Procedure for Data Analysis

Responses to the items of final CBMAT instrument and their corresponding response times constituted the data for the study. The CBMAT software is programmed in such a way that once the key option is clicked out of the four (4) options provided, the system automatically awards 1 mark for such correct response while 0 is awarded to any of the distracters picked for wrong choice of response. Automatic scoring of items had been built-in into the CBMAT program where answers to each item of the scale had been supplied to the system. Recording of time as soon as the respondent clicks on the item was also feasible. Both response and response time were harvested from Excel database.

Meanwhile, data analysis procedure was preceded by data preparation where the elicited data in the files were cleaned to take care of missing or incomplete data to avoid biased parameter estimation and reduction in sample representation (Hyun Kang, 2013). The check for missing data was accomplished with the aid of SPSS software using Missing Data Analysis with Multiple Imputation approach of the Maximum Likelihood method. Both response and response times in Excel files were converted to different formats in Notepad so that the statistical software used could easily accessed and processed.

Data analysis was however carried out using the open source software programming of R-foundation for statistical computing platform through its user friendly interface R-studio of version 3.5.3. R environment embeds the Full Information Item Factor Analysis (FIFA) of the Multidimensional Item Response Theory (MIRT) analysis package used in the study. This package allows the analysis of either dichotomous or polytomous response data of uni-dimensional (4, 3, 2 and 1PL) and multidimensional latent trait models under IRT paradigm (Chalmers, 2012). R programing language is widely used among statisticians and data miners for developing statistical software and data analysis through the application of various statistical methods (Gilbert Duy Doan, 2017).

Meanwhile, De mars (2010) is of the opinion that when using IRT models to examine test items, tests and item responses are only binding if the IRT assumptions hold. Part of the analytical procedure for this study was that the theoretical assumptions of IRT were examined such that effective usage and appropriate interpretations of the useful results thereof are not jeopardized. The first analytical procedure was the check on trait dimensionality and item local independence assumptions so as to appropriate the choice of the right model for calibrating the CBMAT response data. Dimensionality assessment of the pooled CBMAT response data was imvestigated by Stout's test of essential unidimensionality that is implemented in Dimtest 2.0 software (Stout, 2005). A statistical procedure for testing the hypothesis that an essentially unidimensional latent trait model fits observed binary item response data from a test. On the other hand, Thiessen Yen Q3 statistic was used to assess local independence assumption.

While monotonicity assumption using Item Characteristic Curves of the MIRT package was also used.

However, Lognormal Item Response Theory (LNIRT) package of the same R programming was also employed to calibrate the parameters of the response time model. A Bayesian procedural approach with the help of Markov Chain Monte Carlo (MCMC) process that is known as Gibbs sampling was adopted to compute model's test statistics and parameters. The Gibbs sampling is an iterative estimation. However, the technical know-how of this iteration process is given by Klein, Fox and van der Linden (2009) and van der Linden (2007).

Table 3.7 shows the different analytical methods that were used in analysing the research questions posed.

Table 3.7: Research Questions and Statistical Method of Analysis

Research Questions	Method of Analysis
1. Which of the four IRT models for dichotomous test best fit the pooled Computer-Based Mathematics Achievement Test (CBMAT) response data?	Full Information Factor Analysis (FIFA) of the Multidimensional Item Response Theory (MIRT) package of R programing environment. With consideration for model-fit statistic values of -2Loglikelihood, Akaike and Bayesian Information Criteria
2. What is the quality of the pooled CBMAT items under other dichotomous IRT models and the model that best fit the test data?	Calibration with 4, 3, 2 and 1-Parameter Logistic Models of the MIRT package in R-software environment.
3. Is there any significant mean difference in the item parameter estimates of the other IRT models and the model that fits the pooled CBMAT response-data at the developmental stage?	Descriptive Statistics, Related Sample Friedman's Q and Wilcoxon Signed Rank Test Statistics
4. How consistent is the model used in calibrating the pooled CBMAT response-data at the development stage to the model used in calibrating the final CBMAT response-data at the real study stage?	Full Information Factor Analysis (FIFA) of the Multidimensional Item Response Theory (MIRT) package of R programing environment. With consideration for model-fit statistic values of -2Loglikelihood, Akaike and Bayesian Information Criteria
5. Is there any significant mean difference in the examinee's parameter estimates of the other dichotomous IRT models and the model that fits the final CBMAT response data?	Related Sample Friedman's Q Statistic
6. Is there any significant mean difference in the item parameter estimates of the other dichotomous IRT models and the model that fits the CBMAT data at the final stage?	Descriptive Statistics, Related Sample Friedman's Q and Wilcoxon Signed Rank Test Statistics
7. What are the estimates of item and examinee's parameters of the Response Time IRT model using the final CBMAT response and response time data?	Bayesian Estimation Method (Expected a posteriori) of the MCMC algorithm with Gibbs Sampling Approach of the joint LNIRT response time model

8. Is there any significant relationship between item and examinee’s parameters of the LNIRT response time model?	Pearson Product Moment Correlation Coefficient
9. What are the patterns of the person-fit statistics for detection of aberrant response behaviour in the CBMAT response time data?	Bayesian significance test procedure of the person-fit statistics in the joint LNIRT response time model
10. How comparable are the item and examinees parameter estimates of the traditional IRT model to the LNIRT response time model?	Descriptive Statistics and Mann-Whitney U test

3.7 Methodological Challenges

A major challenge that was evident to this study was the dearth of empirical literature on the research subject-matter (4PLM and Response Time Models) especially on studies carried out within Nigeria context. Nevertheless, this challenge was overcome by relying on foreign online journals for those who gave their full articles and contacting some authors through mailing where no full text was made available online as well as some local textbooks and journals on the background information on IRT.

Accessibility of schools with computers was a serious challenge in this study as the state ministry of education in both Lagos and Oyo states did not have the correct number of schools with functional computer laboratories. Some of the listed schools given to the researcher had their computer rooms vandalized or not functioning at all. The researcher made individual effort to search for schools with functioning systems. Availability of sufficient and functional numbers of computers in schools became another problem. The researcher however supported the number of computers available in the selected schools with a large number of laptops to complement the available systems so as to have a sizeable number of sample size for the study.

Another challenge was that even when the hardware components of the available systems were alright, most of the software compatibility that could make the developed CBMAT program work were absent in the schools' systems. Frantic effort was made to install necessary and basic files into computers to make the CBMAT software function appropriately. The threat of viruses in several systems did not allow the researcher to complete data collection procedures on time.

Apart from the aforementioned methodological challenges, epileptic supply of electricity across the schools was obvious. Hence, alternate provision in place of government power supply was made in terms of the supply of fuel and power generating sets. Some schools had provision for solar power while others had standby generating set.

Lastly, alteration of the normal school time-table was another problem the researcher was confronted with since disruption in school time-table and the several holidays declared by the government caused a form of setback in school activities. The researcher therefore persuaded the school authorities such that time allotted for mathematics subject was used to administer the instrument.

CHAPTER FOUR
RESULTS AND DISCUSSIONS

Chapter four displays the findings obtained after data analysis and the discussion from findings. Discussion is made as each of the research questions is answered. Preliminary data analysis is done to inquire into how the distributions of response and response time (RT) data look like so as to appropriate the right statistical tools to the questions posed in the research work. The overall description of the data at both trial-testing and real study stages using the pooled CBMAT (computer-based mathematics achievement test) and the final CBMAT instruments is also presented.

4.1 Preliminary Data Analysis

Response and response time (RT) data that were automatically recorded from the examinees' responses to the pooled CBMAT instrument (731 test takers; 114 items) and the final CBMAT instrument (874 test takers; 40 items) were analyzed beside the background information of each examinee that was available (student's name, age, class and location of school). A preliminary analysis was then done to show the descriptive statistics for responses and response time data as well as their distributions.

Table 4.1: Summary Statistics for Response and Response Time data of the pooled CBMAT instrument (n=731, item=114)

	Mean	SD	Min	Max	Skewness	Kurtosis
Total Score	44.42	13.32	17	104	1.22	1.67
Total Time	5562.56	1684.41	361	10900	-0.015	0.106

Table 4.1 shows the average score(mark)and time (secs) it took the examinees to complete the pooled CBMAT scale and their respective standard deviations values ($\bar{x} = 44.42,\ \sigma = 13.32; \overline{x} = 5562.56 \approx 2\text{hrs}, \sigma = 1684.41$). Analysis shows that an average time of around 1hr and 55mins was spent on 114 items by the test takers. While the lowest and highest scoreand time ($\text{Score}_{\text{min}}$=17, $\text{Score}_{\text{max}}$=104) ($\text{Time}_{\text{min}}$=361secs (6.02mins); Time_{max}=10900secs (181.67mins≈ 3hrs) were recorded in the pooled CBMAT scale.

By implication, the observed average time spent by the examinees to respond to 114-item CBMAT scale is 1hour and 55minutes while average score amounts to 44.

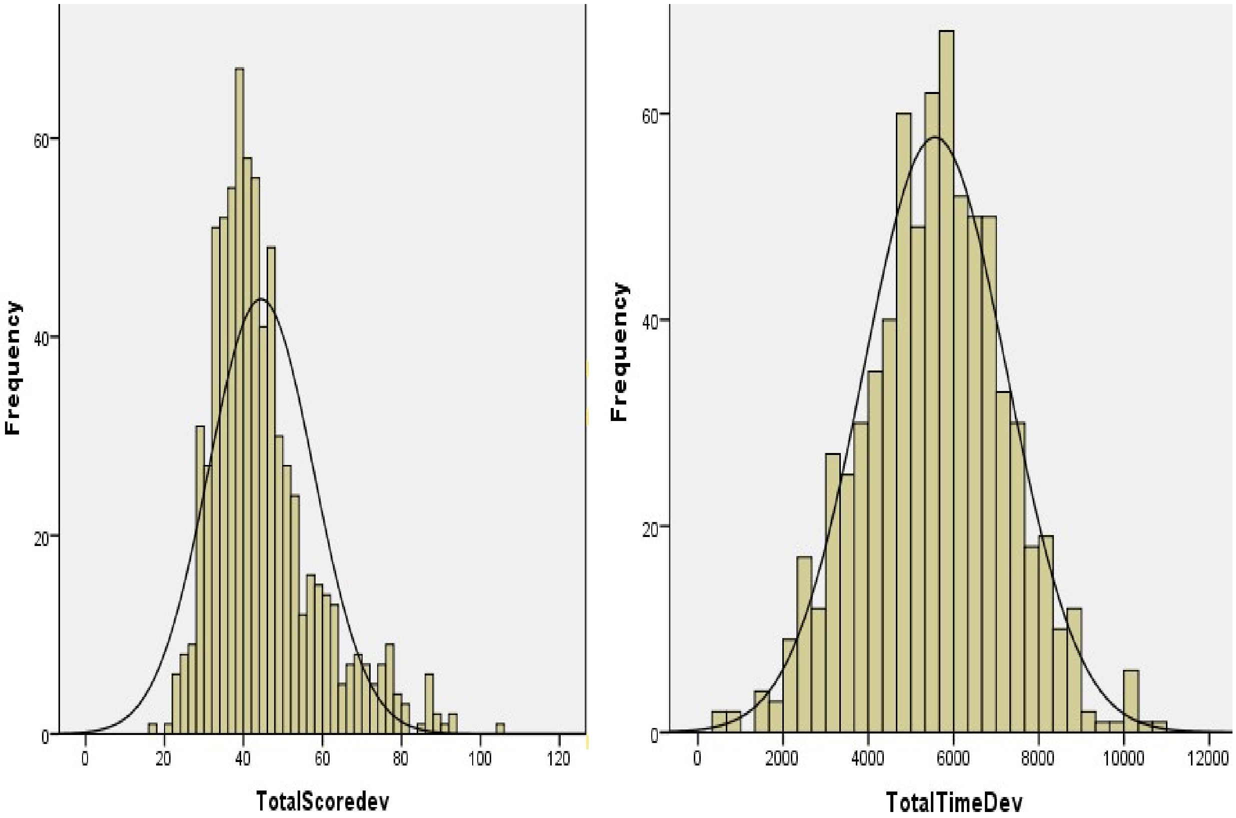


Figure 4.1: Distribution of the Aggregate Score and Aggregate Response Time for Pooled CBMAT data

In Figure 4.1, response data indicates a slight positively skewed distribution. This means that the mean score for the distribution is greater than of the median and modal scores. An indication ofthe fact that majority of the respondents' scores clustered around the lower marks. Meanwhile, response time data gives a normal/unskewed distribution for the pooled CBMAT data. Although RTs data are continuous and more informative, and easier to evaluate statistically, approximately symmetrical distribution is always assumed for them. In the analysis for the RTs data, the mean, median and mode response time are the same because of the normal shaped distribution that is oobserved.

Table 4.2: Summary Statistics for Response and Response Time data of the final CBMAT instrument (n=874, item=40)

	Mean	SD	Min	Max	Skewness	Kurtosis
Total Score	17.81	6.10	2	40	0.945	0.713
Total Time	2356.72	621.43	133	3640	-0.662	0.203

The result presented in Table 4.2 shows the descriptive statistics obtained from investigating the response and RT data for the final CBMAT items. On the average, the examinees had around 17 marks out of 40 while average time spent was 2356.72secs (66mins) to complete the final 40-item CBMAT scale. Analysis also displays the lowest and highest scores/time ($\text{Score}_{\text{min}}$=2,$\text{Score}_{\text{max}}$=40; Time_{min}=133secs (2.2mins), Time_{max}= 3640secs (around 61mins≈1hr)) respectively.

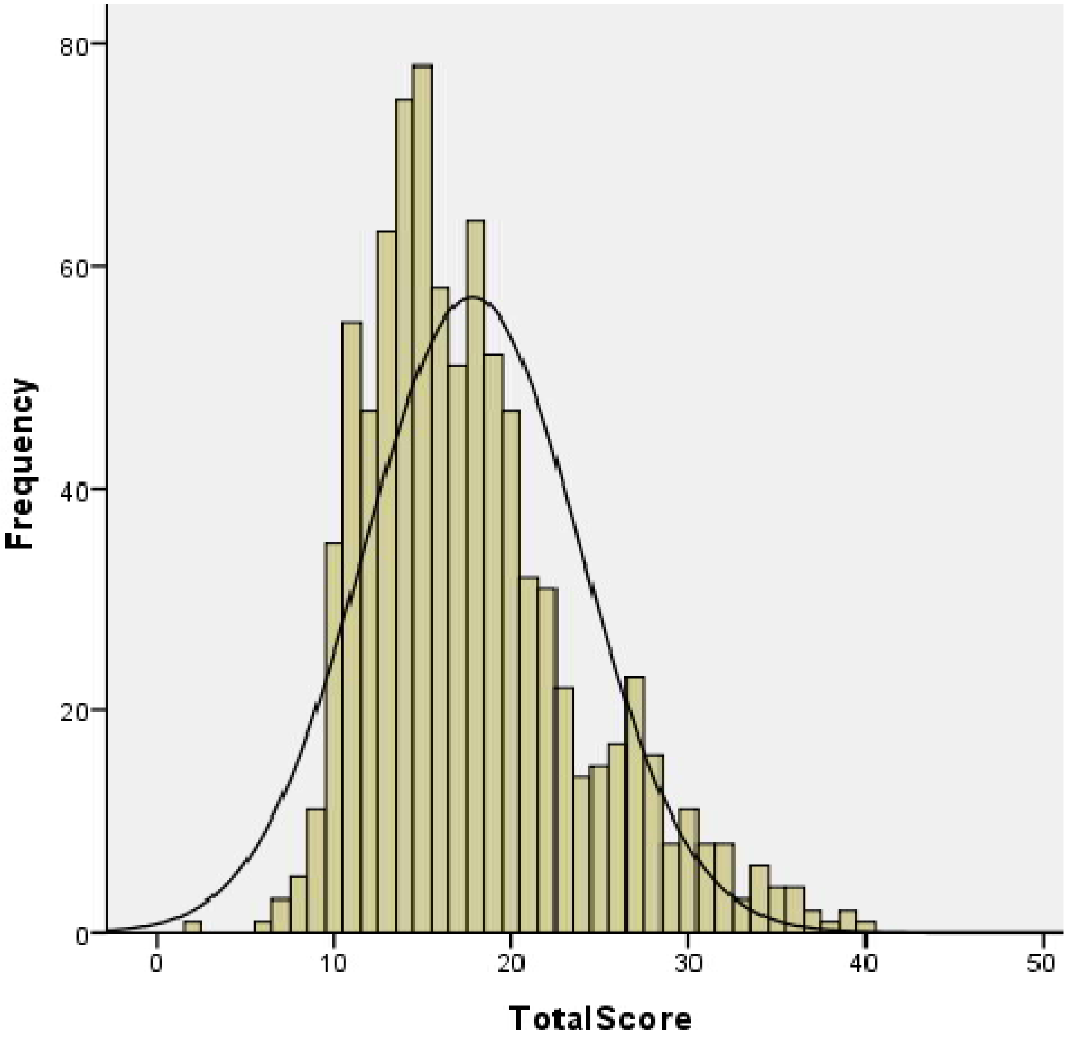


Figure 4.2a: Distribution of the Aggregate Score for the FinalCBMAT data

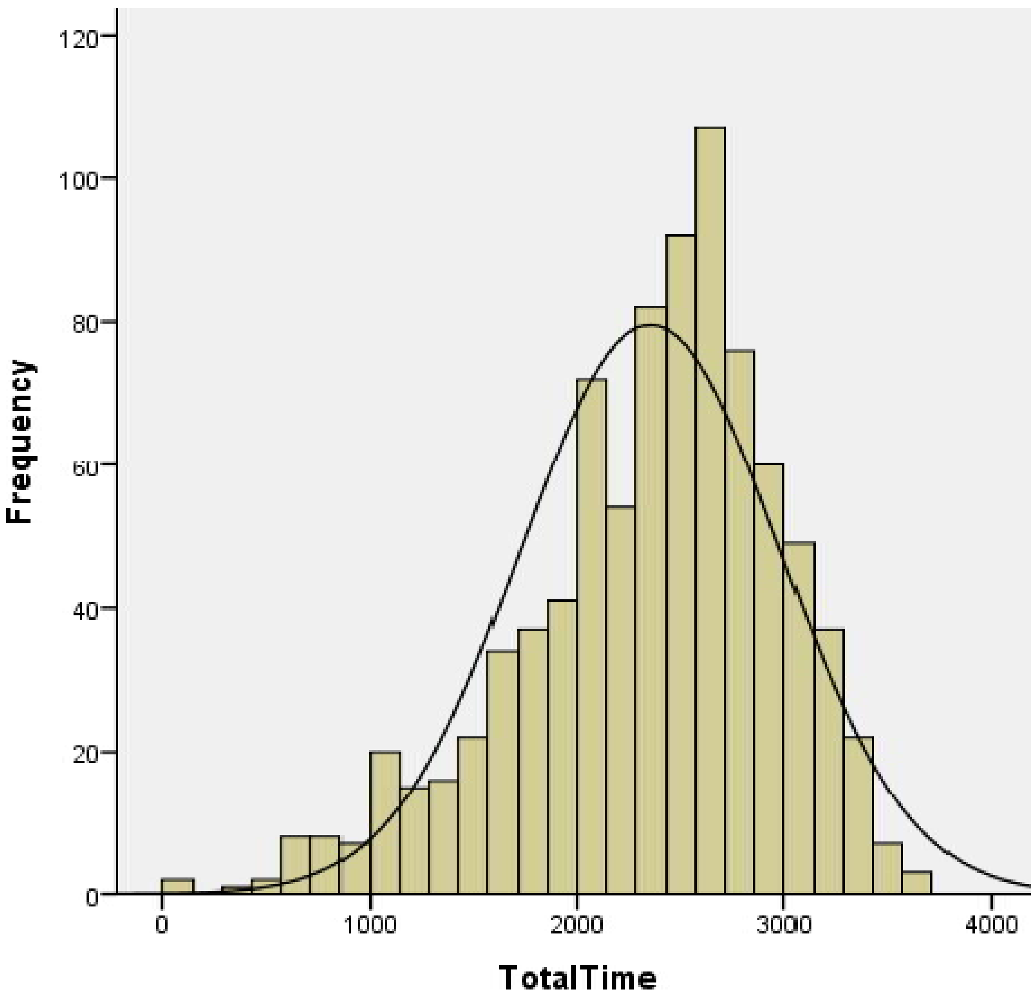


Figure 4.2b: Distribution of the Aggregate Response Time for the FinalCBMAT data

Figures 4.2a and 4.2b describe the response and RT data to be slightly skewed distributions, with a positive skewness for responses while response time depicts a distribution that is negatively skewed. It is observed from the preliminary data analysis that the nature of the distribution of the response data obtained with the CBMAT instruments at both trial-testing and main research stages showed a slight positively skewed distribution. While that of the RT data indicated a normal distribution at the development stage and a negatively skewed distribution at the final stage.

In all, approximate normal distributions were assumed for the response and RT data as supported by the studies of Suh (2016) and Fox and Marianti (2017). Schnipke and Scramps (1997; 2002) were also of the view that the nature of these distributions is not uncommon when response time is involved in testing.

4.2: Research Question 1: Which of the four IRT models for dichotomous test best fits the pooled Computer-Based Mathematics Achievement Test (CBMAT) response data?

4.2.1 Ascertaining Dimensionality and Local Independence Assumptions for pooledCBMAT response data

Prior to answering this research question, it is essential that assumptions (trait dimensionality and local independence) underlying item response theory (IRT) are tested to ascertain the appropriate model that will best explain the pooled CBMAT response data. To assess the dimensionality of the pooled CBMAT instrument, the test-data was subjected to Stout's test of essential unidimensionality which was implemented in DIMTEST 2.0 (Stout,2005) software. Stout's tests the null hypothesis (H_o) "the pooled test-data is essentially unidimensional". This means that failure to reject H_o indicates that the test is unidimensional. However, if H_o is rejected, multidimensionality is evident.

In order to achieve this, the test was divided into two distintparts:the Assessment Subtest (AT) and the Partitioning Subtest (PT). To test H_o, AT (the items that are likely to be measuring a secondary dimension were empirically selected either using the clustering procedure of the package or the item content) was compared with PT (the primary dimension the test assessed which are the remaining items after which AT has been removed).The result is presented in the Table 4.2a.

Table 4.2a: Essential Dimensionality Statistic of the Pooled CBMAT Instrument

TL	TGbar	Test Statistic	p-value
7.426	6.392	1.0285	0.1518

It is indicated in the analysis that 55 items formed the Assessment Subtest (AT) which was selected using 30% of the examinees who sat for the pooled CBMAT with the clustering procedure of the DIMTEST package. The remaining 59 items formed the Partitioning Subtest (PT). Table 2a indicates that the abilities measured by the AT were not significantly different in dimension from the abilities measured by the PT (Test Statistic = 1.0285, $p > 0.05$). Result shows that there is no significant variation between the abilities measured by the primary and secondary dimension subtest of the test measured. The implication of the result is that the pooled CBMAT instrument reveals that only one dominant dimension accounted for the variation observed among the examinees. Thus, the assumption of uni-dimensionality is tenable. Appendix XV (page 336) shows the full DIMTEST result.

Item local independence assumption was also assessed by the use of Yen Q_3 statistic. According to Thiessen (1997), Yen Q_3 statistic value of $\leq |0.2|$ for a pair of item indicates that the items are locally independent while value for any pair of items> $|0.2|$ will show that one of the paired items is locally dependent on each other. The result is shown as follows:

Table 4.2b: Summary of Assessment of Local Independence of the Pooled CBMAT Instrument

	Item	
Item	23	59
11	**---**	0.378
13	0.312	**---**

Table 4.2b presents the summary of the correlation item residual of local independence of the pooled CBMAT instrument. By comparison, pairs of items that violate local independence assumption are items 11 and 59 (0.378) and items 13 and 23 (0.312) respectively with values greater than|0.2|. The remaining items fulfil local independence assumption. Appendix XVI presents a representative part of the output obtained from running Yen Q_3 statistic. The result on the summarized Table 4.4 showed that either items 11and 13 or item 23 and 59 were locally dependent on oneanother. Therefore, items 13 and 23 were deleted from the pool of items. The implication of this is that the pair of items depends on each other as a result of their correlation item residuals that were greater than |0.2|.

Result from the DIMTEST, statistic showed that the pooled CBMAT scale is uni-dimensional since the p-value is greater than 0.05. The examinees' ability only portrayed a dominant trait which is an indication that the abilities measured by the items of Assessment subtest (AT) are not dimensionally different from the ones measured by the items of the Partitioning subtest. This suggests that only mathematics ability in the respondents accounts to the response made to the items of the pooled CBMAT instrument. Moreover, two pairs of item were seemed to violate the assumption of local independence which made either pair of the item to be deleted from the pool of items. It also means that the responses made on the rest of the items in the scale do not depend or inform the ones made on some other items of the same scale.

Suh (2016) and Umobong and Tommy (2017) pointed out that if IRT assumptions are violated, inferences resulting from the estimates generated may likely be erroneous,which could jeopardize the potential advantage of the theory. This result is in support of the findings of Ayanwale (2019) and Okwilagwe and Ogunrinde (2017) when IRT assumptions were checked on the Draft Multiple Choice Mathematic Test and NECO Geography Achievement Test respectively.

However, since the pooled CBMAT scale have been proved undimensional and the rest items ascertained to be locally independent knowing fully well that the CBMAT is dichotomous in nature, it is clear that the test could be calibrated using either the 1-, 2-, 3- or 4-parameter logistic (PL) model. To answer research question one, the pooled CBMAT response-data is exposed to the four models.

4.2.2 Model-Data Fit Assessment

Assessment of model-data fit was achieved with the multidimensional item response theory (MIRT) Package of the R foundation for Statistical Computing Platform via R-Studio of version 3.5.3. The pooled CBMAT test-data was subjected to the four unidimensional IRT models for dichotomously scored response data and model convergence was attained for each of the iteration processes. Thereafter, model-fit result of each of the four models was compared with different information criteria that were available for model fit.

These include the Akaike Information Criteria (AIC) (Burnham and Anderson (2002), Corrected Akaike Information Criteria (AICc; Burnham and Anderson (2004), Sample-Size Adjusted Bayesian Information Criteria (SABIC; Enders and Tofihi, 2008) and Bayesian Information Criteria (BIC; Schwarz, 1978; Yang, 2005). The model that produced the best fit to the test-data is considered the most appropriate. To achieve this feat, several measures were appllied. According to Vrieze (2012); Burnham and Anderson (2002); Konish and Kitagawa (2008); Finch and French (2015), the noticeable ones among them are -2loglikelihood test and the use of information indices.

Information indices are the most frequently used standards for statistical interpretation and comparison of models to adjudge which model best explains the data-set. They are measures of variance not explained by a model, with an added penalty for model difficulty. These indices were computed using the-2loglikelihood chi-square value, interpreted in a way that any model that is having the smallest estimate reveals the best fitting of the response data. For this study, -2loglikelihood, Akaike or first-Order and Bayesian Information Criteria are used. Table 4.2c presents the result of the model-data fit assessment for the pooled CBMAT instrument.

Table 4.2c: Model-data fit Assessment of the Pooled CBMAT Instrument

IRT Model	-2Loglike-lihood	AIC	BIC
1PL	99411.90	99413.80	99418.50
2PL	98028.92	98032.92	98042.11
2PL	98028.92	98032.92	98042.11
3PL	97429.32	97435.32	97449.10
3PL	97429.32	97435.32	97449.10
4PL	97274.92	97282.92	97301.30

Table 4.2c summarises the model-data fit assessment results, indicating the IRT model that best explained the pooled CBMAT data. In the course of analysis at the first convergence stage for 1PL model, the program found a fitting with -2loglikelihood=99411, AIC=99413 and BIC=99418 values as against 2PL model fitting (-2loglikelihood= 98028; AIC=98038 and BIC=98042). The difference between the likelihood values for the two models shows that 2PL model statistically fits the data significantly better than the 1PL model.

In the search for a better model for the response data, calibrating 2PL model to the pooled CBMAT data was in turn compared to that of 3PL model. The result shows that the IRT 3PL model fits the test data better than the 2PL model (3PL model's values -2loglikelihood= 97429; AIC=97435 and BIC=97449). These values were respectively less than its 2PL model's fitting values, which indicates that values for 3PL model-fit were lesser and statistically significant.

The last stage of iteration was the convergence in the 4PL model where comparison of the fitness of 3PL model to 4PL model was made. Table 4.5 reveals that 4PL model fits the pooled CBMAT data better than 3PL model (4PL model: -2loglikelihood=97274; AIC=97282 and BIC=97293). This specifies that the values obtained for -2loglikelihood and the information criteria for 4PL model were respectively less than those of the 3PL model (-2loglikelihood= 97429; AIC=97435 and BIC=97301) and appeared the smallest compared to other values in the 3-, 2- and 1PL models. At the end of the whole process, the result shows that 4PL model fits the pooled CBMAT response data better than the other dichotomously scored response 1-, 2- and 3PL models.

According to Ojerinde, Popoola and Ariyo (2015), the main advantage of an IRT model is its capability to describe and predict respondents' performance on items accurately. If such information provided would be accurate as it relate to performance, it means that the right model that best explains the response data should be employed. This is one of the reasons why model fit assessment is essential to present the most appropriate model needed to predict examinees' true ability. This study employs 4PL model as it seems the best in explaining the pooled CBMAT response data.

This result supports the claim of Magis (2013) that the core advantage of 4PL model is its ability to allow a non-zero probability of responding to an item incorrectly by highly able examinees. 4PL model is capable of handling guessing error that 3PL model could as well take care of.It also helps to leverage any mistake (due to stress for instance) high-ability students might incur in the cause of responding to test items. Loken and Rulison (2010) in their study in a computerized adaptive testing (CAT) environment displayed how the effect of early mistakes committed by brilliant students was highly lessened by applying 4PL model.

Anotherstudy in support of the effectiveness of 4PL model is the work of Liao, Ho, Yen and Cheng (2012) in which the performance of 4PL model, as a robust mechanism model, was examined and compared with 3PL model. Their finding indicated that 4PL model gave a more effective and strong estimation method than the 3PL model.

Research Question 2: What is the quality of the pooled CBMAT items under other dichotomous IRT models and the model that best fit the test data?

4.2.3 Calibrating Item Parameter Estimates

To answer research question two, the quality of the 114- pooled CBMAT items was assessed by calibrating the instrument using MIRT (Full Information Item Factor Analysis of Multidimensional Item Response Theory) package of R-platform via R-Studio with the appropriate command in order to estimate the item parameters of the different models used.Calibration was done with all the existing IRT models in the dichotomously scored models (3-, 2- and 1PL models) and 4PL model that best fits the test data. These estimates are: ***a*** (item discriminating parameter), ***b*** (item difficulty parameter), ***c*** (pseudo-guessing parameter/lower asymptote) as well as ***u*** (the carelessness parameter/upper asymptote).

The values of the parameter estimates are determined by given consideration to the following criteria. Discriminating parameter value is taken as $a > 0.50$ while difficulty parameter range is specified as $-3 \leq b \leq 3$.The pseudo-guessing parameter is taken as $c < 0.35$ and carelessness parameter is considered to be $u > 0.5$ for higher parameterised models like the 3- and 4PL IRT models (This is because the value of the item difficulty parameter becomes the point on the ability scale where the probability of correct response is halfway between the value of parameter c and 1).

These range of values according to Baker (2001) and Kline (2005) are often used by researchers. The calibration results of the quality of the pooled CBMAT response data for the four (1-, 2-, 3- and 4PL) uni-dimensional models are presented in Tables 4.3a, 43b7, 4.3c and 4.3d consecutively. Meanwhile, Table 4.3a expresses the result of calibrating with 4-parameter logistic model that appears to be a better fit of the pooled CBMAT response data.

Table 4.3a: Item parameters of the pooled CBMAT scale calibrated by 4PL mode

Item No	*a*	Remark	*b*	Remark	*c*	Remark	*u*	Remark	Overall	Decision
1	3.42	Good	1.6	Good	0.37	**Poor**	1.00	Good	**POOR**	Delete
2	11.02	Good	0.7	Good	0.19	Good	0.38	**Poor**	**POOR**	Delete
3	2.64	Good	0.6	Good	0.27	Good	0.90	Good	GOOD	**Retain**
4	1.38	Good	2.0	Good	0.00	Good	0.55	Good	GOOD	**Retain**
5	2.54	Good	1.3	Good	0.25	Good	1.00	Good	GOOD	**Retain**
6	1.14	Good	0.2	Good	0.28	Good	1.00	Good	GOOD	**Retain**
7	3.09	Good	0.5	Good	0.19	Good	0.88	Good	GOOD	**Retain**
8	1.61	Good	1.0	Good	0.20	Good	0.98	Good	GOOD	**Retain**
9	1.28	Good	1.1	Good	0.37	**Poor**	1.00	Good	**POOR**	Delete
10	1.78	Good	0.3	Good	0.31	Good	0.56	Good	GOOD	**Retain**
11	2.84	Good	1.1	Good	0.10	Good	1.00	Good	GOOD	**Retain**
12	2.38	Good	2.1	Good	0.17	Good	1.00	Good	GOOD	**Retain**
13	2.30	Good	0.2	Good	0.35	**Poor**	1.00	Good	**POOR**	Delete
14	2.33	Good	1.2	Good	0.10	Good	0.56	Good	GOOD	**Retain**
15	3.77	Good	2.9	Good	0.21	Good	1.00	Good	GOOD	**Retain**
16	2.01	Good	1.8	Good	0.24	Good	0.77	Good	GOOD	**Retain**
17	2.06	Good	1.4	Good	0.4	**Poor**	1.00	Good	**POOR**	Delete
18	2.56	Good	1.3	Good	0.25	Good	1.00	Good	GOOD	**Retain**
19	3.51	Good	1.9	Good	0.34	Good	1.00	Good	GOOD	**Retain**
20	-0.54	**Poor**	3.0	**Poor**	0.13	Good	1.00	Good	**POOR**	Delete
21	3.44	Good	0.3	Good	0.31	Good	0.82	Good	GOOD	**Retain**
22	0.90	**Poor**	0.0	Good	0.00	Good	0.29	**Poor**	**POOR**	Delete
23	1.90	Good	0.0	Good	0.44	**Poor**	1.00	Good	**POOR**	Delete
24	45.73	Good	0.7	Good	0.29	Good	0.79	Good	GOOD	**Retain**
25	2.34	Good	1.0	Good	0.50	**Poor**	1.00	Good	**POOR**	Delete
26	4.32	Good	1.5	Good	0.18	Good	0.87	Good	GOOD	**Retain**
27	17.28	Good	1.6	Good	0.13	Good	0.52	Good	**POOR**	Delete
28	1.47	Good	0.4	Good	0.19	Good	1.00	Good	GOOD	**Retain**
29	2.35	Good	-0	Good	0.28	Good	0.93	Good	GOOD	**Retain**
30	4.54	Good	0.2	Good	0.44	**Poor**	0.86	Good	**POOR**	Delete
31	2.32	Good	0.5	Good	0.31	Good	0.97	Good	GOOD	**Retain**
32	0.51	Good	-0	Good	0.00	Good	0.86	Good	GOOD	**Retain**
33	3.29	Good	1.4	Good	0.20	Good	1.00	Good	GOOD	**Retain**
34	2.43	Good	1.3	Good	0.13	Good	1.00	Good	GOOD	**Retain**
35	38.52	Good	0.5	Good	0.27	Good	0.61	Good	GOOD	**Retain**
36	2.27	Good	2.1	Good	0.22	Good	1.00	Good	GOOD	**Retain**
37	2.06	Good	2.2	Good	0.27	Good	1.00	Good	GOOD	**Retain**
38	6.83	Good	1.5	Good	0.14	Good	0.80	Good	GOOD	**Retain**
39	4.75	Good	2.0	Good	0.00	Good	0.79	Good	GOOD	**Retain**
40	1.24	Good	2.1	Good	0.23	Good	1.00	Good	GOOD	**Retain**
41	2.71	Good	0.6	Good	0.26	Good	0.98	Good	GOOD	**Retain**
42	0.57	Good	0.3	Good	0.00	Good	1.00	Good	GOOD	**Retain**

43	3.69	Good	1.3	Good	0.28	Good	1.00	Good	GOOD	**Retain**
44	1.25	Good	2.3	Good	0.16	Good	1.00	Good	GOOD	**Retain**
45	1.09	Good	3.2	**Poor**	0.22	Good	1.00	Good	**POOR**	Delete
46	2.98	Good	1.6	Good	0.18	Good	1.00	Good	GOOD	**Retain**
47	28.1	Good	0.7	Good	0.34	Good	0.75	Good	GOOD	**Retain**
48	1.55	Good	1.0	Good	0.41	**Poor**	0.89	Good	**POOR**	Delete
49	4.75	Good	1.0	Good	0.37	**Poor**	0.83	Good	**POOR**	Delete
50	6.89	Good	0.4	Good	0.24	Good	0.56	Good	**POOR**	Delete
51	1.56	Good	2.0	Good	0.00	Good	0.93	Good	GOOD	**Retain**
52	13.01	Good	2.3	Good	0.31	Good	1.00	Good	GOOD	**Retain**
53	0.71	Good	0.3	Good	0.09	Good	0.96	Good	GOOD	**Retain**
54	16.34	Good	-0	Good	0.38	**Poor**	0.64	Good	**POOR**	Delete
55	2.02	Good	1.1	Good	0.33	Good	1.00	Good	GOOD	**Retain**
56	0.85	Good	-0	Good	0.24	Good	0.55	Good	**POOR**	Delete
57	4.25	Good	0.9	Good	0.21	Good	0.93	Good	GOOD	**Retain**
58	30.4	Good	1.1	Good	0.39	**Poor**	0.64	Good	**POOR**	Delete
59	3.97	Good	1.0	Good	0.13	Good	1.00	Good	GOOD	**Retain**
60	2.99	Good	0.0	Good	0.31	Good	0.92	Good	GOOD	**Retain**
61	31.72	Good	0.0	Good	0.51	**Poor**	0.79	Good	**POOR**	Delete
62	-1.19	**Poor**	3.0	Good	0.04	Good	1.00	Good	**POOR**	Delete
63	1.20	Good	0.0	Good	0.02	Good	1.00	Good	GOOD	**Retain**
64	3.67	Good	1.7	Good	0.08	Good	1.00	Good	GOOD	**Retain**
65	1.98	Good	0.5	Good	0.40	**Poor**	0.96	Good	**POOR**	Delete
66	1.67	Good	1.1	Good	0.28	Good	0.96	Good	GOOD	**Retain**
67	1.22	Good	2.1	Good	0.42	**Poor**	1.00	Good	**POOR**	Delete
68	14.86	Good	2.0	Good	0.00	Good	0.70	Good	GOOD	**Retain**
69	2.74	Good	1.6	Good	0.24	Good	1.00	Good	GOOD	**Retain**
70	2.42	Good	1.5	Good	0.31	Good	1.00	Good	GOOD	**Retain**
71	2.15	Good	0.8	Good	0.19	Good	1.00	Good	GOOD	**Retain**
72	-12.7	**Poor**	3.0	Good	0.16	Good	1.00	Good	**POOR**	Delete
73	1.11	Good	0.9	Good	0.21	Good	1.00	Good	GOOD	**Retain**
74	4.32	Good	1.9	Good	0.19	Good	0.75	Good	GOOD	**Retain**
75	1.25	Good	3.2	**Poor**	0.17	Good	1.00	Good	**POOR**	Delete
76	8.73	Good	0.7	Good	0.32	Good	0.53	Good	GOOD	**Retain**
77	2.92	Good	1.0	Good	0.33	Good	0.80	Good	GOOD	**Retain**
78	1.40	Good	3.3	**Poor**	0.26	Good	1.00	Good	**POOR**	Delete
79	2.44	Good	1.8	Good	0.22	Good	1.00	Good	GOOD	**Retain**
80	3.52	Good	2.2	Good	0.20	Good	1.00	Good	GOOD	**Retain**
81	23.08	Good	1.2	Good	0.19	Good	0.37	**Poor**	**POOR**	Delete
82	1.98	Good	2.2	Good	0.18	Good	1.00	Good	GOOD	**Retain**
83	0.33	**Poor**	1.0	Good	0.00	Good	1.00	Good	**POOR**	Delete
84	1.94	Good	0.2	Good	0.22	Good	0.98	Good	GOOD	**Retain**
85	0.62	Good	0.4	Good	0.01	Good	1.00	Good	GOOD	**Retain**
86	2.68	Good	0.7	Good	0.41	**Poor**	0.89	Good	**POOR**	Delete
87	2.87	Good	2.3	Good	0.16	Good	1.00	Good	GOOD	**Retain**

88	-12.4	**Poor**	1.0	Good	0.19	Good	0.29	**Poor**	**POOR**	Delete
89	1.98	Good	1.7	Good	0.28	Good	0.81	Good	GOOD	**Retain**
90	2.37	Good	1.9	Good	0.3	Good	1.00	Good	GOOD	**Retain**
91	1.73	Good	2.0	Good	0.23	Good	1.00	Good	GOOD	**Retain**
92	1.08	Good	1.8	Good	0.25	Good	1.00	Good	GOOD	**Retain**
93	25.84	Good	1.4	Good	0.23	Good	0.91	Good	GOOD	**Retain**
94	0.12	**Poor**	7.7	**Poor**	0.00	Good	0.99	Good	**POOR**	Delete
95	5.33	Good	0.7	Good	0.34	Good	0.59	Good	**POOR**	Delete
96	0.06	**Poor**	10	**Poor**	0.00	Good	0.99	Good	**POOR**	Delete
97	0.33	**Poor**	3.8	**Poor**	0.00	Good	1.00	Good	**POOR**	Delete
98	21.31	Good	0.1	Good	0.31	Good	0.57	Good	**POOR**	Delete
99	2.01	Good	3.0	Good	0.24	Good	1.00	Good	GOOD	**Retain**
100	2.52	Good	0.2	Good	0.31	Good	0.82	Good	GOOD	**Retain**
101	1.21	Good	2.0	Good	0.13	Good	1.00	Good	GOOD	**Retain**
102	1.15	Good	1.4	Good	0.33	Good	1.00	Good	GOOD	**Retain**
103	1.49	Good	1.6	Good	0.15	Good	1.00	Good	GOOD	**Retain**
104	0.01	**Poor**	60	**Poor**	0.00	Good	1.00	Good	**POOR**	Delete
105	2.06	Good	1.7	Good	0.26	Good	1.00	Good	GOOD	**Retain**
106	4.42	Good	1.6	Good	0.22	Good	0.72	Good	GOOD	**Retain**
107	3.15	Good	1.3	Good	0.20	Good	1.00	Good	GOOD	**Retain**
108	2.89	Good	1.0	Good	0.22	Good	0.89	Good	GOOD	**Retain**
109	3.82	Good	1.4	Good	0.33	Good	1.00	Good	GOOD	**Retain**
110	0.87	Good	0.6	Good	0.28	Good	0.95	Good	GOOD	**Retain**
111	25.84	Good	0.0	Good	0.43	**Poor**	0.81	Good	**POOR**	Delete
112	30.34	Good	1.1	Good	0.38	**Poor**	0.84	Good	**POOR**	Delete
113	27.58	Good	1.3	Good	0.27	Good	0.66	Good	GOOD	**Retain**
114	6.73	Good	1.1	Good	0.22	Good	0.84	Good	GOOD	**Retain**

Table 4.3a gives the parameter values that were estimated while calibrating with 4PL model. Each of the item parameters was adjudged according to some set criteria stated above and decision was taken separately concerning individual estimates as good or bad and whether such item should be retained or deleted from the pooled CBMAT items. Table 4.6 shows that for ***a***-parameter, 10 items were found as poor; 8 items were considered poor for ***b***-parameter; ***c***-parameter had 17 poor items; while upper asymptote parameter gave the least number of poor items to be 4.

Therefore, items that were adjudged good having met the four conditions as '**good**' all through were retained in the decision column. Out of the 114 items that constituted the pooled CBMAT instrument, 77 items were found to fit 4PL model and survived the calibration process.Thess were retained. The remaining 37 items were discarded. These items include item 1, 2, 9, 13, 17, 20, 22, 23, 25, 27, 30, 45, 48, 49, 50, 54, 56, 58, 61, 62, 65, 67, 72, 75, 78, 81, 83, 86, 88, 94, 95, 96, 97, 98,104,111 and item112.

This study is in agreement with the work of Rulison and Loken (2009) where comparison of calibration results of a Computerized Adaptive Testing (CAT) with 2-, 3- and 4PL models was done. Their finding was that 4PL model gave better estimates in terms of the mean slope, difficulty, discrimination and lower asymptote parameters. Also, Loken and Rulison (2010) used a Bayesian approach to effectively recover parameter estimates under 4PL model and found that the overall fit was greatly improved. The work of Tavares, Andrade and Pereira (2004) also argued in favour of the usage of 4PL model.

However, Table 4.3b presents the parameter estimates of the calibration made with 3PL model.

Table 4.3b: Item parameters of the pooled CBMAT Instrument calibrated with 3PL Model

Item No	***a***	**Remark**	***b***	**Remark**	***c***	**Remark**	**Overall**	**Decision**
1	3.76	Good	1.6	Good	0.37	**Poor**	**Poor**	Delete
2	2.69	Good	1.4	Good	0.26	Good	Good	**Retain**
3	1.35	Good	1.2	Good	0.38	**Poor**	**Poor**	Delete
4	3.06	Good	0.4	Good	0.41	**Poor**	**Poor**	Delete
5	2.29	Good	1.4	Good	0.40	**Poor**	**Poor**	Delete
6	1.39	Good	0.6	Good	0.25	Good	Good	**Retain**
7	2.40	Good	1.0	Good	0.50	**Poor**	**Poor**	Delete
8	1.46	Good	-0.2	Good	0.18	Good	Good	**Retain**
9	3.85	Good	1.4	Good	0.20	Good	Good	**Retain**
10	2.66	Good	2.0	Good	0.27	Good	Good	**Retain**
11	2.20	Good	0.6	Good	0.24	Good	Good	**Retain**
12	1.61	Good	2.8	Good	0.23	Good	Good	**Retain**
13	0.85	Good	-1.0	Good	0.00	Good	Good	**Retain**
14	0.61	Good	0.3	Good	0.03	Good	Good	**Retain**
15	2.79	Good	1.1	Good	0.19	Good	Good	**Retain**
16	0.53	Good	-1.3	Good	0.01	Good	Good	**Retain**
17	1.43	Good	0.4	Good	0.34	Good	Good	**Retain**
18	3.84	Good	1.6	Good	0.25	Good	Good	**Retain**
19	1.18	Good	0.9	Good	0.22	Good	Good	**Retain**
20	1.97	Good	1.4	Good	0.33	Good	Good	**Retain**
21	2.91	Good	2.4	Good	0.20	Good	Good	**Retain**
22	1.44	Good	1.3	Good	0.30	Good	Good	**Retain**
23	1.65	Good	2.0	Good	0.28	Good	Good	**Retain**
24	6.21	Good	1.5	Good	0.22	Good	Good	**Retain**
25	0.89	Good	3.2	**Poor**	0.16	Good	**Poor**	Delete
26	1.35	Good	2.0	Good	0.13	Good	Good	**Retain**
27	2.65	Good	1.7	Good	0.27	Good	Good	**Retain**
28	4.26	Good	1.5	Good	0.33	Good	Good	**Retain**
29	2.31	Good	1.9	Good	0.26	Good	Good	**Retain**
30	0.18	**Poor**	0.4	Good	0.04	Good	**Poor**	Delete
31	1.36	Good	-0.7	Good	0.21	Good	Good	**Retain**
32	2.83	Good	2.0	Good	0.17	Good	Good	**Retain**
33	1.79	Good	2.1	Good	0.23	Good	Good	**Retain**
34	0.32	**Poor**	-3.3	**Poor**	0.01	Good	**Poor**	Delete
35	2.07	Good	1.1	Good	0.25	Good	Good	**Retain**
36	1.30	Good	0.4	Good	0.14	Good	Good	**Retain**
37	0.55	Good	1.1	Good	0.15	Good	Good	**Retain**
38	3.71	Good	2.0	Good	0.23	Good	Good	**Retain**
39	1.15	Good	2.1	Good	0.22	Good	Good	**Retain**

40	1.08	Good	2.4	Good	0.14	Good	Good	**Retain**
41	0.67	Good	-1.6	Good	0.00	Good	Good	**Retain**
42	7.63	Good	2.2	Good	0.31	Good	Good	**Retain**
43	0.25	**Poor**	1.6	Good	0.01	Good	**Poor**	Delete
44	1.19	Good	-0.8	Good	0.04	Good	Good	**Retain**
45	3.82	Good	1.7	Good	0.08	Good	Good	**Retain**
46	0.34	**Poor**	-2.0	Good	0.02	Good	**Poor**	Delete
47	0.06	Poor	39.4	**Poor**	0.09	Good	**Poor**	Delete
48	0.81	Good	2.3	Good	0.26	Good	Good	**Retain**
49	3.55	Good	2.1	Good	0.20	Good	Good	**Retain**
50	1.46	Good	0.0	Good	0.13	Good	Good	**Retain**
51	1.23	**Poor**	-3.3	**Poor**	0.19	Good	**Poor**	Delete
52	0.98	Good	1.7	Good	0.22	Good	Good	**Retain**
53	0.01	**Poor**	81.8	**Poor**	0.01	Good	**Poor**	Delete
54	1.13	Good	0.5	Good	0.22	Good	Good	**Retain**
55	0.06	**Poor**	10.7	**Poor**	0.01	Good	**Poor**	Delete
56	2.38	Good	1.2	Good	0.22	Good	Good	**Retain**
57	1.88	Good	1.4	Good	0.34	Good	Good	Retain
58	1.77	Good	0.8	Good	0.24	Good	Good	**Retain**
59	1.91	Good	0.7	Good	0.16	Good	Good	**Retain**
60	2.91	Good	1.2	Good	0.10	Good	Good	**Retain**
61	1.59	**Poor**	-3.2	**Poor**	0.20	Good	**Poor**	Delete
62	5.06	Good	1.8	Good	0.34	Good	Good	**Retain**
63	2.21	Good	0.1	Good	0.49	**Poor**	**Poor**	Delete
64	2.44	Good	2.3	Good	0.13	Good	Good	**Retain**
65	2.11	Good	0.6	Good	0.31	Good	Good	**Retain**
66	1.47	Good	1.4	Good	0.24	Good	Good	**Retain**
67	0.52	Good	-2.1	Good	0.01	Good	Good	**Retain**
68	3.65	Good	1.3	Good	0.28	Good	Good	**Retain**
69	1.63	Good	1.2	Good	0.29	Good	Good	**Retain**
70	0.83	Good	-1.8	Good	0.01	Good	Good	**Retain**
71	2.11	Good	1.2	Good	0.33	Good	Good	**Retain**
72	3.77	Good	1.1	Good	0.13	Good	Good	**Retain**
73	1.43	Good	0.3	Good	0.15	Good	Good	**Retain**
74	1.61	Good	1.9	Good	0.44	**Poor**	**Poor**	Delete
75	2.24	Good	0.8	Good	0.20	Good	Good	**Retain**
76	3.35	Good	2.4	Good	0.19	Good	Good	**Retain**
77	2.85	Good	1.8	Good	0.22	Good	Good	**Retain**
78	0.32	Good	-0.3	Good	0.05	Good	Good	**Retain**
79	4.27	Good	2.1	Good	0.16	Good	Good	**Retain**
80	1.84	Good	1.9	Good	0.23	Good	Good	**Retain**
81	0.96	Good	2.0	Good	0.29	Good	Good	**Retain**
82	4.04	Good	2.5	Good	0.25	Good	Good	**Retain**
83	1.69	Good	1.6	Good	0.16	Good	Good	**Retain**

84	3.43	Good	1.3	Good	0.21	Good	Good	**Retain**
85	0.92	Good	-0.3	Good	0.14	Good	Good	**Retain**
86	0.91	Good	2.6	Good	0.14	Good	Good	**Retain**
87	1.43	Good	0.5	Good	0.38	**Poor**	**Poor**	Delete
88	0.35	**Poor**	1.4	Good	0.05	Good	**Poor**	Delete
89	1.28	Good	2.2	Good	0.08	Good	Good	**Retain**
90	2.84	Good	1.3	Good	0.25	Good	Good	**Retain**
91	0.48	**Poor**	-4.0	**Poor**	0.00	Good	**Poor**	Delete
92	3.48	Good	1.7	Good	0.18	Good	Good	**Retain**
93	1.29	Good	0.2	Good	0.33	Good	Good	**Retain**
94	2.76	Good	1.4	Good	0.14	Good	Good	**Retain**
95	4.39	Good	1.8	Good	0.13	Good	Good	**Retain**
96	0.56	Good	0.5	Good	0.03	Good	Good	**Retain**
97	3.15	Good	1.6	Good	0.18	Good	Good	**Retain**
98	0.82	Good	1.6	Good	0.14	Good	Good	**Retain**
99	0.46	**Poor**	-0.3	Good	0.00	Good	**Poor**	Delete
100	2.17	Good	2.1	Good	0.39	**Poor**	**Poor**	Delete
101	0.91	**Poor**	-3.2	**Poor**	0.03	Good	**Poor**	Delete
102	1.29	Good	1.1	Good	0.24	Good	Good	**Retain**
103	3.70	Good	1.5	Good	0.32	Good	Good	**Retain**
104	2.71	Good	2.2	Good	0.19	Good	Good	**Retain**
105	1.96	Good	2.7	Good	0.26	Good	Good	**Retain**
106	2.65	Good	2.1	Good	0.19	Good	Good	**Retain**
107	1.44	Good	1.0	Good	0.36	Good	Good	**Retain**
108	3.70	Good	1.8	Good	0.30	Good	Good	**Retain**
109	0.10	**Poor**	10.5	**Poor**	0.03	Good	**Poor**	Delete
110	0.69	Good	1.5	Good	0.20	Good	Good	**Retain**
111	1.08	Good	1.3	Good	0.31	Good	Good	**Retain**
112	2.77	Good	2.0	Good	0.21	Good	Good	**Retain**
113	0.95	Good	1.0	Good	0.34	Good	Good	**Retain**
114	2.87	Good	1.3	Good	0.20	Good	Good	**Retain**

Table 4.3b displays the result of the calibration process for the pooled CBMAT responses to the 3PL model. The quality of the parameter (discrimination, difficulty and lower asymptote) indices were ascertained and adjudged either good or poor with the same criteria stated above. It is observed that for *a*-parameter, 100 items were considered good while 14 items were poor. 104 items were accepted as good and 10 items poor for the estimate of *b*-parameter;*c*-parameter had 105 good and 9 poor items.

In all of the 114 items of the pooled CBMAT items, 90 good items survived and fitted the model while 24 items were considered poor and deleted. It appears that when 3PL and 4PL model parameter estimates were compared as regards the number of items which survived calibration process, more good items were retained in the 3PL model. This made the number of discarded items to be reduced by 13 items. Thismight be because only the contribution of guessing was accounted for while carelessness estimate (mistake, anxiety, inattention and so on) was not measured, thereby producing more items than 4PL model.

Ojerinde, Onoja and Ifewulu (2013) calibrated 2012/2013 UTME Use of English test and found that 95% of the items had a better fit with 3PL model. The same was with Fakayode (2018) who calibrated June/November 2015 WAEC Mathematics test items as having a better fit with 3PL model. Other studies that showed that the items of their scales were calibrated with 3PL model are the works of Ani (2014) whose scale found that 49 (98%) out 50 items were reliable and Kpolovie and Emekene (2016) who validated a global non-verbal mental ability scale in Nigeria with 3PL model.
Then, item parameter estimates of 2-parameter logistic model are given in Table 4.8 when calibration was done.

Table 4.3c: Item parameters of the pooled CBMAT Instrument calibrated with 2PL Model

Item no	***a***	***Remark***	***b***	***Remark***	***Overall***	***Decision***
1	0.39	**Poor**	0.9	Good	**Poor**	Delete
2	0.71	Good	0.9	Good	Good	**Retain**
3	0.54	Good	-0.2	Good	Good	**Retain**
4	1.31	Good	-0.5	Good	Good	**Retain**
5	0.44	**Poor**	0.2	Good	**Poor**	Delete
6	0.84	Good	-0.1	Good	Good	**Retain**
7	0.65	Good	-0.7	Good	Good	**Retain**
8	1.24	Good	-0.6	Good	Good	**Retain**
9	0.82	Good	1.3	Good	Good	**Retain**
10	0.27	**Poor**	3.0	Good	**Poor**	Delete
11	1.16	Good	0.0	Good	Good	**Retain**
12	0.20	**Poor**	5.5	**Poor**	**Poor**	Delete
13	0.86	Good	-1.0	Good	Good	**Retain**
14	0.58	Good	0.2	Good	Good	**Retain**
15	1.06	Good	0.7	Good	Good	Delete
16	0.47	**Poor**	-1.5	Good	**Poor**	Delete
17	0.88	Good	-0.6	Good	Good	**Retain**
18	0.61	Good	1.4	Good	Good	**Retain**
19	0.69	Good	0.3	Good	Good	**Retain**
20	0.55	Good	0.6	Good	**Good**	**Retain**
21	0.14	**Poor**	9.2	**Poor**	**Poor**	Delete
22	0.64	Good	0.3	Good	Good	**Retain**
23	0.37	**Poor**	1.9	Good	**Poor**	Delete
24	0.67	Good	1.5	Good	Good	**Retain**
25	0.30	**Poor**	4.2	**Poor**	**Poor**	Delete
26	0.55	Good	2.3	Good	Good	**Retain**
27	0.52	Good	1.4	Good	**Good**	**Retain**
28	0.53	Good	0.9	Good	Good	**Retain**
29	0.41	**Poor**	2.0	Good	**Poor**	Delete
30	0.15	**Poor**	-0.1	Good	**Poor**	Delete
31	1.38	Good	-1.0	Good	**Good**	**Retain**
32	0.44	**Poor**	3.2	**Poor**	**Poor**	Delete
33	0.35	**Poor**	2.8	Good	**Poor**	Delete
34	0.30	**Poor**	-3.5	**Poor**	**Poor**	Delete
35	0.90	Good	0.5	Good	Good	**Retain**
36	1.04	Good	0.0	Good	Good	**Retain**
37	0.43	**Poor**	0.4	Good	**Poor**	Delete
38	0.43	**Poor**	2.6	Good	**Poor**	Delete
39	0.42	**Poor**	1.9	Good	**Poor**	Delete
40	0.43	**Poor**	2.9	Good	**Poor**	Delete
41	0.69	Good	-1.6	Good	Good	**Retain**
42	0.13	**Poor**	5.6	**Poor**	**Poor**	Delete

43	0.23	**Poor**	1.7	Good	**Poor**	Delete
44	1.27	Good	-0.9	Good	Good	**Retain**
45	0.85	Good	2.5	Good	**Good**	**Retain**
46	0.32	**Poor**	-2.2	Good	**Poor**	Delete
47	0.03	**Poor**	63.4	**Poor**	**Poor**	Delete
48	0.31	**Poor**	1.7	Good	**Poor**	Delete
49	0.33	**Poor**	3.9	**Poor**	**Poor**	Delete
50	1.38	Good	-0.3	Good	Good	**Retain**
51	0.02	**Poor**	67.5	**Poor**	**Poor**	Delete
52	0.49	Good	1.1	Good	Good	**Retain**
53	0.03	**Poor**	21.3	**Poor**	**Poor**	Delete
54	0.79	Good	-0.2	Good	Good	**Retain**
55	0.06	**Poor**	10.7	**Poor**	**Poor**	Delete
56	0.84	Good	0.8	Good	Good	**Retain**
57	0.59	Good	0.4	Good	Good	**Retain**
58	0.96	Good	0.1	Good	Good	**Retain**
59	1.23	Good	0.3	Good	Good	**Retain**
60	1.27	Good	1.1	Good	Good	**Retain**
61	0.03	**Poor**	41.3	**Poor**	**Poor**	Delete
62	0.25	**Poor**	2.3	Good	**Poor**	Delete
63	1.26	Good	-1.0	Good	Good	**Retain**
64	0.38	**Poor**	4.6	**Poor**	**Poor**	Delete
65	1.11	Good	-0.2	Good	Good	**Retain**
66	0.60	Good	0.9	Good	Good	**Retain**
67	0.50	Good	-2.2	Good	Good	**Retain**
68	0.69	Good	0.9	Good	Good	**Retain**
69	0.64	Good	0.4	Good	Good	**Retain**
70	0.87	Good	-1.8	Good	Good	**Retain**
71	0.72	Good	0.3	Good	Good	**Retain**
72	1.35	Good	0.9	Good	Good	**Retain**
73	1.16	Good	-0.1	Good	Good	**Retain**
74	0.35	**Poor**	0.1	Good	**Poor**	Delete
75	1.16	Good	0.3	Good	Good	**Retain**
76	0.25	**Poor**	5.7	**Poor**	**Poor**	Delete
77	0.43	**Poor**	2.4	Good	**Poor**	Delete
78	0.27	**Poor**	-0.7	Good	**Poor**	Delete
79	0.29	**Poor**	5.4	**Poor**	**Poor**	Delete
80	0.45	**Poor**	2.0	Good	**Poor**	Delete
81	0.37	**Poor**	1.1	Good	**Poor**	Delete
82	0.17	**Poor**	6.6	**Poor**	**Poor**	Delete
83	0.70	Good	1.5	Good	**Good**	**Retain**
84	0.83	Good	1.1	Good	Good	**Retain**
85	0.85	Good	-0.7	Good	Good	**Retain**

86	0.38	**Poor**	3.2	Good	**Poor**	Delete
87	0.80	Good	-0.6	Good	Good	**Retain**
88	0.27	**Poor**	1.4	Good	**Poor**	Delete
89	0.63	Good	2.7	Good	Good	**Retain**
90	0.72	Good	0.9	Good	Good	**Retain**
91	0.59	**Poor**	-3.3	**Poor**	**Poor**	Delete
92	0.54	Good	2.4	Good	Good	**Retain**
93	0.85	Good	-0.7	Good	Good	**Retain**
94	1.00	Good	1.3	Good	Good	**Retain**
95	0.60	Good	2.7	Good	Good	**Retain**
96	0.52	Good	0.4	Good	Good	**Retain**
97	0.70	Good	1.7	Good	Good	**Retain**
98	0.51	Good	1.3	Good	Good	**Retain**
99	0.40	**Poor**	-0.3	Good	**Poor**	Delete
100	0.21	**Poor**	1.5	Good	**Poor**	Delete
101	0.57	**Poor**	-4.2	**Poor**	**Poor**	Delete
102	0.68	Good	0.4	Good	Good	**Retain**
103	0.52	Good	1.0	Good	Good	**Retain**
104	0.27	**Poor**	5.0	**Poor**	**Poor**	Delete
105	0.16	**Poor**	6.2	**Poor**	**Poor**	Delete
106	0.40	**Poor**	3.3	**Poor**	**Poor**	Delete
107	0.63	Good	-0.3	Good	Good	**Retain**
108	0.35	**Poor**	1.9	Good	**Poor**	Delete
109	0.07	**Poor**	13.6	**Poor**	**Poor**	Delete
110	0.41	**Poor**	0.7	Good	**Poor**	Delete
111	0.50	Good	0.2	Good	Good	**Retain**
112	0.37	**Poor**	3.1	**Poor**	**Poor**	Delete
113	0.54	Good	-0.4	Good	Good	**Retain**
114	0.83	Good	1.1	Good	Good	**Retain**

Table 4.3c gives the result of the calibration process done with the pooled CBMAT response data using 2PL model. For discrimination and difficulty parameter estimates,

50 and 21 items were found not fitting. While 63 items that were found to fit the model survived and were retained in the last column of decision making, leaving a total of 51 items to be discarded.

The 2PL model produced the least number of survived and retained items when compared with 4- and 3PL models. The reason could be that pseudo-guessing and mistake parameters were not accounted for since both could be factors that could affect examinees response which in-turn will after examinees ability (Amarnani, 2009). The calibration result here supported the work of Metibemu (2016) whocalibrated the100 physics achievement test (PAT) items and was found that 98% of the PAT items fit the 2PL model.

Table 4.3d, however presentestimates of the result of calibration with one-parameter logistic model (1PLM).

Table 4.3d: Item parameter of the pooled CBMAT Instrument calibrated with 1PL Model

Item No	*b*	Remark	Decision	Item No	b	Remark	Decision
1	0.7	Good	**Retain**	44	-1.7	Good	**Retain**
2	1.2	Good	**Retain**	45	3.8	**Poor**	Delete
3	-0.2	Good	**Retain**	46	-1.4	Good	**Retain**
4	-1.0	Good	**Retain**	47	3.2	**Poor**	Delete
5	0.1	Good	**Retain**	48	1.0	Good	**Retain**
6	-0.1	Good	**Retain**	49	2.5	Good	**Retain**
7	-0.9	Good	**Retain**	50	-0.5	Good	**Retain**
8	-1.1	Good	**Retain**	51	2.6	Good	**Retain**
9	1.9	Good	**Retain**	52	1.0	Good	**Retain**
10	1.6	Good	**Retain**	53	1.1	Good	**Retain**
11	0.2	Good	**Retain**	54	-0.2	Good	**Retain**
12	2.2	Good	**Retain**	55	1.2	Good	**Retain**
13	-1.5	Good	**Retain**	56	1.3	Good	**Retain**
14	0.2	Good	**Retain**	57	0.5	Good	**Retain**
15	1.3	Good	**Retain**	58	0.3	Good	**Retain**
16	-1.3	Good	**Retain**	59	0.7	Good	**Retain**
17	-0.8	Good	**Retain**	60	2.3	Good	**Retain**
18	1.6	Good	**Retain**	61	2.6	Good	**Retain**
19	0.3	Good	**Retain**	62	1.1	Good	**Retain**
20	0.6	Good	**Retain**	63	-2.0	Good	**Retain**
21	2.6	Good	**Retain**	64	3.4	**Poor**	Delete
22	0.4	Good	**Retain**	65	-0.3	Good	**Retain**
23	1.3	Good	**Retain**	66	1.0	Good	**Retain**
24	1.9	Good	**Retain**	67	-2.1	Good	**Retain**
25	2.5	Good	**Retain**	68	1.2	Good	**Retain**
26	2.4	Good	**Retain**	69	0.5	Good	**Retain**
27	1.4	Good	**Retain**	70	-2.7	Good	**Retain**
28	0.9	Good	**Retain**	71	0.4	Good	**Retain**
29	1.6	Good	**Retain**	72	2.0	Good	**Retain**
30	0.0	Good	**Retain**	73	-0.1	Good	**Retain**
31	-2.1	Good	**Retain**	74	0.1	Good	**Retain**
32	2.7	Good	**Retain**	75	0.7	Good	**Retain**
33	1.9	Good	**Retain**	76	2.8	Good	**Retain**
34	2.1	Good	**Retain**	77	2.0	Good	**Retain**
35	0.8	Good	**Retain**	78	-0.4	Good	**Retain**
36	0.0	Good	**Retain**	79	3.1	**Poor**	Delete
37	0.3	Good	**Retain**	80	1.8	Good	**Retain**
38	2.1	Good	**Retain**	81	0.8	Good	**Retain**
39	1.6	Good	**Retain**	82	2.2	Good	**Retain**
40	2.4	Good	**Retain**	83	1.9	Good	**Retain**
41	-2.0	Good	**Retain**	84	1.7	Good	**Retain**
42	1.5	Good	**Retain**	85	-1.1	Good	**Retain**
43	0.8	Good	**Retain**	86	2.4	Good	**Retain**
				87	-0.9	Good	**Retain**

88	0.7	Good	**Retain**
89	3.2	**Poor**	Delete
90	1.3	Good	**Retain**
91	3.7	**Poor**	Delete
92	2.4	Good	**Retain**
93	-1.0	Good	**Retain**
94	2.3	Good	**Retain**
95	3.1	**Poor**	Delete
96	0.4	Good	**Retain**
97	2.2	Good	**Retain**
98	1.3	Good	**Retain**
99	-0.3	Good	**Retain**
100	0.6	Good	**Retain**
101	4.5	**Poor**	Delete
102	0.6	Good	**Retain**
103	1.0	Good	**Retain**
104	2.6	Good	**Retain**
105	1.9	Good	**Retain**
106	2.5	Good	**Retain**
107	-0.3	Good	**Retain**
108	1.3	Good	**Retain**
109	1.9	Good	**Retain**
110	0.6	Good	**Retain**
111	0.2	Good	**Retain**
112	2.2	Good	**Retain**
113	-0.4	Good	**Retain**
114	1.7	Good	**Retain**

Table 4.3d displays the outcome for calibrating the pooled CBMAT instrument with 1PL model. In this model, items only differ in how hard they are to be answered but not by what means they estimate the latent attribute. So the number of examinees that correctly answer a question is taken as a sufficient statistic for estimating difficulty index. Table 4.9 shows that only 8 items (7%) were deleted while 106 items (93%) were retained. More items seemed good with calibration with 1PL model which might be because of its limited capability to take care of other systematic variance in the cause of testing. After all, the model assumes one for discriminating parameter for all examinees and 0 for the guessing parameter.

This result is in line with the study of Umobong and Tommy (2017) that employed Rasch/1PL model with the help of Winsteps software to assess the infit and outfit statistics with a specified range to determine good and bad items of the NECO biology test. Their finding was that 95% of 2014 and 2015 NECO biology items had infit statistic within the specified range with the Rasch model. However, the result of calibration with 1PL model gives the highest number of retained items in the pooled CBMAT items for this study. This could be as a result of the model's simplicity that it is known for.

In all, when calibration was done with all the four models, 77 items survived calibration with 4PL model, while 90 were with 3PL model, the 2PL model produced 63 items while the highest surviving and retained items were from 1PL with 106 items. It was expected that out of all the models through which calibrations were done, the one with the minimum model-fit assessment value would best explain the pooled CBMAT data. This happened to be the 4PL model with 37 mis-fitting items that were not able to fit the model.

Research Question 3: Is there any significant mean difference in the item parameter estimates of the other IRT models and the model that fits the pooled CBMAT response-data at the developmental stage?

To answer research question three, descriptive statistics for the other dichotomous IRT models and the model that fit the pooled CBMAT data for each of the item parameter estimate were estimated. The distributions of each of the parameter estimates for the different models (1-, 2-, 3- and 4PL models) were subjected to hypothesis testing using

Related Samples Friedman's Two-Way Analysis of Variance by Ranks and Related Samples Wilcoxon Signed Rank Test. This was done in order to establish the significant difference in the parameter estimates when matched.

4.2.4 Comparison of Discrimination Parameter Estimates (*a*)

Table 4.4a: Descriptive Statistics for Discriminating Parameter Estimates of 2-, 3- and 4PL Models

Models	N	Mean	SD
2PL	114	0.56	0.38
3PL	114	1.89	1.45
4PL	114	5.30	9.03

Table 4.4a displays the descriptive statistics of the discrimination parameter estimates of 2-, 3- and 4PL models for the pooled CBMAT. It is discovered that test items under 4PL model were able to differentiate more between higher and lower achievers in the pooled CBMAT items (Mean = 5.3, SD = 9.03) than items under 2PL and 3PL models (2PL: Mean = 0.56, SD = 0.38 and 3PL: Mean = 1.89, SD = 1.45). Thus, to test how significant the difference in the estimates among the models was the hypothesis was tested using Related Samples Friedman's Two-Way Analysis of Variance by Ranks. The result is shown as follows:

Table 4.4b: Hypothesis Test Summary for Discrimination Parameter Estimates

Null Hypothesis	N	Test Statistic	Df	Sig.	Decision
The distributions of discrimination parameter estimates for 2PL, 3PL and 4PL are the same	114	122.68	2	0.00	Reject the null hypothesis

Table 4.4b displays the distribution of the discrimination parameter estimates of 2-, 3- and 4PL models. A null hypothesis that there is no significance difference in the discrimination parameter estimates for 2-, 3- and 4PL models was conducted. The result specifies that the H_0 was rejected which implies that the discrimination estimates were distinctly different from one another (Test Statistic = 122.68, $p < 0.05$). This signifies that there is statistical significance difference in the discrimination estimates of 2-, 3- and 4PL models for the pooled CBMAT data. The implication of the result is that in assessing how well the test items were able to differentiate between less and highly-able respondents, various models in dichotomous category might fair distinctly from one another. Thereby, appropriate model is suggested for a particular response data.

4.2.5 Comparison of Difficulty Parameter Estimates (*b*)

Moreover, the distributions of 1-, 2-, 3- and 4PL models with respect to item difficulty parameter estimates were also compared and subjected to hypothesis testing using Related Samples Friedman's Two-Way Analysis of Variance by Ranks. Table 4.4c offers the result.

Table 4.4c: Mean and Standard Deviation for Difficulty Parameter Estimates

Models	N	Mean	SD
1PL	114	0.98	1.44
2PL	114	0.66	10.13
3PL	114	1.50	8.69
4PL	114	1.54	5.75

Table 4.4c displays the descriptive statistics of the difficulty parameter estimates of 1-, 2-, 3- and 4PL models for the pooled CBMAT. The result shows that test items under 4PL model appeared to have the highest mean (Mean = 1.54, SD = 5.75). This indicate that items calibrated under 4PL model are more difficult than items under 1PL, 2PL and 3PL models (1PL: Mean = 0.98, SD = 1.44: 2PL: Mean = 0.66, SD = 10.13: 3PL: Mean = 1.50, SD = 8.69). Although items of CBMAT in 3PL model show less difficulty than 4PL model, test items of the pooled CBMAT in 1PL model seem more difficult than 2PL model.

To test whether the differences observed in the difficulty estimates of the CBMAT items among the four models were statistically significant, Related Samples Friedman's Two-Way Analysis of Variance by Ranks was conducted. Table 4.4d shows the outcome.

Table 4.4d: Hypothesis Test Summary for Difficulty Parameter Estimates

Null Hypothesis	N	Test Statistic	df	Sig.	Decision
The distributions of difficulty parameter estimates for 1PL, 2PL, 3PL and 4PL are the same	114	24.45	3	0.00	Reject the null hypothesis

Table 4.4d reveals that the distribution of difficulty parameter estimates of pooled CBMAT items under the four models was noticeably different from one another (Test Statistics=24.45, $p < 0.05$). It means that the null hypothesis that there is no significant difference among the difficulty parameter estimates was rejected. Difficulty parameter estimates which is an indication of location index of examinee ability on the ability continuum display dissimilarity of its estimated value among the four dichotomously-scored response IRT models for the pooled CBMAT items.

4.2.6 Comparison of Estimates of Guessing Parameter (*c*)

Also, the distributions of the lower asymptote estimates for 3- and 4PL models were as well compared and subjected to hypothesis testing using Related Samples Wilcoxon Signed-Rank test. The result is hereby presented in Table 4.4e.

Table 4.4e: Descriptive Statistics for Pseudo-Guessing Parameter Estimates

Models	N	Mean	SD
3PL	114	0.20	0.12
4PL	114	0.23	0.12

Table 4.4e presents the mean and standard deviation of the guessing parameter estimates of 3- and 4PL models of the CBMAT items. The outcome shows that the percentage of guessing to items by the examinees under 4PL model is more in the CBMAT items (Mean = 0.23, SD = 0.12) than the items under 3PL models (3PL: Mean = 0.2, SD = 0.12),although the difference of 0.03% seems small. To confirm whether the acclaimed negligible difference observed between the two models was statistically noteworthy, Related-Samples Wilcoxon Signed Rank Test was carried out. A null hypothesis that the median of difference between the guessing parameter estimate of 3PL and that of 4PL models is equal to zero was tested. The outcome is offered in Table 4.4f.

Table 4.4f: Hypothesis Test Summary for Pseudo-Guessing Parameter Estimates

Null Hypothesis	N	Test Statistic	Standard Error	Standardized Test Statistic	Sig.	Decision
The distributions of Pseudo-Guessing parameter estimates for 3- and 4PL are the same	114	3752.5	335.22	2.09	0.04	Reject the null hypothesis

The aftermath of the investigation on the hypothesis tested specifies that the distributions of guessing parameter estimates of the pooled CBMAT items for 3- and 4PL models are not the same (Standardized Test Statistics=2.09, $p < 0.05$). The subjected distributions give a seemingly significant difference. This means that guessing parameter estimate will fare differently for different models.

However, from the results obtained in the different analysis carried out in the cause of answering Research Question Three, it is observed that the estimates of item parameter obtained from each the models (1-, 2-, 3- and 4PL models) display statistical significant difference. This could connote that if there had not been a new development in the psychometric parlance, where the fourth model could be explored, there is a tendency of the continual usage of the three most popular uni-dimension IRT models (1-, 2- and 3PL) in calibrating test items that are of dichotomous-response format.

Georgiev (2008) was of the view that the most appropriate means of approaching model assessment is by analysing response data with all relevant models. It implies that if all available models are not exhausted, there might be estimation bias/error in the process of estimating the true ability of the respondents.

This study agrees with Adegoke (2013) who compared the item statistics obtained from classical test theory and 2PL IRT models of Physics Achievement Test (PAT) that was developed and administered to SSII students in Ibadan Zone I of Oyo State. It was discovered that the two models produced comparable results but 2PL IRT model gave statistics that were more stable and reliable. Other studies that compared item/person's parameters and found significant difference are the works of Fakayode (2018), Metibemu (2016) and Adewale (2015). The findings of Courville (2004) and Ojerinde (2013) revealed in their studies that significant differences were not evident in parameter estimates of the models used.

Research Question 4: How consistent is the model used in calibrating the pooled CBMAT response-data at the development stage to the model used in calibrating the final CBMAT response-data at the real study stage?

4.3 Phase II:

Having understood from research question one that 4PL model fitted the pooled CBMAT response data at the development stage, calibration of both item and examinee parameters were also done. It was also revealed in the second research question that 77 (68%) items out of 114 that fitted 4PL model survived validation process. This was what led to reconstructing another table of specification (Table 3.5, page 130 in Chapter three), where the very best 40 items were selected based on how steep or informative their individual item characteristic curves (ICCs) were. The 40-item CBMAT that made up the final instrument was used to collect data for the real study.

4.3.1 Assessing Trait Dimensionality Assumption of the Final CBMAT Response Data

Research question four was however answered by subjecting response data from examinees who answered the final CBMAT instrument to analysis.Calibration was done to see whether the model that fitted the test data of the pooled CBMAT instrument at the development stage was consistent with the model that fitted the final CBMAT response data.

Assessment of the number of dimensions of the final CBMAT instrument was done by subjecting the responses of the examinees to Stout's test of essential dimensionality of DIMTEST 2.0 (Stout, 2005) software. A null hypothesis (Ho) "test data is essentially unidimensional" was tested and a decision of whether to reject or not to reject the null hypothesis was taken. The finding shows that a dominant trait (ability) was exhibited by the examinees in the course of responding to the items of the scale. The result is shown in Table 4.4g.

Table 4.4g: Essential Dimensionality of the Final CBMAT Instrument

Assessment Subtest (AT)	Partitioning Subtest (PT)
2 5 8 9 10 12 13 15 18 19 29	1 3 4 6 7 11 14 16 17 21 23 24 25
22 27 29 30 33 35 36 39	26 28 31 32 34 37 38 40

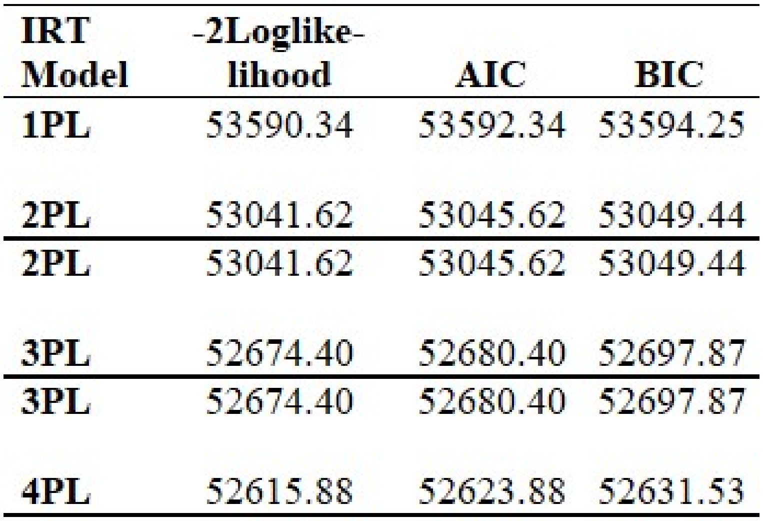
DIMTEST Statistic			
TL	TGbar	T	p-value
6.2187	6.2818	-0.0628	0.5250

Table 4.4g reveals that the abilities measured by the assessment subset (AT) (primary dimension) were not significantly different from the abilities measured by the partitioning subset (PT) (secondary dimension) (T= -0.0628, p > 0.05). This indicates that the final CBMAT instrument also measured only one dominant trait among the examinees in their response to the instrument. Then, the assumption of unidimensionality was tenable in the response data for the final CBMAT instrument.

4.3.2 Assessing Model-Data fitof the final CBMAT response data

Having ascertained that the final CBMAT instrument is unidimensional, the model-data fit was also assessed. This is done through the MIRT package of the R software through R-Studio environment. Comparison of the different values of -2loglikelihood, Akaike and Bayesian information criteria (AIC and BIC) were made. The result is given in table 4.4h.

Table 4.4h: Model-data fit Assessment of the Survived CBMAT Instrument



IRT Model	-2Loglike-lihood	AIC	BIC
1PL	53590.34	53592.34	53594.25
2PL	53041.62	53045.62	53049.44
2PL	53041.62	53045.62	53049.44
3PL	52674.40	52680.40	52697.87
3PL	52674.40	52680.40	52697.87
4PL	52615.88	52623.88	52631.53

Table 4.4h presents the model-data fit assessment showing the model that produced the best fit when calibration was done for the final scale. For response data, result shows that when the values obtained for 1- and 2PL models was compared, an indication that 2PL had -2Loglikelihhod = 53041.62, AIC = 53045.62 and BIC = 5353594.25 values that were less than -2Loglikelihhod = 53590.34, AIC = 53592.34 and BIC = 53049.44 values of 1PL model was evident. Also, the same decision applies to the comparison of 2PL and 3PL models as well as that of 3PL and 4PL models consecutively. The result eventually revealed that four-parameter logistic model fitted the final CBMAT response data.

The analysis above shows that the final CBMAT and pooled CBMAT scales were unidimensional and both were found to fit 4PL dichotomously scored response format IRT model. Result shows that both instruments (pooled and final CBMAT) were consistent in their calibration processes with the 4PL model.

Research Question 5: Is there any significant mean difference in the examinee's parameter estimates of the other dichotomous IRT models and the model that fits the final CBMAT response data?

The distribution of each of the examinees' parameter estimates (ability) in the four dichotomous (1-, 2-, 3- and 4PL) models were subjected to hypothesis testing under Related-Samples Friedman's Two-Way Analysis of Variance by ranks. This is to check if there was any observed difference among the estimated abilities and to see if the differences were significant. The results are presented in Figure 4.3 and Table 4.5.

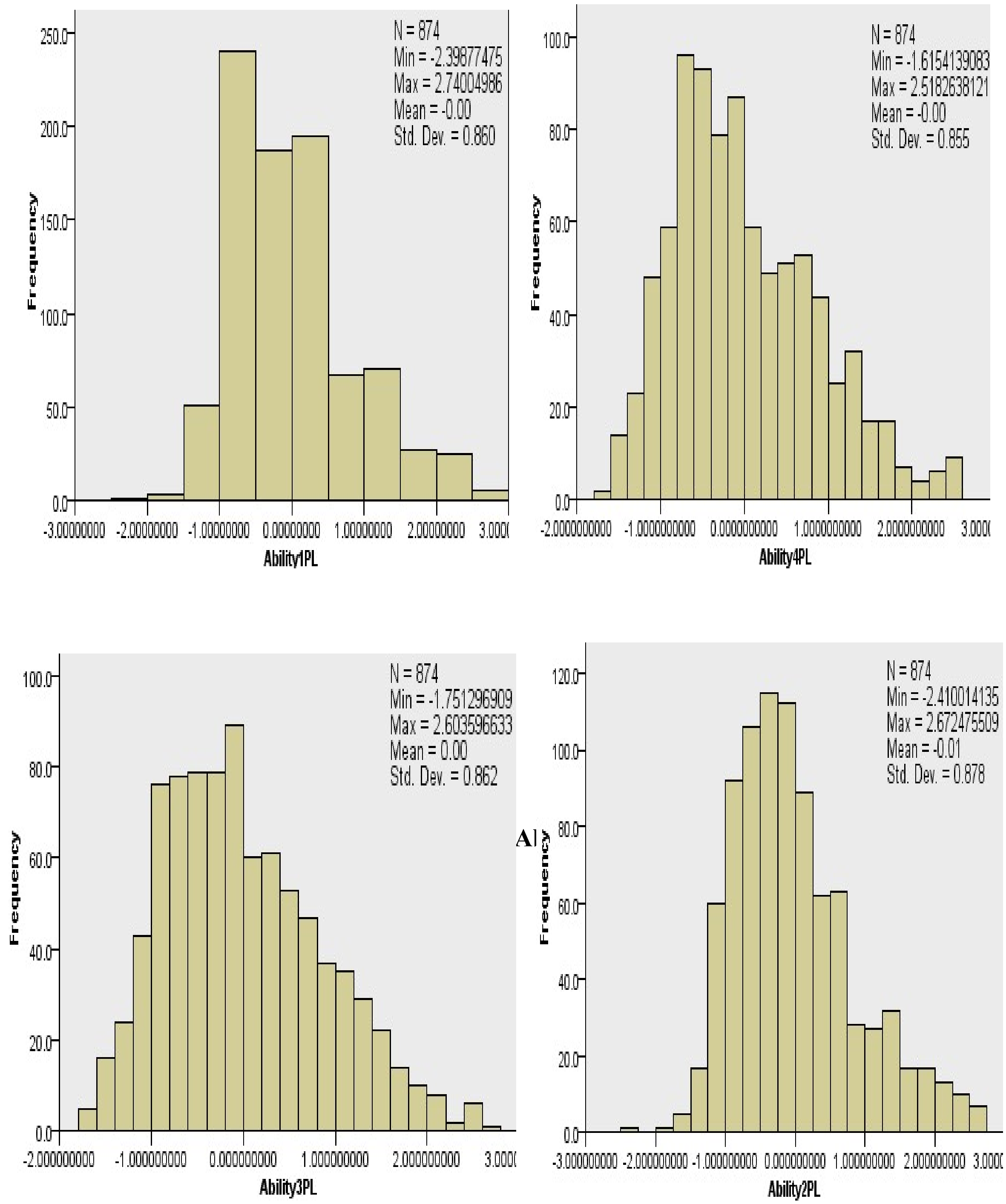


Figure 4.3: Mean Rank Distributions of Ability Parameter Estimates under 1-,2-,3- and 4PL Models

Table 4.5: Hypothesis Test Summary for Examinees' Parameter Estimates of the Final CBMAT instrument

Null Hypothesis	N	Test Statistic	Df	Sig.	Decision
The distribution of ability parameter estimates for 1-, 2-, 3- and 4PL are the same	874	16.89	3	0.01	Reject the Null Hypothesis

Figure 4.3 as well as Table 4.5 present the results obtained from Freidman's Q statistic showing the difference observed in ability estimates of 1-, 2-, 3- and 4PL models of the final CBMAT instrument. Friedman's test statistic indicates that ability scores from the four different models were evaluated differently. This is because the null hypothesis that the distributions of the ability parameter estimates for 1-, 2-, 3- and 4PL models are the same was rejected (Test Statistic = 16.89, $p < 0.05$). Therefore, the significance test tells how statistically confident it can be that there is truly a difference in the ability estimates of the four models. The result affirms that in the estimated ability parametergotten, a significance difference existed with the different available models.

By implication, the model that best fits a data-set is expected to be used in the calibration process since different models will produce different results. Before now, calibration was done only with the 1-, 2- and 3PL models (Amarnani, 2009: Ojerinde *et al*, 2014) which could mean that true ability estimates of the examinees might not be accurately estimated. This is because of some systematic variances that were not taken care of in the previous models. But because of the development of new sophisticated statistical packages that could estimate the parameters of highly parameterised model that takes care of some other estimation error, some estimation errors incurred are now taken care of when done with 4PL model (Loken and Rulison, 2010; Liao *et al*, 2012).

Research Question 6: Is there any significant mean difference in the item parameter estimates of the other dichotomous IRT models and the model that fits the CBMAT data at the final stage?

Research question six was answered by runninganalysis with the distributions of item parameters estimated from the response data of the final CBMAT instrument. Descriptive statistics of each of the estimates was found to see if there was any observed difference in their means. In the same vein, the distributions of the item parameter estimates for the different models (1-, 2-, 3- and 4PL) were subjected to hypothesis testing to observe any significant difference when compared. The results are hereby presented in Table 4.6a

4.4.1 Discrimination Parameter (*a*)

Table 4.6a: Descriptive Statistics for Discriminating Parameter Estimates

Models	N	Mean	SD
2PL	40	0.69	0.31
3PL	40	2.41	2.56
4PL	40	5.01	7.04

Table 4.6a shows the mean and standard deviation of the discrimination parameter estimate of 2-, 3- and 4PL models of the final CBMAT items. The result shows that test items under 4PL model differentiated the better amid well-able and less-able examinees in the final CBMAT items (Mean = 5.01, SD =r7.04) than items under 2PL and 3PL models (2PL: Mean = 0.69, SD = 0.31and 3PL: Mean = 2.41, SD = 2.56). Thus, hypothesis testing was carried out using Related Samples Friedman's Two-Way Analysis of Variance by Ranks to find whether there is significant difference in the discriminating parameters of 2-, 3- and 4PL models. The result is given as follows:

Table 4.6b: Hypothesis Test Summary for Discrimination Parameter Estimates

Null Hypothesis	N	Test Statistic	Df	Sig.	Decision
The distribution of discrimination parameter estimates for 2PL, 3PL and 4PL are the same	40	51.35	2	0.00	Reject the null hypothesis

Table 4.6b offers result for discrimination estimates of 2-, 3- and 4PL models when the final CBMAT data was used. A null hypothesis that there is no significance difference in the discrimination parameter estimates for 2-, 3- and 4PL models was tested. The null hypothesis was rejected. This implied that the discrimination estimates were distinctly different from one another (Test Statistic = 51.35, $p < 0.05$). This signifies that there is statistical significance difference in the discrimination estimates of 2-, 3- and 4PL models. The implication of the result was that test items under the different models discriminated differently amidst high and low achieving students. This finding was the same when the discrimination parameters of the different models in the pooled CBMAT items were compared.

4.4.2 Difficulty Parameter (*b*)

Also, the distributions of 1-, 2-, 3- and 4 parameter logistic models with respect to their item difficulty parameter estimates were compared and subjected to the Related Samples Friedman's Two-Way Analysis of Variance by Ranks statistic. The outcome is given in Table 4.6c.

Table 4.6c: Descriptive Statistics for Difficulty Parameter Estimates

Models	N	Mean	SD
1PL	40	0.36	1.18
2PL	40	0.52	1.43
3PL	40	0.94	1.07
4PL	40	1.06	1.75

Table 4.6c displays the descriptive statistics of the difficulty parameter estimates of 1-, 2-, 3- and 4PL models of the final CBMAT instrument. The result shows that test items under 4PL model appeared more difficult (Mean = 1.06, SD = 1.75) than items under 1PL, 2PL and 3PL models (1PL: Mean = 0.36, SD = 1.18: 2PL; Mean=0.52, SD = 1.43: 3PL: Mean = 0.94, SD = 1.07).

However, Related Samples Friedman's Two-Way Analysis of Variance by Ranks was carried out to see whether the differences observed in the difficulty estimates of the final CBMAT items among the four models were statistically important. The result is shownon Table 4.6d.

Table 4.6d: Hypothesis Test for Difficulty Parameter Estimates

Null Hypothesis	N	Test Statistic	df	Sig.	Decision
The distributions of difficulty parameter estimates for 1PL, 2PL, 3PL and 4PL are the same	40	27.27	3	0.00	Reject the null hypothesis

Table 4.6d reveals that the distributions of difficulty parameter estimates of the final CBMAT items under the four dichotomous IRT models were noticeably different from one another (Test Statistics=27.27, $p<0.05$). It means that the null hypothesis that states that the distributions of the difficulty parameter estimates are the same was rejected. The difficulty parameter, an indication of where examinees are located on the ability scale, displays that there is divergence of its estimated value among the four dichotomously-scored response IRT models for the final CBMAT response data. The finding of the difficulty parameter in the final CBMAT items is in the support of what was found with the pooled CBMAT response data.

4.4.3 Pseudo-guessing Parameter (*c*)

Table 4.6e: Descriptive Statistics for Pseudo-Guessing Parameter Estimates

Models	N	Mean	SD
3PL	40	0.24	0.14
4PL	40	0.27	0.16

Table 4.6e presents the mean and standard deviation of the guessing parameter (lower asymptote) estimates of 3- and 4PL models of the final CBMAT. The result shows that the percentage of guessing to test items by the examinees when 4PL model was used in calibration was more (27%) than those who were subjected to guessing when calibration was carried out with 3PL models (24%) using the final CBMAT scale. The difference in estimates of the lower asymptote parameter for both models seems small by 0.03%.

To check whether the difference observed in the lower asymptote estimates of final CBMAT items between the two models was truly minimal according to the mean result, Related-Samples Wilcoxon Signed Rank Test was conducted. A null hypothesis that the median of difference between the guessing parameters of 3PL and that of 4PL models is equal to zero was tested. The outcome is shown in Table 4.6f.

Table 4.6f: Hypothesis Test Summary for Pseudo-Guessing Parameter Estimates

Null Hypothesis	N	Test Statistic	Standard Error	Standardized Test Statistic	Sig.	Decision
The distributions of Pseudo-Guessing parameter estimates for 3- and 4PL are the same	40	508	71.64	1.65	0.1	Do not reject the null hypothesis

The subjected estimates of guessing parameters of both 3- and 4PL models to hypothesis testing gives the result in Table 4.6f and a seemingly insignificant difference is observed. It therefore shows that the hypothesis that the distribution of guessing parameter estimates of final CBMAT items for 3- and 4PL models are the same (Standardized Test Statistics=1.68, $p > 0.05$) is tenable. This means that the guessing parameters display that there is no distinction in their estimated values in 3- and 4PL IRT models for the final CBMAT response data. This finding is not in agreement with the result obtained when the distribution of the guessing parameter estimates for the pooled CBMAT at the development stage were compared.

Research Question 7: What are the estimates of item and examinee's characteristics of the Response Time model when the final CBMAT response and response time data are used?

Having transformed both response and response-time data of 874 test takers into the appropriate format for analysis, the datawas then subjected to analysis with the help of a Gibbs sampler approach (A Bayesian Approach that had been programmed into R-package functions of version 0.3.5 called LNIRT (Log-normal response time IRT modelling) (a modified version of the *CIRT* R-package of Fox, Klein Entink and van der Linden, 2007). This technique helps in the concurrent analysis of responses and response times in an IRT modelling framework. This is done by applyingLognormal model to the response time data, and normal IRT model to the response data. The joint LNIRT model was also used to estimate person-fit statistics for the for the CBMAT response data.

The instrument used for this study (CBMAT) enables examinees' responses and response time (time spent on each item) to be automatically recorded. This is what made the usage of LNIRT model possible such that the two data sets were analysed with the hierarchical/joint model approach. Fox and Marianti (2017) as well as Fox (2018) applied the usage of the joint model approach (LNIRT) in their studies. Schnipke and Scrams (2002) found from their study that a significant interpretation was given to the response time data because the data found a better fit with LNIRT model. Markov Chain Monte Carlo (MCMC) diagnostics (Klein, Fox and Van der Linden, 2009; Suh, 2016) was utilized in assessing the convergence of the chains.

A burn-in time of 1000 repetitions with an aggregate of 5000 MCMC iterations was achieved in the diagnosis when estimating the model parameters and their respective means and variances. The estimated parameters of LNIRT model include: (a) item discrimination; a_k(b) item difficulty; b_k (c) guessing; c_k (d) time discrimination; Φ_k (e) time-intensity (λ_k) for item, while person's parameters include (a) ability; θ_i and (b) speed; ζ_i estimates. The code for R adopted in the estimation procedures is found in the R-package for LNIRT in Appendix XVII.

Table 4.7a presents the item parameter estimates evaluated from the LNIRT model while Table 4.7b gives the estimates of the test takers' ability (θ_i) and speed (ζ_i) parameters (Full table can be seen in Appendix XVIII; pages 281-287). Meanwhile, Table 4.7c displays the descriptive statistics for parameter estimates of LNIRT model.

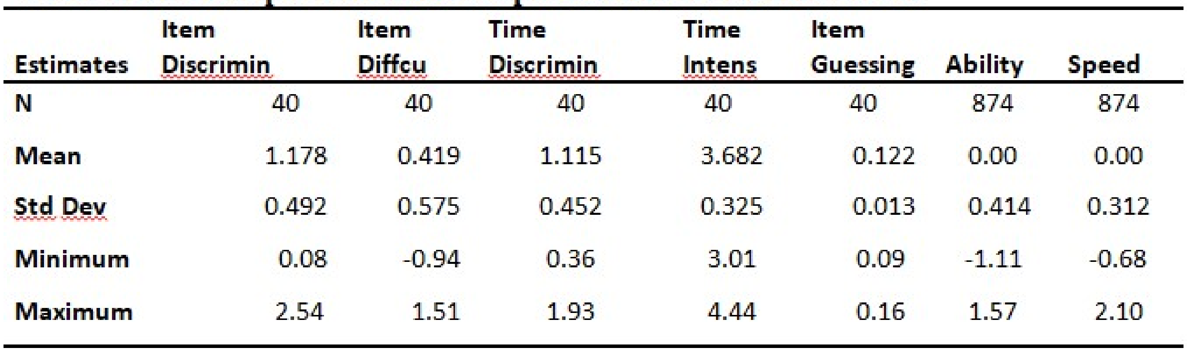
Table 4.7a: Parameter Estimates of the Final CBMAT Scale Calibrated with LNIRT

Item S/N	Item Discri	Item Difficu	Time Discri	Time Inten	Item Guessing
1	0.95	-0.81	0.9	3.60	0.15
2	1.38	-0.14	1.40	3.70	0.13
3	0.08	-0.002	1.45	3.70	0.13
4	0.61	0.54	1.65	4.01	0.12
5	2.07	0.76	1.72	4.11	0.10
6	0.91	1.25	1.60	3.86	0.11
7	0.98	0.51	1.93	4.44	0.12
8	0.62	-0.03	1.46	3.98	0.13
9	1.28	-0.41	1.3	3.75	0.14
10	1.46	-0.06	1.73	3.57	0.13
11	0.14	0.16	1.3	3.32	0.12
12	0.92	0.64	1.14	3.36	0.12
13	1.40	0.83	1.33	3.43	0.13
14	0.79	-0.81	1.26	3.59	0.16
15	1.01	0.5	1.83	4.14	0.12
16	1.53	0.6	1.42	3.74	0.13
17	1.44	1.11	1.44	3.72	0.12
18	1.60	0.85	1.43	3.87	0.12
19	1.76	0.26	1.63	3.91	0.13
20	0.95	0.15	1.06	3.82	0.13
21	1.29	0.1	1.23	3.64	0.13
22	1.73	0.42	1.11	3.37	0.13
23	0.92	-0.94	1.14	4.06	0.16
24	2.54	1.51	1.43	3.88	0.09
25	0.51	0.19	1.04	3.8	0.12
26	0.8	0.58	1.17	3.79	0.12
27	1.66	0.45	0.91	3.35	0.11
28	1.35	1.38	1.08	4.06	0.11
29	1.14	1.11	0.82	3.69	0.12
30	0.86	0.71	0.94	4.42	0.11
31	1.23	0.18	0.53	3.01	0.12
32	1.63	0.93	0.42	3.06	0.11
33	1.55	0.38	0.50	3.84	0.12
34	0.66	1.05	0.59	3.57	0.11
35	0.87	0.11	0.61	3.93	0.13
36	1.00	0.51	0.58	3.5	0.12
37	1.32	0.75	0.42	4.03	0.11
38	1.5	0.58	0.36	3.48	0.11
39	0.97	-0.16	0.38	3.13	0.13
40	1.71	1.02	0.37	3.57	0.11

Table 4.7b: Representative of the examinee parameters of the final CBMAT scale calibrated with LNIRT model

S/N	ABILITY	SPEED	S/N	ABILITY	SPEED	S/N	ABILITY	SPEED
1	0.99	0.13	30	0.34	-0.33	59	0.32	-0.09
2	-0.07	0.37	31	0.16	1.06	60	0.10	0.05
3	-0.23	0.04	32	-0.28	0.67	61	-0.37	-0.08
4	0.19	-0.12	33	0.24	0.11	62	-0.17	0.68
5	0.09	0.95	34	0.47	-0.17	63	0.18	-0.14
6	-0.32	0.34	35	0.39	-0.05	64	-0.38	-0.38
7	0.30	-0.06	36	-0.31	0.82	65	0.12	0.06
8	0.32	-0.23	37	0.45	0.30	66	0.03	-0.26
9	0.06	-0.23	38	-0.10	0.77	67	-0.18	-0.20
10	0.16	-0.15	39	0.61	0.08	68	-0.06	0.10
11	-0.05	-0.23	40	0.40	-0.01	69	-0.20	0.72
12	0.13	0.07	41	0.30	0.93	70	-0.24	0.32
13	-0.16	0.44	42	-0.20	0.06	71	0.35	-0.19
14	0.51	-0.21	43	-0.26	-0.25	72	-0.42	-0.01
15	0.11	0.45	44	0.50	1.08	73	0.14	-0.01
16	-0.11	0.06	45	-0.27	0.11	74	-0.50	0.87
17	-0.44	-0.06	46	-0.06	-0.09	75	-0.01	0.61
18	0.24	-0.04	47	0.10	0.05	76	-0.39	1.11
19	-0.10	1.12	48	0.66	-0.08	77	-0.18	-0.20
20	-0.30	-0.12	49	0.02	0.68	78	-0.09	0.77
21	0.60	0.08	50	-0.13	-0.14	79	-0.21	0.51
22	0.40	-0.01	51	-0.16	-0.38	80	-0.39	0.69
23	-0.05	0.12	52	-0.44	0.06	81	-0.04	0.00
24	-0.05	-0.38	53	0.29	-0.26	82	0.34	-0.18
25	-0.19	-0.42	54	0.68	-0.20	83	-0.33	0.06
26	0.18	0.02	55	0.07	0.10	84	-0.35	0.35
27	-0.25	-0.33	56	-0.10	0.72	85	-0.64	-0.08
28	-0.13	-0.39	57	-0.34	0.32	86	0.23	0.64
29	-0.13	-0.14	58	-0.07	-0.19	87	0.19	0.02
----	-----	-----	----	-----	-----	----	-----	----
----	-----	-----	----	-----	-----	----	-----	----
----	-----	-----	----	-----	-----	----	-----	----
863	0.61	-0.28	867	-0.45	0.28	871	-0.76	0.79
864	-0.11	0.42	868	-0.47	-0.28	872	-0.42	-0.08
865	-0.04	-0.19	869	-0.11	-0.22	873	-0.10	-0.36
866	-0.49	0.36	870	-0.54	0.42	874	-0.72	-0.02

Table 4.27: Descriptive statistics for parameters for LNIRT model



Estimates	Item Discrimin	Item Diffcu	Time Discrimin	Time Intens	Item Guessing	Ability	Speed
N	40	40	40	40	40	874	874
Mean	1.178	0.419	1.115	3.682	0.122	0.00	0.00
Std Dev	0.492	0.575	0.452	0.325	0.013	0.414	0.312
Minimum	0.08	-0.94	0.36	3.01	0.09	-1.11	-0.68
Maximum	2.54	1.51	1.93	4.44	0.16	1.57	2.10

Table 4.7a and 4.7b give the estimated item and person's parameters while Table 4.7c provides their estimated descriptive statistics. The estimated mean and standard deviation for item difficulties (*b*) value are 0.419 and 0.575 while difficulties range from -0.94 to 1.51. It is seen that the range in item difficulties is relatively large.This is an indication that the examinees 'ability on the whole range of the scale continuum is accurately estimated (Fox and Marianti, 2017; Fox, 2016 and Suh, 2016). The estimated mean for time intensities (λ_k) is around 3.7 with a variance of 0.325. Time intensity mean value of exp (3.7) on a logarithmic scale which is approximately 40seconds is the average time needed for test takers to complete an item of the CBMAT scale.

Meanwhile, time intensity estimate ranges from *exp* (3.01) ≈ 20 secs to *exp* (4.44) ≈ 85 secs (1min 25secs) for each item of the scale. This implies that for the items of the CBMAT scale, each examinee is expected to spend a minimum of 800secs (13.33mins) and a maximum of 3400secs (56.67mins) to correctly respond to the whole of the 40 items of the scale according to individual examinee's ability on the continuum. However, in the observed response time, it was shown that it took 15 examinees to answer 40 items in less than 800secs (minimum estimated time by the model) and 21 examinees spent more than 3400secs (maximum estimated time by the model). In all, 36 out of 874 test takers did not fall within the estimated time the LNIRT model evaluated. This shows that about 4% of the examinees response time did not follow the estimated time while 96% are able to work within the estimated time. the major differences that ensued in RTs are accounted for the variation in time intensity.

For item discrimination parameter (*a*), only 2 items of the scale (3 and 11) discriminate poorly with values 0.08 and 0.14 while item 24 (2.54) seems the most discriminating item that is able to differentiate appropriately between low and high ability respondents. Time discrimination parameter (Φ_k) on the other hand, differentiates between item time intensity and test taker's speed with a term (λ_k - ζ_i). The mean value of 1.115 indicates that on the average, at every 3secs, time discrimination parameter works on the sensitivity of an item for the different speed level of the examinees. The time residual variance is 0.452 and it ranges from 0.36 to 1.93. The Figures in 4.7 give the scatter plots of item parameter estimates of the 40-CBMAT scale to further show the patterns of variability among the test takers.

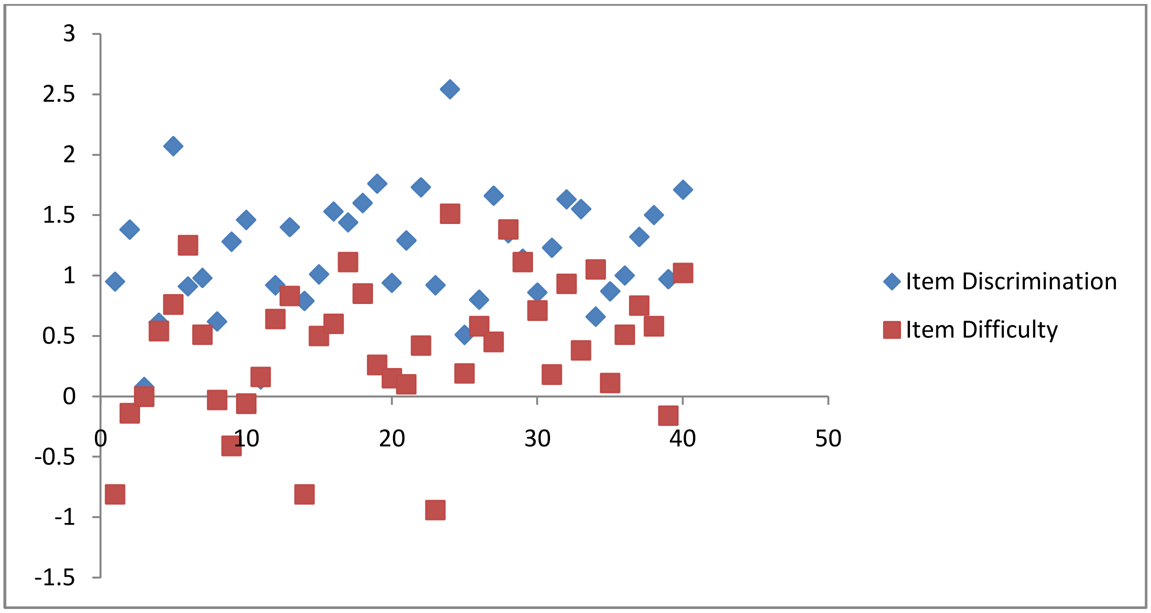


Figure 4.4a: Scatter Plot of Item Discrimination versus Item Difficulty

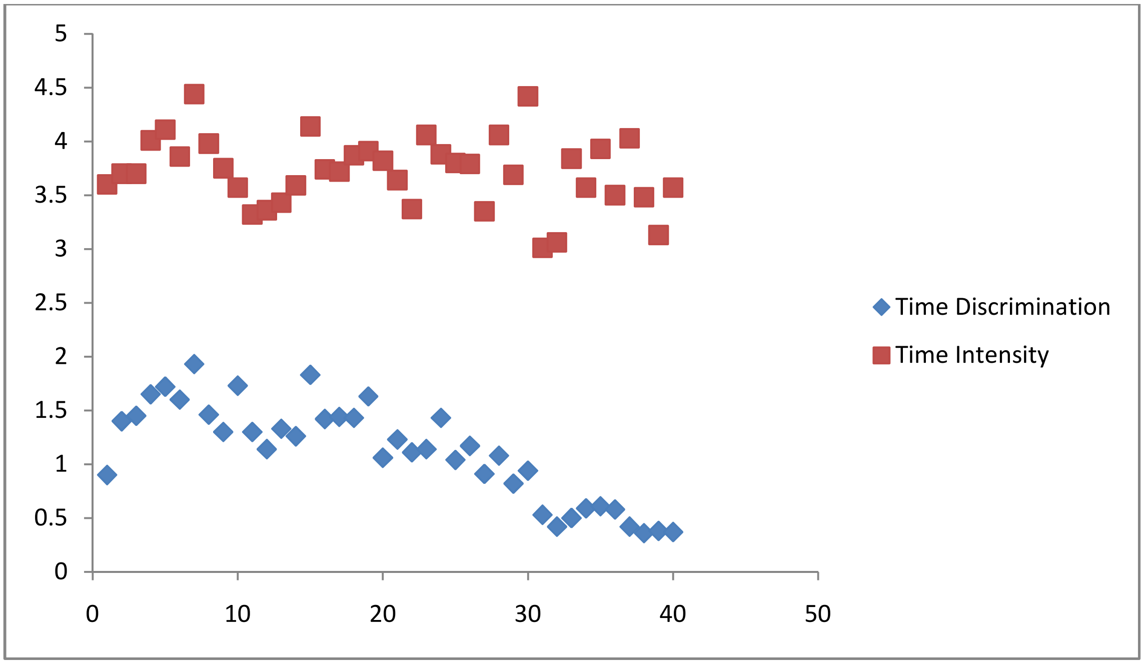


Figure 4.4b: Scatter Plots of Time Discrimination versus Time Intensity

In the plot of figure 4.4a, most items discriminate satisfactorily among examinees' proficiency levels, where just two items had high discriminating estimates of above 2.0. The time discrimination patterns in figure 4.4b are higher for the first 20 items which is an indication that responses to those items show more variation between slow and fast working examinees. Then, it retrogressed as the items get halfway and to the finishing point. However, it is typical of some studentswhen they realise that their time is no longer sufficient to beginning to exhibit some surprizingresponse time when assessment process is about finishing. This implies that the time discrimination power became lessened between the working speeds of the respondents. This finding about the pattern of variability of the item parameters is in line with the findings of Marianti (2015).

For persons' parameters (ability and speed) in Table 4.27, analysis shows that the mean estimates for both ability and speed levels are the same (0.00) while their variation values across examinees are 0.414 and 0.312 and 1.57 to 2.1 respectively. The estimated means for person parameters is pointing to the fact that examinees operated under average ability and speed throughout the entire test. This finding laid credence to the work of Fox and Marianti (2017) while Marianti (2015) found a little high mean in the person parameters estimated from the 40-item chess test, which was contrary to the finding of this study.

Research Question 8: Is there any significant relationship between item and examinee's parameters of the LNIRT response time model?

In order to answer research question eight, the relationship between item characteristics with examinee ability parameters of the LNIRT joint model were investigated with Pearson Product Moment Correlation and results are presented in Table 4.8.

Table 4.8: Relationship Estimates among Parameters of the LNIRT Model

Variance Components		LNIRT	
		Correlation Coeff.	**P-value**
Person Covariance Matrix			
Ability & Speed	$\rho_{\theta\zeta}$	-0.061	0 .069
Item Covariance Matrix			
Discrimination	$\sum_{12}$	0.387*	0.014
	$\sum_{13}$	-0.025	0.878
	$\sum_{14}$	-0.044	0.786
Difficulty	$\sum_{23}$	-0.400	0.808
	$\sum_{24}$	0.103	0.529
Time Discrim & Intensity	$\sum_{34}$	0.479^{**}	0.002

*Correlation is significant at 0.05 level (2-tailed)

**Correlation is significant at 0.05 level (2-tailed)

For all the examinees in Table 4.8, the estimated correlation coefficient between accuracy and speed gives a negative value of -0.061. This is a within-person phenomenon that is called speed-accuracy trade-off. The negative value implies that as ability of the examinee increases, the working speed decreases and vice-versa. It indicates that as high-ability examinees were working faster to beat down time (at a reduced speed), the low-ability ones were working slower within their ability with much more speed. The result also depicts a low and insignificant relationship between ability and speed estimates. This result is in support of the work of Luce (1986) who found a negative correlation but the studies of Fox and Marianti (2017), Molenaar, Tuerlinckx and van der Maas (2015a) and van der Linden (2007) discussed a hierarchical framework that allows ability and speed to be positively correlated where high ability examinees tended to use relatively more time in their responses.

For correlation coefficient values among item parameter estimates, only item discrimination and difficulty as well as time-discrimination and intensity show moderately high, positive and significant relationships (r_{ab}= 0.387; $r_{\Phi\lambda}$= 0.479, p-value<0.05). For covariance estimate ($\sum_{12}$) between item discrimination and difficulty, an indication that difficult items are being answered by high-ability examinees is evident and vice-versa. This shows that test takers with higher ability tend to work faster than these with low ability. For the covariance estimate ($\sum_{34}$)between time discrimination and time intensity, the model explains that for high time-intensive items, the speed factor explains much variation in the examinees response times as it did for low time-intensive items. The relationship between item difficulty and time intensity is positive but low and insignificant ($r_{b\lambda}$= 0.103, p-value>0.05). It shows that the difficult items were apparently taking much more time to be solved than the easy ones but a pointer to the fact that questions that are time-intensive are seen as the more difficult ones (Fox and Marianti, 2017).

Research Question 9: What are the patterns of the person-fit statistics for detection of aberrant response behaviour in the CBMAT response time data?

Just as response on test gives information about examinee’s ability/performance which can be defined in the traditional IRT model, response times also reveal information about the working speed of the examinee which could be modelled by the Lognormal response time item response theory (LNIRT) model. Research question nine was answered by allowing LNIRT model to compute person-fit statistic when analysis was run. The person-fit statistics has the capability of detecting aberrant response behaviour in examinees if the response and response time data are subjected to Bayesian significance test. Fox and Marianti (2017); Dimitrov and Smith (2006) were of the opinion that Bayesian significance test computes the extremeness of response accuracy (RA) pattern in testing. Figure 4.8 shows the graphs of the values assessed from the person-fit statistic drawn against the probability of posterior significance. These graphs depict aberrant patterns exhibited by the test takers in terms of their response times and response accuracy while responding to the CBMAT instrument. Appendix XIX gives the complete results of the person-fit test through the use of Bayesian significance test embedded in the R language of the LNIRT package.

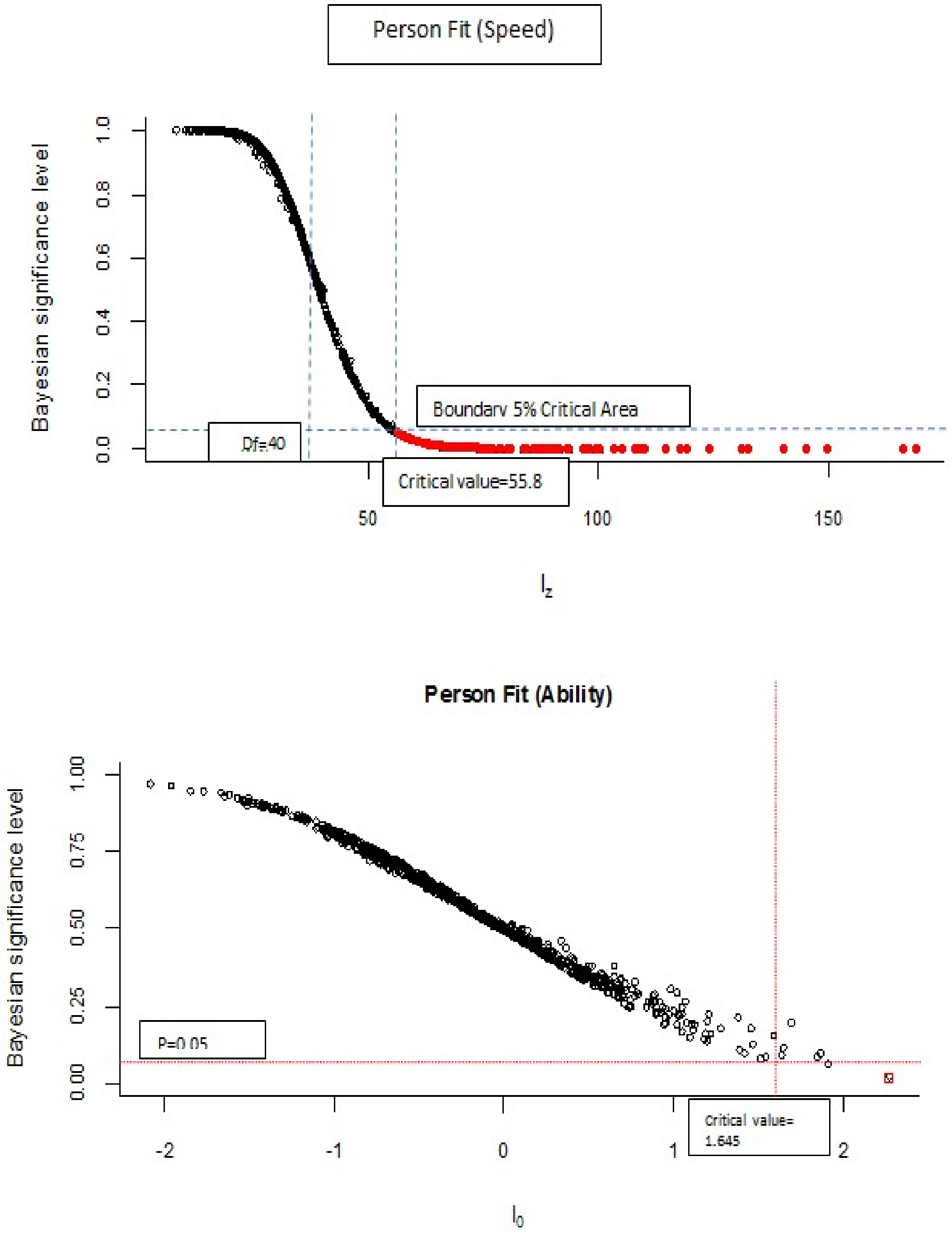


Figure 4.5a: Estimates of person-fit statistic versus posterior probability of significance for response time and response patterns

Person-fit test is known as a statistical tool for checking and detecting irregular response behaviours when response times and response accuracy data are available. Both graphs in Figure 4.9a are plotted with some R codes in the LNIRT modelling window (Appendix XX; page 349) with the estimates from the LNIRT model. The upper plot displays the deviant patterns for response times where the estimated values assumed a chi-square distribution with 40 degrees of freedom at significant level of 0.05. A critical statistic value of 55.8 when α-level is 0.05 is obtained. Then, estimated statistical values that are greater than 55.8 are located in the critical region demarcated by the dotted line at the critical value. At this significant level, the number of unusual patterns that is estimated for response times is 127 which are 14.53% of the examinees.

Also, for the response accuracy pattern in the lower graph, the person-fit statistic values I_o that were estimated are also plotted versus the posterior probability of significance (p-value). The critical region is found above the statistic value of 1.645 at 0.05 level of significant. This means that examinees with a statistic value that is more than 1.645 can be found in this region. It is observed from the study that 1 or almost no person (0.11%) was recognised with aberrant RA patterns.
In this study 127 examinees were observed to have displayed aberrant behaviour as far as their response times were concerned while very few if not any wereseen with deviance when it came to their response patterns.Van der Linden and Guo (2008) opine that the connection between RT and RA patterns undoubtedly increase the power of noticing inappropriate behaviour in test settings.

Also, Wise, Pastor and Kong (2009) whose study captured 329 test takers with a 65-item computer-based version of the Natural World Assessment test (NAW-8) found out that 25% of the students exhibited abnormal behaviour on 10% of the items. Other studies on aberrant response behaviour patterns are the works of Mariant (2017) and Schnipke and Scrams (2002) where some unexpected small RTs towards finishing a test were observed.

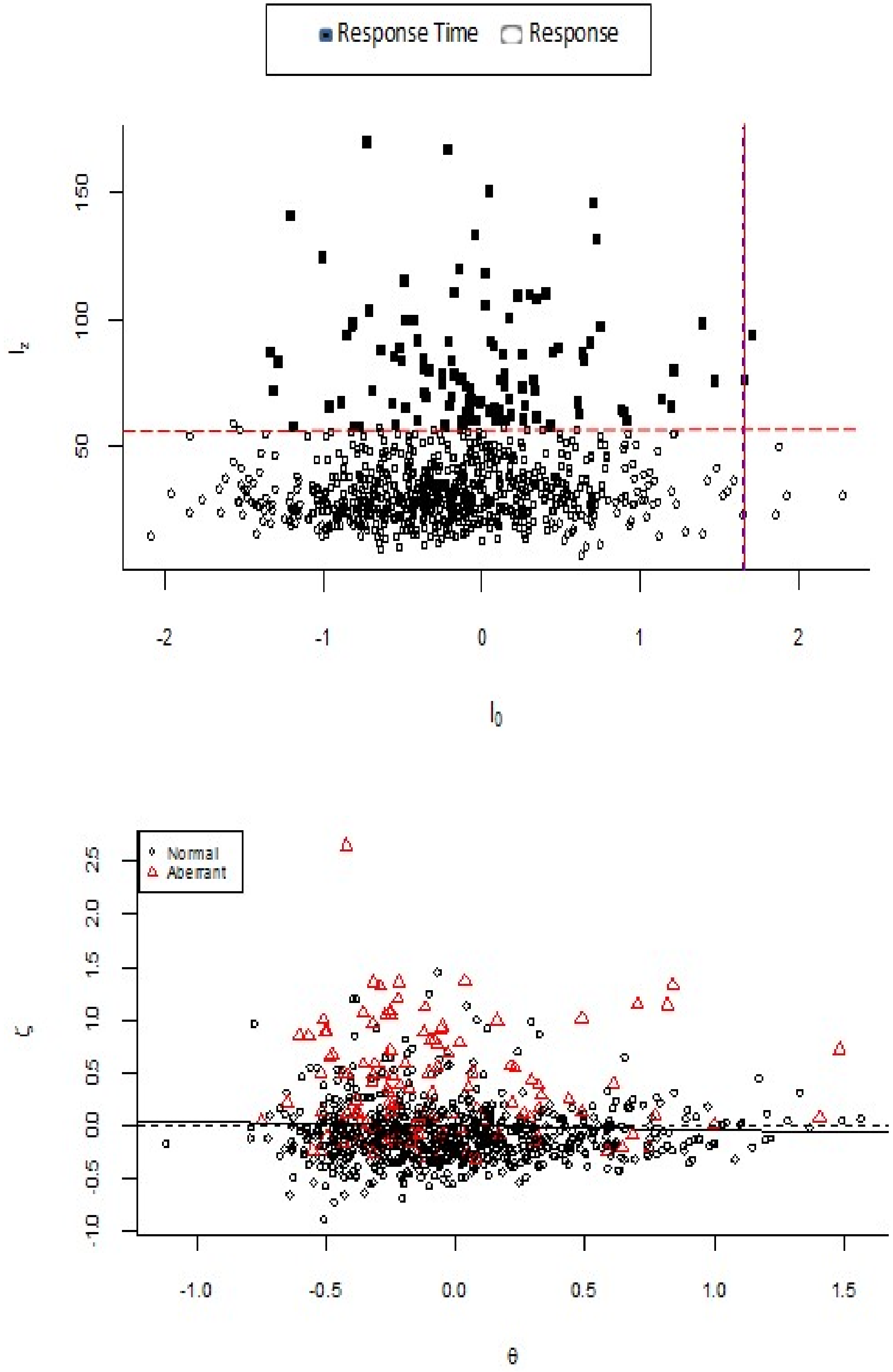


Figure 4.5b: The Ability versus speed values of the identified normal and irregular Examinees' behaviour

In an attempt to further investigate the association between speed and ability for patterns of abnormal and normal response behaviour among the examinees, Figure 4.9b shows the estimated ability values (x-axis) plotted against the speed values (y-axis). The upper plot reveals that for aberrant response behaviour with respect to RTs and response, patterns above the horizontal and vertical dotted lines (55.8, 1.645 at significance level of 0.05) indicate a large number of students who fall in the critical region with respect to RTs,while a few with respect to response accuracy showed abnormality. But for both RT and response patterns, only 2 examinees show aberrant patterns. This corroborate the very low negative relationship (-0.06) that ensues between speed and ability as indicated in the lower part of the plot.

Research Question 10: How comparable are the item and examinees parameter estimates of the traditional IRT model to the LNIRT response time model?

Table 4.9a: Descriptive Statistics for Ability Parameter Estimates

Models	N	Mean	SD
TRAD IRT	874	0.000294	0.862
LNIRT	874	0.001478	0.412

Table 4.9b: Mann-Whitney U test of the Traditional IRT and LNIRT ability parameters.

Null Hypothesis	Test Statistic	Asymp. Sig (2-tailed)	Decision
The distribution of the ability parameter estimate for 3PL and LNIRT are the same	354822.5	0.01	Reject Ho

To provide answer to reasech question 10, the mean and standard deviation values (0.000294, 0.862; 0.00148, 0.412) of the examinees ability for both conventional 3PL and LNIRT response time models were found. The result shows that LNIRT model had a higher mean, an indication that a better proficiencyparameter was estimated by the model when comparedto the traditional model. Hypothesis testing was also carried out using Mann-Whitney U test to compare ability estimates of the traditional IRT (3PL) model to that of the Lognormal response time IRT model. From the U-test, the mean rank for ability parameter for 3PL was 843.48 while the mean rank of 905.52 was obtained for the ability parameter of LNIRT model. It is observed that ability mean rank for LNIRT model showed a better estimate with a gain difference of 62.04. A null hypothesis that the distributions of the examinees parameter for the two models are the same was posed and the result showed that H_o was rejected (test statistic =354822.5, $p < 0.05$).

It implies that proficiency estimations generated by 3PL and LNIRT functions remained statistically different from each other. Therefore, the examinees' ability produced by the LNIRT model was statistically and significantly higher than the one estimated by the traditional 3PL IRT model. This further buttresses the fact that the presence of collateral information such as response times in testing could give more information on the performance of students which could mean that LNIRT model support a more objective way of estimating examinees true abilities. This finding supports the work of Suh (2016) who found contrasting results when parameter estimates of two different response time models were compared. Marianti (2015), van der Linden (2006) and van Zandt (2000) affirmed that a good model fit with LNIRT model was presented in general.

4.5: General Discussion on the Findings of the Study

When examinees are assessed on an achievement test, a general belief that correct response is made on each item of the test in accordance to the ability such examinees possess is assumed (Liao *et. al*, 2012). This assumption is somewhat erroneous in the sense that it might not always hold in real life situation because other factors could account for the correct/incorrect response which in turn may jeopardize the measurement of true ability that is intended. It is on this note that IRT framework, a new methodology to measuring student learning outcomes more objectively (Adedoyin, 2010 and Ojerinde, 2013) was considered in this study as an approach of usage.

The development of more sophisticated models in their different capacities, as the need arises, enables researchers to continually search for a more appropriate way of measuring students' true abilities. The inculcation of 1-, 2- and 3PL models in calibrating dichotomous response format data is said to have been over flogged in their usage. This gave room for the formulation and development of other parameterized models that could lessen some estimation errors/biases incurred by calibrating with the earlier three prominent models.

This study however explored two new noticeable models in the IRT framewor (4PL and LNIRT) which have been previously suggested by some research works (Loken and Rulison, 2010) to estimate models'parameters more correctly so as to make appropriate inferences about the measured traits. Comparisons were made among the different models posed in the study and relationships between item characteristics and examinees parameters of the response time model were evaluated.

4.51: Phase I

At the preliminary stage of data analysis, the data collected from the field (response and response time dataset) were subjected to analysis to quantitatively describe the basic features and distributions of each dataset and simple summaries about the sample were provided.

However, model-fit assessment was carried out with the four available dichotomous (1-, 2-, 3- and 4PL) IRT models on the pooled CBMAT response data at the trial testing stage of the study. The finding was that 4PL model gave the most appropriate/best model goodness of fit to the given data. This is an indication that the pooled CBMAT data was better explained by 4PL model which showed the lowest AIC, BIC and DIC values among the four models. The finding of this study lays credence to the several suggestions made by the studies of Osgood*et. al*. (2002), Reise and Waller (2003) and Tavares *et. al*. (2004) who advocated more usage of 4PL model as a way to further establish the model's utility in assessing true examinees performance.

Calibrations were done to estimate both item and examinee parameters with the four models to see how good and bad each item of the pooled CBMAT instrument was, according to some criteria. It was discovered that 77 items survived 4PL model, 90 survived 3PL model, 63 survived 2PL model while 106 items survived 1PL model. But because the pooled CBMAT response data fitted 4PL, it became the core model adopted for the study. Meanwhile, significant mean difference among the distributions of the same parameters for different models was established. Results showed that the distribution of the parameter estimates (item discrimination, item difficulty and lower asymptote) for the different models are not the same.

This result agrees with Ayanwale (2019) and Fakayode (2018) that noticeable mean variance existed in the person parameters calibrated from the 3PL model used for NECO and June and November versions of WAEC mathematics achievement tests in 2015. Meanwhile Metibemu (2016) findings was contrary in the sense that the developed physics achievement test and 2014 WAEC physics objective test showed the same mean difference in the distribution of their person ability.

The advocate for more usage of 4PL model and its capability to reducing overestimation errors or lessening biasses in measurement have become a reality in this study. Thereby error as right guessing to difficult items on the part of low-ability students and the ones that arise from high-ability students (carelessness such as mistake, tiredness, anxiety, inattention, lack of familiarity to computer usage and techniques, misrread of questions) has been well taken care of with 4PL model. All these assure better measurements that promote accurate/good representation of the required ability and skill the student is truly possessing as at the time of assesment.

4.52: Phase II

At the main study stage, after validation of the pooled CBMAT instrument had been done, some bad items were discarded, while the good items constituted the new instrument tagged final CBMAT instrument. DIMTEST of essential undimensionality and another data-model fit assessment weretested out to find out whether the scale would be consistent with having one dimension or fitting 4PL model. The findings revealed the same result as the trial testing stage. Significant mean differences were also found in the item characteristics estimate of the four models for the final CBMAT instrument except for the distribution of the guessing parameters where significant mean difference in the estimates for the 3- and 4PL models was recorded.

Then, because of the consideration for the use of computer-based test in this study, another new approach in the IRT parlance (Lognormal Response Time IRT model; LNIRT) was explored. This model was used to produce interpretations around examinees' ability and speed as they relate to students' performances/achievements and aids to assess examinee's suitability in the response time and response arrays. Examinee fit helps to identify test takers with aberrant response behaviour in terms of RTs and responses. This study estimated the parameters of the joint modeling of responses and response times (LNIRT). The relationship between speed and ability parameters also provided evidence about the examinees and items of the final CBMAT instrument.

The finding of this study as regards the joint model (LNIRT) showed that a negative low correlation between examinees ability and speed existed. Although the direction of association was opposite, a statistical insignificant relationship was observed. For the item response-time related parameters, item discrimination and item difficulty as well as time discrimination and time intensity established high, positive and significant relationships. Difficult items discriminate well between low and high ability respondents. The associations among these parameters; item and time discriminations, item and time difficulty and discrimination as well as item difficulty and time discrimination showed negative and insignificant relationships.

Examinees pattern of aberrant response behaviour in the final CBMAT response and response time data was also identified by computing person fit statistic with the help of Bayesian significance testing that is embedded in the LNIRT package. It was discovered in this study that for RT patterns, 14.53% of the examinees, which was 127 persons were observed with aberrant response in their RT patterns. While very few examinees with 0.11% were recognised with aberrant patterns in their responses.

In the last research question for this study, the comparability of examinee parameter estimates for the traditional 3PL and LNIRT models were made to discover which model fared well in estimating examinees' trait. The outcome of the study presented LNIRT model as a better model for estimating person ability than the usual IRT models. This could be as a result other collateral information the joint model (LNIRT) supplies (computation of person fit statistic) in terms of ability to detect/identify aberrant examinees with response time (RT), response accuracy (RA) or both RT and RA patterns (Fox and Marianti, 2017). This is possible since person-fit statistic is capable of differentiating test takers with abnormal item response patterns or that of response time from those with non-aberrantl item response patterns or RT patterns.

According to Fox and Marianti (2016), LNIRT model is as well capable of investigating the effects of time limits, a situation where test takers are running out of time and a change in their current strategy to work faster is adopted. LNIRT model has the capacity to compute response time residuals using Kolmogorov-Smirnov (KS) test thatallows the percentage of extremeness of the irregular pattern on items and examinee's ability to be known.

CHAPTER FIVE
SUMMARY, CONCLUSION AND RECOMMENDATIONS

This chapter provides an overview of all that the research work entails.It itemises the key findings, implications, recommendations from the study, limitations,conclusion and suggestions for future research.

5.1 Summary of Findings

The campaign for innovations in the way assessment is done is a pointer to having more approaches to objectively measuring students' learning outcomes so as to depict the true ability of examinees in terms of their performances. IRT method is one of the modern-day approaches of assessing impartially because of the many flexible models its principles are associated with. This has projected it as a more robust tool for testdevelopers, testusers, psychometricians and researchers in combatting the very many challenges the world is facing in measurement and assessment in educational settings.

The commonly used IRT models (1-, 2- and 3PL) in the dichotomous domain have been said to be capable of taking care of some systematic variances CTT model could not care for. However, the 3PL model was identified with some underestimation errors which led to certain improvement in the 3PL model to accommodate the 4th item characteristic known as carelessness parameter.This allowed the newly formulated 4PL model to become functional.

On the other hand, advancement in science and technology has enabled the possibility of administering CBT. This innovation created a significant development in psychometrics by allowing more sophisticated approaches to measuring some variables that were termed difficult to measure in the pre-CBT era. One of the advantages of CBT is the benefit of modelling response time as collateral information into estimating learner's ability correctly. Then, the need to explore the LNIRT model arose.

This study was therefore anchored on exploring the applicability of 4PL and the Lognormal response time IRT (LNIRT) models in the calibration of CBMAT scale. These models were applied with empirical data generated from the response and response time of the CBMAT items. Sample of the research study was made up of mathematics students in senior secondary II of schools with functional computer systems in the Southwestern states of Nigeria. The following are seen as the key findings:

- At the trial-testing stage (Phase I), the pooled 114-item of the computer-based mathematics achievement test was unidimensional, which was an indication that only mathematics ability trait was dominantly exhibited by the examinees in the CBMAT.

- Two pairs of items were observed to be locally dependent and were discarded while the remaining 98% of the items were certified as not being locally independent.

- Model-fit assessment result specified that 4PL IRT model adequately fitted the pooled CBMAT response data. This implies that 4PL model happened to be the most appropriate model capable of explaining the CBMAT response data.

- Based on the acceptable range for each of the item parameters criterion, calibration with all the available models in the uni-dimensional category was done. The pooled CBMAT items were deleted as unfit due to their nonconforming values generated as their parameter estimates that were below or above the set criteria:
  - 4PL model; Discrimination (10), Difficulty (8), Guessing (17) and Carelessness (4). In all, 77 items were considered good and retained while 37 items were bad and discarded.
  - 3PL model; Discrimination (14), Difficulty (10) and Guessing (9). In all, 90 items were considered good and retained while 24 items were bad and discarded.
  - 2PL model; Discrimination (50) and Difficulty (21). In all, 63 items were considered good and retained while 51 items were bad and discarded.
  - 1PL model; Difficulty (8). In all, 106 items were considered good and retained while 8 items were bad and discarded.

- Statistical significant mean difference was observed when the distributions of discrimination, difficulty and guessing parameter estimates were compared under 1-, 2-, 3- and 4PL models. 4PL model produced better estimates of all its item parameters than the other three models.

- At Phase II stage of the study, the 40-item final CBMAT response and response time data' assessment of trait dimensionality and model-data fit indicated that the scale was uni-dimensional and fitted 4PL model. Therefore, the 4PL model was evident to fit both pooled and final CBMAT instruments. This showed how consistent the pooled CBMAT response data at the development stage was with the final CBMAT response data at the real study stage.

- Indication of noticeable statistical mean variance was recorded in the observed scores of the examinees (mathematics ability) when comparisons of their different distributions were made among 1-, 2-, 3- and 4PL models for the final CBMAT response data. The 4PL model estimated examinees' ability better than the other models in the dichotomous category.

- For discrimination parameter estimates of 2-, 3- and 4PL models, there was statistical significant mean deviation when the distributions of each estimates were compared with one another. Item discrimination parameter for 4PL model fared the highest among the three models. This means that discrimination parameter of 4PL model differentiated appropriately well between high and low ability examinees.

- For difficulty parameter estimates of 1-, 2-, 3- and 4PL models of the final CBMAT response data, a significant mean difference was noticed when the distributions of each of the estimates was related with one another.

- The distributions of guessing parameter of the 3PL and 4PL models showed that there was no mean significant difference in their estimates when comparison was made. An implication that examinees' probability of guessing when the final CBMAT response data was calibrated with different models was the same. This finding was different from the one obtained when the pooled CBMAT response data was used at the validation stage.

- ❖ Item (discrimination, difficulty, guessing, time discrimination and time intensity) and examinee (ability and speed) parameter estimates of the LNIRT were calibrated using the final CBMAT response and response time data. The following findings were recorded:
    - For person parameter of the LNIRT model, examinees' mean ability and speed at which they solved the CBMAT items were the same. These estimated parameters are pointing to the fact that examinees operated under equal (constant and average) ability and speed throughout the testing period.
    - Findings on the mean and range estimates of item difficulty parameter (*b*) for the LNIRT model was considered to be relatively large, an indication that the examinees' ability on the whole range of the scale continuum is accurately estimated.
    - For item discrimination parameter (*a*), two items of the final CBMAT discriminated poorly while only one item had the most discriminating power between low and high ability respondents.
    - Time discrimination parameter (Φ) of the LNIRT model revealed that on the average, at every 3secs in the course of responding to the items, the parameter worked on the sensitivity of an item for the different speed levels examinees exhibited.
    - The finding on time intensity parameter (λ) displayed that the range of time each examinee was expected to spend in supplying correct response to the whole items of the CBMAT scale was 13.33mins. < λ< 56.67mins. However, in the observed response time, 15 examinees fell below the estimated minimum time while 21 examinees fell above the estimated maximum time. Therefore 36 examinees (4%) fell outside the estimated time range of the LNIRT model.
- ❖ Correlation coefficient estimate between ability and speed parameters showed a negative, low and insignificant relationship. This implies that as the ability of the examinee increases, his/her working speed decreases and vice-versa.

- A positive, moderately high and statistical significant relationship existed between discrimination and difficulty of parameter estimates of LNIRT model.
- Time discrimination and time intensity parameter estimates demonstrated a positive, moderately high and significant relationship in the correlation coefficient value.
- Also, the finding on the association that existed concerning difficulty and time intensity estimates revealed positive but low and minor association. It means that the difficult items were apparently taking much more time to be solved than the easy ones.This is a pointer to the fact that the time-intensive items are the more difficult items.
- The findings on pattern of aberrant response behaviour of examinees showed that 127 examinees displayed irregular behaviour in their response time patterns while very few examinees revealed abnormality in their response patterns.
- There was significant mean difference between ability estimates of the traditional IRT model and Lognormal response time model (LNIRT). Therefore, higher estimated mean value of the LNIRT model connotes a better estimate than the traditional 3PL IRT model.

5.2 Implications and Recommendations

The findings of this study as summarised above have beneficial and educational implications for the various stakeholders in the education sector. Such stakeholders as test developers, test givers, psychometricians, researchers in the field of education, both public and private examining bodies, school administrators, policy makers, teachers and even students should consider the following:

- The observance of stochastic interpretations of examinees scores/performances in various assessment settings has defaulted true representation of students' ability either in placement or selection. However, IRT approach should be imbibed by test developers for a careful and comprehensive process of ascertaining items that contribute to the measured traits.
- The formulation and development of several new models in the IRT framework is in tandem with the positive shift modern-day research is experiencing. Therefore,

the affordability of model-data fit analysis aids the right choice of a model. IRT method is recommended as a robust mechanism that can cater for any data in various categories of the measurement scale.

- Another implication of the findings of this study is in the act of making appropriate decisions on behalf of students either for formative or summative purpose. It is therefore recommended to the concerned stakeholders that 4PL model which is considered to lessen estimation error should be sufficiently and adequately explored.
- The implication of this study is to shift attention from the paper-pencil type of assessment to computer-based test in estimating students' ability that will depict their true performances. A suggestion of focus on CBT for a more secure assessment by teachers, school administrators, test developers, and examining body is advocated.

5.3 Limitations

- Representative parts of the two states considered out of the six stratified states in the southwest geopolitical zone of Nigeria could actually serve as a limit in this research work.This might not infer generalisation of the inferences thereof.
- The researcher was limited in assessing so much sample size for the study because of the limited number of schools with functioning computer laboratories as well as the restricted number of batches the researcher assessed during test administration.
- Response time analysis that was carried out in this thesis was meant to identify examinees aberrant response patterns. However, statistical irregularities were identified. Other types of abnormality in response in terms of physical activities were not observed. Also, the type of aberrant patterns and RT between groups of test takers were not investigated.
- This study was also limited to the usage of cognitive data got from computer-based mathematics achievement test only.

5.4 Conclusion

The use of multidimensional item response theory (MIRT) and Lognormal response time packages of the R programming language via R-studio was employed by the researcher to explore the applicability of 4-parameter logistic (4PL) and LNIRT models on computer-based mathematics achievement test administered to senior secondary school II students. After convergence of various iterations on several models in the MCMC algorithms, the 4PL was found fitting on the response data while Lognormal function for response time was used to calibrate parameters of the model with response time data. There was an indication that the 4PL model performed better than the previously used 1-, 2- and 3PL models that were known to be famous and widely used. Also, the calibration and estimation of the item and examinees parameters from all the models showed that estimates of 4PL model performed well.

This contributed an exceptional input to assessing students' true capability in educational measurement because errors incurred in the cause of calibrating with especially 3PL model had been drastically reduced. This was done such that a brilliant or highly-able student's ability or performance will no longer be wrongly estimated. While provision for adequate estimation of average or less-able respondents who might rightly guess some items has been satisfactorily catered for in the adoption of 4PL model.

In practice, labelling a test-taker as aberrant is a serious judgement that can affect further the validity of a test. LNIRT model used in this study constituted a major contribution by utilizing item response times through computer-based tests. The examinees' ability estimated with LNIRT model was better when compared with the ones from conventional IRT models. RTs could be used as tools for creating improved inferences about the competencies of test takers.

5.5 Contributions to Knowledge

Some of the great contributions the study made to knowledge include the following:

- The utility of the most recent 4PL model in the unidimensional category for calibrating and estimating model parameters for true representation of examinees proficiency in mathematics was explored as suggested by previous research. The study thereby becomes one of the factual evidences that 4PL model's usage indeed supports a more objective measurement in assessment setting.
- The Bayesian approach adopted with the use of the MCMC algorithm of the MIRT package in the R environment eased convergence at every iteration stage, which aided model parameters estimation processes. The study provides a theoretical knowledge for potential researchers, psychometricians and public examining bodies on how IRT models are fitted and parameters are estimated.
- The self-developed and validated CBT mode of assessment used seems the first of it kind because of the inculcation of response time that affords respondents behavioural/response pattern monitored in the cause of responding to the items of the scale. Therefore, the CBMAT instrument becomes a tool to gather empirical data in the hands of school teachers, researcher, assessors and test givers not only to know students' performances but also to detect both normal and aberant responses.
- The study afforded the benefit of reinforcing computer awareness usage in the classroom setting among students at the secondary school level. Some schools' principals pleaded that the CBMAT package should be left on the school computer systems for students in the terminal classes to practise with.

5.6 Suggestion for further studies

- Further empirical studies should be carried out to affirm more utility of 4PL IRT model as literature reviewed indicated very few or no research work has reflected the underlying advantages the model has over others; model-fit analysis revealed 4PL model as best fit. Therefore, the model's utility in Africa, especially Nigeria, requires that further studies should give it more considerations.

- In the same vein, foreign research studies on the usage of item response time models have shown some exemplary suggestions that could alleviate failure rate as far as test taker's academic performances are concerned. These were viewed from the perspective of the kind of behaviours students' exhibit while responding to items of a scale. In this thesis, a joint model (LNIRT) was adopted. Therefore, envisaged local studies in our clime should be encouraged to explore of the various response time models available. The likes of CUSUM-based technique, Lognormal response times (LNRT) and Gamma and Weibull distributions as well as Lognormal response times moving average (LNRTMA) models should be used to combat irregular/aberrant response patterns in students while assessing them.

- Even though statistical substantiation is advantageous, it is still not an enough evidence to determine that abnormal behavioural pattern transpired. Empirical indication is advised in accompanying further bases of evidence in attaining convincing proof for abberant response behaviour patterns in the examinees.

- In this thesis, computer-based mathematic achievement test which happened to be cognitive was the main instrument. Further studies could adopt non-cognitive test data like personality or attitude to generate better inferences about proficiencies of test takers.
- The researcher could not go into attempting to detect differential item functioning (DIF) of item parameters of either 4PL or LNIRT models across the different groups/locations that existed in the study. This aspect could be of interest to upcoming researchers in the nearest future.

REFERENCES

Abiodun, R.F.A. 1997. The challenges of mathematics in Nigeria's economic goals ofvision 2010; Keynote address presented at the 34th annual national conference of the mathematical association of Nigeria, Sept. 1-6.

Adedoyin, O.O. 2010. Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Educational Science.*2(2): 107-113.

Adegoke, B.A. 2013. Comparison of item statistics of physics achievement test using classical test and item response theory frameworks: *Journal of Education and Practice 4, No. 2.*

.2014.The role of item analysis in detecting and improving faulty physics1735 objective test items: *Journal of Education and Practice.5, No.21.*

Adeleke, J.O. 2009. The basics of research and evaluation tools.Somerest Ventures.

Adewale, J.G. 2015. Equating two year BECE results in basic science and technology in Oyo state, Nigeria.Bindura University of Science Education.

.2018. Item construction. One-day workshop in the Faculty of Science, University of Ibadan, Centre of Excellent for Teaching and Learning in collaboration with The Institute of Education. October 24, 2018.

Aitkin, M. 2016. *Expectation maximization algorithm and extensions*. Handbook of Item Response Theory. 2, pg 217-236. Taylor and Francis Group.6000 Broken Sound Parkway NW, Suite 300 Boca Raton.

Alordiah, C.O. 2015. A progressive step in educational measurement: An Application of the Rasch Model on Mathematics Achievement Test.*Nigerian journal of educational Research and Evaluation*.14. No.3.

Amarnani, R. 2009. Two theories, one theta: A gentle introduction to item response theory as an alternative to classical test theory. *The International Journal of Educational and Psychological Assessment,3, 104–109.*

Ambali, A.G. 2014. Real deal: Mathematics education for sustainable development. Text of the keynote Address delivered at the opening ceremony of the 51st annual conference of the mathematical association of Nigeria at the University of Ilorin, Nigeria.

Anamezie, R.C. and Nnadi, F.O. 2018. Parameterization of teacher-made physics achievement test using deterministic-input-noisy-and-gate model. *Journal of Education and Practice.9, No.32.*

Anderson, L.W. (Ed), Krathwohl, D. R.,Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P.R., Raths. J. and Wittrock, M. C.2001.A taxonomy for learning,teaching and assessing. A Revision of Bloom's Taxonomy of Educational Objectives (Complete edition). New York: Longman

Ani, E.N. 2014. Application of item response theory in the development and validation of multiple choice test in economics. M.Ed. Thesis. Department of Science Education, University of Nigeria, Nsukka. Source: University of Nigeria Virtual Library.

Ariyo S.O. 2017. Diagnostic assessment and effect of after-school programmes on the Mathematics learning outcomes of low achieving students in Oyo State Secondary Schools.A Ph.D Post-Field Seminar Presentation.Institute of Education, University of Ibadan.

Ariyo A.O. 2015. An overview of classical test theory and item response theory in test development.*Nigerian Journal of Educational Research and Evaluation 14,No 3.*

Ariyo, A.O and Lemut, T.I. 2015. Ensuring quality in the test development process through innovations in item calibration: A comparison of classical test theory and item response theory eras in Jamb, Nigeria. 3rd AEAA Conference, Accra, Gbana, 28th August – 2nd September, 2015.

Asikhia O. A. 2010. Students and teachers' perception of the causes of poor academic performance in Ogun State secondary schools (Nigeria): Implications for counseling for national development. *European Journal of Social sciences*13(2).

Ayanwale, M.A. 2019.Efficacy of item response theory in score ranking and concurrent validity of dichotomous and polytomous response mathematics achievement test in Nigeria.An Unpublished Ph.DThesis.International Centre for Educational Evaluation (ICEE),Institute of Education, University of Ibadan.

Azuka, B.F. and Kurume, M.S. 2015. Mathematics Education for sustainable development: implications to the production and retention of mathematics teachers in Nigerian schools. *British Journal of Education,3, No.1, pp. 44-51*.

Baker, F.B. 1985. *The basics of item response theory*. Portsmouth, NH: Heinemann.

, 1998. An investigation of item parameter recovery characteristics of a Gibbs sampling procedure.*Applied Psychological Measurement 22, 153-169*.

, 2001. The basics of item response theory. College Park, MD: ERIC Clearing House on Assessment and Evaluation. Original work published in 1985. Retrieved from http://echo.edres.org:8080/irt/baker/.

Baker, F.B. and Kim, S. 2004. *Item response theory: Parameter estimation techniques. 2nd ed*. New York Marcel Dekker.

Balov, N. and Marchenko, Y. 2016. Bayesian binary item response theory models using bayesmh.http://blog.stata.com/2016/01/18/bayesian-binary-item-response-theory-models using-bayesmh/.

2016. In the spotlight: Bayesian IRT–4PL model. *Stata News, Quarter 1, 31 No 1.*

Barton, M.A. and Lord, F.M. 1981.An upper asymptote for the three-parameter logistic itemresponse model Princeton, NJ Educational Testing Service.

Baud M, and Masseyeff, R.F. (Eds). 1993. Data analysis, mathematical modeling, pp 656-671 in Methods of Immunological Analysis 1.*Fundamentals,* VCH Publishers, Inc., New York, NY.

Becker, J. D. 2006. Digital equitynin education: A multilevel examination of differencesin and relationships between computer access, computer use and state-level technology policies. *Education Policy Analysis Archives*, 15(3), 1-38.

Beguin, A.A and Glas, C.A.W. 2001.MCMC estimation and some model-fit analysis of multi-dimensional IRT model. *Psychometrika, 66, 541-561.*

Birnbaum, A.1957. Efficient design and the use of tests of ability for various decision-making problems (Series Report No. 58-16, Project No.7755-23). Randolp Air Force Base, TX: USAF School of Aviation Medicine.

.1958. On the estimation of mentak ability (Series Report No. 15, Project No.7755-23). Randolp Air Force Base, TX: USAF School of Aviation Medicine.

. 1968. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley.

Black, P.J. and William, D. 2009.Assessment and classroom learning.Assessment in Education.5, 7-7.

Bock, R.D. and Mislevy, R.J. 1982.Adaptive EAP estimation of ability in a microcomputer environment.*Applied Psychological Measurement, 6, 431-444.*

Bradlow, E. T., Wainer, H., and Wang, X. 1999.A Bayesian random effects model for Testlets.*Psychometrika, 64, 153-168.*

Braun, H. A., Kanjee, A., Bettinger, E. and Kremer, M. 2006. Improving education through assessment, innovation and evaluation. Cambridge, MA: American Academy of Arts and Sciences.

Bridgeman, B. and Cline, F. 2000. Variations in mean response times for questions on computer-adaptive GRE general test: Implications for fair assessment; ETS RR-00-7; Educational Testing Service: Princeton, NJ, USA.

. 2004. Effects of differentially time-consuming tests on computer-adaptive test scores. J. Educ. Meas. 41, 137–148.

Burgos, J.G. 2010. Bayesian methods in psychological research.The case of IRT.*International Journal of Psychological Research.*Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile.

Burnham, K. P. and Anderson, D. R. 2002.Model selection and multimodel inference. A practical information-theoretic approach (2ndед.). New York: Springer.

. 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33(2), 261-304.*

Carroll, J.B. 1993. Human Cognitive Abilities A Survey of Factor Analytic Studies; Cambridge University Press: New York, NY, USA.

Cees, A. W. Glas and Rob, R. Meijer.2003.A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, Sage Publications 27 No. 3, pp. 217–233.

Cella, M., Dymond, S., Cooper, A. and Turnbull, O. 2007. Effects of decision-phase time constraints on emotion-based learning in the Iowa Gambling Task. *Brain and Cognition 64, 164–169.*

Cengiz, M. A. and Ozturk, Z. 2013. A Bayesian for item response theory in assessing the progress test in medical students.*International Journal of Research in Medical and Health Sciences.3, No. 3.*

Chang, H.H. and Yin, Z. 2008. To weight or not to weight?Balancing influence of initial items in adaptive testing.*Psychometrika, 73, 441-450.*

Chalmers, R., P. 2012. MIRT: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48* (6), 1-29.

Courville, T.G. 2004. An empirical comparison of item response theory and classical test item /person statistics. Unpublished Doctoral Thesis,Texas A and M University.

Crocker, L. and Algina, J. 1986. *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich College Publishers: Fort Worth, 527.

2008. Introduction to classical and modern test theory. New York: Holt Rinehart and Winston.

Daniel B.W. 2016. Treating all rapid responses as errors (TARRE) Improves estimates of ability (slightly).Test and Assessment Modelling, 58, 1, 15-31.

Davis D., Zhang A., Etienne C., Huang I. and Malit M., 2000. Principles of curve fitting for multiplex sandwich immunoassays. Bio-Rad Laboratories, Inc., Alfred Nobel Drive, Hercules, CA 94547 USA.

De Ayala, R.J. 2009. *The theory and practice of item response theory*. The Guilford Press,New York.NY 10012. www.guilford.com.

De Boeck, P. and M. Wilson, (Eds.) 2004.*Explanatory item response models:* A generalized linear and nonlinear approach (Statustics for Social and Behavioural Sciences).2004th Edition.New York: Springer.

De Boeck, P. and Partchev, I. 2012. IRTrees: Tree-based item response models of the GLMM family. *J. Stat. Softw. 2012, 48, 1–28.*

De Champlain, A.F. 2010. A primer on classical test theory and item response theory for assessments in medical education.*Med Educ.44(1):109-17.*

De Mars, C. 2010. *Item response theory.Understanding statistics measurement.*Oxford University Press.

Dimitrov, D. M. and Smith, R. M. 2006.Adjusted Rasch person-fit statistics. *Journal of Applied measurement, 7(2), 170-183.*

Eccles, H., Haigh, M., Richards, M., Mei, T.H. and Choo, Y.W. 2012.Implementing e-assessment In Singarepore: The Student Experience.IAEA Conference.

Eggen, T.J.H.M. 2000. On the loss of information in conditional maximum likelihood estimation of item parameters.*Psychometrika, 65 (3), 337-362*.

Eggen,T. J.H.M and Verhelst, N.D. 2011. Item calibration in incomplete testing designs.*Psicológica, 32, 107-132*.

Embretson, S. E. 1991. A multidimensional latent trait model for measuring learning and change.*Psychometrika, 56:495–516.*

. 1996. The new rules of measurement. *Psychological Assessment,8(4), 341–349.*

. 2000. Multidimensional measurement from dynamic tests: Abstract reasoning under stress. *Multivariate Behavioral Research, 35:505–542.*

Embretson, S. E. and Reise, S. P. 2000.*Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum Associates.

Enu, V.O. 2015. The use of item response theory in the validation and calibration of mathematics and geography items of joint command schools promotion examination in Nigeria.A Ph.D thesis.International Centre for Educational Evaluation (ICEE), Institute of Education.University of Ibadan.

Ercikan, K.and Koh, K. 2005. Construct comparability of the English and French versions of TIMSS. *International journal of testing(5), 23-35.*

Fakayode, O. T. 2018. Relative Effectiveness of CTT and IRT in equating WAEC Mathematics test scores for June and November 2015. A Ph.D Thesis.International Centre for Educational Evaluation (ICEE), Institute of Education.University of Ibadan.

Famoroti, A.A. 2019. Relative effectiveness of computerize adaptive and linear computer based-testing in estimating examinees' performances using social studies test in Osun State, Nigeria. A Ph.D Prefield Research Proposal, Institute of Education, University of Ibadan.

Ferrando, P. J. 1994. Fitting item response models to the EPI-A impulsivity subscale. *Education and Psychological Measurement 54118–127.*

Finn, B. 2015.Measuring motivation in low-stakes assessments; research report. No. RR-15; Educational Testing Service: Princeton, NJ, USA.

Finch, W.H. and French, B.F. 2015.*Latent variable modelling with R*. New York: Routledge.

Fox, J.P. 2018. Course LNIRT: Modelling response accuracy and response times. A pdf document retrieved online on 15th April, 2019.

. 2010. *Bayesian item response modeling: Theory and Applications*. New York: Springer

Fox, J.P. and Marianti, S. 2017. Person-fit statistics for joint models for accuracy and speed.*Journal of Educational Measurement, 54(2), 243–262. ISSN 1745-3984.*

. 2016. Joint modeling of ability and differential speed using responses and response times. *Journal of Multivariate Behavioral Research, 51(4), 540-553*.

Fox, J.P., Klein EntinK, R.H and van der Linden, W.J. 2007. Modeling of responses and response times with the package cirt, *Journal of Statistical Software.*

Fox, J. P. and Glas, C. A. W. 2001.Bayesian estimation of a multilevel IRT model using Gibbs sampling.*Psychometrika, 66, 271-288.*

Fraley, R. C., Waller, N. G. and Brennan, K. A. 2000.An item response theory analysis of self-report measures of adult attachment.*Journal of Personality and Social Psychology* 78 350–365.2000.

Fulcher, G. and Davidson, F. 2007. *Language testing and assessment*: An advanced resource book. London and New York: Routeledge.

Galdin, M. and Laurencelle, L. 2010. Assessing parameter invariance in item response theory's logistic two item parameter model: A Monte Carlo investigation. *Tutorials in Quantitative Methods for Psychology. 6, 2, p. 39-51.*

Gaviria, J. L. 2005. Increase in precision when estimating parameters in computer assisted testing using response time. *Quality and Quantity, 39, 45-69.*

George, A. C. and Robitzsch, A. 2015. Cognitive diagnosis in R: A didactic. *The Qualitative Methods for Psychology, 11(3), 189-295.*

Georgiev, N. 2008. Item analysis of C, D and E series from Raven's standard progressive Matrices with item response theory 2PL model.*Europe's Journal of Psychology.*

Gill, J., Heeringa, S., van der Linden, W.J, Long J.S and Snijders, T. 2016.Chapman and Hall/CRC Statistics in the Social and Behavioural Sciences Series. Aims and Scope: Series Editors.

Gilbert Duy Doan, M. S. 2017. Why R is a language of choice for data scientists. https//www.quora.com/

Glas, C.A.W and Van der Linden, W.J. 2010.Marginal likelihood inference for a model for item responses and response times.*Br. J. Math. Stat. Psychol. 63, 603–626*

Glas, C.A.W. 2016. *Maximum- likelihood estimation.Handbook of item response theory.* 2,pp. 197-216. Taylor and Francis Group.6000 Broken Sound Parkway NW, Suite 300 Boca Raton.

González B.J. 2010. Bayesian Methods in Psychological Research: The case of IRT. *International Journal of Psychological Research. 3,1, 164-176.*

Gottschalk P.G. and Dunn J.R. 2005.The five-parameter logistic: A characterization and comparison with the four-parameter logistic. *Analytical Biochemistry 343, 54–65*

Gray-Little, B., Williams, V. S. L. and Hancock, T. D. 1997.An item response theory analysis of the Rosenberg Self-Esteem Scale. Personality and Social Psychology Bulletin 23, 443–451.

Green, B. F. 2011. A comment on early student blunders on computer-based adaptive tests. *Applied Psychological Measurement, 35, 165-174.*

Haberman, S.J. 2016. Models with nuisance and incidental parameters.*Handbook of Item Response Theory.*2. Pp. 151-168.

Hambelton, R. K..1989.Principles and selected applications of item response theory. University of Massachusettes at Amherst.

Hambleton, R. K., and Swaminathan, H. 1985. *Item response theory: Principles and Applications.*Norwell, MA Kluwer Academic Publishers.

Hambleton, R. K and Jones, R. W. 1993.Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development.An NCME Instructional Module in Educational Measurement.University of Massachusetts at Amherst.

Hambleton, R. K, Swaminathan, H. and Rogers, H. J. 1991.*Fundamentals of item response theory.* Newbury Park, California: Sage Publications

Hambleton, R. K. and Jones, R.W. 1993.Comparison of CTT and IRT and their applications to test development. An NCME Instructional Module 16, Fall 1993. Retrieved on 17th July,2017 from http://www.ncme.org/pubs/items/24.pdf.

Hays, R.D., Morales, L.S., MPH, M.D and Reise, S.P. 2000.Item response theory and health outcomes measurement in the 21st Century.*Med Care.2000 Sep; 38(9).*

Hedeker, D., Mermelstein, R.J. and Flay, B.R. 2006. Application of item response theory models for intensive longitudinal data. University of Illinois at Chicago.Retrived August 30, 2017, from https://www.researchgate.net/publication/228957583.

Heitz, R. P. 2014. The speed-accuracy tradeoff: History, physiology, methodology and behavior. *Frontiers in Neuroscience*, *8, 1-19*.

Henard, D.H. 2000. Item response theory.In L.G. Grimm and P.R. Yarnold Eds. Reading and understanding more multivariate statistics (pp. 67-97). Washington, DC: America Psychological Association. Reserve Desk at Lee Library.

Hoijtink, H. and Molenaar,I.W.1997. A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. Psychometrika, 62, 171189.

http://www.languageinindia.com/july2010/teachersqualification.pdf. Retrived on18/10/2016).

http://www.nigerianeye.com/2014/08/mass-failure-as-waec-releases-results.html.

http://www.tribuneonlineng.com/school-governing-board-solution-malaise-oyo-public schools/

http:www.africasti.com/headlines/Nigeria laments poor performance of students in mathe matics. Students' L2 performance at the secondary level. Strength for Today and Bright Hope For Tomorrow, 10(7), July 2010.

https://oyostate.gov.ng/oyo-teaching-service-commission-tescom/

https://sustainabledevelopment.un.org/sdg4.

https://www.http://educationdistrict1.lagosstate.gov.ng/. Retrieved 08/02/2018.

https://www.mailman.columbia.edu/research/population-health-methods/item-response theory**.**

Huba, M. E. and Freed, J. E. 2000. Learner-centered assessment on college campuses: Shifting the focus from teaching to learning. Boston, MA: Allyn and Bacon.

Hyun Kang, M. D. 2013. The prevention and handling of the missing data.*Korean Journal of Anesthesiology. 64(5): 402-406.*

Iji, C. O. 2007. Challenges of primary mathematics for universal basic education (UBE).ABACUS.*The Journal of Mathematics Association of Nigeria*, *32, 1, 10 – 16.*

Imam, J.D., Onyeneho, P., Onoja, G.O., and Ifewulu, C.B. 2015.Assuring quality of UTME through application of modern test theory in test production process. 33rd AEAA Conference, Accra.

Impara, J. C. and Plake, B. S. 1998. Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement,35, 69-81.*

Ingrisone, S. J. (2008b). An extended item response theory model incorporating item response time.Unpublished doctoral dissertation, Florida State University, Tallahassee.

Janssen, R., Tuerlinckx, F., Meulders, M. and de Boeck, P. 2000. A hierarchical IRT model for Criterion referenced measurement. *Journal of Educational and Behavioral Statistics*, 25, 285-306.

Khairani, A.Z. and Nordin, M.S. 2011. The development and construct validation of the mathematics proficiency test for 14-year-old students. *Asia Pacific Journal of Educators and Education*, 26, No. 1, 33–50.

Kilpatrick, J. and Larborde, C.(Eds.). 1996. *International handbook of mathematics education*. pp 11-47. Dordrecht: kluwer Academic Publisher.

Kim, J. S. and D. M. Bolt. 2007. Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice. 26: 38–51.*

Kim, S.-H. 2001. An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25, 163-176*.

Klein Entink, R.H, Fox, J.-P.and van der Linden, W.J. 2009. A multivariate multilevel Approach to the modeling of accuracy and speed of test takers.*Psychometrika, 74 (1), 21-48.*

Kline, T. J. B. 2005. *Psychological testing: A practical approach to design and evaluation*.Thousand Oaks, CA: Sage.

Konish, S. and Kitagawa, G. 2008.*Informaton criteria and statistical modeling*.Springer Series in Statistics Book Series.Springer-Vwelag New York.

Kpolovie, P. J. and Emekene, C. O. 2016.Psychometric advent of advanced progressive matrices- smart version (APM-SV) for use in Nigeria.*European Journal of Statistics and Probability*, 4(3), 20-30.

Kuku, A.O. 2012. Mathematics as a time-tested resource for scientific, technological,socioeconomic and intellectual development.Distinguished Mathematics Lecture delivered at the University of Ibadan. Ibadan: Ibadan University Press.

Kyllonen, P. C. and Zu, J. 2016. Use of Response Time for Measuring Cognitive Ability.*Journal of Intelligence, Educational Testing Service*, Princeton, NJ.08541,USA.

Lacouture, Y. and Cousineau, D. 2008. How to use MATLAB to fix the ex-Gaussian and other probability functions to a distribution of response times. *Tutorials in Quantitative Method for Psychology.4 (1), p. 35-45.*

Lagos EKO secondary education project.2009 Project Information Document (PID).http://wwwwds.worldbank.org/external/default/wdscontentserver/wdsp/ib/2009/05/05.

Lanza, S. T., Foster, M., Taylor, T. K. and Burns, L. 2005. Assessing the impact of measurement specificity in a behaviour problems checklist: An IRT analysis. Technical Report 05-75. University Park, PA: The Pennsylvania State University, the methodology centre.

Lawal, R.O. 2009.Home factor as determinants of science achievement among selected senior secondary school students in Ibadan, An unpublished M.Ed. dissertation, Institute of Education, Olabisi Onabanjo University.

Lawley, D. N. 1943. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61:273–287.

Lawrence I. A. and Kolawole O. Usman 2007. Mathematics Education for Dynamic Economy in Nigeria in the 21st Century.J.*Soc. Sci., 15.3: 293-296.*

Lee, Y. H. and Chen, H. 2011. A review of recent response-time analyses in educational testing.*Psychol Test Assess Model, 53(3), 359-379.*

Lee, Y. H. and Haberman, S.J. 2015.Investigating test-taking behaviors using timing and process data.*Int. J. Test. 16, 240–267.*

Lee, Y. H. and Jia, Y. 2014. Using response time to investigate students' test-taking behaviors in a NAEP computer-based study.*Large-Scale Assess. Educ. 2014, 2, 8.*

Liao, W-W., Ho, R-G., Yen, Y-C.and Cheng, H-C. 2012. The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behaviour and Personality, 40, 1679-1694.*

Linacre, J. M. 2003. Winsteps. Beaverton Oregon: Winsteps.com. National Council on Measurement in Education, New Qrleans.Network.

.2004. Discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions*, 18, 959-960.

Lindquist, E. F. 1953. The theory of test construction.In H. E Hawkes, E. F. Lindquist and C. Mann (Eds.).*The construction and use of achievement examinations. Boston:* Houghton Mifflin.

Loken, E. and Rulison, K. L. 2010.Estimation of a 4-parameter item response theory model.*The British Journal of Mathematical and Statistical Psychology*, 63(3), 509-525.

Lord, F. M. and Novick, M. R. 1968. *Statistical theories of mental health scores.*Addison Wesley, Reading, MA.

Lord, F. M. 1980. *Applications of item response theory to practical testing problems*.Erlbaum, Hillside, NJ.

Luce, R.D. 1986. *Response times: Their role in inferring elementary mental organisation*. NewYork: Oxford University Press.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. 2000. WinBUGS—A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10, 325-337.*

Magis, D. 2012. Some robust ability estimators with logistic item response models. Unpublished manuscript, Department of Education, University of Lie`ge, Lie`ge, Belgium.

. 2013. A Note on the Item Information Function of the Four-Parameter Model. Applied *Psychological Measurement; 37(4) 0415.*

Magis, D. and Raıˆche, G. 2012. Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software,*48, 1-19.

Malcolm, T. 2003. An achievement test. Retrieved November, 20, 2013, from http://www.wisegeek.com/what-is-an- achievement-test.htm

Mamman, M. and Eya, S.2014.Trends analyses of students' mathematics performance in West African Senior Secondary Certificate Examination from 2004 to 2013: Implications for Nigeria's vision 20:2020. *British Journal of Education, 2, No.7, pp. 50-64.*

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables and their application as psychometric models for response times. *Psychometrika, 58(3), 445-469.*

Marianti, S. 2015. Contributions to the joint modeling of responses and response times. A Ph.D Thesis, University of Twente, Enschede, The Netherlands.

Martelli, I. 2014. Multidimensional item response theory models with general and specific latent traits for ordinal data. An unpublished Ph.D thesis. Università di Bologna

Masters, G.N. 1982. A Rasch model for partial credit scoring.*Psychometrika.47,149-174*

McDonald, R. P. 1981. The dimensionality of tests and items.*Br J Math Stat Psychol. 34:100–117.*

. 1999. Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates.

McMorris, B. J. and Potenza, M. T. 2002.Analyzing multiple-item measures of crime and deviance.Item response theory scaling.*Journal of Quantitative Criminology.18, 267-296.*

Mellenbergh, G. J.1995. Conceptual notes on models for discrete polytomous item responses.*Applied Psychological Measurement, 19:91–100.*

Meng, H. 2007. A comparison study of IRT calibration methods for mixed-format tests in vertical scaling.A Ph.D thesis, University of Iowa.http://ir.uiowa.edu/etd/338.

Metibemu, M.A. 2016. Comparison of classical test theory and item response theory frameworks in the development and equation of physics achievement tests in Ondo State, Nigeria**.** A Ph.D thesis, Institute of Education, University of Ibadan.

Michael U. 2011.A survey of factors responsible for students' poor performance in mathematics in senior secondary school certificate examination (SSCE) in Idah local government area of Kogi State, Nigeria.B.Sc (Ed) Mathematics, University of Benin, Benin City, Nigeria.

Michael, O. 2016. School Governing Board: Solution to malaise of Oyo public schools? http://whatsupibadan.com/?p=21295.

Mislevy, R.J. and Bock, R.D. 1997. BILOG 3: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

. 1982. Biweight estimates of latent ability. *Educational and Psychological Measurement 42 725–737.*

Molenaar, D., Tuerlinckx, F. and van der Maas, H. L. (2015a).A generalized linear factor model approach to the hierarchical framework for responses and response times.*British Journal of Mathematical and Statistical Psychology, 68(2), 197-219.*

Morizot, J., Ainsworth, A. T. and Reise, S. P. 2007. Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, and R. F. Krueger (Eds.), Handbook of Research Methods in Personality Psychology. pp. 407-423. New York: Guilford.

Muraki, E. 1992. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16, 159-176.*

NCE, 2013.Strengthening the Institutional Management of Education for Quality Service Delivery.NUC Bulletin.

NCERT, 2006.National Focus Group on Teaching of Mathematics. Executive Summary, National Council of Educational Research and Training. First Edition, Sri Aurobindo Marg, New Delhi.

Nenty, H. J. 1998. Introduction to item response theory.*Global Journal of Pure and Applied Sciences.4(1), 93-100*.

. 2004. From classical test theory (CTT) to item response theory (IRT): An introduction to a desirable transition. In O. A. Afemikhe and J. G. Adewale (Eds.), Issues in educational measurementand evaluation in Nigeria (in honour of 'Wole Falayajo) (Chapter 33, pp.371-384Yaoundé, Cameroon: Educational Assessment and Research Network in Africa.

Nering, M. L. and Ostini, R. 2010.*Handbook of polytomous item response theory models*. Routledge/Taylor & Francis Group, UK.

NERDC-Nigerian Educational Research and Development Council. 2007. The think tank of education: The 9-year basic education curriculum structure. Retrieved from http.//www.needsNigeria.org.

Newell, A. 1973.*You can't play 20 questions with nature and win*. In Visual Information Processing; Chase, W.G., Ed.; Academic Press: New York, NY, USA.

Niss, M. 1996. *Goals of mathematics in teaching*. In A.J. Bishop, K. Clements, C. Keitel, J. International handbook of mathematics education, 11-47.

Novick, M. R. 1966. The axioms and principal results of classical test theory.*Journal of mathematical psychology, 3, 1-18.*

Novick, M.R. and Jackson, P.H. 1974. *Statistical methods for educational psychological research.* New York: McGraw-Hill.

Odinko, M.N. 2014. *Evaluation research theory and practice*.Giraffe Books.A survey of institutional needs of primary school teachers in Nigeria.

Ogunsakin, I.B. and Shogbesan, Y.O. 2018. Item response theory (IRT): A modern statistical theory for solving measurement problem in 21st century. *International Journal of Scientific Research in Education.11(3B), 627-635.*

Ojerinde, D. 1999. Mathematics in technological development focus on the next millennium: Implication for secondary education in Nigeria. A lead paper at the 36th Annual Conference of Mathematics Association of Nigeria (MAN), Nigeria.

. 2012. From paper-based test to computer-based test for UTME: A desirable transition. A paper presented at the 4th Oyo State ICTs in National Development Summit, Ibadan.

. 2013. Classical test theory (CTT) vs Item response theory(IRT): An Evaluation of the comparability of item analysis results: Lecture presentation at the Institute of Education, University of Ibadan.

. 2013. Implementing and sustaining ICT-based assessment and evaluation in the Nigerian education system. National conference on ICT in education, National University Commission (NUC) Auditorium, Maitama, Abuja, 19th -20th November.

. 2015. Innovations in Assessment: Jamb Experience: *Nigerian Journal of Educational Research and Evaluation.14(3),1-9*.

. 2016. *The preface ofvital issues in the introduction of computer-based testing in large-scale assessment.* A compilation of papers presented at Local and international conferences. Joint Admission and Matriculation Board (JAMB). ISBN:978-978-953-759-4.

Ojerinde, D., Popoola, K, Ojo, F. and Onyeneho, P. 2012. *Introduction to item response theory: parameter models, estimation and application*. Goshen Print Media Ltd.

Ojerinde, D., Anyaegbu, G., Onoja, G.O and Adelakun, E.O. 2013. The impact of e-Testing on examination security: A case study of the UTME. Paper presented at Arusha, Tanzania at the 31st Association for Educational Assessment in Africa (AEAA) 12th-16th August.

Ojerinde, D., Onoja, G.O and Ifewulu, C.B. 2013. Comparative analysis of candidates' performances in the pre and post eras in JAMB: Case study of the use of English in the 2012 and 2013 UTME. 39th IAEA Conference, Tel Aviv, Isreal. 20th-25th October.

Ojerinde, D., Popoola, K., Onyeneho, P. and Akintunde, A. 2013. Item response function: A systematic tool for enhancing test item quality. 39th IAEA conference, Tel Aviv, Isreal, 20th - 25th October.

Ojerinde, D., Popoola, K, Ojo, F. and Ariyo, A. 2014.*Practical applications of item response theory in large-scale assessment*. Nigeria: Marvelous Mike Press Limited.

Ojerinde, D., Popoola, K., Onyeneho, P. and Egberongbe, A. 2015. Achieving examination security through the deployment of computer-based test in Nigeria: JAMB Experience: 7th International conference on education and new learning technology, Barcelona, Spain.6th-8th July.

Ojerinde, D., Popoola, K. and Ariyo, A. 2015.Comparison of item calibration data using item response theory modelling software.41st IAEA Conference, Kansas, USA.11th-16thOctober.

Okorie, B.N. and Mojiboye, E. J. 2015. The computer alternative and the logistics of the conduct of unified tertiary matriculation examinations: a comparative analysis of assessment requirements for the years 2011 and 2015.National conference on ICT in education, National University Commission (NUC) Auditorium, Maitama, Abuja, 19th -20th November.

Okoro, O.M. 2006. *Measurement and evaluation in education*. Uruowulu-Obosi: Pacific Publishers Ltd.

Okpala, P. N., Onocha, C. O. and Oyedeji, O. A. 1993.*Measurement in Education*. Jattu Uzairue: Edo State, Stirling-Horden Publishers (Nig) Ltd.

Okwilagwe, E.A and Ogunrinde, M.A. 2017.Assessment of unidimensionality and local independence of WAEC and NECO 2013 geography achievement tests.*African Journal of Theory and practice of Educational Assessment. 5, 31-45.*

Olonade, P. O. 2017. Equating 2014 senior school certificate mathematics examinations of West African Examinations Council and National Examinations Council in Lagos State, Nigeria.A Ph.D Thesis, Institute of Education, University of Ibadan.

Olonade, P. O., Metibemu, M. A. and Adewale, J.G. 2017. Unidimensional item response theory versus multidimensional item response theory: Evaluating the similarity of item calibration results in mathematics test in Lagos State. *African Journal of Theory and practice of Educational Assessment. 5, 73-86.*

Omotayo. S.A. 2017. Effects of dynamic geometry software and 5e-insructional model on students’ achievement, interest and retention in senior secondary school geometry in Ibadan. A Ph.D Thesis, Institute of Education, University of Ibadan.

Okpala, N. P. and Onocha, C. O. 1995.Effect of systematic assessment procedures on students’ achievement in mathematics and science subjects.*UMESCO AFRICA,* 10.55-61.

Onunkwo, G .I .N. 2002.*Fundamentals of education measurement and evaluation*. Owerri: Cape Publishers Int’l Ltd.

Osgood, D. W., McMorris, B. J. and Potenza, M. T. 2002. Analyzing multiple-item measures of crime and deviance: Item response theory scaling.*Journal of QuantitativeCriminology,*18, 267-296.

Osterlind, S. J. and Wang, Z. 2012. Item response theory in measurement, assessment and evaluation for higher education.Routledge.

Osuji, U.S. A, Okonkwo, C.A and Nnachi, R.O. 2006.Measurement and Evaluation, EDU 726.School of Education, National Open University of Nigeria.

Partchev, I. 2004. An article on a visual guide to item response theory Friedrich- Schiller-Universitat Jena.

Partchev, I. and De Boeck, P. 2012. Can fast and slow intelligence be differentiated? *Intelligence, 40, 23–32.*

Patz, R. J. and Junker, B. W. 1997. Applications and extensions of MCMC in IRT: Multiple Item types, missing data, and rated responses -Technical Report No. 670. Pittsburgh, PA: Carnegie Mellon University, Department of Statistics.

. 1999. A straightforward approach to Markov chain monte carlo methods for item response theory models. *Journal of Educational and Behavioural Statistics,24, 146-178.*

Peterson, J. L. 2014. Multidimensional item response theory observed score equating methods for mixed-format tests. Ph.D thesis, University of Iowa.Iowa Research Online.

Petrillo, J., Cano, S. J., McLeod, L. D. and Coon, C. D. 2015. Using classical test theory, item response theory and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples.Volume 18, Issue 1, 25–34.

Quellmalz, E. S. and Pellegrino, J. W. 2009.Technology and testing. Science, 323, 75-78.

R Core Team R. 2014. A language and environment for statistical computing; R Foundation for Statistical Computing: Vienna, Austria. Available online: http://www.R-project.org/ (accessed on 10 September 2016).

Ranger, J. and Ortner, T. 2011.Assessing personality traits through response latencies using Item response theory.*Educ. Psychol. Meas. 71, 389–406.*

Ranger, J. 2013.Modeling responses and response times in personality tests with rating scales.*Psychol. Test Assess. 55, 361–382.*

Rasch, G. 1960. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research.

Ratcliff, R. and Tuerlinckx, F. 2002. Estimating the parameters of the diffusion model" approaches to dealing with contamitant reaction times and parameter variability. *Psychon. Bull. Rev. 9, 438-482.*

Ratcliff, R., Thapar, A., Gomez, P. and McKoon, G. 2004. A diffusion model analysis of the effects of aging in the lexical decision task.*Psychology and Aging. 19, (2),278.*
'

Ratcliff, R., Smith, P.L. Brown, S.D. and McKoon, G. 2016. Diffusion decision model: Current issues and history. *Trends Cogn. Sci. 20, 260–281.*

Reckase, M. D. 1997. A linear Logistic Multidimensional Model for dichotomous item response data. In W.J van der Linden and R.K Hambleton (Eds.), Handbook of modern IRT. (pp,271- 286), New York. Springer –Verlag.

.2009.*Multidimensional item response theory*. New York, NY: Springer.

Reeve, B. B. 2000.Item and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory. Unpublished Doctoral Dissertation, University of North Carolina at Chapel Hill.

. 1990. Fitting the two-parameter model to personality data. *Applied Psychological Measurement.14,45–58*.

Reise, S. P. and Waller, N. G. 2003. How many IRT parameters does it take to model psychopathology items?*Psychological Methods,8. 2.164–184.*

. 2009. Item response theory and clinical measurement. Annual Review of Clinical Psychology, 5, 27-48.

Reise, S.P, Widaman, K.F and Pugh, R.H. 1993.Confirmatory factor analysis and item Response theory: Two approaches for exploring measurement invariance. *Psycho Bull; 114:552–566.*

Ripley, M. 2009. Transformational computer-based testing.In F. Scheurmann and J. Bjornsson (Eds.).*The Transition to Computer-Based Assessment* (pp. 89-91).

Rivkin, S. G., Hanushek, E. A and Kain, J. F 2005: Teachers, Schools and academic achievement. Econometrics, 73, 417-458**.**

Roskam, E.E. 1987. Toward a psychometric theory of intelligence. Progress in Mathematical Psychology; Roskam, E.E., Suck, R., Eds.; North Holland: Amsterdam, The Netherlands. pp. 151–171.

.1997. Models for speed and time-limit tests. In Handbook of Modern Item Response Theory; van der Linden, W.J., Hambleton, R.K., Eds.; Springer: New York, NY, USA, pp. 187–208.

Rouse, S. V., Finger, M. S. and Butcher, J. N. 1999. Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment,* 72 282–307.

Rudner, L.M. 2012. Demystifying the GMAT: Computer-based testing terms. Retrieved on 29/05/2017 from www.gmat.com

Rufa'i, A. R. 2012. Nigeria: Laments poor performance of students in Mathematics. Daily Champion, Newspaper of March 2nd.

Rulison, K. L. and Loken, E. 2009. I've fallen and I can't get up: Can high ability students recover from early mistakes in computer adaptive testing?*Applied Psychological Measurement,* 33, 83– 101.

Rupp, A. A. 2003. Item response modeling with BILOG-MG and MULTILOG for Windows.*International Journal of Testing,3, 365-384*.

Salend, S. J. 2009. Classroom testing and assessment for ALL students: Beyond standardization. Thousand Oaks, CA: Corwin Press.

Santor, D.A, Ramsay, J.O. and Zuroff, D.C. 1994. Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychol Assess. 6:255–270.*

Schmiedek, F., Oberauer, K., Wilhelm, O., Susss, H-M and Wittmann, W.W. 2007.Individual differences in components of reaction time distributions and their relations to working memory and intelligence.*J Exp Psychol Gen*. 136(3): 414-429.

Schnipke, D. L. and Scrams, D. J. 1999. Response-time feedback on computer administered tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Scheiblechner, H. 1979. Specific objective stochastic latency mechanisms.*Journal of Mathematical Psychology*, 19, 18-38.

Schnipke, D. L.and Scrams, D. J. 1997. Modeling item response times with a two-state mixture model: A new method of measuring speededness.*Journal of Educational Measurement*, 34(3), 213-232.

Schnipke, D. L. and Scrams, D. J. 2002. Exploring issues of examinee behaviour: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), Computer-based testing: Building the foundation for future assessments (pp. 237-266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwarz, G. 1978. Estimating the dimension of a model.*Annals of Statistics, 6, 461-464.*

Scrams, D. J. and Schnipke, D. L. 1997. Making use of response times in standardized tests: Are accuracy and speed measuring the same thing? Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Schwarz, G. 1978. Estimating the dimension of a model.*Annals of Statistics*, 6: 481-464.

Sidhu, K. S. 1995. The Teaching of Mathematics New Delhi: Sterling Publisher private Limited.

Sijtsma, k. and Junker, B.W. 2006. Item Response Theory: Past performance, presentdevelopments and future expectation. *Behaviourmetrika.33, No1, 75-102.*

Spearman, C. 1904.The proof and measurement of association between two things.*American Journal of Psychology,15, 72 – 101.*

Steinberg, L., Thissen, D. 1995. Item response theory in personality research. Shrout, P. E., Fiske S.T. Personality research, methods, and theory: A festschrift honoring Donald W. Fiske 161–181. Hillsdale, NJ Erlbaum.

Stout, W. 2005.*DIMTEST (Version 2.0) Computer software manual.* Champaig, IL. The William Stout Institute for Measurement.

Suh, H. 2016. A study of Bayesian estimation and comparison of response time models in item response theory. Unpublish Ph.d thesis. Department of Psychology and Research in Education. University of Kansa, USA.

Swaminathan, H. and Gifford, J. A. 1986.Bayesian estimation in the three-parameter logistic model. Psychometrika, 51(4), 589-601.

Swygert, K. A. 1998. An examination of item response times on the GRE-CAT. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

Tavares, H. R., Andrade de, D. F. and Pereira, C. A. 2004.Detection of determinant genes and diagnostic via item response theory.*Genetics and Molecular Biology,27 679–685.*

Templin, J. 2011. ICPSR Summer Item Response Theory Workshop. July 11-15, 2011. University of Georgia. jonathantemplin.com/files/irt/irt11icpsr_lecture01-07.pdf

.2013. Obtaining diagnostic classification model estimate using Mplus. https|doi.org/10.1111/emip.12010.

Tendeiro, J. N. and Meijer, R. R. 2012. A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement, 36, 420-442.*

Thissen, D. 1983. Timed testing: An approach using item response testing. In D.J. Weiss Ed., New horizons in testing: Latent trait theory and computerized adaptive testing pp. 179-203. New York: Academic Press.

. 1991. MULTILOG: Multiple category item analysis and test scoring using item response theory - Computer software. Chicago: Scientific Software International.

Thissen, D. and Steinberg, L. 1986. A taxonomy of item response models. *Psychometrika, 51:567–577.*

Thompson, S. J., Johnstone, C. J. and Thurlow, M. L. 2002. Universal design applied to large-scale assessments (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Centre on Educational Outcome.

Thorndike, E. L. 1918. *The nature, purposes and general methods of measurements of educational products*.In the Measurement of educational products.(The seventeenth yearbook of the National Society for the study of Education, Part II). Bloomington, Illinois: Public School Publishing.

Thorndike, E. L., Bregman, E.O., Cobb, M. V., Woodyard, E., and Inst of Educational Research Div of Psychology. 1926. Teachers Coll, Columbia U. *The measurement of intelligence.*Teachers College Bureau of Publications.

Thorpe, G. L. and Favia, A. 2012. Data analysis using item response theory methodology: An Introduction to Selected Programs and Applications. PsychologyFacultyScholarship.http://digitalcommons.library.umaine.edu/psy_facpub/20.

Thorpe, G. L., McMillan, E., Sigmon, S. T., Owings, L. R., Dawson, R. and Bouman, P. 2007. Latent trait modelling with the common beliefs survey: Using item response theory to evaluate an irrational beliefs inventory. *Journal of Rational- Emotive and Cognitive Behaviour Therapy,* 25, 175-189.

Troy-Gerard, .C.2004.An empirical comparison of item response theory and classical test theory item/person statistics.Unpublished doctoral dissertation, University Texas A and M.

Umar-Ud-Din Khan, M. and Mohamood, S. 2010.Effects of teachers' academic qualification on students' L2 performance at the secondary level.Strength for Today and Bright Hope for Tomorrow, 10(7), July 2010 http://www.languageinindia.com/july2010/teachersqualification.pdf.

Umobong, M.E. and Tommy, U.E. 2017. Dimensionality of national examinations' council's biology examinations: Assessing test quality in modern trend approach. *African Journal of Theory and practice of Educational Assessment. 5,14-30.*

Uwadiae, I. 2017. Quality education in Nigeria: Challenges and way forward. Keynote address at the 5th international conference of the Institute of Education, University of Ibadan, Nigeria.

van der Linden, W.J. 2005. *Linear models for optimal test assembly*; Springer: New York, NY, USA.

. 2006. A lognormal model for response times on test items. *J. Educ. Behav. 31, 181–204.*

. 2007. A hierarchical framework for modelling speed and accuracy on test items. *Psychometrika, 72, 287–308.*

. 2008. Using response times for item selection in adaptive testing. *J. Educ. Stat. 33, 5–20.*

. 2009. Conceptual issues in response-time modeling. *Journal of Educational Measurement,46, No. 3, pp. 247–272.*

. 2011. Modelling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modelling, 53, 334–358*

. 2016. Handbook of item response theory. Volume 2. Taylor and Francis Group.6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL, 33487-2742.

van der Linden, W. and Guo, F. 2008. Bayesian procedures for identifying aberrant response time patterns in adaptive testing. Psychometrika, 73, 365–384.

van der Linden, W. J. and Hambleton, R. K. 1997. *Handbook of modern item response theory*.Springer, New York.

van der Linden, W. J. and Williams R. 2017. The personality project.An introduction to psychometric theory. Chapter 8:The "New Psychometrics" – Item Response Theory. https://personality-project.org/r/book.

van der Linden, W.J, Entink, R. H. K and Fox, J. P. 2010. IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement 34(5), 327–347* sagepub.com/journals Permissions.

van der Linden, W.J., Scrams, D.J. and Schnipke, D.L. 1999. Using response-time constraints to control for differential speededness in computerized adaptive testing.*Appl. Psychol. 23, 195–210.*

van Zandt, T. 2000. How to fit a response time distribution.*Psychonomic Bulletin and Review, 7(3), 424-465*

Verhelst, N.D., Verstralen, H.H.F.M. and Jansen, M.G.H.1997.A logistic model for time limit tests.*In Handbook of Modern Item Response Theory*; van der Linden, W.J., Hambleton, R.K., Eds.; Springer: New York, pp. 169–186.

Vrieze, S. J. 2012. Model selection and psychological theory: A discussion of the differences netween the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psycho Methods*. 17(2), 228-243.

Wainer, H. 2000. CATs: Whither and Whence. Retrieved on 29/05/2017 from http://www.ets.org/Media/Research/pdf/RR00-17-Wainer.pdf.

Wainer, H., Bradlow, E. T. and Du, Z. 2000.Testlet response theory: An analogue for the 3 PL useful in testlet-based adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), Computer adaptive testing: Theory and practice. Boston: Kluwer-Nijhoff.

Wallace, C. S., and Bailey, J. M. 2010. Do concept inventories actually measure anything? Astronomy Education Review, 9, 010116.

Waller, N. G., Reise, S. P.2009. Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI Embretson, S., RobertsJ. S. New directions in psychological measurement with model-based approaches Washington, DC. *American Psychological Association.*

Wang, T. and Hanson, B.A. 2005.Development and calibration of an item response model that incorporates response time.*Appl. Psychol. Meas. 29, 323–339*.

Weiss, D.J. and Minden, S.V. 2012. A comparison of item parameter estimates From X-calibre4.1andBilog-MG.TechnicalReport by Assessment Systems Corporation (ASC).

Wheeler,D. 1983. Mathematisation matters for the learning of mathematics, 3,1; 45- 47. https://flm-journal.org

Wiberg, M. 2004. Classical test theory vs Item response theory.An evaluation of the theory test in the Swedish driving-licence test. EM No. 50. Umea University, Sweden. Retrieved 17th July, 2017.

Wikipedia, 2013.E-assessment retrieval on Saturday, 14th October, 2017 from http://en.m.wikipedia.org.

Wise, S., Pastor, D.A. and Kong, X. 2009.Correlates of rapid-guessing behavior in low stakes testing. Implications for test development and measurement practice. *Appl. Meas. Educ. 22, 185–205.*

Wright, B.D. 2016. Treating all rapid responses as errors - TARRE improves estimates of ability slightly. *Psychological Test and Assessment Modeling, 58, 2016 (1), 15-31.*

Wright, B.D. and Mead, R.J. 1976. Calibrating rating scale with the Rasch model (Reseach memorandum n^0 23). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

www.personality-project.org/r/book/The Personality Project: An introduction to psychometric theory: CHAPTER 8.

Yang, Y. 2005. Can the strength of AIC and BIC be shared? A conflict between model identification and regression estimation.*Biometrika*, 92: 937-950.

Yen, T.Y. 1993.A comparison of three statistical procedures to identify clusters of items with local dependency.Huynh University of Carolina.

Yen, Y.C., Ho, R.G., Liao, W.W., Chen, L.J. and Kuo, C.C. 2012. An empirical evaluation of the slip correction in the four parameter logistic models with Computerized adaptive testing. *Applied Psychological Measurement, 36, 75-87.*

Zakariya, Y. F and Bamidele, E. F, 2015. Investigation into the causes of poor academic performance in mathematics among Nigerian undergraduate students.*World Journal of Social Sciences and Humanities.* 1, No. 1, pp 1-5.

Zeng, L. 1997. Implementation of marginal Bayesian estimation with four-parameter beta prior distributions.*SAGE Journal*, *21, Issue 2.*

Zenisky, A. l., and Baldwin, P. 2006. Using response time data in test development and validation: research with beginning computer users. Paper presented at the annual meeting of the national council on measurement in education, san Francisco.

Zickar, M.J. and Broadfoot, A. A. 2009.The Partial Revival of a Dead Horse? Comparing Classical Teat Theory And Item Response Theory. In C. E. Lance and R. J. Vandenberg (Eds.), Statistical and Methodological Myths and Urban Legends: Doctrine, Variety and Fable in the Organizational and Social Sciences (pp. 37-60). New York: Routledge.

Zimowski, M. F., Muraki, E., Mislevy, R. J. and Bock, R. D. 1996. BILOG-MG: Multiple group IRT analysis and test maintenance for binary items – Computer software. Chicago: Scientific Software International.

APPENDIX I

SUGGESTED SCHOOLS WITH COMPUTER LABOURATORIES IN OYO STATE

S/n	Names of schools	LGA
1.	Methodist Grammar School, Bodija	Ibadan North
2.	Oba Akinyele Memmorial High School, Mokola	Ibadan North
3.	Abadina College, UI	Ibadan North
4.	Loyola College Ibadan	Ibadan NorthEast
5.	Lagelu Grammar School, Agugu	Ibadan NorthEast
6.	Oba Abass Grammar School, Eleyele Roasd	Ibadan NorthWest
7.	Jericho High School	Ibadan NorthWest
8.	Wesley College of Science, Ibadan	Ibadan SouthEast
9.	Ibadan City Academy, Ibadan	Ibadan SouthEast
10.	Ibadan Grammar School, Ibadan	Ibadan SouthEast
11.	Government Secondary School, Orita Aperin	Ibadan SouthEast
12.	St Anne's School, Molete	Ibadan SouthEast
13.	Ansarudeen (ADS) Grammar School, Oke Ado	Ibadan SouthWest
14.	Our Lady of Apostle Secondary School, Odo Ona	Ibadan SouthWest
15.	St Teresa College, Oke Ado	Ibadan SouthWest
16.	Obasieku High School, Eruwa	Ibarapa East
17.	Awe High School, Awe	Afijio
18.	Fiditi Grammar School, Fiditi	Afijio
19.	Government College, Ogbomosho	Ogbomoso North
20.	Baptist Academy Ogbomoso	Ogbomosho South
21.	Baptist High School, Okeho	Iseyin
22.	Isabatudeen Girls Grammar School, Ibadan	Lagelu
23.	Prospect High School, Abanla	Lagelu
24.	Ojongbolu Grammar School, Oyo	Oyo West
25.	Koso Community Grammar School, Iseyin	Iseyin
26.	Muslim Secondary School, Saki	Saki West
27.	Igboora Grammar School, Igboora	Ibarapa Central
28.	Elekuro High School, Oke Ogbere	Ona Ara
29.	School of Science, Oyo	Atiba
30.	School of Science, Ogbomoso	Ogbomoso North
31.	Community High School, Ido	Ido
32.	Irepo Grammar School, Igboho	Oorelope
33.	Igbo Elerin Grammar School, Ibadan	Lagelu
34.	Okaka Community Grammar School, Okaka	Itesiwaju
35.	Komu-Babaode High School	Itesiwaju
36.	Community Grammar School, Akanran	Ona-Ara
37.	Muslim Grammar School, Ighoho	Orelope
38.	Igbojaye Community High School, Igbojaye	Itesiwaju
39.	Ogbomoso High School, Ogbomoso	Ogbomoso South
40.	Ibarapa Central Ayelogun Grammar School, Idere	Ibarapa Central
41.	Akolu Grammar School, Eruwa	Ibarapa East
42.	Olivet Baptist High School, Oyo	Oyo East
43.	Federal Girls College, Oyo	Oyo
44.	Igangan High School, Igangan	Ibarapa North

Source: Planning, Research and Statistics Unit, Ministry of Education (2018)

APPENDIX II

LIST OF SCHOOLS USED FOR THE TRIAL-TESTING STAGE IN OYO STATE

S/n	Names of schools	LGA
1.	Abadina College, University of Ibadan	Ibadan North
2.	Islamic High School, Basorun	Ibadan North
3.	Islamic Day Secondary School, Basorun	Ibadan North
4.	Loyola College Ibadan	Ibadan North-East
5.	Lagelu Grammar School, Agugu	Ibadan North-East
6.	Oba Abass Grammar School, Eleyele Roasd	Ibadan North-West
7.	St Anne's School, Molete	Ibadan South-East
8.	Ibadan Grammar School, Ibadan	Ibadan South-East
9.	Ibadan Boys High School, Oke Ado	Ibadan South-West
10.	Queens' School, Apata	Ibadan South-West
11.	Government College, Apata,Ibadan	Ibadan South-West
12.	Oranyan Grammar School, Oyo	Atiba
13.	Bishop Philips Academy, Iwo Road	Egbeda
14.	Abiodun Atiba Memorial Institute, Oyo	Oyo East
15.	Oliveth Baptist High School, Oyo	Oyo East

APPENDIX III

INTERNATIONAL CENTRE FOR EDUCATIONAL EVALUATION
INSTITUTE OF EDUCATION
UNIVERSITY OF IBADAN

THE POOLED COMPUTER–BASED MATHEMATICS ACHIEVEMENT TEST (CBMAT)

PLEASE READ THIS INSTRUCTION CAREFULLY. RESPONDENT SHOULD ANSWER ALL THE QUESTIONS BY PICKING THE RIGHT ANSWER FROM THE OPTIONS A-D PROVIDED. AN ANWSER SHOULD BE PROVIDED TO A QUESTION BEFORE CLICKING THE NEXT BUTTON THAT WILL BRING THE FOLLOWING QUESTION. ALSO NOTE THAT, YOU CANNOT GO BACK TO THE PREVIOUS QUESTION(S). THE SUBMIT AND OK BUTTONS MUST BE CLICKED AFTER ATTEMPTING ALL THE QUESTIONS TO SUCCESSFULLY COMPLETE YOUR EXERCISE.

1. Find 39 ⊕ 29 in modulo 6 (a) 1 (mod 6) (b) 2 (mod 6) (c) 3 (mod 6 (d) 4 (mod 6)

2. Out of 25 teachers, 16 are married and 15 are women, if 6 of the men are married, how many of the women are not married? (a) 15 (b) 10 (c) 5 (d) 3

3. Solve the equation (x-2) (x+7) = 0; x = (a) 2 or -7 (b) -2 or -7 (c) 2 or 7 (d) -2 or 7

4. Solve 8x = 10 (mod 3) (a) x = 0 (mod 3) (b) x = 1 (mod 3) (c) x = 2 (mod 3) (d) x = 3 (mod 3)

5. If sin P = $\frac{3}{5}$ and P is an acute angle, what is the value of tan P.

   (a) $\frac{2}{5}$(b) $\frac{3}{4}$ (c)$\frac{3}{5}$ (d) $\frac{2}{3}$

6. Find 567 in standard form (a) 5.67x10^2 (b)56.7x10^2 (c) 567x10^3 (d) 0.567x10^2

7. Given that y = 4+3x-x^2, complete the table of values for the given equation.

   Table 1

X	-1	1	2	3
Y	0	6		

   (a) 6, 2 (b) 2, 6 (c) 6, 4 (d) 4, 6

8. The members of a set of even numbers less than 15 are

(a) {2,3,4,6,8,11,13} (b) {2,4,6,8,10,12,14} (c) {3,4,6,8,9,10,14} (d) {2,4,6,8,9,10,14}

9. The expression pq^{-2} can be rewritten as (a) $^{p}/_{q}$(b) $^{p^2}/_{q^2}$ (c) $^{p}/_{q^2}$ (d)$^{p^2}/_{q}$

10. Evaluate $\sin 137^0$ (a) + $\sin 43^0$ (b) – Sin 43^0 (c) + Cos 43^0 (d) – Cos 43^0

11. If M = 314_5 and N = 24_5, calculate M+N (a) 234_5 (b) 334_5 (c) 342_5 (d) 343_5

12. What is M÷N in base 5, if M = 314_5 and N = 24_5 (a) 11_5 (b) 13_5 (c) 15_5 (d) 17_5

13. Express 0.00562 in standard form (a) $5.62\text{x}10^{-3}$ (b) $5.62\text{x}10^{-2}$ (c) $5.62\text{x}10^{2}$ (d) $5.62\text{x}10^{3}$

Use the tables 2 and 3 to answer Q14

Table 2: Logarithm table **Difference**

								Difference				
X	0	1	2	3	4	5	6	7	1	2	3	4	5
53	7243	7251	7259	7267	7275	7284	7292	7300	1	2	2	3	4

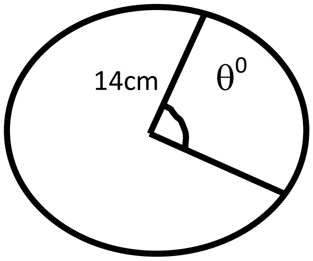
Table 3: Antilogarithm table **Difference**

X	0	1	2	3	4	1	2	3	4
19	1549	1552	1556	1560	1563	0	1	1	1
73	5370	5383	5395	5408	5420	1	3	4	5
91	8128	8147	8166	8185	8204	2	4	6	8

14. Evaluate 53.75^3 (a) 15.53 (b) 1553 (c) 15530 (d) 155300

15. Calculate in terms of π, the total surface area of a cone of base diameter 12cm and height 10cm. (a) 6π ($\sqrt{136}$ + 6) cm^2 (b) 3π ($\sqrt{136}$ + 6) cm^2 (c) 2π ($\sqrt{136}$ + 6) cm^2 (d) 6π ($\sqrt{136}$ + 3) cm^2

16. What angle does an arc 6.6cm in length subtend at the centre of a circle of radius 14cm? use π = $\frac{22}{7}$



6.6cm (a) 8^0 (b) 15^0 (c) 18^0 (d) 27^0

Figure 1

17. Convert 2077ten to base eight (a) 4305_8 (b) 4035_8 (c) 4503_8 (d) 5034_8

18. Find the square of 111_2 (a) 110001 (b) 100011 (c) 111000 (d) 000111

19. Convert this bicimal 1.101 to decimal (a) 16.25 (b) 162.5 (c) 16.25 (d)1.625

20. A car's petrol tank is 0.8m long, 25cm wide and 20cm deep. How many litres ofpetrol can it hold? (a) 4000 litres (b) 400 litres (c) 40 litres (d)4 litres

21. The shorter hand of a clock points to 5 while the longer points to 12 on a clock face. What number does it point to after 15hours (a) 10 (b) 9 (c) 8 (d) 7

22. Calculate 12 $\ominus$5 in modulo 4 (a) 3(mod 4) (b) 4(mod 4) (c) 5(mod 4(d) 6(mod 4)

23. This decimal number 0.0078 can be expressed in standard form as (a) 7.8 x 10^{-4}(b) 7.8 x 10^{-3} (c) 7.8 x 10^{-2} (d) 7.8 x 10^{-1}

24. Calculate the area in hectares of a rectangular field 126m long and 97m wide (1 hectare = 10,000m^2) (a) 2.44 (b) 24.4 (c) 1.22 (d) 12.2

25. Solve $4^{c-1} = 64$; c = (a) 4 (b) 3 (c) 2 (d) 1

26. Reduce $\left\{\frac{8}{50}\right\}^{-1/2}$ to its lowest term (a) $\frac{4}{25}$ (b) $\frac{25}{4}$ (c) $\frac{5}{4}$ (d) $2\frac{1}{2}$

27. If the area of a square field is 3.95 hectares calculate the length of a side of the field in metres (1 hectare = 10,000m^2) (a) 158m (b) 159m (c) 198m (d) 199m

Use this information to answer Q28-31

Given that μ = {a,b,c,d,e,f}, X = {a,b,c,d} Y = {c,d,e} and Z = {b,d,f}

28. The set XUY is (a) {a,b,c,d,e,f} (b) {a,b,c,d} (c) {a,b,c,d,e} (d) {a,b,c,e,f}

29. μ∩ Z is (a) {a,b,c,d} (b) {c,d,f} (c) {b,d,f} (d) {a,c,d}

30. ZUØ is (a) {a,b,d} (b) {b,d,f} (c) {c,d,e} (d) {a,b,c,}

31. XU (YUZ) is (a) {a,b,c,d,e,f} (b) {c,d,e,f} (c) {a,b,d,e,f} (d) {a,b,c.f}

Given this Venn diagram below to answer Q32

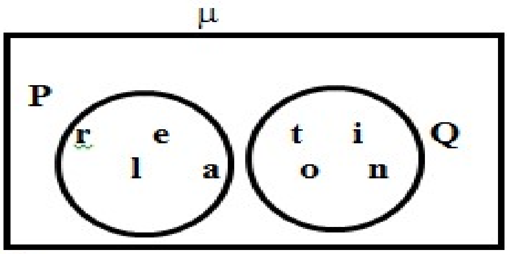


Figure 2

32. The set PUQ is (a) a proper subset of P (b) the universe set (c) an empty set (d) an intersection set

Given a Universal Set μ= {1,2,3,4,5}, A = {1,3} and B = {3,4}. Use this information to answer Q33 and 34

33. The set A′ is (a) {3,4} (b) {1,3} (c) {2,3,5} (d) {2,4,5}

34. The set (A∩B)′ is (a) {1,2,3} (b) {4,5} (c) {1,3,4,5} (d) {1,2,4,5}

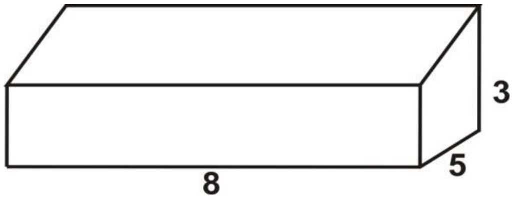
35. A company employs 100 people, 65 of whom are men, 60 people including all the women, are paid weekly. How many of the men are paid weekly?
(a) 40 (b) 35 (c) 25 (d) 15

36. Find -2(mod 9) in its simplest form (a) 4 (mod 9) (b) 5 (mod 9) (c) 6 (mod 9) (d) 7 (mod 9)

37. Given a set {y: 1 < y ≤ 6} y ε N, the members of the set are (a) {1,2,3,4} (b) {2,3,4,5,6} (c) {1,2,3,4,5} (d) {1,2,3,4,5,6}

38. Make x the subject of the equation $a = \frac{b+x}{b-x}$ (a) $\frac{a(b-1)}{a+1}$ (b) $\frac{a(b-1)}{a-1}$ (c) $\frac{b(a+1)}{a-1}$ (d) $\frac{b(a-1)}{a+1}$

39. Given the solid shape in figure 3, what is its volume in cm^3?



(a) 120 (b) 150 (c) 240 (d) 400

Figure 3

40. When travelling between two towns, the time taken varies inversely with the average speed. When the average speed is 42km/h, the journey takes 4hours. Find the average speed if the journey takes 2hours 20minutes.
(a) 60km/h (b) 65km/h c) 72km/h (d) 79km/h

41. Factorize x^2-8x-20 (a) (x+10) (x+2) (b) (x-10) (x+2) (c) (x+4) (x-5) (d) (x-4) (x+5)

42. Evaluate 2 ⊗ 2 in modulo 4 (a) 0 (mod 4) (b) 1(mod 4) (c) 2(mod 4) (d) 3 (mod 4)

43. Find the quadratic equation whose roots are 3 and 4
(a) x^2-7x+12=0 (b) x^2+7x+12=0 (c) x^2+7x-12=0 (d) x^2-7x-12=0

44. The perimeter of a rectangle is 20m and the length is xm. Find the area of the rectangle in terms of x (a) 10(x+5) (b) x^2(x-10 (c) x(10-x) (d) x^2(10+x)

45. Remove bracket from 3- (a-$\overline{5-6a}$) (a) -8-7a (b) 8+7a (c) 8-7a (d) 7a-8

46. Calculate the angle marked with x in figure 4

(a) 28^0 (b) 56^0 (c) 100^0 (d)124^0

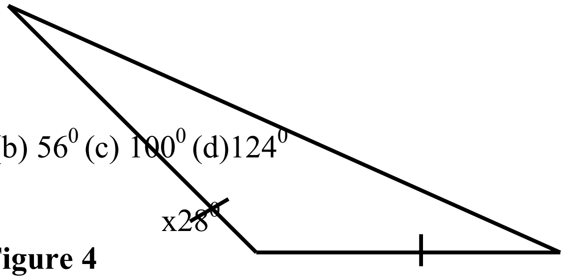


Figure 4

47. Calculate the area in hectares of a rectangular field 126m long and 97m wide (1 hectare = 10,000m^2) (a) 2.44 (b) 24.4 (c) 1.22 (d) 12.2

Use the following expressions in logical reasoning to answer Q48 -Q54

A: Lagos is a city in Nigeria.
B: A square is a rectangle.
C: God willing.
D: They are lovely people.
E: Some students in SS1 study physics and some don’t.
F: He ran at a constant speed.
G: If you press the switch, the door will open.

48. Which of the following is right of an expression A (a) it is a command (b) it is not true (c) it is a statement (d) it is a negative statement

49. Expression B is (a) true (b) false (c) it is not a statement (d) it is a compound statement

50. Expression C is (a) true (b) false (c) it is not a statement (d) it is a sentence

51. Expression D is (a) it is a command (b) it is a sentence (c) true (d) false

52. Expression E is (a) an implication statement (b) a conditional statement (c) a compound statement (d) an equivalence statement

53. The negation of expression F is (a) He will run at a constant speed (b) He ran at a speed that is somehow constant (c) He ran at a varied speed (d) He does not mean to run at a constant speed.

54. The equivalence of expression G is (a) if you do not press the switch, the door will not open (b) if you press the switch, the door will not open (c) if you do not press the switch, the door will open (d) if the door is not open, you should press the switch

Use this information to answer Q55 and 56.
For the set of positive whole numbers, let
P: X is exactly divisible by 2
Q: X is an even number.
This symbol $\Rightarrow$ means imply and ~ means negation

55. Which of the following is true about statement P and Q
(a) $P \Rightarrow Q$ (b) $\sim P \Rightarrow Q$ (c) $\sim Q \Rightarrow P$ (d) $P \Rightarrow \sim Q$

56. P and Q are (a) disjoint statement (b) conjoint statement (c) equivalent statement (d) compound statement

57. Log_{10} 10000 is equal to (a) 1 (b) 2 (c) 3 (d) 4

Given the figure below;

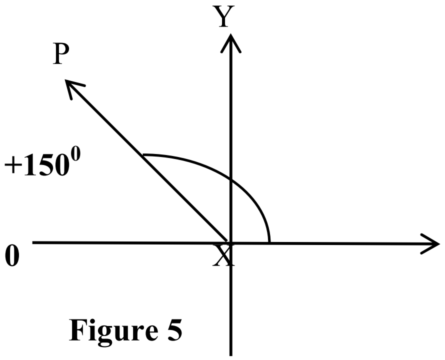


Figure 5

58. Which quadrant does P lies in, given that the angle between $\overline{OP}$ and $\overline{OX}$ is $+150^0$?
(a) 1st quadrant (b) 2nd quadrant (c) 3rd quadrant (d) 4th quadrant

59. If $M = 314_5$ and $N = 24_{5,}$ calculate M+N (a) 234_5 (b) 334_5 (c) 342_5 (d) 343_5

The table below gives the ages of student in SS1C who were born in April. Use the information to answer Q 60 – 63.

Table 4

Age (year)	13	14	15	16	17
Number of students	8	5	6	3	8

60. How many students were born in April?
(a) 10 (b) 20 (c) 30 (d) 40

61. The average age of the students in the class is
(a) 10 (b) 15 (c) 14 (d) 13

62. The median age of the students is
(a) 16 (b) 15 (c) 14 (d) 13

63. Students who fall to the range of ages 13 to 15 are
(a) 16 (b) 17 (c) 18 (d) 19

64. If sin θ = $\frac{5}{13,}$ find the value of $\frac{1 + \tan\theta}{1 - \tan\theta}$

(a) $\frac{17}{12}$(b) $\frac{5}{12}$ (c) $\frac{15}{7}$ (d) $\frac{17}{7}$

Use the following data to answer Q65 – Q67

The shoe sizes of a group of 24 students in a class are

8 6, 7, 5, 4, 6, 5, 7

6, 5, 7, 6, 8, 5, 4, 6

5, 5, 6, 7, 8, 8 6 7

65. The frequency of the shoe size 5 has the tally of (a) **III**(b) **IIII** (c) **II~~II~~**~~ (d~~) **I~~IIII~~**
66. The tally **~~IIII~~ II** is the frequency of the shoe size that appear most among the group. This shoe size is (a) 5 (b) 6 (c) 7 (d) 8
67. The average shoe size in the class is (a) 6 (b) 7 (c) 8 (d) 9
68. In a right–angled triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides (a) true (b) false (c) none of the option (d) All of the option.

Use the two diagrams as applicable in figure 6a and 6b to answer Q69–71

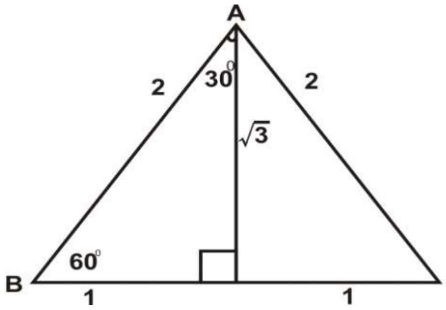


Figure 6a

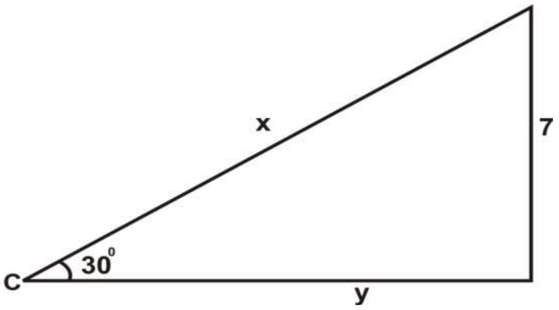


Figure 6b

69 What is the side marked x in figure 6b? (a) 12 (b) 13 (c) 14 (d) 15

70. What is the side marked y? (a) $6\sqrt{3}$ (b) $7\sqrt{3}$ (c) $5\sqrt{3}$ (d) $2\sqrt{3}$

71. Find sin 30^0 in figure 6b above (a) $\frac{1}{2}$ (b) $\frac{2}{3}$ (c) $\frac{3}{\sqrt{3}}$ (d) $\frac{1}{\sqrt{3}}$

72. An isosceles triangle is such that one of the base angles is twice the third angle. Find the value of one of its base angles. (a) 72^0 (b) 60^0 (c) 45^0 (d) 36^0

73. Convert 1264eight to base ten (a) 629 (b) 692 (c) 962 (d) 296

74. Calculate the perimeter of a sector of a circle of radius 7cm, the angle of the sector being 108^0. If $\pi = \frac{22}{7}$ (a) 13.2cm (b) 17.4cm (c) 22.3cm (d) 27.2cm

75. Calculate the area of the shaded segment in figure 7, if $\pi = \frac{22}{7}$

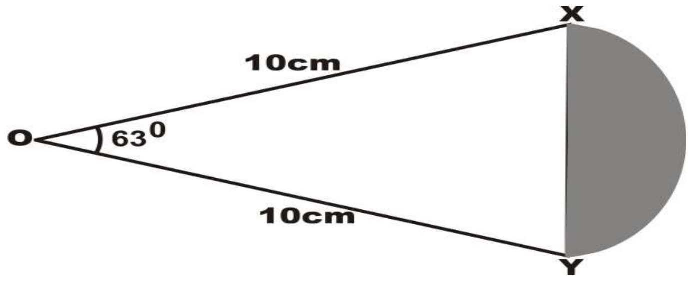


Figure 7

(a) 55cm^2 (b) 44.55cm^2 (c) 20cm^2 (d) 10.45cm^2

76. Find 2⨸3 in modulo 4 (a) 2 (mod 4) (b) 3 (mod 4) (c) 4 (mod 4) (d) 5(mod 4)

77. From a point P on level ground, the angle of elevation of the top of a tree is 60^0. If the tree is 39m high, how far is its base from P.

(a) $\frac{39}{\sqrt{3}}$m (b) $\frac{25}{\sqrt{3}}$ m (c) $\frac{27}{\sqrt{2}}$ m (d) $\frac{20}{\sqrt{3}}$ m

78. The chord $\overline{AB}$ of a circle whose centre 0 is 10cm long, and $A\hat{O}B = 140^0$. Calculate the radius of the circle (a) 3.42cm (b) 4.24cm (c) 5.32cm (d) 5.87cm

79. Simplify $\left\{\frac{8}{27}\right\}^{-2/3}$ (a) $4^{1/4}$ (b) $3^{1/4}$ (c) $2^{1/4}$ (d) $1^{1/4}$

80. If the volume of a rectangular-based pyramid is $70cm^3$ and its base area is $28cm^2$, calculate the height of the pyramid. (a) 10.5cm (b) 9.5cm (c) 8.5cm (d) 7.5cm

81. The sum of the interior angles of and n-sided convex polygon is
(a) (2n+4) right angles (b) (2n-4) right angles (c) (n+2) right angles (d) (n-2) right angles

82. Two triangles are congruent, if
(a) two sides and the included angle of one are respectively equal to two sides and the included angle of the other (SAS).

(b) two angles and a side of one are respectively equal to two angles and the corresponding side of the other (ASA).

(c) the three sides of one are respectively equal to the three sides of the other (SSS).

(d) all of the above hold.

83. One of the following is true of an isosceles triangle.

(a) The base angles are equal.
(b) The equal sides do not meet at the vertex.
(c) The bisector of the vertex angle does not meet the base at right angle.
(d) The two triangles formed by the bisector are not equivalent.

84. All these are quadrilaterals **EXCEPT** (a) rhombus (b) trapezium (c) cylinder (d) flying kite

85. All of these properties are true of a rectangle and a square **EXCEPT**
(a) the diagonal are equal (b) all the form angles are right angles (c) opposite sides are longer than one another (d) opposite sides are parallel

86. A boy's age is x years and his father is four times as old. Find the father's age in y years' time (a) (4x-y) (b) (4x+y) (c) (y-4x) (d) (-4x-y)

87. The angle marked u in figure 8 is (a) 124^0 (b) 100^0 (c) 56^0 (d) 28^0

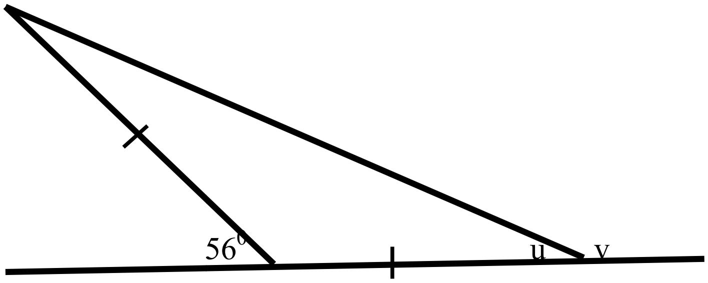


Figure 8

88. The angle marked v is in figure 8 is (a) 156^0 (b) 152^0 (c) 124^0 (d) 116^0

Table 5: Logarithm Table. **Use the tables 5 and 6 to answer Q89**

								Difference				
X	0	1	2	3	4	5	6	7	1	2	3	4	5
53	7243	7251	7259	7267	7275	7284	7292	7300	1	2	2	3	4

Table 6: Antilogarithm Table

					Difference			
X	0	1	2	3	4	1	2	3	4
19	1549	1552	1556	1560	1563	0	1	1	1
73	5370	5383	5395	5408	5420	1	3	4	5
91	8128	8147	8166	8185	8204	2	4	6	8

89. Evaluate $\sqrt[3]{537.5}$ (a) 8130 (b) 81.30 (c) 8.13 (d) 0.813

90. If you are to construct angle 30^0, which of the following angles should bisection be made (a) 90^0 (b) 60^0 (c) 45^0 (d) 30^0

91. On the 3rd quadrant of the Cartesian plane
(a) All ratios are positive (b) only sine is positive (c) only tangent is positive (d) only cosine is positive

92. Calculate this bicimal, 11.01 - 1.11 (a) 11.01 (b) 11.1 (c) 1.10 (d) 0.110

93. If x α y and x = 3 when y = 12, find the relationship between x and y
(a) x = 4y (b) $4x=\frac{1}{y}$ (c) $x = \frac{y}{4}$ (d) $x = \frac{4}{y}$

94. Calculate in terms of π, the total surface area of a solid cylinder of radius 3cm and height 4cm (a) $62\pi cm^3$ (b) $52\pi cm^2$ (c) $42\pi cm^3$ (d) $32\pi cm^3$

95. Calculate the area of this trapezium.

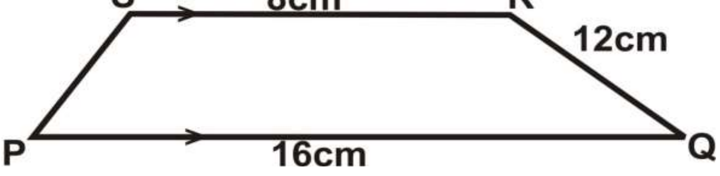


Figure 9

(a) 72cm^2 (b) 62cm^2 (c) 52cm^2 (d) 42cm^2

96. Solve 8 cos θ - 1 = 0 (a) 56.4^0, 123.6^0 (b) 82.8^0, 277.2^0 (c) 115.4^3, 244.6^0 (d) 26.4^0, 153.6^0

97.

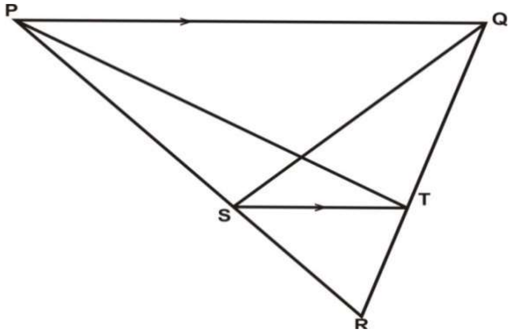


Figure 10

In figure 10,$\overline{STP\|Q}$, which triangle is equal in area to ΔPQS

(a) ΔPQT (b) ΔPQS (c) ΔQST (d) ΔSTR

Use figure 11 to answer Q98 - Q101

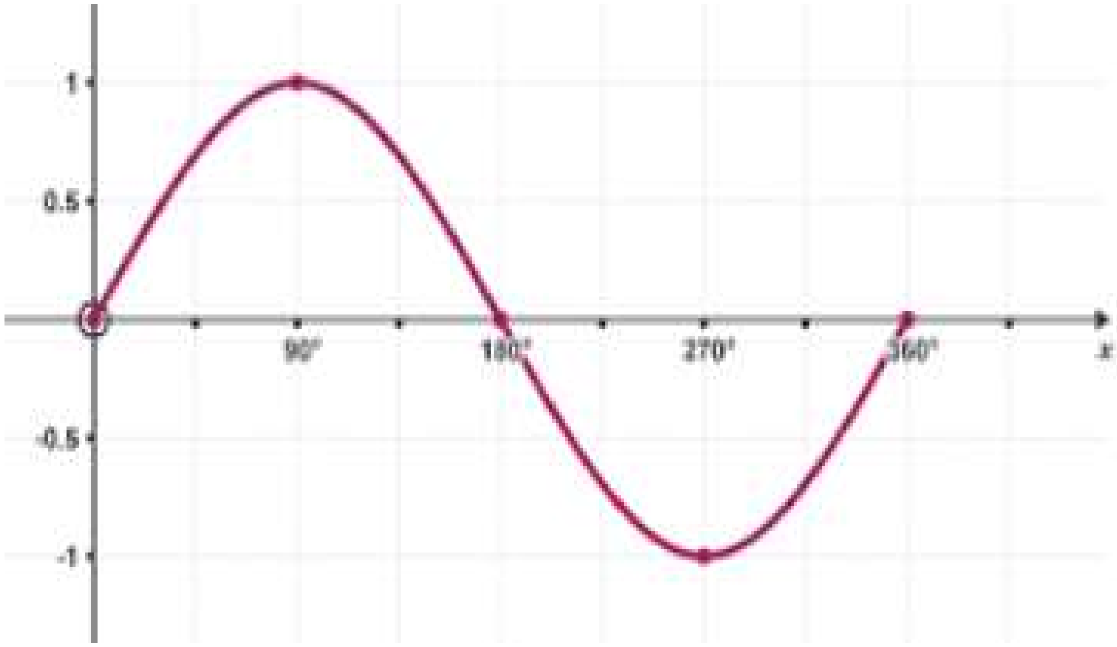


Figure 11

98. Find the value of sin 210^0 (a) 0.5 (b) -0.5 (c) 1.0 (d) -1

99. Solve the equation 5 sin θ = 4 using the graph above, θ =
(a) 54^0 or 126^0 (b) 64^0 or 116^0 (c) 74^0 or 106 (d) 84^0 or 96^0

100. Find the value of sin 270^0 (a) 1 (b) 0.5 (c) -1 (d) -0.5

101. The values for sins 0^0 and 360^0 are the same, therefore the value is (a) 0 (b) 0.5 (c) -1 (d) 1

Use the following information to answer Q102 - 106
The Pie Chart in figure 12 represents 24 hours in the life of a student

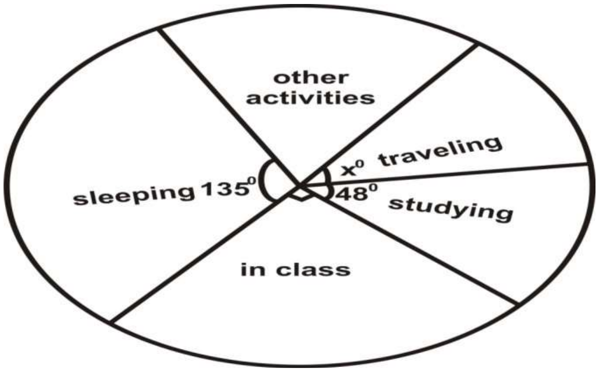


Figure 12

102. What fraction of the time is spent sleeping
(a) $\frac{1}{8}$(b) $\frac{1}{4}$ (c)$\frac{3}{8}$ (d) $\frac{1}{2}$

103. What percentage of time is spent studying
(a) $13\frac{1}{3}$% (b) 23 $\frac{1}{3}$ % (c) 33 $\frac{1}{3}$% (d) 43 $\frac{1}{3}$ %

104. How much time is spent studying?
(a) 4hrs 12min (b) 3hrs 12min (c) 2hrs 15mins (d) 1hr 15mins

105. What is the ratio of the time of studying to the time of sleeping

(a) 14:30 (b) 15:40 (c) 16:45 (d) 20:50

106. In its simplest form, what fraction of time is the student spending in class
(a) $\frac{1}{4}$(b) $\frac{1}{3}$ (c)$\frac{1}{2}$ (d) $\frac{1}{5}$

The examination result of a class is given by the bar chart in figure 13. Use it to answer questions 107 – 111

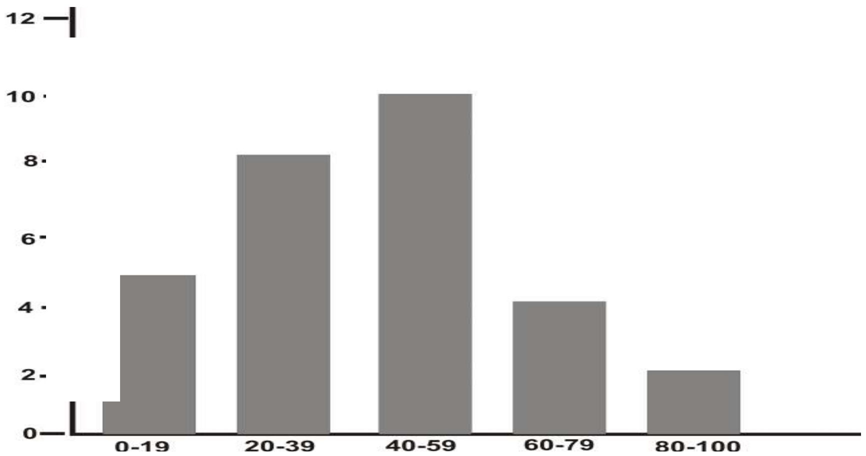


Figure 13 **Marks**

107. How many students took the examination (a) 49 (b) 39 (c) 29 (d) 19

108. If the pass mark is 40, how many students passed the examination? (a) 20 (b) 18 (c) 16 (d) 14

109. How many students failed the examination (a) 13 (b) 14 (c) 15 (d) 16

110. Which of the range of marks is the modal class? (a) 20 – 39 (b) 40 – 59 (c) 60 – 79 (d) 80 – 100

111. Which of the range of marks is the median class (a) 20 – 39 (b) 40 – 59 (c) 60 – 79 (d) 80 – 100

Use figure 14 to answer Q112 – Q114

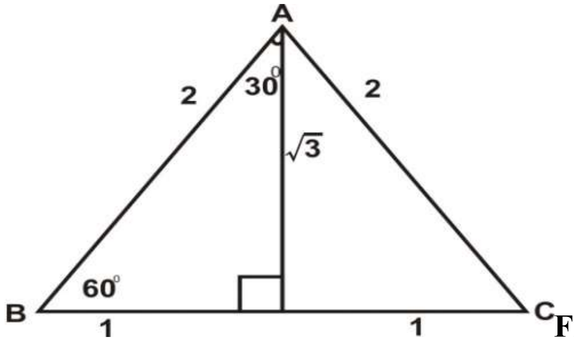


Figure 14

112. From the triangle above, sin60^0 is (a) $\frac{1}{2}$ (b) $\frac{2}{\sqrt{3}}$ (c) $\frac{\sqrt{3}}{2}$ (d) $\frac{1}{\sqrt{3}}$

113. Cos30^0 is (a) $\frac{1}{2}$ (b) $\frac{2}{\sqrt{3}}$ (c) $\frac{\sqrt{3}}{2}$ (d) $\frac{1}{\sqrt{3}}$

114. Tan 30^0 is (a) $\frac{1}{2}$ $(b)\,\frac{2}{\sqrt{3}}$ (c) $\frac{\sqrt{3}}{2}$ (d) $\frac{1}{\sqrt{3}}$

APPENDIX IV

INTERNATIONAL CENTRE FOR EDUCATIONAL EVALUATION
INSTITUTE OF EDUCATION
UNIVERSITY OF IBADAN+

PART A: COMPUTER–BASED MATHEMATICS ACHIEVEMENT TEST (CBMAT)

PLEASE READ INSTRUCTION CAREFULLY. RESPONDENT IS MEANT TO ANSWER ALL THE QUESTIONS AND PICK THE RIGHT ANSWER FROM THE OPTIONS A-D. AN ANWSER SHOULD BE PROVIDED TO A QUESTION, AFTER WHICH THE BUTTON NEXT WILL BE CLICKED TO BRING THE FOLLOWING QUESTION. ALSO, THE SUBMIT BUTTON MUST BE CLICKED AFTER ATTEMPTING ALL THE QUESTIONS TO SUCCESSFULLY SUMBIT YOUR EXERCISE.

1. Find 39 $\oplus$ 29 in modulo 6 (a) 1 (mod 6) (b) 2 (mod 6) (c) 3 (mod 6 (d) 4 (mod 6)

2. Solve the equation (x-2) (x+7) = 0; x = (a) 2 or -7 (b) -2 or -7 (c) 2 or 7 (d) -2 or 7

3. If sin P = $\frac{3}{5}$ and P is an acute angle, what is the value of tan P.
   (a) $\frac{2}{5}$(b) $\frac{3}{4}$ (c)$\frac{3}{5}$ (d) $\frac{2}{3}$

4. Given that $y = 4+3x-x^2$, complete the table of values for the given equation.
   Table 1

X	-1	1	2	3
Y	0	6		

   (a) 6, 2 (b) 2, 6 (c) 6, 4 (d) 4, 6

5. The expression pq^{-2} can be rewritten as (a) $^{p}/_{q}$ (b) $^{p\,2}/_{q2}$ (c) $^{p}/_{q}\,2$ (d) $^{p\,2}/_{q}$

6. If M = 314_5 and N = $24_{5,}$ calculate M+N (a) 234_5 (b) 334_5 (c) 342_5 (d) 343_5

7. Express 0.00562 in standard form (a) $5.62\text{x}10^{-3}$ (b) $5.62\text{x}10^{-2}$ (c) $5.62\text{x}10^{2}$ (d) $5.62\text{x}10^{3}$

8. Calculate in terms of π, the total surface area of a cone of base diameter 12cm and height 10cm. (a) 6π ($\sqrt{136}$ + 6) cm^2 (b) 3π ($\sqrt{136}$ + 6) cm^2 (c) 2π ($\sqrt{136}$ + 6) cm^2 (d) 6π ($\sqrt{136}$ + 3) cm^2

9. Convert 2077ten to base eight (a) 4305_8 (b) 4035_8 (c) 4503_8 (d) 5034_8

10. Convert this bicimal 1.101 to decimal (a) 16.25 (b) 162.5 (c) 16.25 (d)1.625

11. The shorter hand of a clock points to 5 while the longer points to 12 on a clock face.
What number does it point to after 15hours (a) 10 (b) 9 (c) 8 (d) 7

12. This decimal number 0.0078 can be expressed in standard form as
(a) 7.8 x 10^{-4} (b) 7.8 x 10^{-3} (c) 7.8 x 10^{-2} (d) 7.8 x 10^{-1}

13. Solve $4^{c-1} = 64$; c = (a) 4 (b) 3 (c) 2 (d) 1

14. If the area of a square field is 3.95 hectares calculate the length of a side of the field in metres (1 hectare = 10,000m^2) (a) 158m (b) 159m (c) 198m (d) 199m

Use this information to answer Q15-16
Given that μ = {a,b,c,d,e,f}, X = {a,b,c,d} Y = {c,d,e} and Z = {b,d,f}

15. μ∩ Z is (a) {a,b,c,d} (b) {c,d,f} (c) {b,d,f} (d) {a,c,d}

16. XU (YUZ) is (a) {a,b,c,d,e,f} (b) {c,d,e,f} (c) {a,b,d,e,f} (d) {a,b,c.f}

Given a Universal Set μ= {1,2,3,4,5}, A = {1,3} and B = {3,4}. Use this information to answer Q17

17. The set A′ is (a) {3,4} (b) {1,3} (c) {2,3,5} (d) {2,4,5}

18. A company employs 100 people, 65 of whom are men, 60 people including all the women, are paid weekly. How many of the men are paid weekly?
(a) 40 (b) 35 (c) 25 (d) 15

19. Given a set {y: 1 < y ≤ 6} y ε N, the members of the set are (a) {1,2,3,4} (b) {2,3,4,5,6} (c) {1,2,3,4,5} (d) {1,2,3,4,5,6}

20. Given the solid shape in figure 3, what is its volume in cm^3?

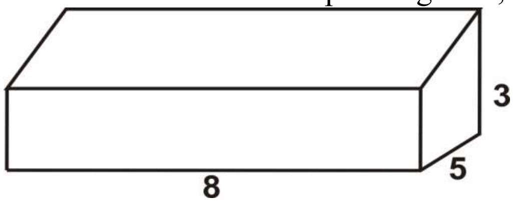


Figure 3 (a) 120 (b) 150 (c) 240 (d) 400

21. Factorize x^2-8x-20 (a) (x+10) (x+2) (b) (x-10) (x+2) (c) (x+4) (x-5) (d) (x-4) (x+5)

22. Find the quadratic equation whose roots are 3 and 4
(a) x^2-7x+12=0 (b) x^2+7x+12=0 (c) x^2+7x-12=0 (d) x^2-7x-12=0

23. Remove bracket from 3- (a-$\overline{5 - 6a}$) (a) -8-7a (b) 8+7a (c) 8-7a (d) 7a-8

24. Calculate the area in hectares of a rectangular field 126m long and 97m wide
(1 hectare = 10,000m^2) (a) 2.44 (b) 24.4 (c) 1.22 (d) 12.2

Use the following expressions in logical reasoning to answer Q25 –Q27

A: Lagos is a city in Nigeria.
B: A square is a rectangle.
C: God willing.
D: They are lovely people.
E: Some students in SS1 study physics and some don't.
F: He ran at a constant speed.
G: If you press the switch, the door will open.

25. Expression B is (a) true (b) false (c) it is not a statement (d) it is a compound statement

26. Expression D is (a) it is a command (b) it is a sentence (c) true (d) false

27. The negation of expression F is (a) He will run at a constant speed (b) He ran at a speed that is somehow constant (c) He ran at a varied speed (d) He does not mean to run at a constant speed.

Use this information to answer Q28.
For the set of positive whole numbers, let
P: X is exactly divisible by 2
Q: X is an even number.
This symbol ⇒ means imply and ~ means negation

28. Which of the following is true about the statements P and Q?
(a) $P \Rightarrow Q$ (b) $\sim P \Rightarrow Q$ (c) $\sim Q \Rightarrow P$ (d) $P \Rightarrow \sim Q$

29. Log_{10} 10000 is equal to (a) 1 (b) 2 (c) 3 (d) 4

30. If M = 314_5 and N = 24_5, calculate M+N (a) 234_5 (b) 334_5 (c) 342_5 (d)343_5

The table below gives the ages of student in SS1C who were born in April. Use the information to answer Q32 – 32.

Table 4

Age (year)	13	14	15	16	17
Number of students	8	5	6	3	8

31. The average age of the students in the class is
(a) 10 (b) 15 (c) 14 (d) 13

32. Students who fall to the range of ages 13 to 15 are
(a) 16 (b) 17 (c) 18 (d) 19

Use the following data to answer Q33 – Q34

The shoe sizes of a group of 24 students in a class are

8 6, 7, 5, 4, 6, 5, 7
6, 5, 7, 6, 8, 5, 4, 6
5, 5, 6, 7, 8, 8 6 7

33. The frequency of the shoe size 5 has the tally of (a) **III** (b) **IIII** (c) **~~IIII~~** (d) **~~IIII~~**

34. The average shoe size in the class is (a) 6 (b) 7 (c) 8 (d) 9

Use the two diagrams as applicable in figure 6a and 6b to answer Q35–36

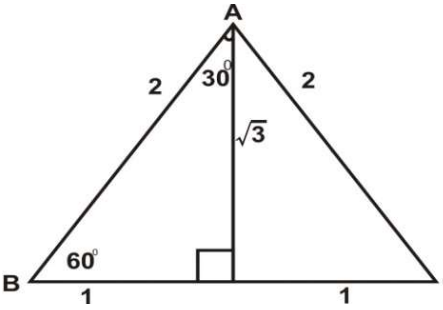


Figure 6a

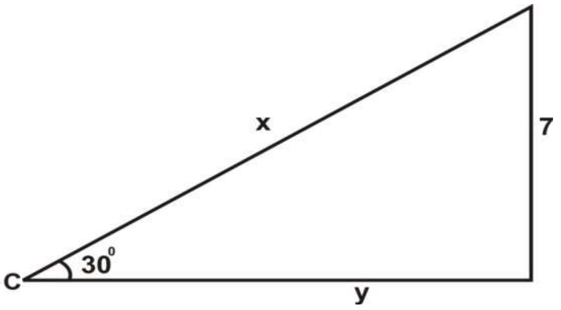


Figure 6b

35. What is the side marked x in figure 6b? (a) 12 (b) 13 (c) 14 (d) 15

36. Find sin 30^0 in figure 6b above (a) $\frac{1}{2}$ (b) $\frac{2}{3}$ (c) $\frac{3}{\sqrt{3}}$ (d) $\frac{1}{\sqrt{3}}$

37. Convert 1264eight to base ten (a) 629 (b) 692 (c) 962 (d) 296

38. Calculate the area of the shaded segment in figure 7, if $\pi = \frac{22}{7}$

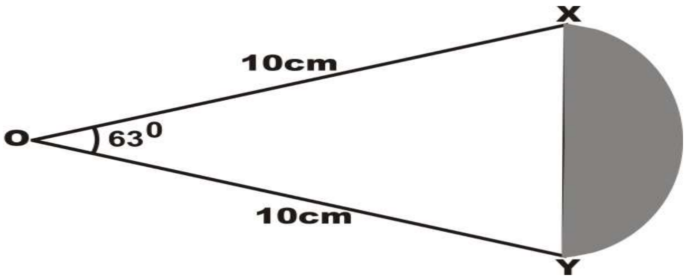


Figure 7

(b) 55cm^2 (b) 44.55cm^2 (c) 20cm^2 (d) 10.45cm^2

39. From a point P on level ground, the angle of elevation of the top of a tree is 60^0. If the tree is 39m high, how far is its base from P.

(a) $\frac{39}{\sqrt{3}}$m (b) $\frac{25}{\sqrt{3}}$ m (c) $\frac{27}{\sqrt{2}}$ m (d) $\frac{20}{\sqrt{3}}$ m

40. Simplify $\left\{\frac{8}{27}\right\}^{-2/3}$ (a) 4$^{1/4}$ (b) 3$^{1/4}$ (c) 2$^{1/4}$ (d) 1$^{1/4}$

41. The sum of the interior angles of and n-sided convex polygon is

(a) (2n+4) right angles (b) (2n-4) right angles (c) (n+2) right angles (d) (n-2) right angles

42. One of the following is true of an isosceles triangle.

(a) The base angles are equal.
(b) The equal sides do not meet at the vertex.
(c) The bisector of the vertex angle does not meet the base at right angle.
(d) The two triangles formed by the bisector are not equivalent.

43. All of these properties are true of a rectangle and a square **EXCEPT**

(a) the diagonal are equal (b) all the form angles are right angles (c) opposite sides are longer than one another (d) opposite sides are parallel

44. The angle marked u in figure 8 is (a) 124^0 (b) 100^0 (c) 56^0 (d) 28^0

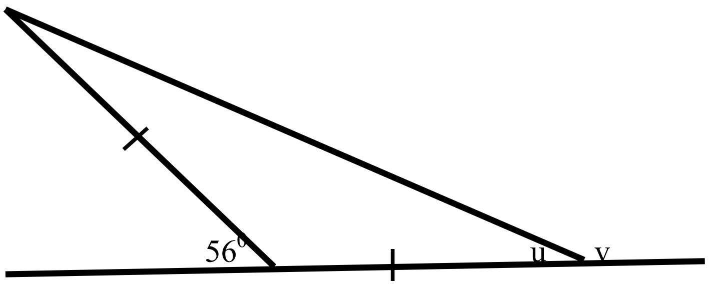


Figure 8

Table 5: Logarithm Table. Use the tables 5 and 6 to answer Q45

									Difference				
X	0	1	2	3	4	5	6	7	1	2	3	4	5
53	7243	7251	7259	7267	7275	7284	7292	7300	1	2	2	3	4

Table 6: Antilogarithm Table **Difference**

X	0	1	2	3	4	1	2	3	4
19	1549	1552	1556	1560	1563	0	1	1	1
73	5370	5383	5395	5408	5420	1	3	4	5
91	8128	8147	8166	8185	8204	2	4	6	8

45. Evaluate $\sqrt[3]{537.5}$ (a) 8130 (b) 81.30 (c) 8.13 (d) 0.813

46. On the 3rd quadrant of the Cartesian plane
(a) All ratios are positive (b) only sine is positive (c) only tangent is positive (d) only
cosine is positive

47. If x α y and x = 3 when y = 12, find the relationship between x and y
(a) x = 4y (b) 4x=$\frac{1}{y}$ (c) x = $\frac{y}{4}$ (d) x = $\frac{4}{y}$

48. Calculate the area of this trapezium.

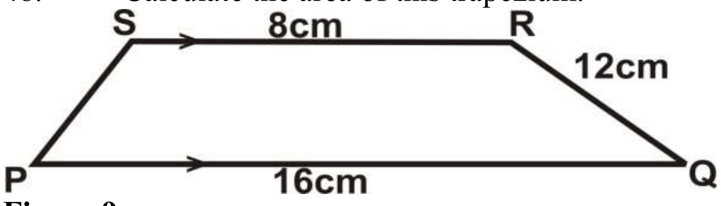


Figure 9

(a) $72cm^2$ (b) $62cm^2$ (c) $52cm^2$ (d) $42cm^2$

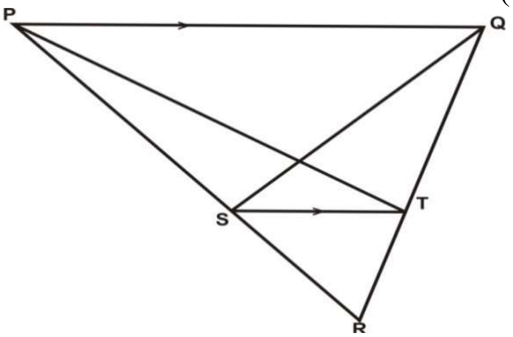


Figure 10

49. In figure 10,$\overline{ST}\|\overline{PQ}$, which triangle is equal in area to ΔPQS
(a) ΔPQT (b) ΔPQS (c) ΔQST (d) ΔSTR

Use figure 11 to answer Q50 – Q51

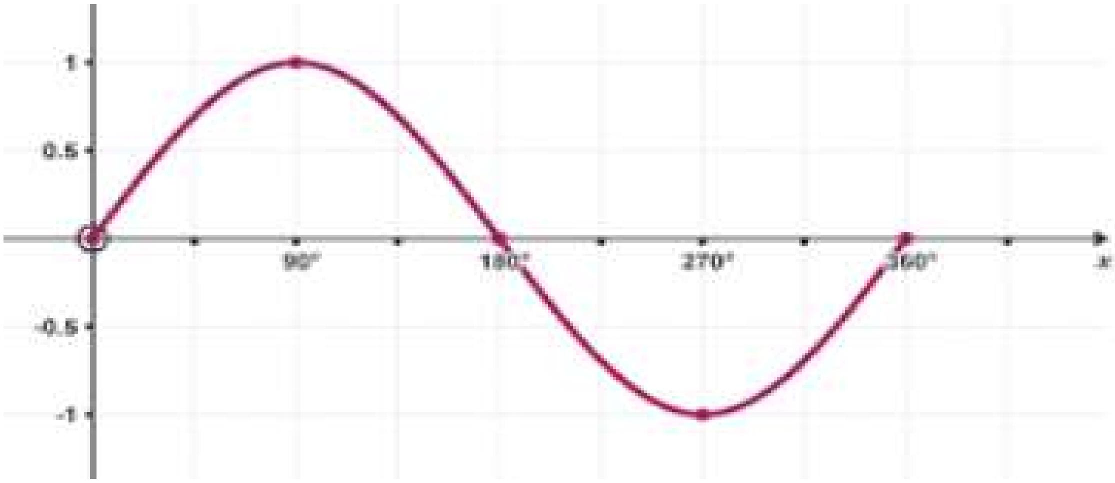


Figure 11

50. Use the graph to solve the equation 5 sin θ = 4 using the graph above, θ =
(a) 54^0 or 126^0 (b) 64^0 or 116^0 (c) 74^0 or 106 (d) 84^0 or 96^0

51. The values for sins 0^0 and 360^0 in the graph are the same, therefore the value is
(a) 0 (b) 0.5 (c) -1 (d) 1

Use the following information to answer Q52 - 53
The Pie Chart in figure 12 represents 24 hours in the life of a student

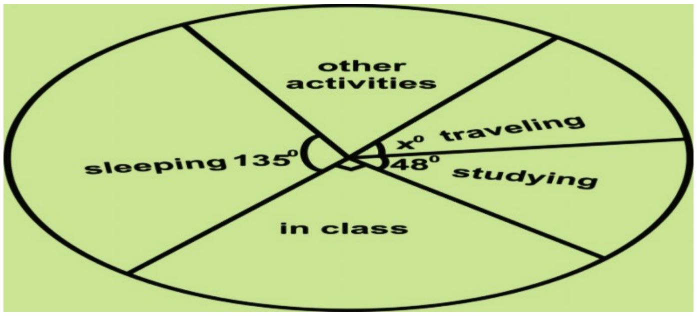


Figure 12

52. What percentage of time is spent studying?
(a) 13$\frac{1}{3}$% (b) 23 $\frac{1}{3}$ % (c) 33 $\frac{1}{3}$% (d) 43 $\frac{1}{3}$ %

53. What is the ratio of the time of studying to the time of sleeping?
(a) 14:30 (b) 15:40 (c) 16:45 (d) 20:50

The examination result of a class is given by the bar chart in figure 13. Use it to answer Q54 – 56

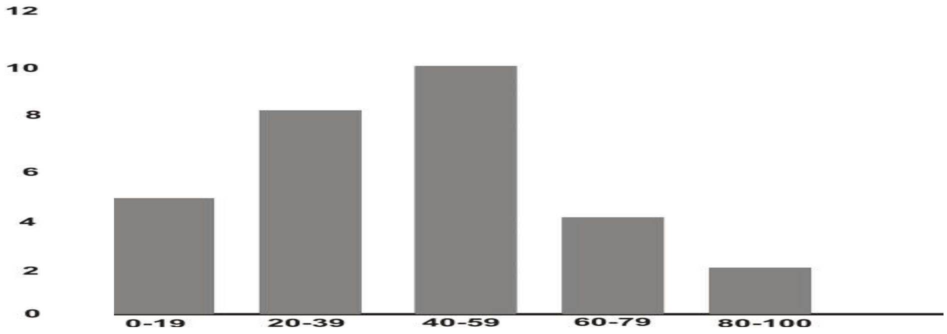


Figure 13 **Marks**

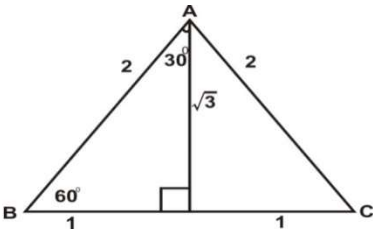
54. How many students took the examination? (a) 49 (b) 39 (c) 29 (d) 19

55. How many students failed the examination? (a) 13 (b) 14 (c) 15 (d) 16

56. Which of the range of marks is the median class?

(a) 20 – 39 (b) 40 – 59 (c) 60 – 79 (d) 80 – 100

Use figure 14 to answer Q57



57. $\cos 30^0$ is (a) $\frac{1}{2}$ (b) $\frac{2}{\sqrt{3}}$ (c) $\frac{\sqrt{3}}{2}$ (d) $\frac{1}{\sqrt{3}}$

APPENDIX V

INTERNATIONAL CENTRE FOR EDUCATIONAL EVALUATION
INSTITUTE OF EDUCATION
UNIVERSITY OF IBADAN

PART B: COMPUTER–BASED MATHEMATICS ACHIEVEMENT TEST (CBMAT)

PLEASE READ INSTRUCTION CAREFULLY. RESPONDENT IS MEANT TO ANSWER ALL THE QUESTIONS AND PICK THE RIGHT ANSWER FROM THE OPTIONS A-D. AN ANWSER SHOULD BE PROVIDED TO A QUESTION, AFTER WHICH THE BUTTON NEXT WILL BE CLICKED TO BRING THE FOLLOWING QUESTION. ALSO, THE SUBMIT BUTTON MUST BE CLICKED AFTER ATTEMPTING ALL THE QUESTIONS TO SUCCESSFULLY SUMBIT YOUR EXERCISE.

1. Out of 25 teachers, 16 are married and 15 are women, if 6 of the men are married, how many of the women are not married? (a) 15 (b) 10 (c) 5 (d) 3

2. Solve 8x = 10 (mod 3) (a) x = 0 (mod 3) (b) x = 1 (mod 3) (c) x = 2 (mod 3) (d) x = 3 (mod 3)

3. Find 567 in standard form (a) $5.67 \text{x} 10^2$ (b)$56.7 \text{x} 10^2$ (c) $567 \text{x} 10^3$ (d) $0.567 \text{x} 10^2$

4. The members of a set of even numbers less than 15 are (a) {2,3,4,6,8,11,13} (b) {2,4,6,8,10,12,14} (c) {3,4,6,8,9,10,14} (d) {2,4,6,8,9,10,14}

5. Evaluate $\sin 137^0$ (a) + $\sin 43^0$ (b) – $\text{Sin } 43^0$ (c) + $\text{Cos } 43^0$ (d) – $\text{Cos } 43^0$

6. What is M÷N in base 5, if M = 314_5 and N = 24_5 (a) 11_5 (b) 13_5 (c) 15_5 (d) 17_5

Use the tables 1 and 2 to answer Q7

Table : Logarithm table **Difference**

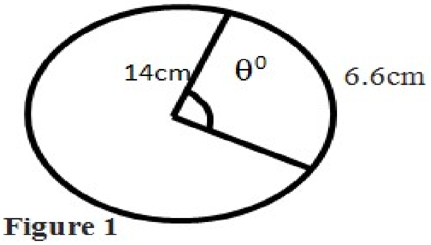
X	0	1	2	3	4	5	6	7	1	2	3	4	5
53	7243	7251	7259	7267	7275	7284	7292	7300	1	2	2	3	4

Table 2: Antilogarithm table **Difference**

X	0	1	2	3	4	1	2	3	4
19	1549	1552	1556	1560	1563	0	1	1	1
73	5370	5383	5395	5408	5420	1	3	4	5
91	8128	8147	8166	8185	8204	2	4	6	8

7. Evaluate 53.75^3 (a) 15.53 (b) 1553 (c) 15530 (d) 155300

8. What angle does an arc 6.6cm in length subtend at the centre of a circle of radius 14cm? use $\pi = \frac{22}{7}$



(a) 8^0 (b) 15^0 (c) 18^0 (d) 27^0

Figure 1

9. Find the square of 111_2 (a) 110001 (b) 100011 (c) 111000 (d) 000111

10. A car's petrol tank is 0.8m long, 25cm wide and 20cm deep. How many litres of petrol can it hold? (a) 4000 litres (b) 400 litres (c) 40 litres (d) 4 litres

11. Calculate 12○ 5 in modulo 4 (a) 3(mod 4) (b) 4(mod 4) (c) 5(mod 4 (d) 6(mod 4)

12. The exterior angle of a triangle is equal to the sum of the opposite interior angles
(a) None of the option (b) All of the options (c) True (d) False

13. Reduce $\left\{\frac{8}{50}\right\}^{-1/2}$ to its lowest term (a) $\frac{4}{25}$ (b) $\frac{25}{4}$ (c) $\frac{5}{4}$ (d) $2\frac{1}{2}$

Use this information to answer Q14-15
Given that μ = {a,b,c,d,e,f}, X = {a,b,c,d} Y = {c,d,e} and Z = {b,d,f}

14. The set XUY is (a) {a,b,c,d,e,f} (b) {a,b,c,d} (c) {a,b,c,d,e} (d) {a,b,c,e,f}

15. ZUØ is (a) {a,b,d} (b) {b,d,f} (c) {c,d,e} (d) {a,b,c,}

Given this Venn diagram below to answer Q16

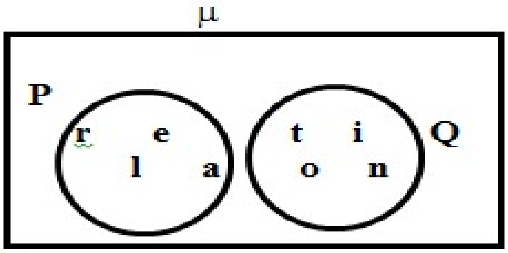


Figure 2

16. The set PUQ is (a) a proper subset of P (b) the universe set (c) an empty set (d) an intersection set

Given a Universal Set μ= {1,2,3,4,5}, A = {1,3} and B = {3,4}. Use this information to answer Q17

17. The set (A∩B)′ is (a) {1,2,3} (b) {4,5} (c) {1,3,4,5} (d) {1,2,4,5}

18. Find -2(mod 9) in its simplest form (a) 4 (mod 9) (b) 5 (mod 9) (c) 6 (mod 9) (d) 7 (mod 9)

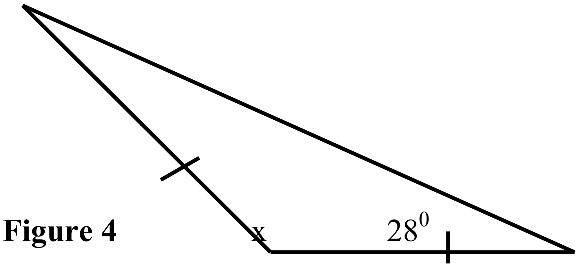
19. Make x the subject of the equation a = $\frac{b+x}{b-x}$ (a) $\frac{a(b-1)}{a+1}$ (b) $\frac{a(b-1)}{a-1}$ (c) $\frac{b(a+1)}{a-1}$ (d) $\frac{b(a-1)}{a+1}$

20. When travelling between two towns, the time taken varies inversely with the average speed. When the average speed is 42km/h, the journey takes 4hours. Find the average speed if the journey takes 2hours 20minutes.
(a) 60km/h (b) 65km/h c) 72km/h (d) 79km/h

21. Evaluate 2 ⊗ 2 in modulo 4 (a) 0 (mod 4) (b) 1(mod 4) (c) 2(mod 4) (d) 3 (mod 4)

22. The perimeter of a rectangle is 20m and the length is xm. Find the area of the rectangle in terms of x (a) 10(x+5) (b) x^2(x-10 (c) x(10-x) (d) x^2(10+x)

23. Calculate the angle marked with x in figure 4



(a) 28^0 (b) 56^0 (c) 100^0 (d) 124^0

Figure 4

Use the following expressions in logical reasoning to answer Q48 -Q54

A: Lagos is a city in Nigeria.
B: A square is a rectangle.
C: God willing.
D: They are lovely people.
E: Some students in SS1 study physics and some don't.
F: He ran at a constant speed.
G: If you press the switch, the door will open.

24. Which of the following is right of an expression A (a) it is a command (b) it is not true (c) it is a statement (d) it is a negative statement

25. Expression C is (a) true (b) false (c) it is not a statement (d) it is a sentence

26. Expression E is (a) an implication statement (b) a conditional statement (c) a

compound statement (d) an equivalence statement

27. The equivalence of expression G is (a) if you do not press the switch, the door will not open (b) if you press the switch, the door will not open (c) if you do not press the switch, the door will open (d) if the door is not open, you should press the switch

Use this information to answer Q28.
For the set of positive whole numbers, let
P: X is exactly divisible by 2
Q: X is an even number.
This symbol ⇒ means imply and ~ means negation

28. Which of the following is true about statement P and Q
(a) disjoint statement (b) conjoint statement (c) equivalent statement (d) compound statement

Given the figure below;

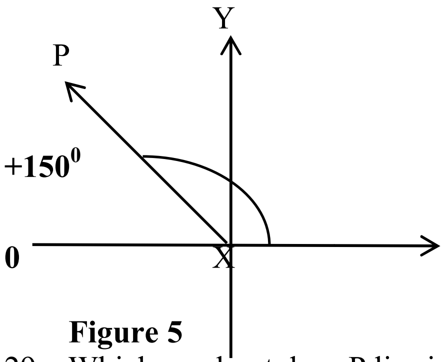


Figure 5

29. Which quadrant does P lies in, given that the angle between $\overline{OP}$ and $\overline{OX}$ is $+150^0$?
(a) 1st quadrant (b) 2nd quadrant (c) 3rd quadrant (d) 4th quadrant

The table below gives the ages of student in SS1C who were born in April. Use the information to answer Q30 – 31.

Table 4

Age (year)	13	14	15	16	17
Number of students	8	5	6	3	8

30. How many students were born in April?
(a) 10 (b) 20 (c) 30 (d) 40

31. The median age of the students is
(a) 16 (b) 15 (c) 14 (d) 13

32. If $\sin\theta = \frac{5}{13}$, find the value of $\frac{1+\tan\theta}{1-\tan\theta}$

(a) $\frac{17}{12}$(b) $\frac{5}{12}$ (c) $\frac{15}{7}$ (d) $\frac{17}{7}$

Use the following data to answer Q33

The shoe sizes of a group of 24 students in a class are

8 6, 7, 5, 4, 6, 5, 7

6, 5, 7, 6, 8, 5, 4, 6

5, 5, 6, 7, 8, 8 6 7

33. The tally **~~HHH~~ II** is the frequency of the shoe size that appear most among the group. This shoe size is (a) 5 (b) 6 (c) 7 (d) 8

34. In a right–angled triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides (a) true (b) false (c) none of the option (d) All of the option.

Use the two diagrams as applicable in figure 6a and 6b to answer Q35

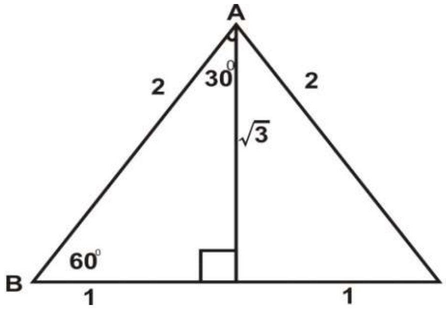


Figure 6a

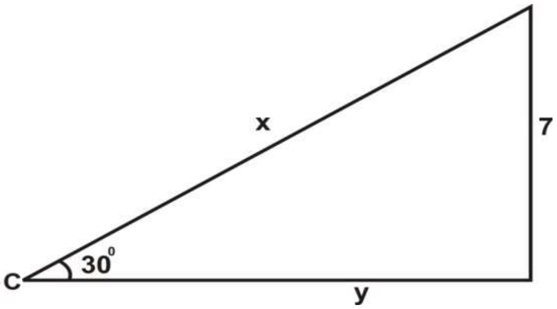


Figure 6b

35. What is the side marked y? (a) $6\sqrt{3}$ (b) $7\sqrt{3}$ (c) $5\sqrt{3}$ (d) $2\sqrt{3}$

36. An isosceles triangle is such that one of the base angles is twice the third angle. Find the value of one of its base angles. (a) 72^0 (b) 60^0 (c) 45^0 (d) 36^0

37. Calculate the perimeter of a sector of a circle of radius 7cm, the angle of the sector being 108^0. If $\pi = \frac{22}{7}$ (a) 13.2cm (b) 17.4cm (c) 22.3cm (d) 27.2cm

38. Find 2$\oplus$3 in modulo 4 (a) 2 (mod 4) (b) 3 (mod 4) (c) 4 (mod 4) (d) 5(mod 4)

39. The chord $\overline{AB}$ of a circle whose centre 0 is 10cm long, and A$\hat{O}$B = 140^0. Calculate the radius of the circle (a) 3.42cm (b) 4.24cm (c) 5.32cm (d) 5.87cm

40. If the volume of a rectangular-based pyramid is $70cm^3$ and its base area is $28cm^2$, calculate the height of the pyramid. (a) 10.5cm (b) 9.5cm (c) 8.5cm (d) 7.5cm

41. Two triangles are congruent, if

(a) two sides and the included angle of one are respectively equal to two sides and the included angle of the other (SAS).

(b) two angles and a side of one are respectively equal to two angles and the corresponding side of the other (ASA).

(c) the three sides of one are respectively equal to the three sides of the other (SSS).

(d) all of the above hold.

42. All these are quadrilaterals **EXCEPT** (a) rhombus (b) trapezium (c) cylinder (d) flying kite

43. The age of a boy is x years while his father's is four times as old as he is. Find the father's age in y years' time (a) (4x-y) (b) (4x+y) (c) (y-4x) (d) (-4x-y)

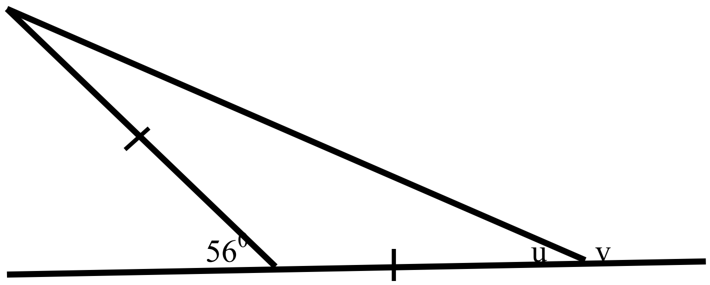


Figure 8

44. The angle marked v is in figure 8 is (a) 156^0 (b) 152^0 (c) 124^0 (d) 116^0

45. If you are to construct angle 30^0, which of the following angles should bisection be made (a) 90^0 (b) 60^0 (c) 45^0 (d) 30^0

46. Calculate this bicimal, 11.01 - 1.11 (a) 11.01 (b) 11.1 (c) 1.10 (d) 0.110

47. Calculate the total surface area of a solid cylinder of radius 3cm and height 4cm in terms of π (a) $62\pi cm^3$ (b) $52\pi cm^2$ (c) $42\pi cm^3$ (d) $32\pi cm^3$

48. Solve $8 \cos \theta - 1 = 0$ (a) 56.4^0, 123.6^0 (b) 82.8^0, 277.2^0 (c) 115.4^3, 244.6^0 (d) 26.4^0, 153.6^0

Use figure 11 to answer Q49 – Q50

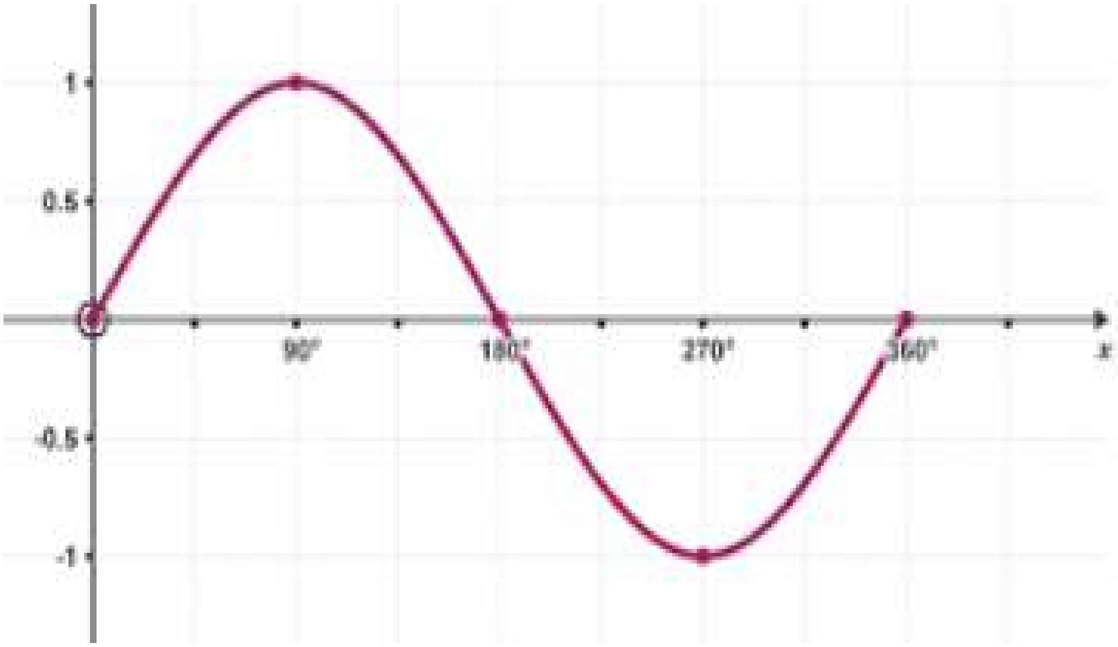


Figure 11

49. Find the value of sin 210^0 (a) 0.5 (b) -0.5 (c) 1.0 (d) -1

50. Find the value of sin 270^0 (a) 1 (b) 0.5 (c) -1 (d) -0.5

Use the following information to answer Q51 - 53

The Pie Chart in figure 12 represents 24 hours in the life of a student

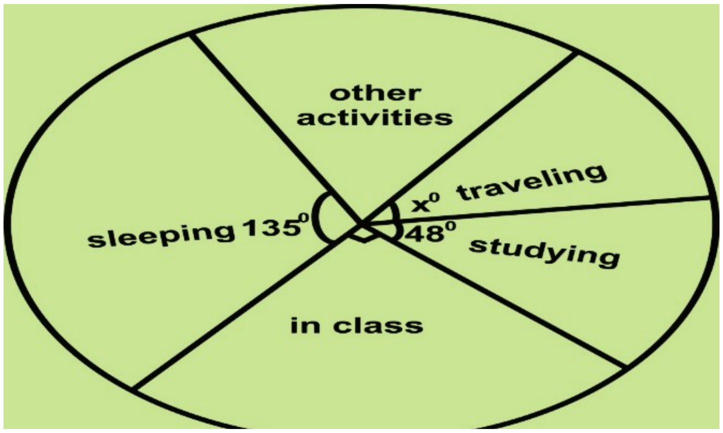


Figure 12

51. What fraction of the time is spent sleeping

(a) $\frac{1}{8}$(b) $\frac{1}{4}$ (c)$\frac{3}{8}$ (d) $\frac{1}{2}$

52. How much time is spent studying?
(a) 4hrs 12min (b) 3hrs 12min (c) 2hrs 15mins (d) 1hr 15mins

53. In its simplest form, what fraction of time is the student spending in class
(a) $\frac{1}{4}$(b) $\frac{1}{3}$ (c)$\frac{1}{2}$ (d) $\frac{1}{5}$

The examination result of a class is given by the bar chart in figure 13. Use it to answer Q54 – 55

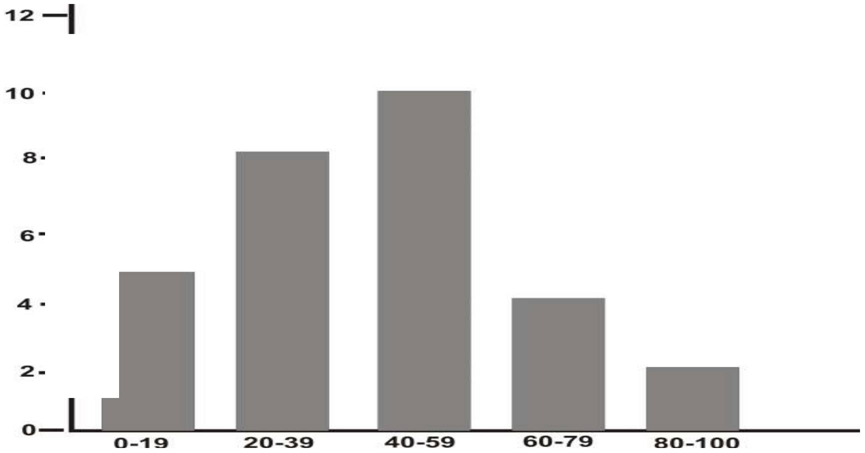


Figure 13 **Marks**

54. If the pass mark is 40, how many students passed the examination?
(a) 20 (b) 18 (c) 16 (d) 14

55. Which of the range of marks is the modal class? (a) 20 – 39 (b) 40 – 59 (c) 60 – 79(d) 80 – 100

Use figure 14 to answer Q56 – Q57

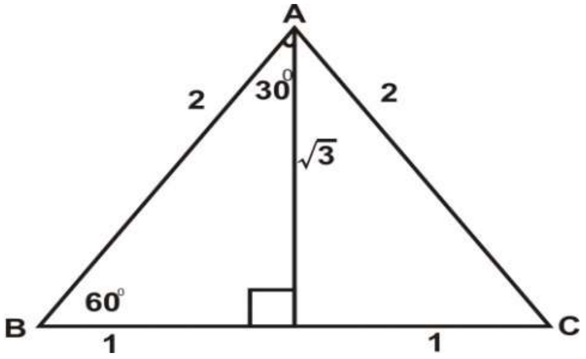


Figure 14

56. From the triangle above, sin60^0 is (a) $\frac{1}{2}$(b) $\frac{2}{\sqrt{3}}$ (c) $\frac{\sqrt{3}}{2}$ (d) $\frac{1}{\sqrt{3}}$

57. Tan 30^0 is (a) $\frac{1}{2}$ (b) $\frac{2}{\sqrt{3}}$ (c) $\frac{\sqrt{3}}{2}$ (d) $\frac{1}{\sqrt{3}}$

APPENDIX VI

9th July, 2018

The Honourable Commissioner,
Ministry of Education, Science and Technology,
The Secretariat, Ibadan.
Oyo State.

Dear Sir/Ma,

LETTER OF INTRODUCTION - LAWAL R. Omonike (MATRIC NO: 125511)

I hereby wish to introduce the bearer who is a Post-Graduate Student (Ph.D) in the Institute of Education, University of Ibadan. Her research work is based on Assessment/Testing in Education which is targeted towards improving students' present performances both in school-based and external examinations in secondary schools.

She has been authorised to collect data on the research topic **"Application of 4-Parameter Logistic and Response-Time IRT Models in the Calibration of Senior Secondary School Computer-Based Mathematics Test in Southwest Nigeria.**

In view of this, I hereby solicit your kind support to ensure that relevant information is collected from your office to grant her access to the appropriate schools. I confirm that the data so collected will be treated with utmost confidentially and used basically for research purpose.

Kindly accord her the necessary assistance.

Yours Faithfully,

Prof. J. G. Adewale
Head of Unit, ICEE,
08033263534

APPENDIX VII

17th October, 2018.

Office of the Head of Service,
The Secretariat, Ibadan.
Oyo State.

Dear Sir/Ma,

LETTER OF INTRODUCTION - LAWAL R. Omonike (MATRIC NO: 125511)

I hereby wish to introduce the bearer who is a Post-Graduate Student (Ph.D) in the Institute of Education, University of Ibadan. Her research work is based on Assessment/Testing in Education which is targeted towards improving students' present level of performance both at school-based and external examinations levels.

She has been authorised to collect data on the research topic **"Application of 4-Parameter Logistic and Response-Time IRT Models in the Calibration of Senior Secondary School Computer-Based Mathematics Test in Southwest Nigeria.**

Due to the nature of her research work that involves the usage of Computer-Based Testing, the researcher will like to access Senior Secondary Schools with Computer Laboratories.

In view of this, I hereby solicit your kind support to ensure that approval is granted from your honourable office to access all the schools she intends collecting data for the study. I confirm that the data so collected will be treated with utmost confidentiality and used for research purpose only.

Kindly accord her the necessary assistance.

Yours Faithfully,

Prof. J. G. Adewale
Head of Unit, ICEE,
Institute of Education.
08033263534

APPENDIX VIII

International Centre for Educational Evaluation,
Institute of Education,
University of Ibadan.
17th October, 2018.

Office of the Head of Service,
The Secretariat, Ibadan.
Oyo State.

Dear Sir/Ma,

Application for Approval to Access Schools

I, Lawal Omonike, a postgraduate student (Ph.D) of the Institute of Education, University of Ibadan with matric number (125511) hereby write to seek approval into accessing government-owned senior secondary schools with computer laboratories/centres in Education District I Lagos State.

My research topic is **"Application of 4-Parameter Logistic and Response-Time IRT Models in the Calibration of Senior Secondary School Computer-Based Mathematics Test in Southwest Nigeria**. Schools with computers are the media for the collection of data because of the mode of assessment delivery that is mainly computer-based while the subject of target and class to be used is mathematics and SSII students respectively.

The purpose of this study is to further positively enhance students' present academic performances at the end of the term, session and ultimately at WAEC level using a more robust approach in constructing and analysing assessment items (questions) and instrument. The researcher will as well be analysing students' response time that will automatically be recorded while responding on the computers. Capturing these variables (Responses and Response-Time) will enable the researcher to analyse students' performances for a more appropriate placement and what their abilities can carry in their further academic studies.

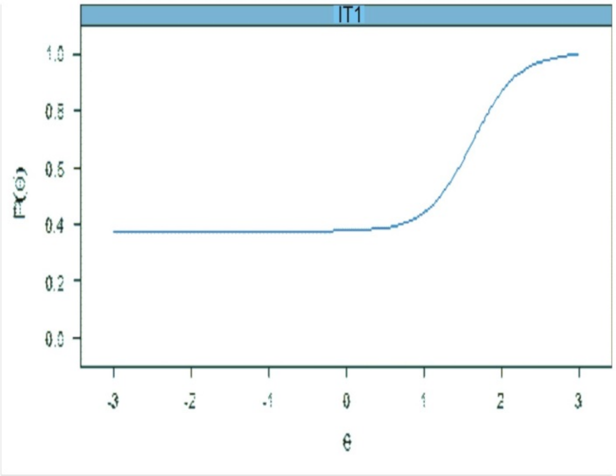
Meanwhile, the researcher intends to furnish this honourable office and the Ministry of Education with the findings of the research and recommendations as appropriate at the end of the research work if granted this opportunity.

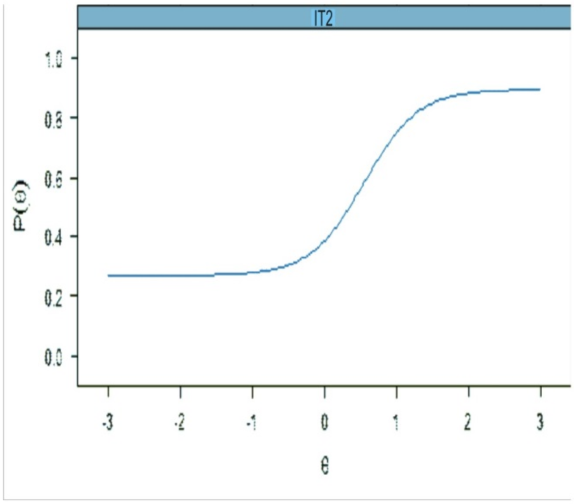
Looking forward to your favourable assistance Sir/Ma.

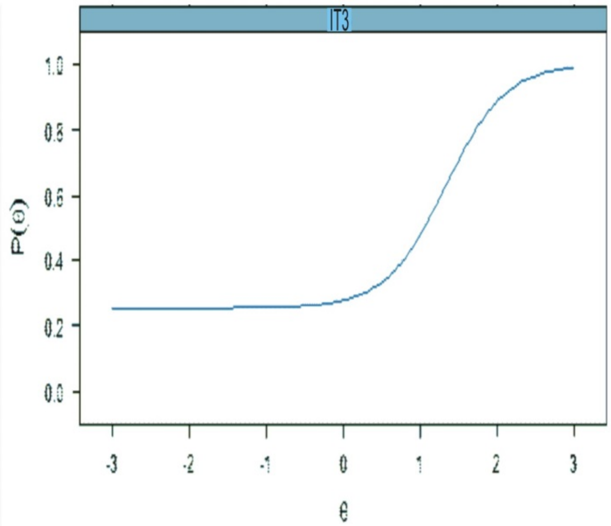
Yours Faithfully.

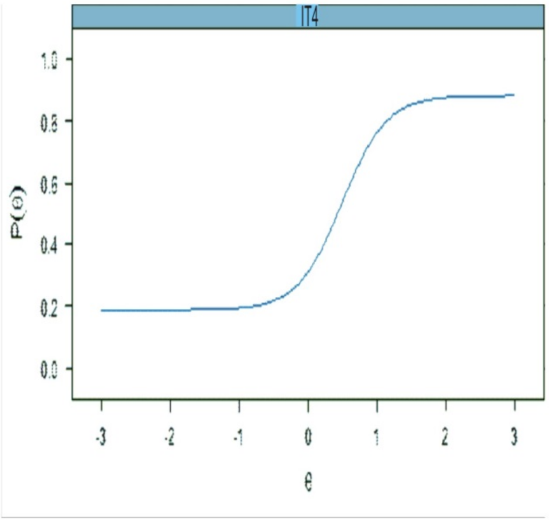
Lawal R. Omonike
08067495500

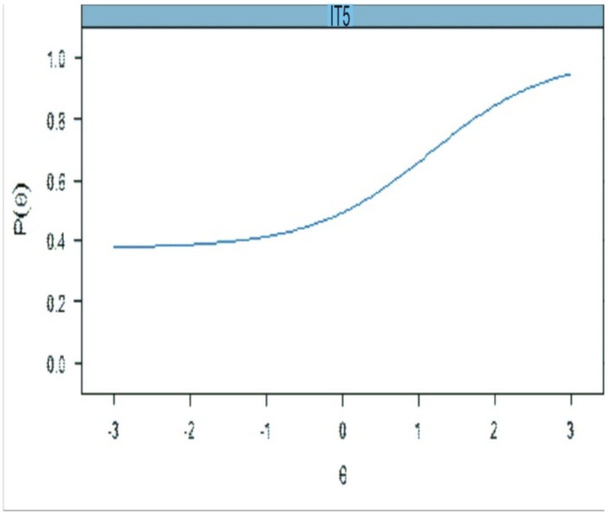
APPENDIX IX (ITEM CHARACTERISTICS CURVES, ICCs)

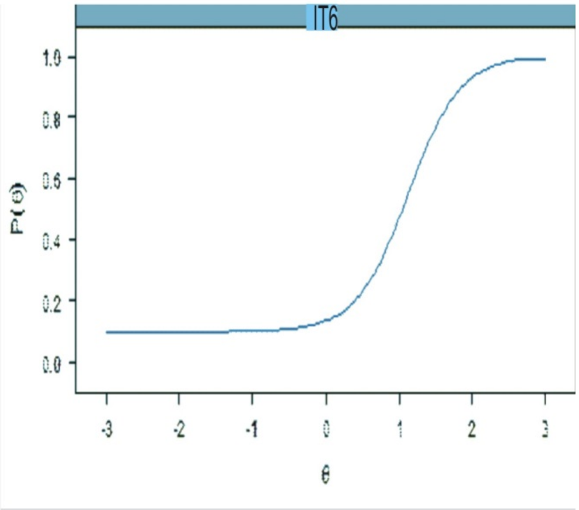


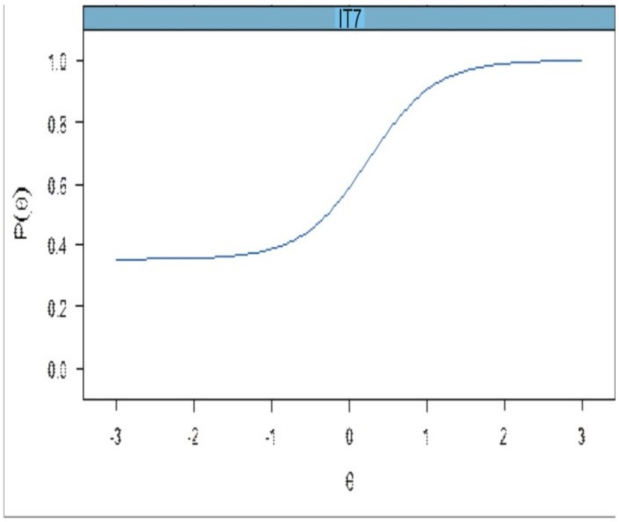


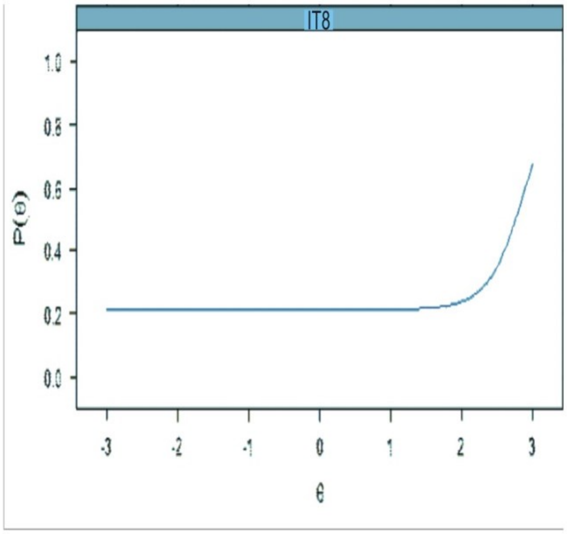


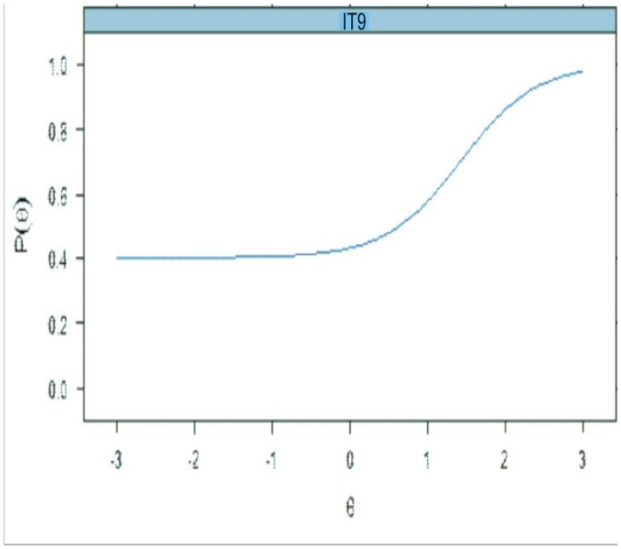


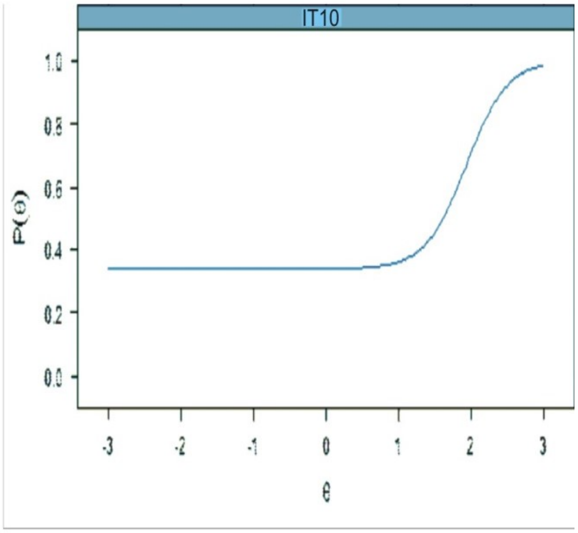


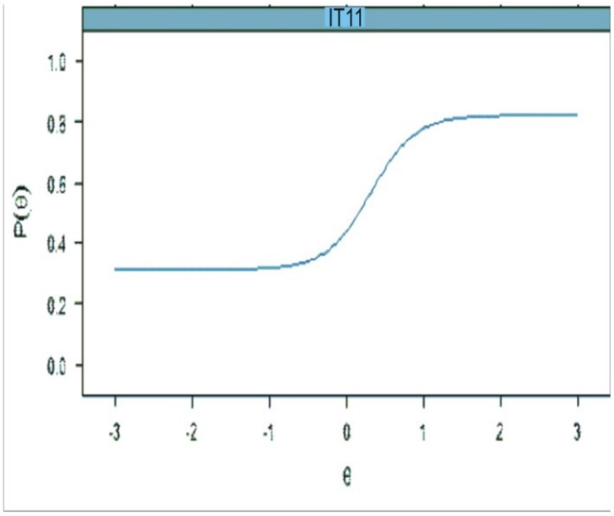


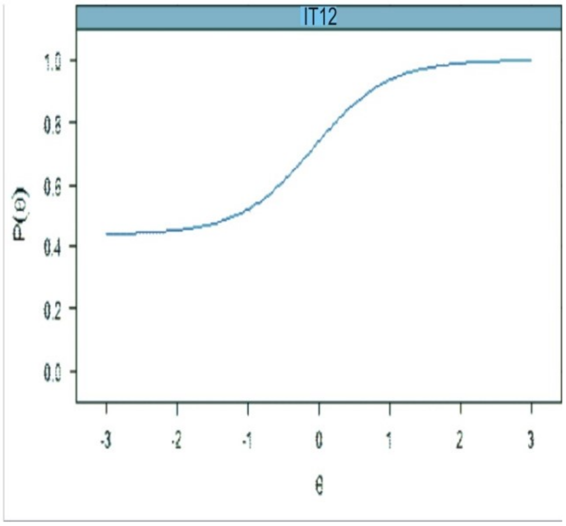

IT7
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

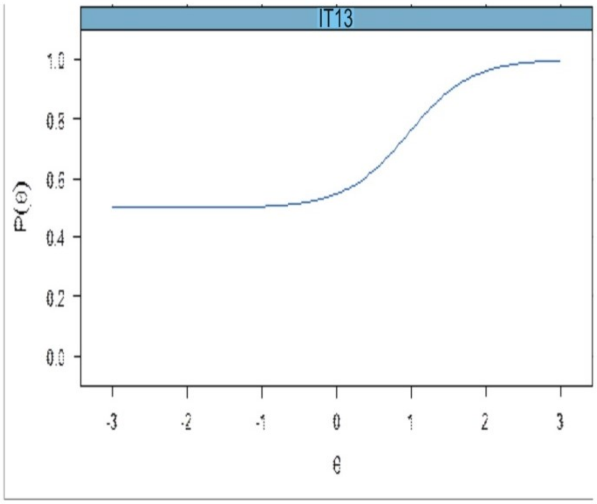

IT8
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

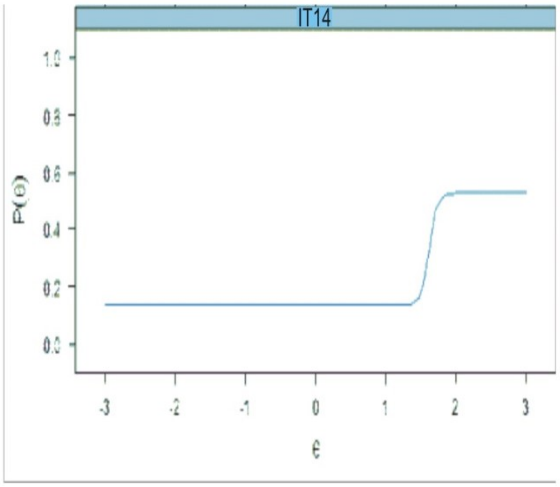

IT9
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

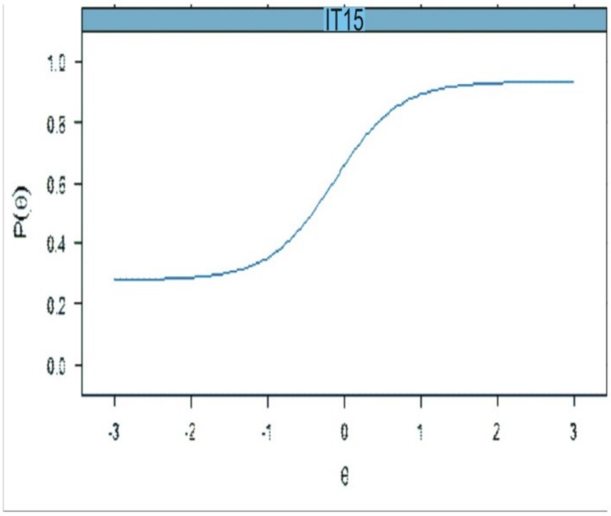

IT10
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

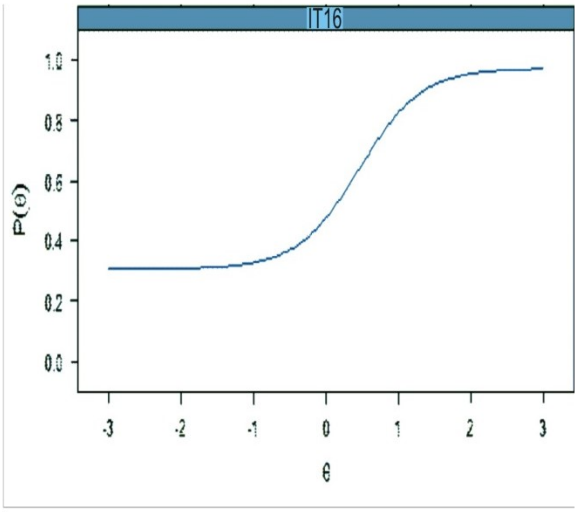

IT11
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

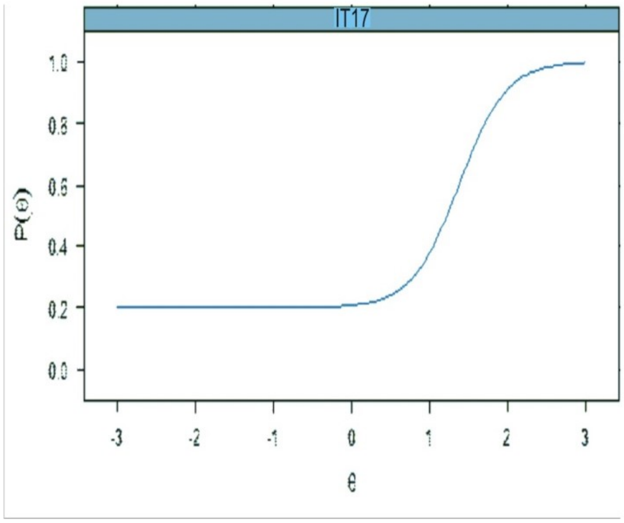

IT12
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

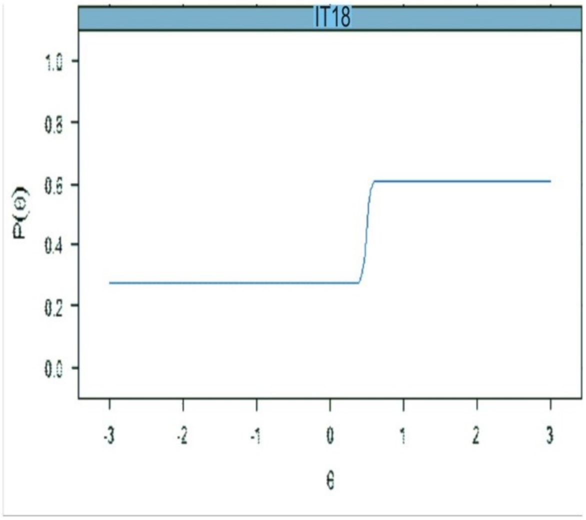

IT13
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

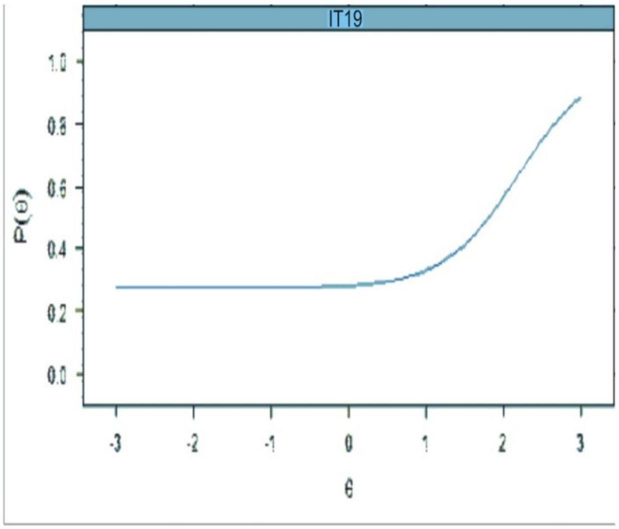

IT14
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

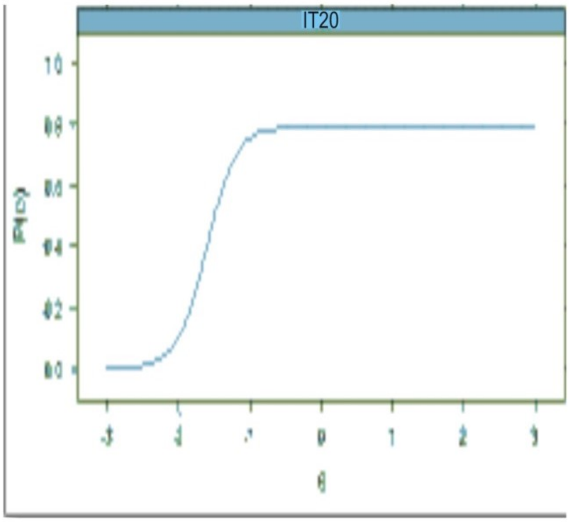

IT15
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

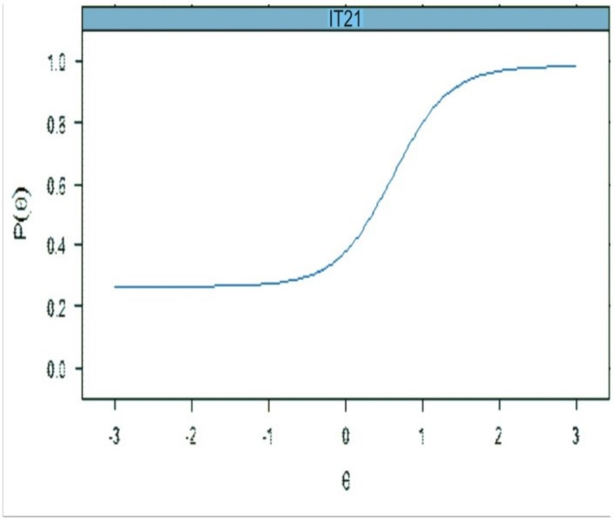

IT16
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

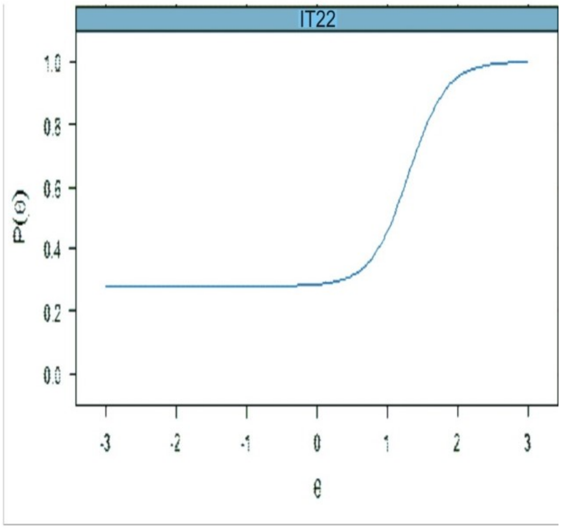

IT17
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

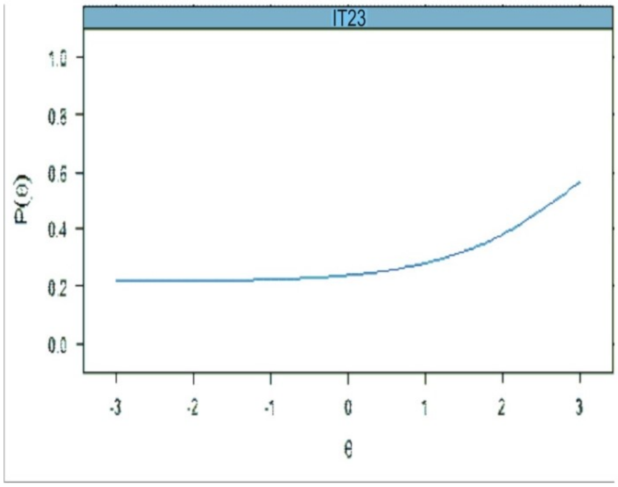

IT18
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

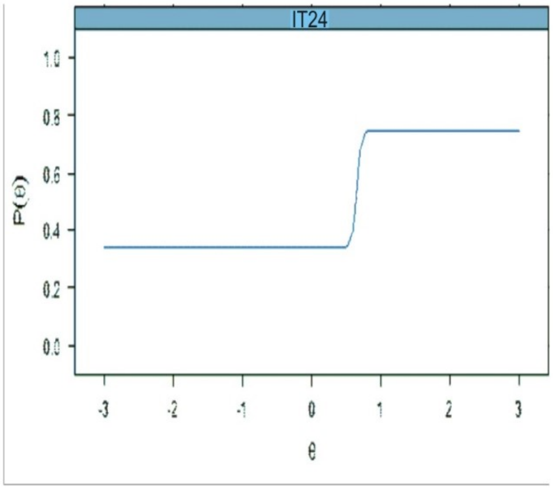

IT19
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

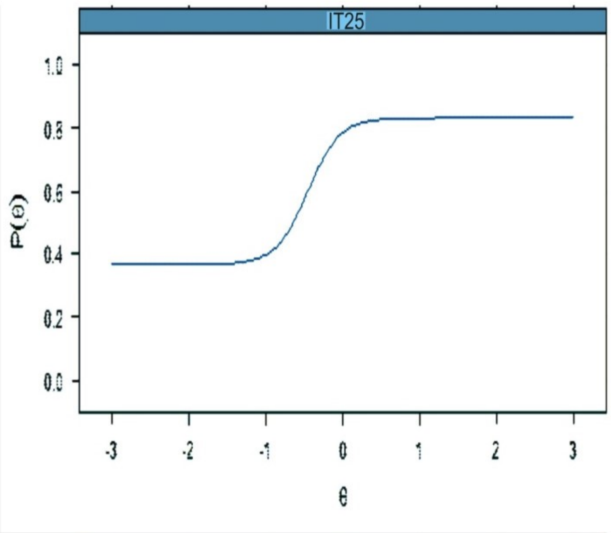

IT20
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

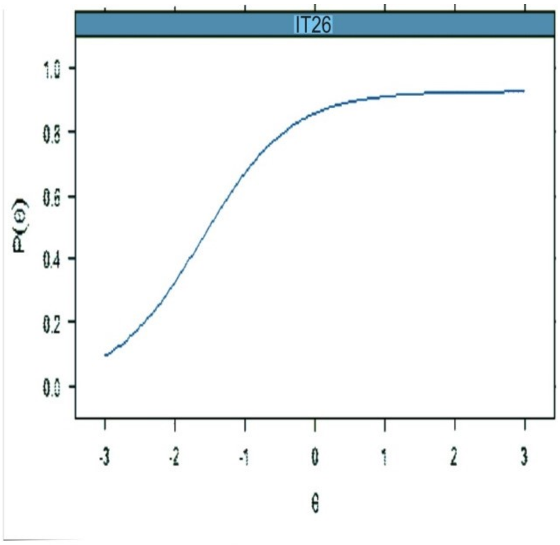

IT21
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

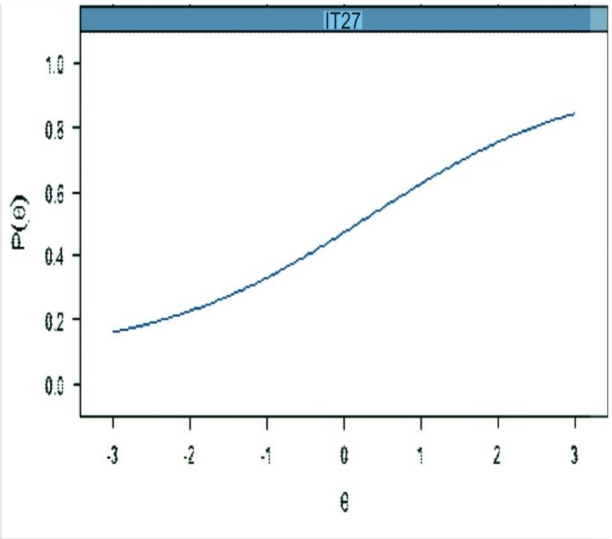

IT22
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

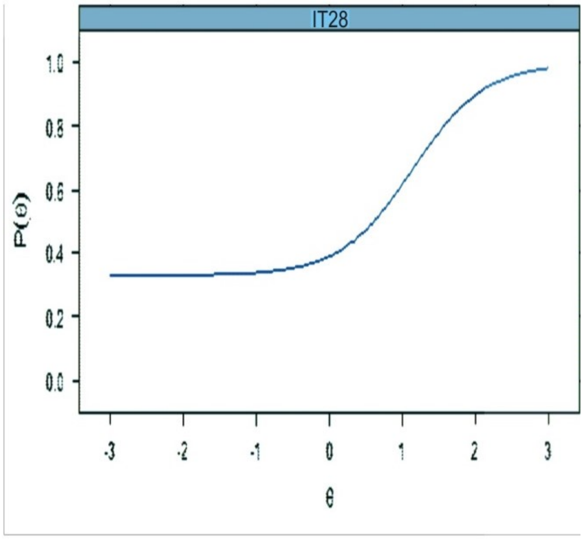

IT23
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

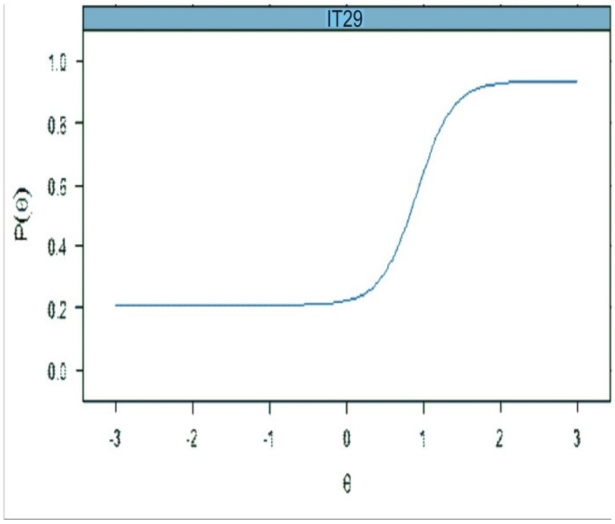

IT24
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

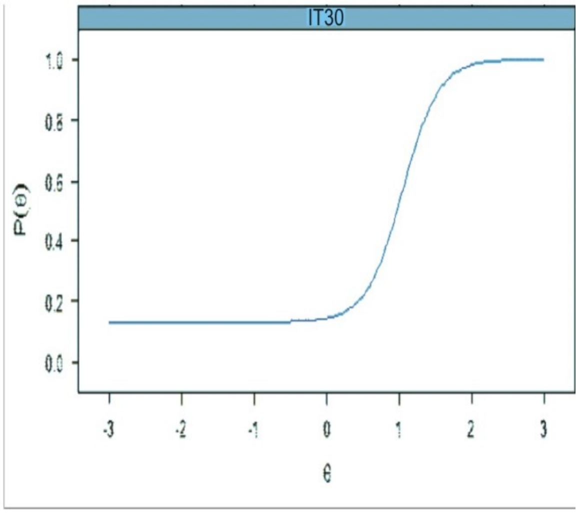

IT25
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

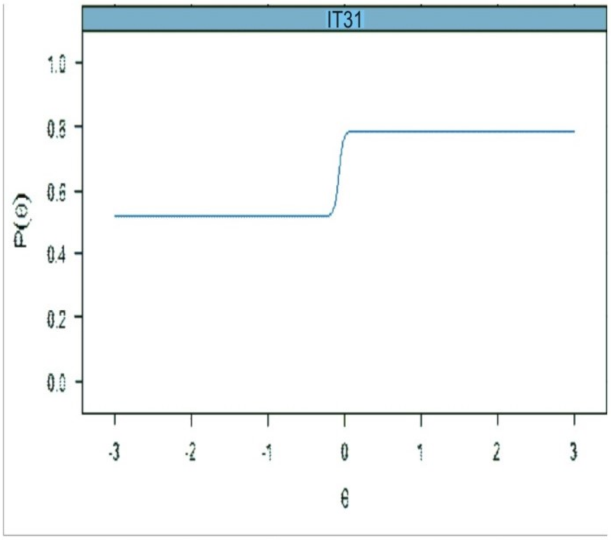

IT26
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

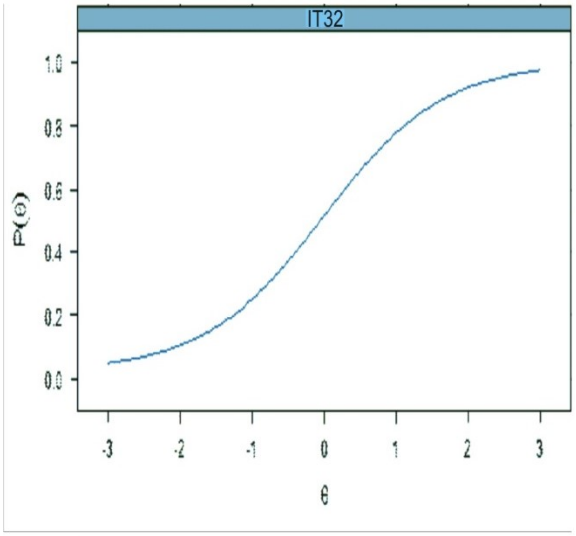

IT27
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

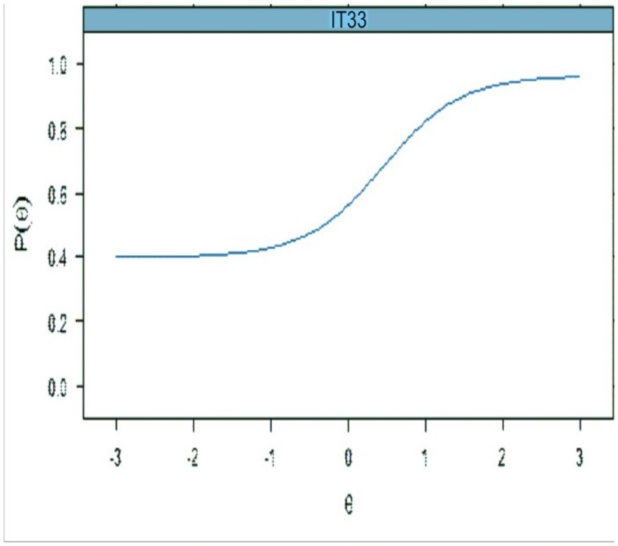

IT28
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

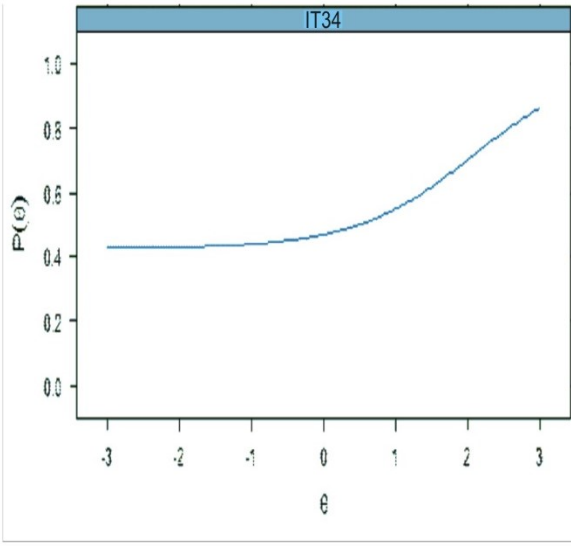

IT29
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

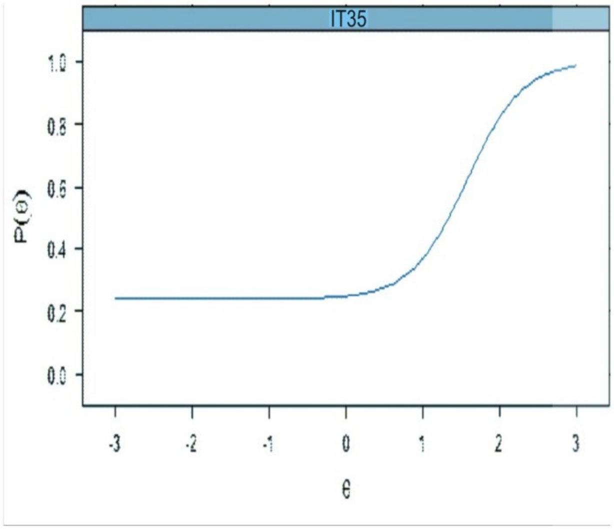

IT30
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

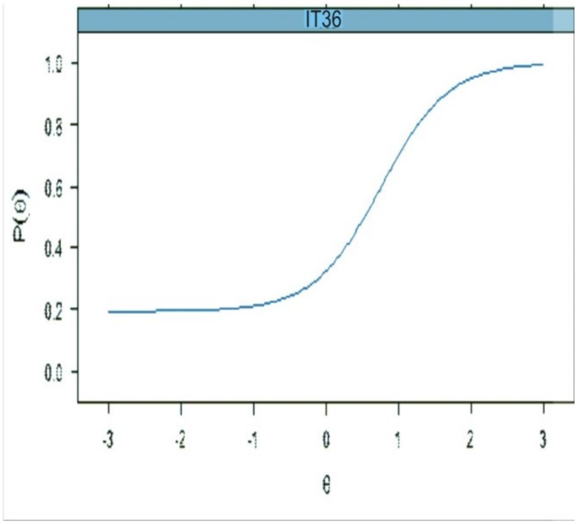

IT31
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

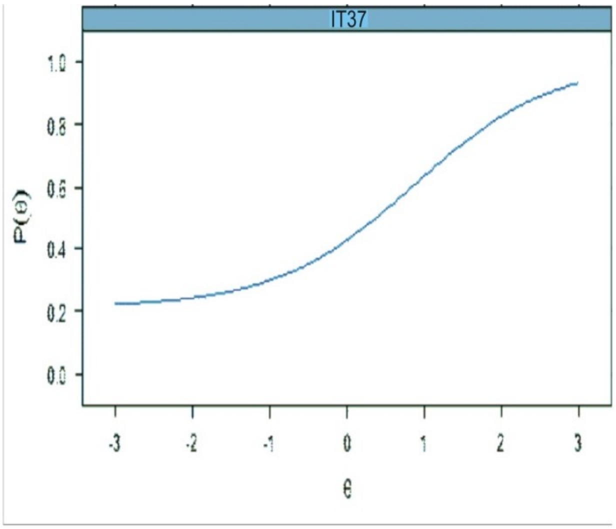

IT32
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

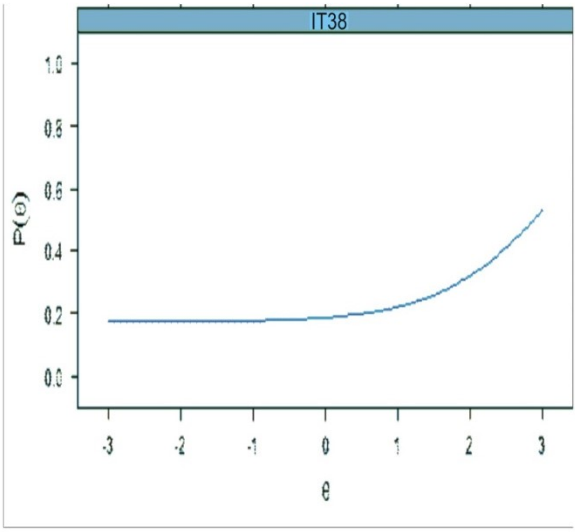

IT33
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

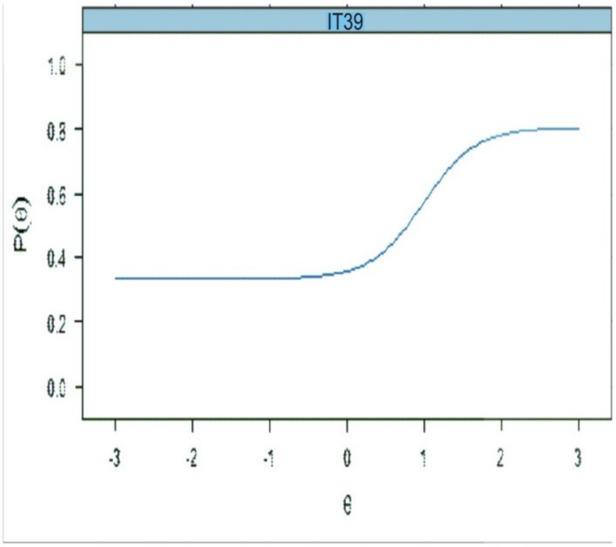

IT34
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

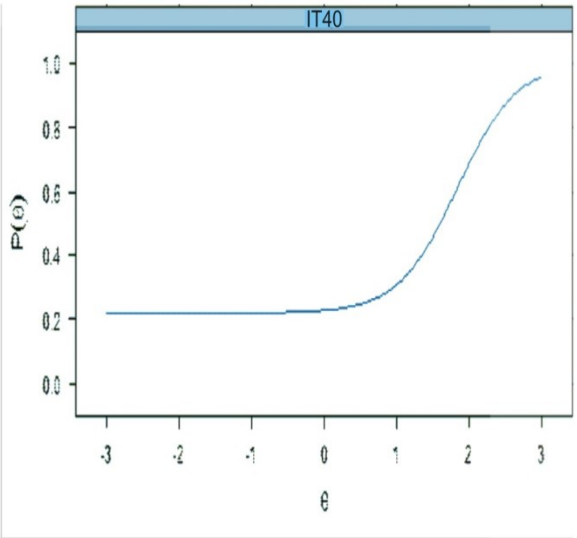

IT35
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

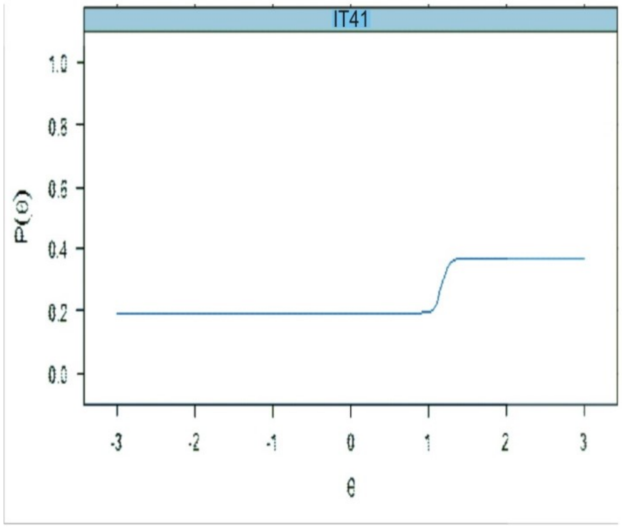

IT36
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

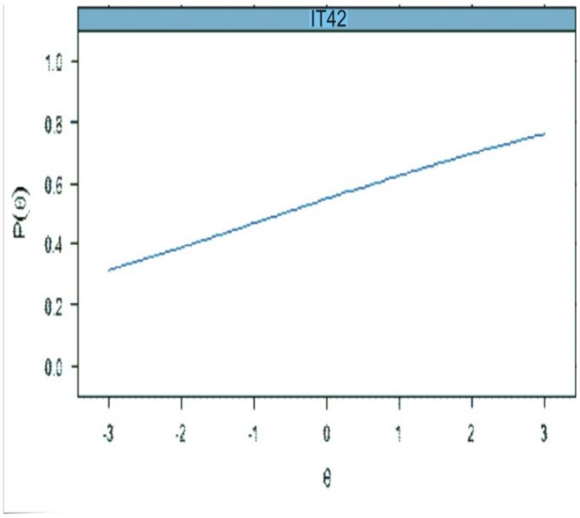

IT37
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

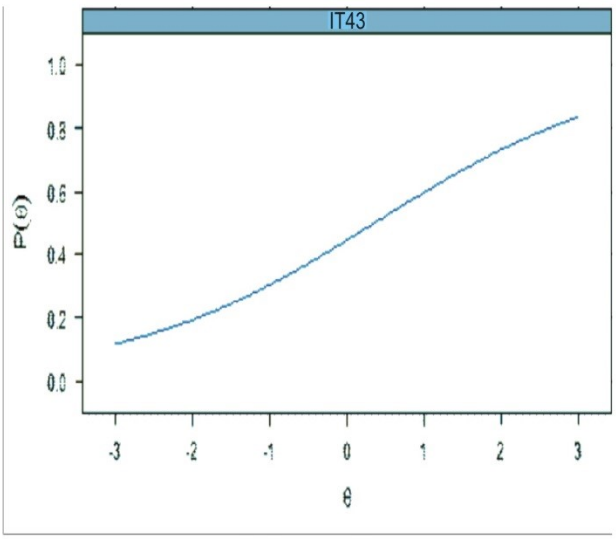

IT38
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

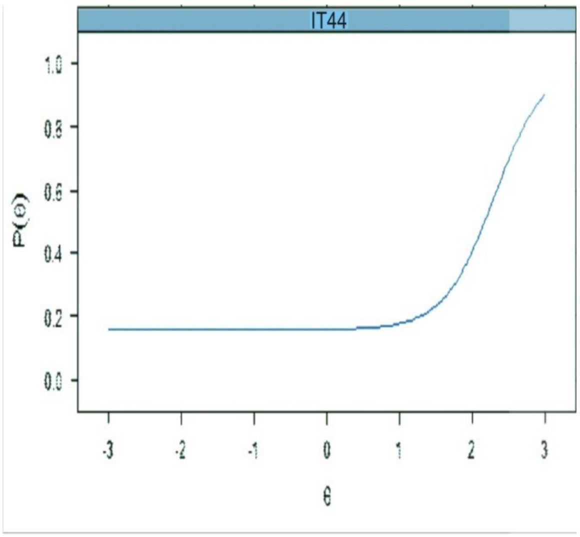

IT39
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

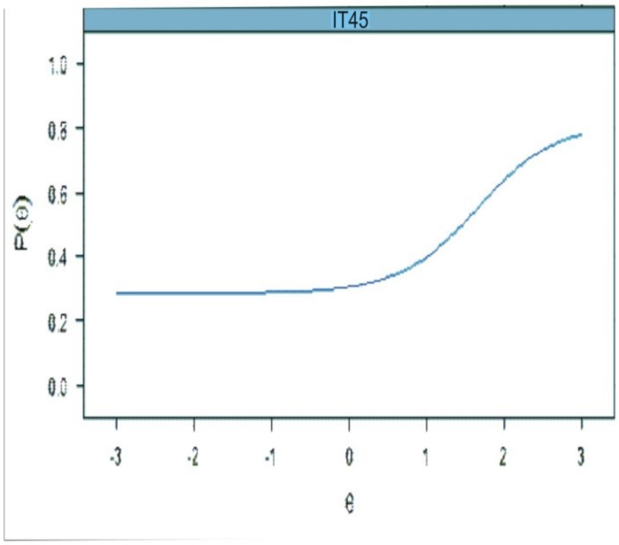

IT40
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

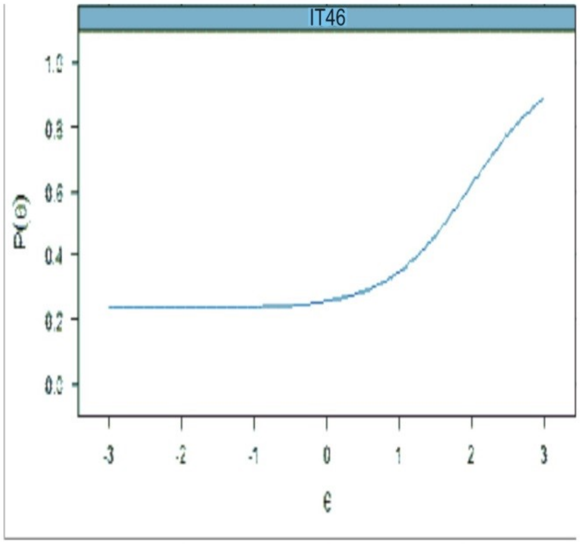

IT41
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

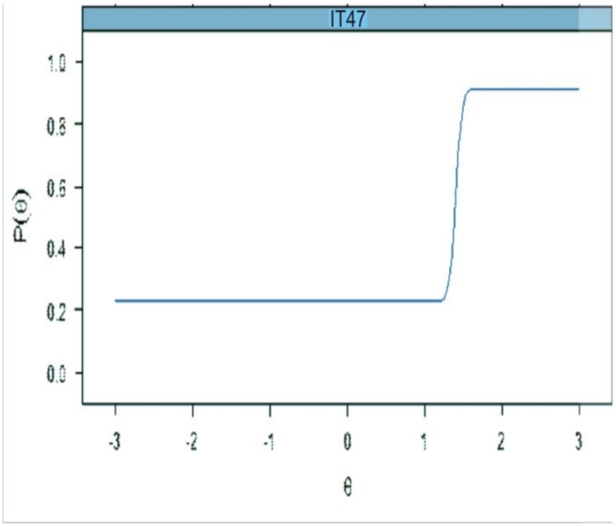

IT42
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

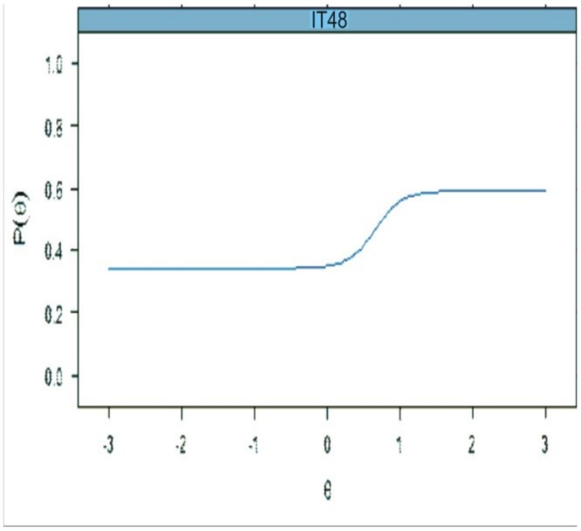

IT43
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

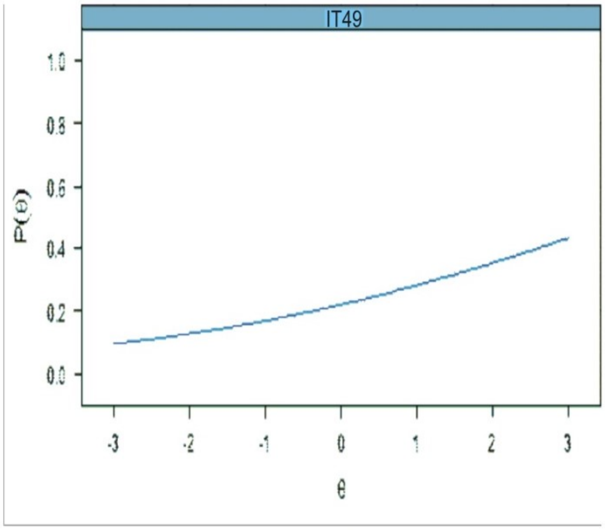

IT44
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

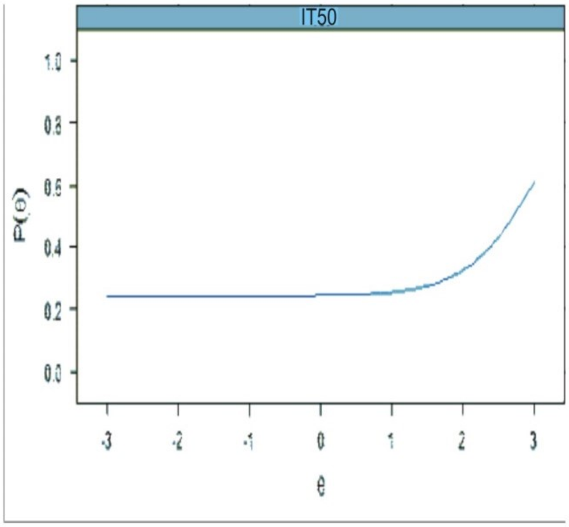

IT45
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

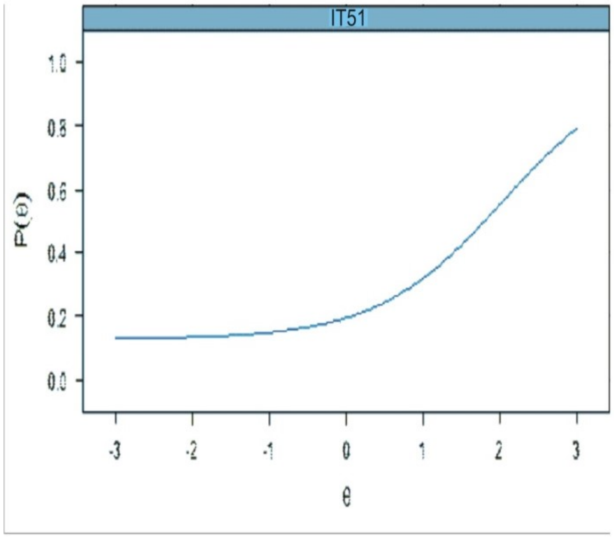

IT46
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

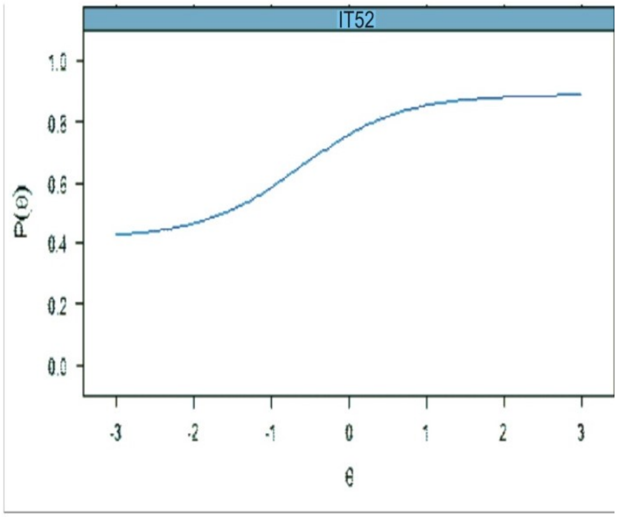

IT47
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

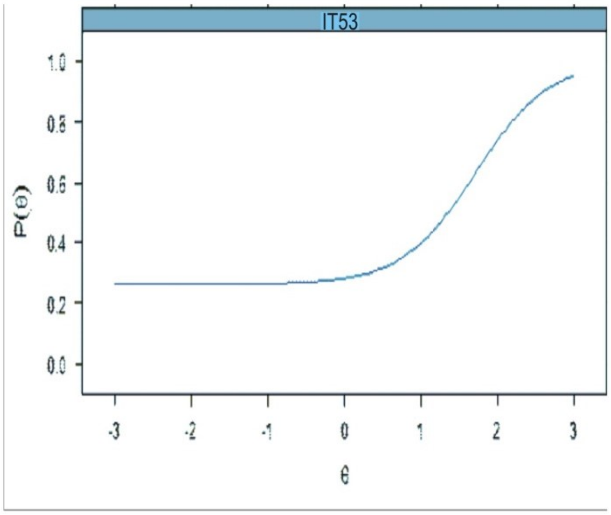

IT48
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

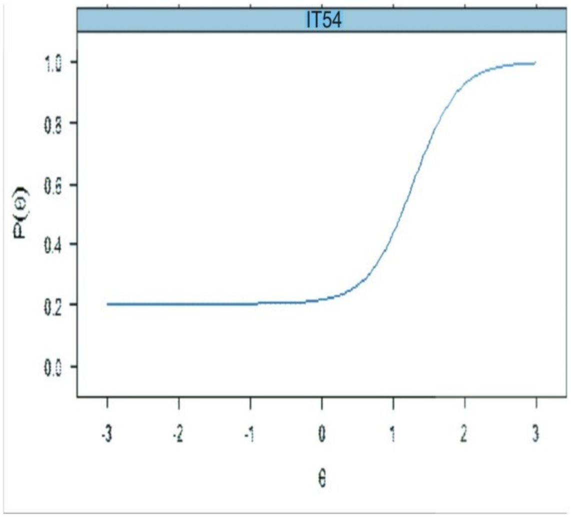

IT49
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

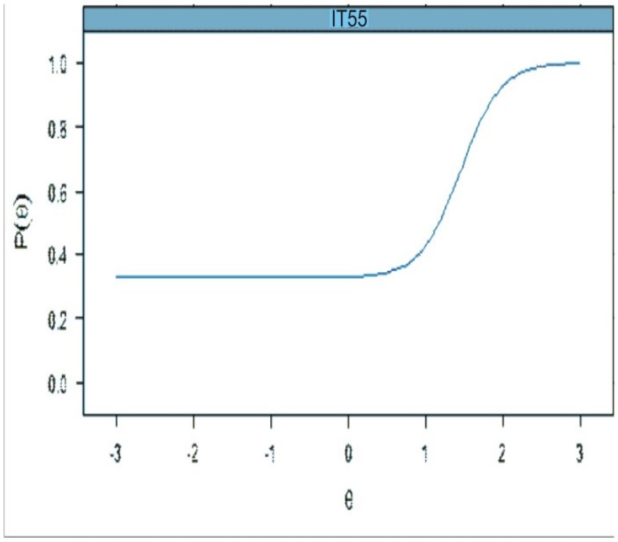

IT50
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

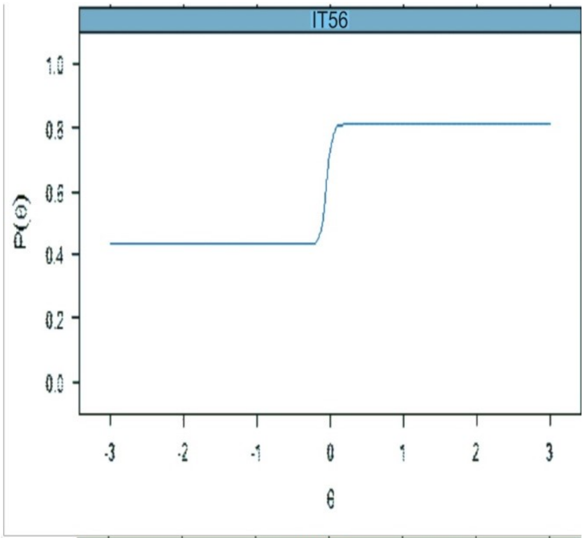

IT51
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

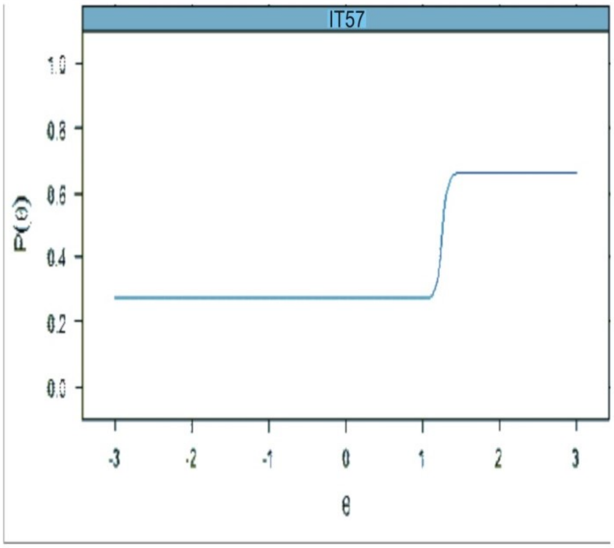

IT52
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

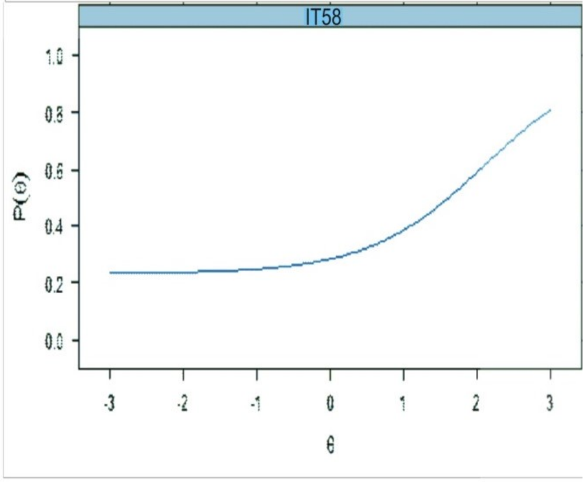

IT53
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

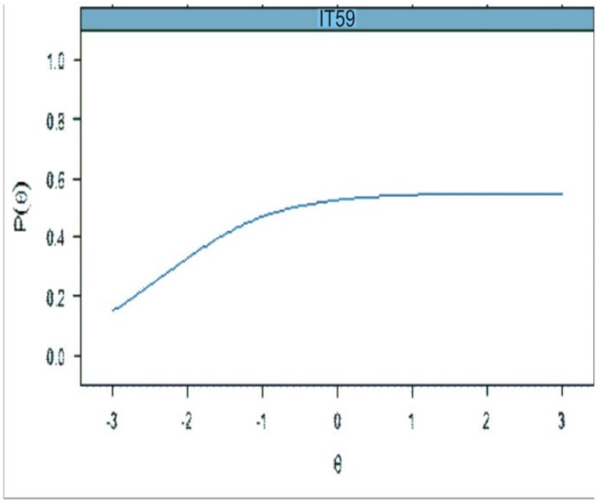

IT54
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

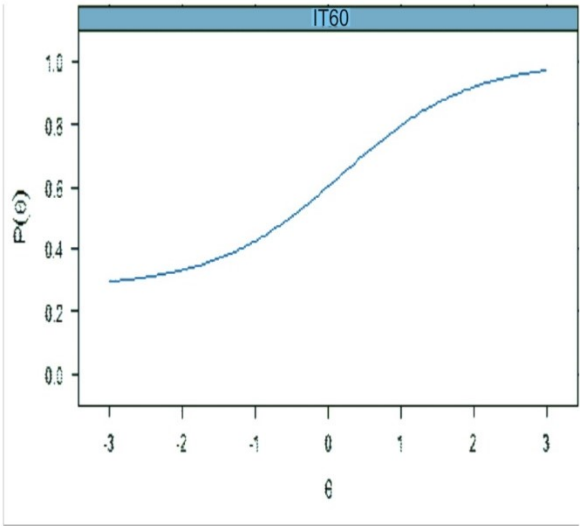

IT55
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

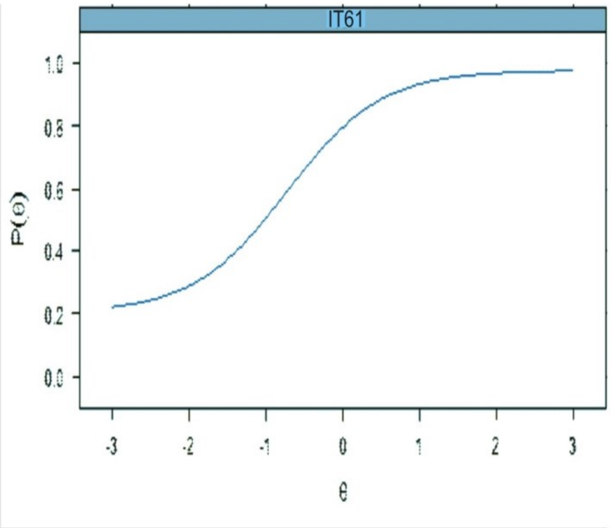

IT56
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

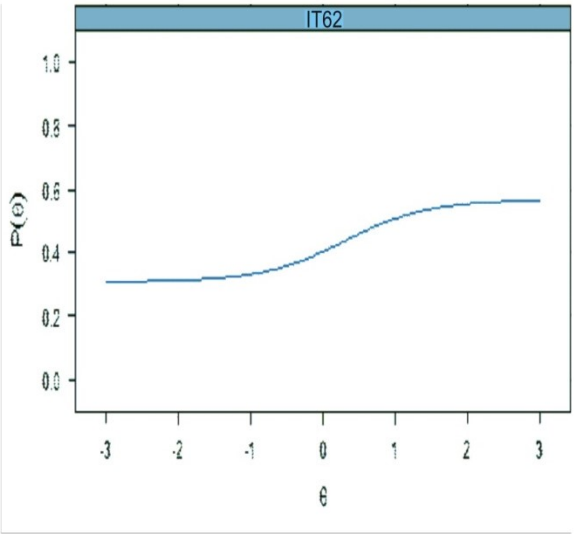

IT57
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

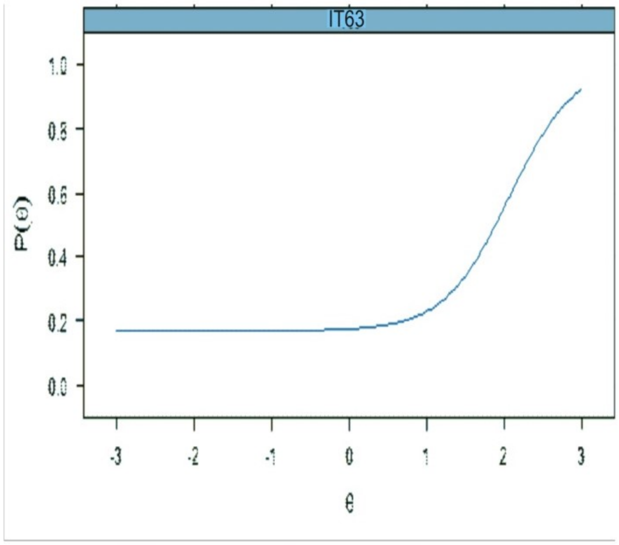

IT58
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

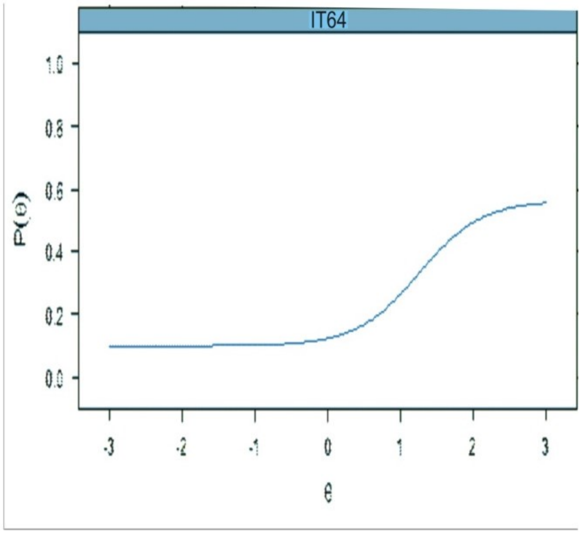

IT59
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

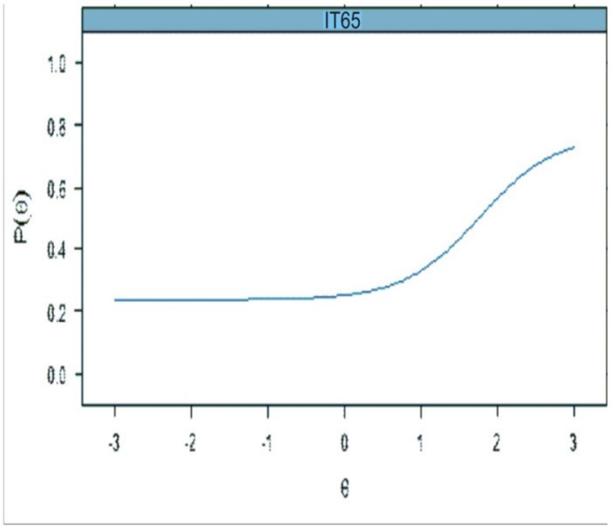

IT60
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

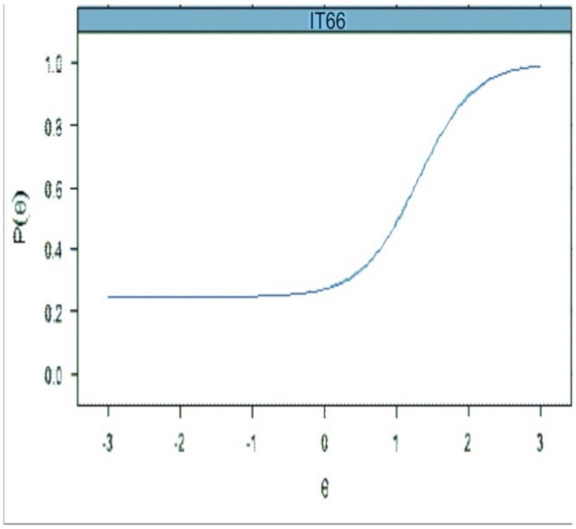

IT61
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

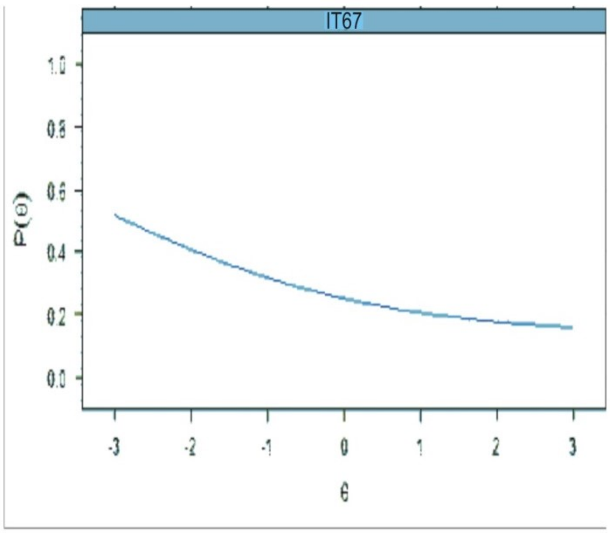

IT62
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

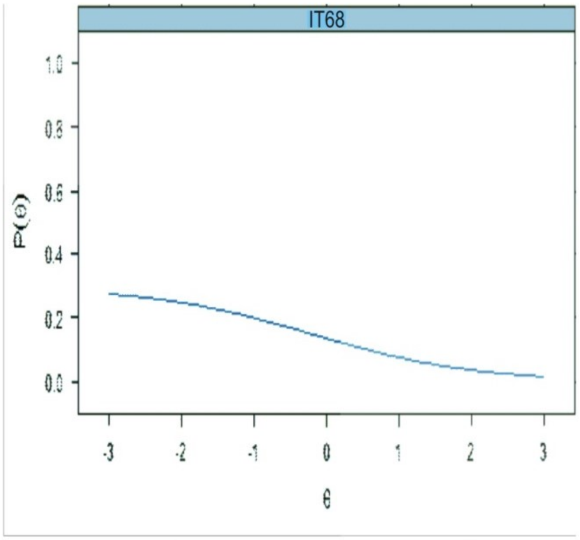

IT63
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

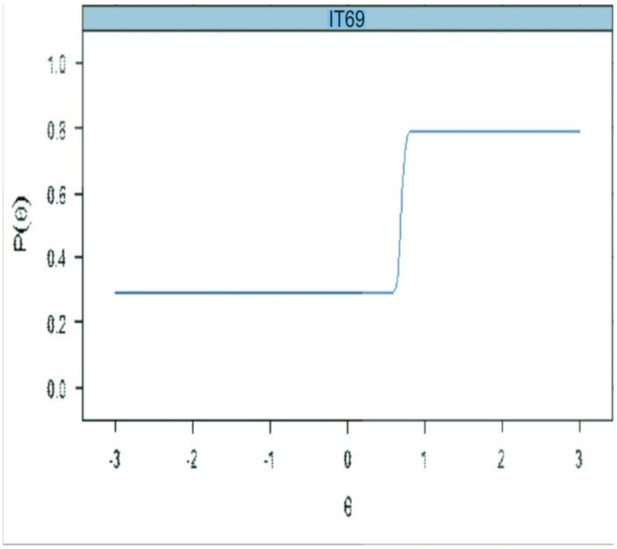

IT64
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

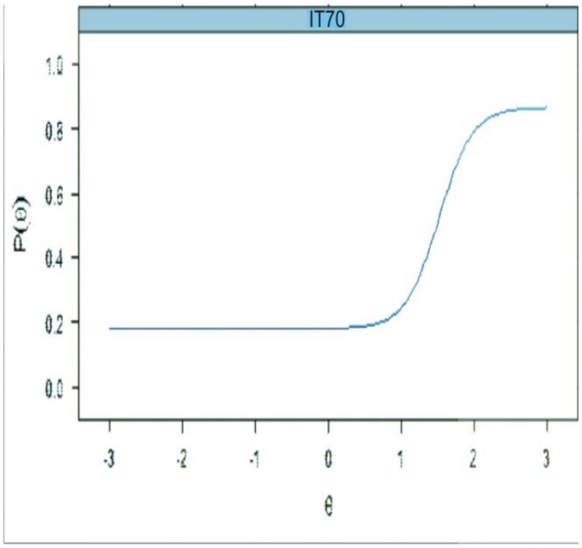

IT65
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

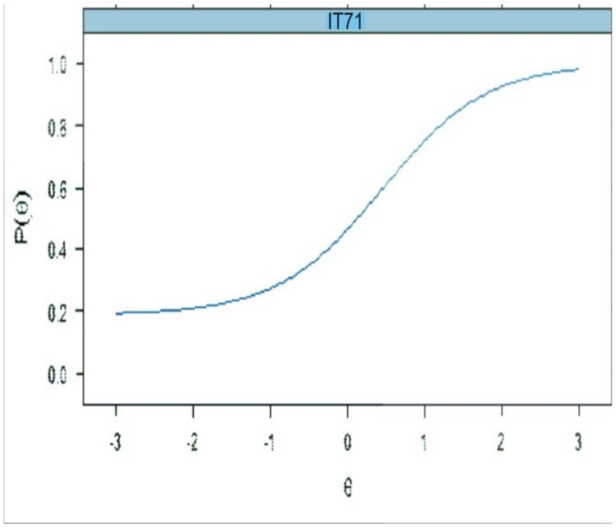

IT66
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

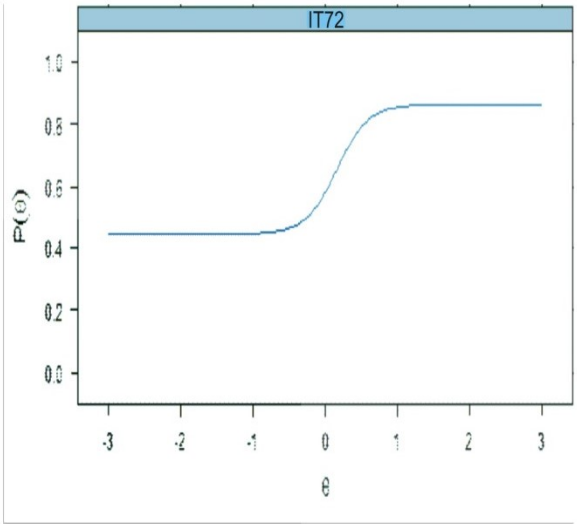

IT67
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

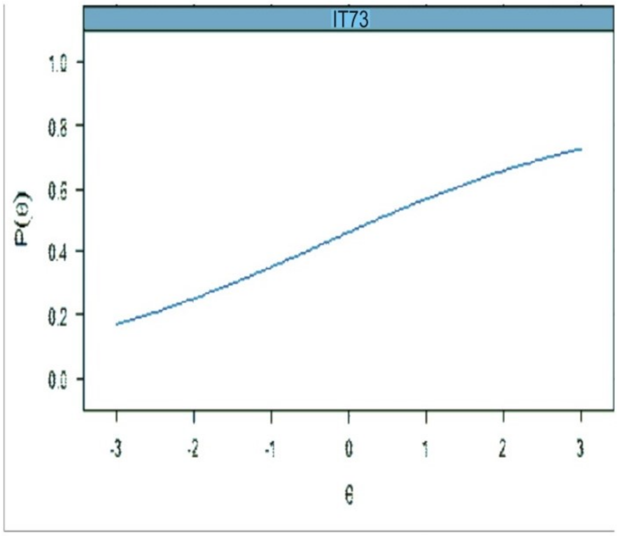

IT68
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

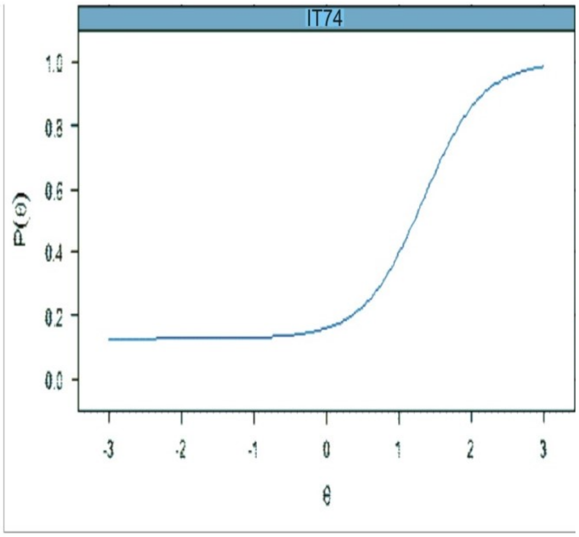

IT69
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

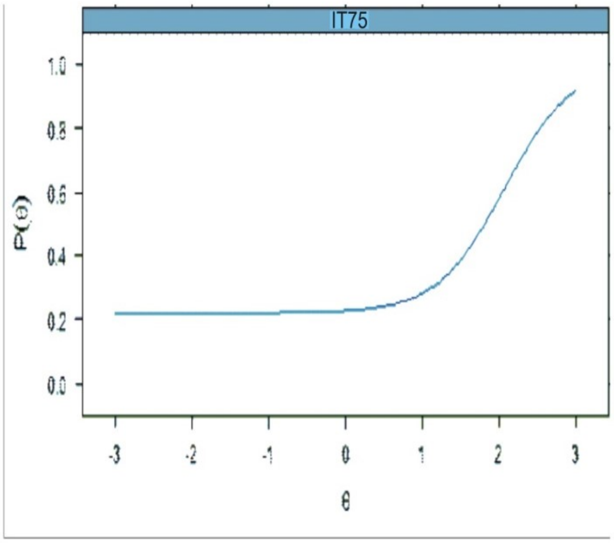

IT70
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

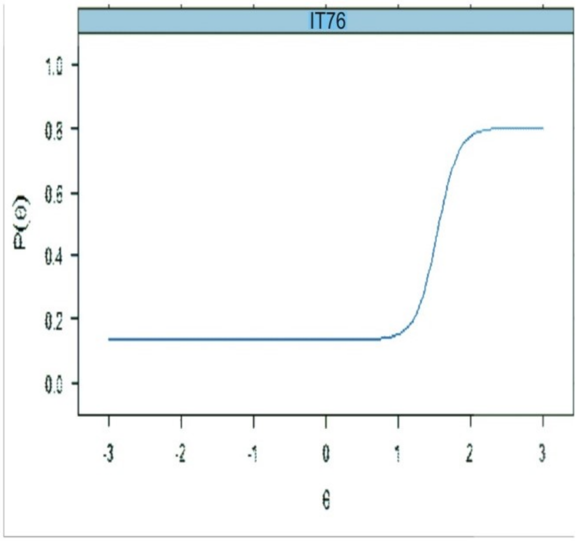

IT71
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

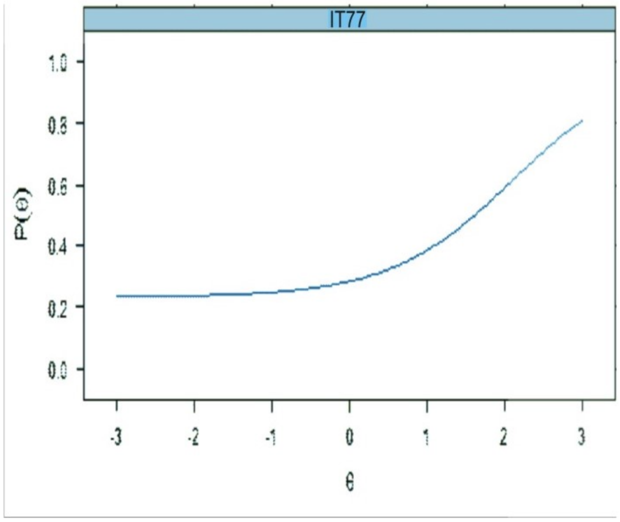

IT72
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

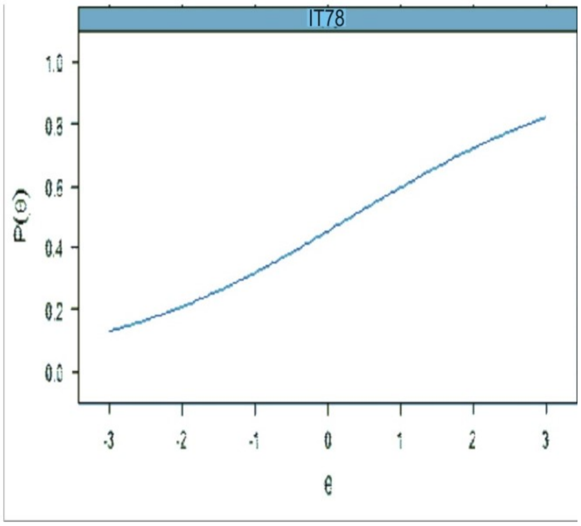

IT73
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

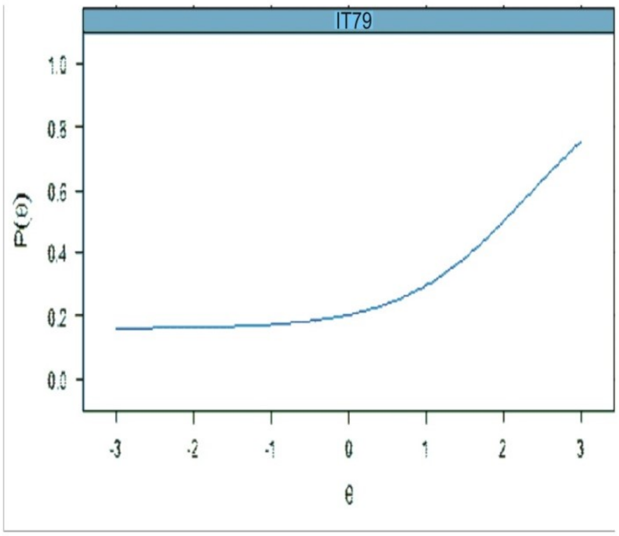

IT74
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

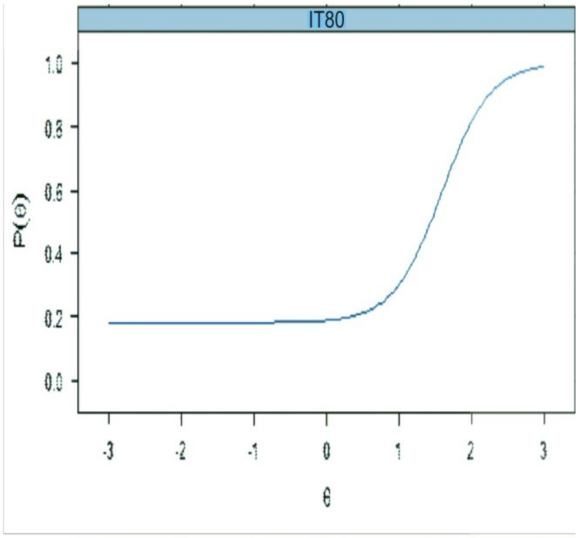

IT75
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

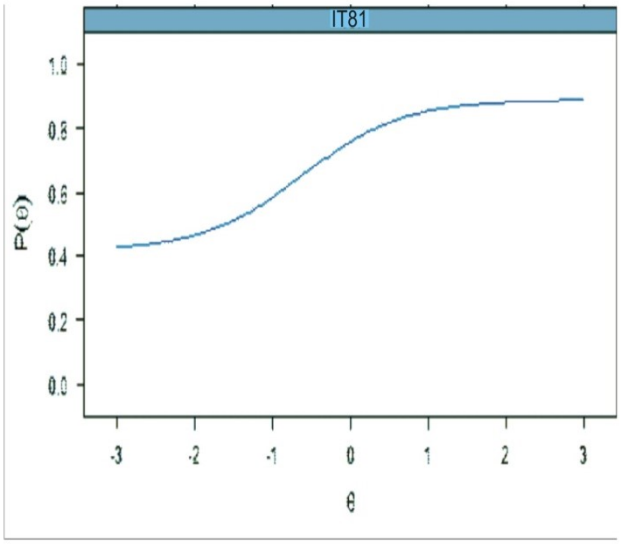

IT76
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

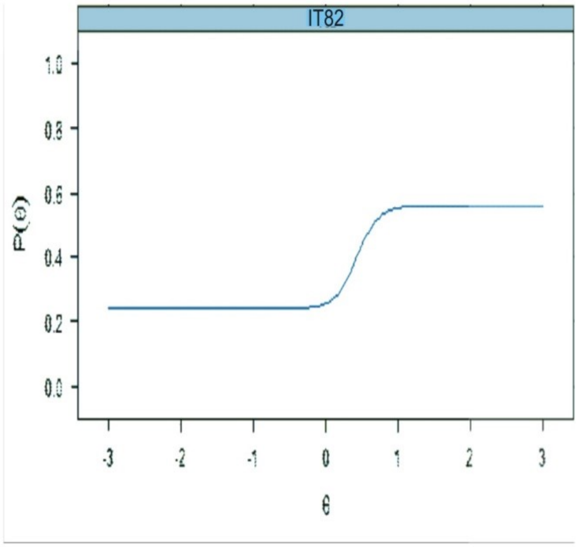

IT77
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

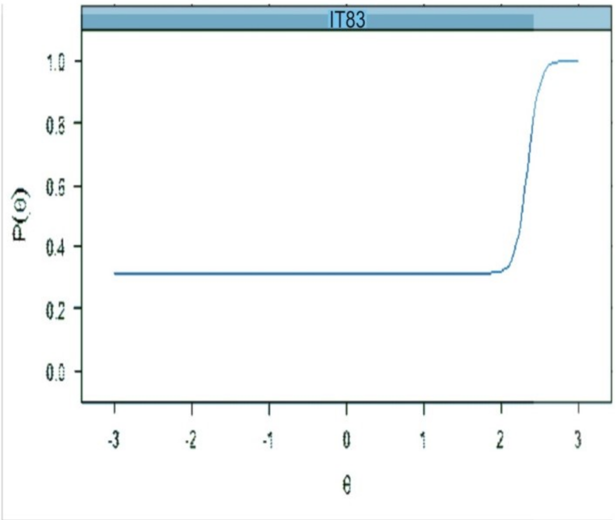

IT78
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

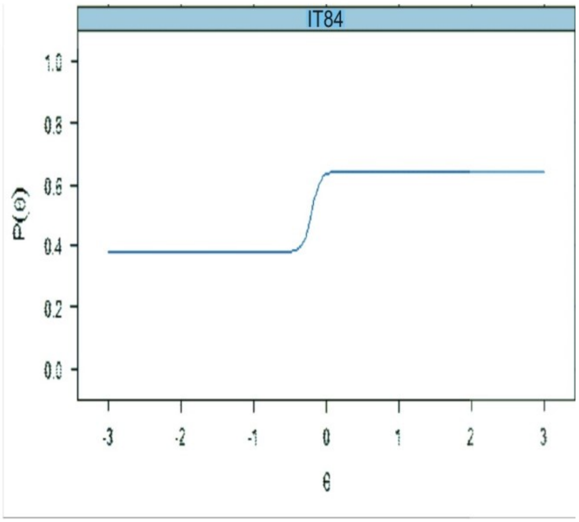

IT79
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

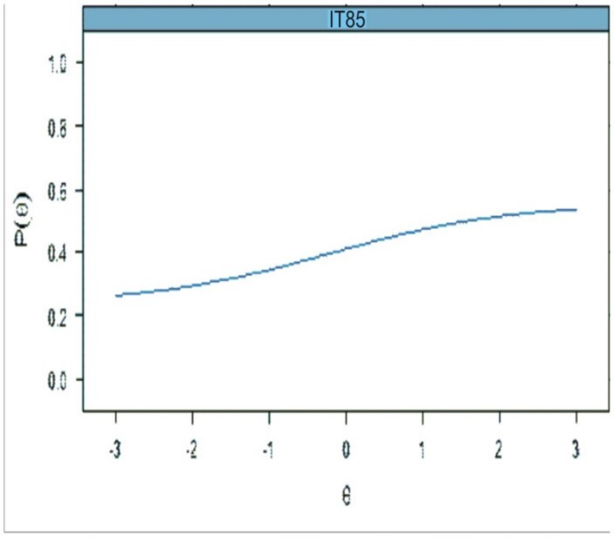

IT80
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

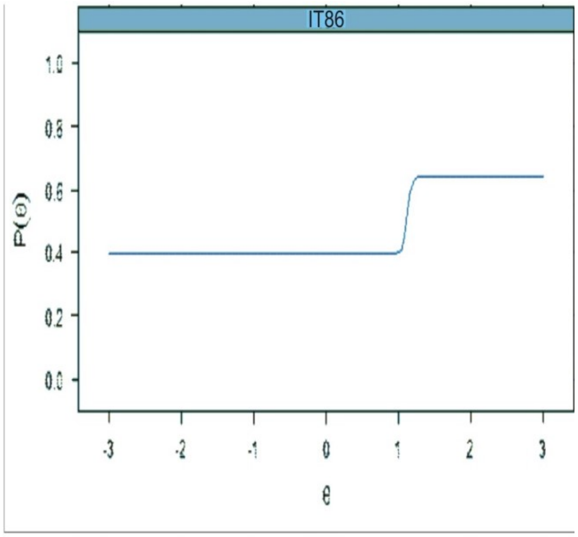

IT81
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

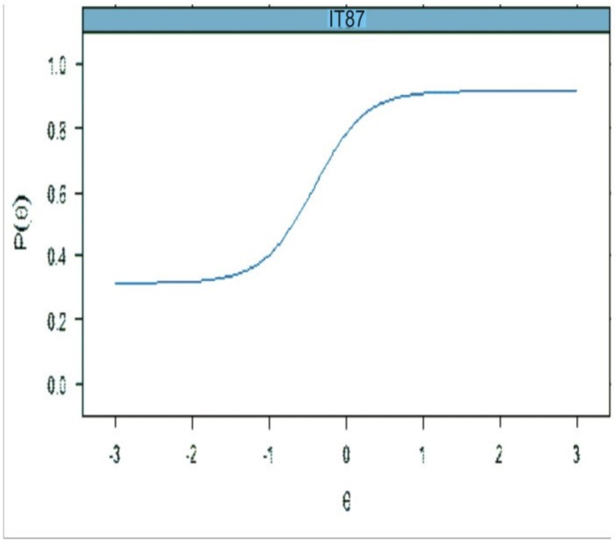

IT82
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

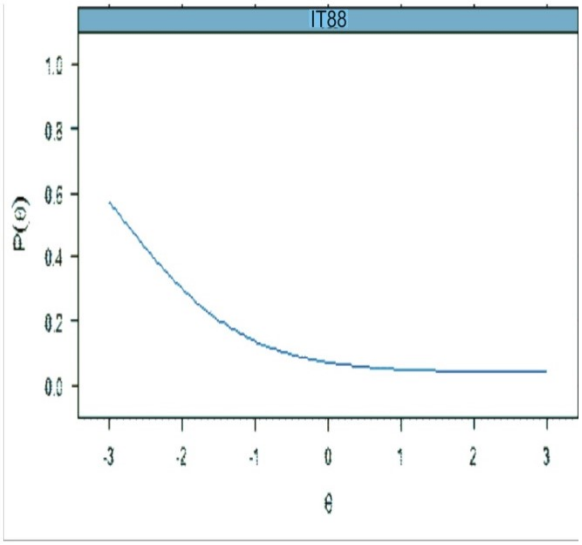

IT83
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

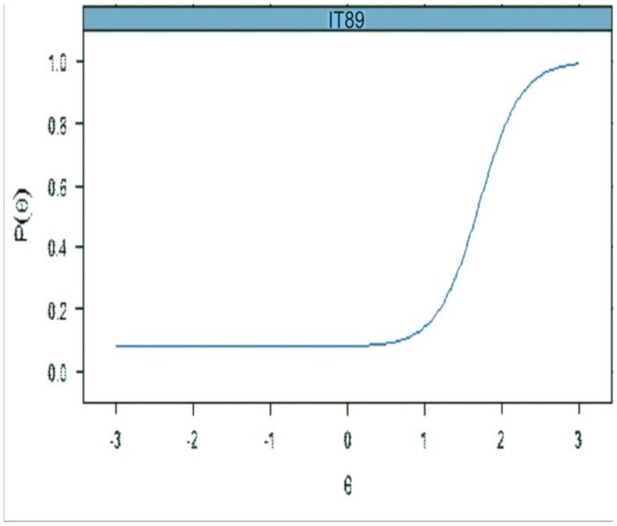

IT84
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

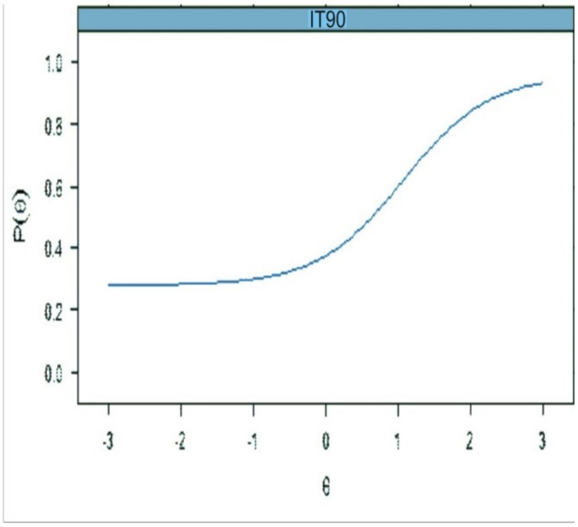

IT85
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

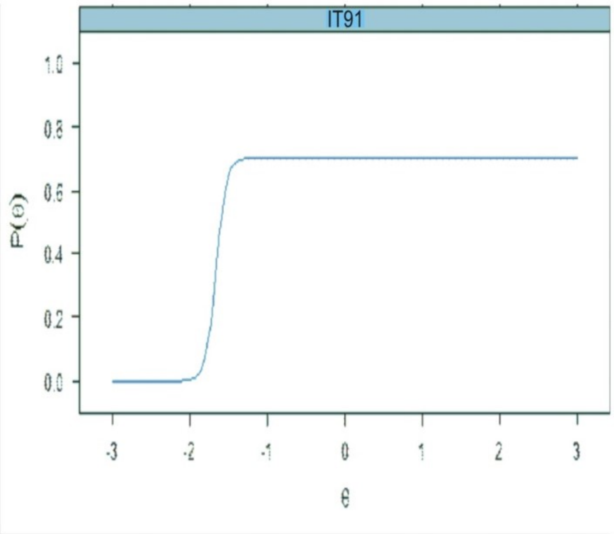

IT86
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

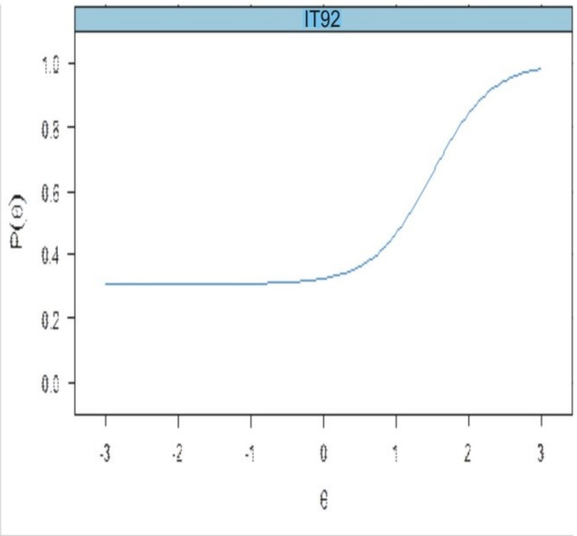

IT87
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

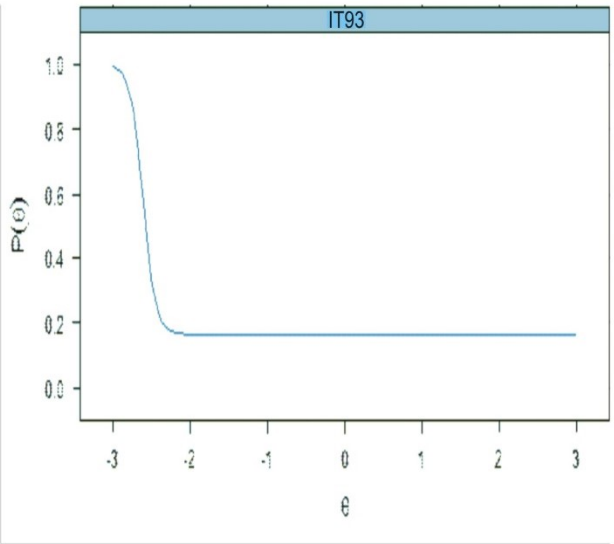

IT88
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

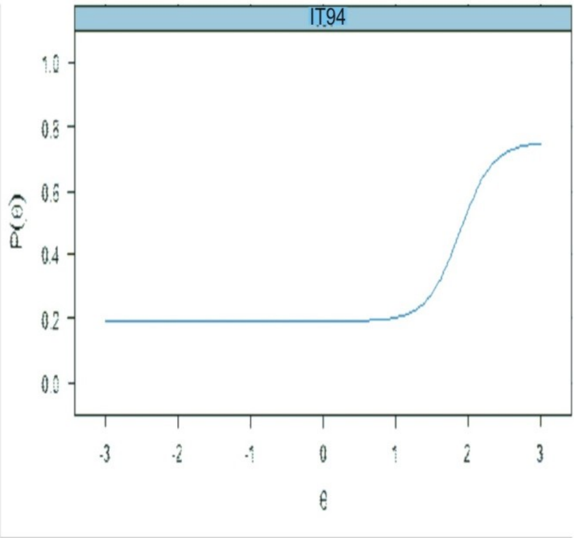

IT89
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

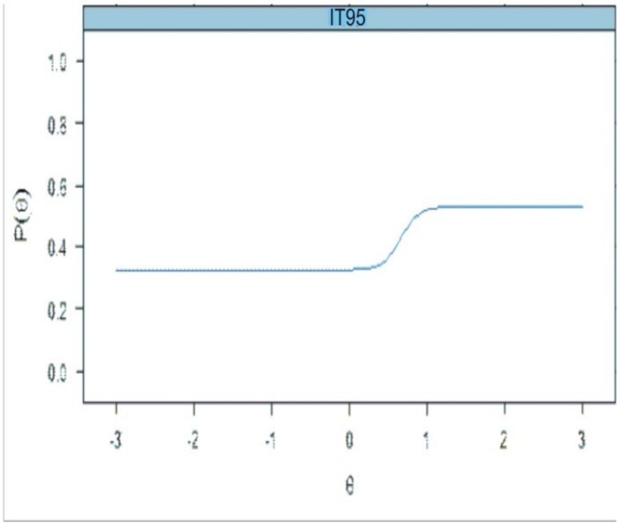

IT90
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

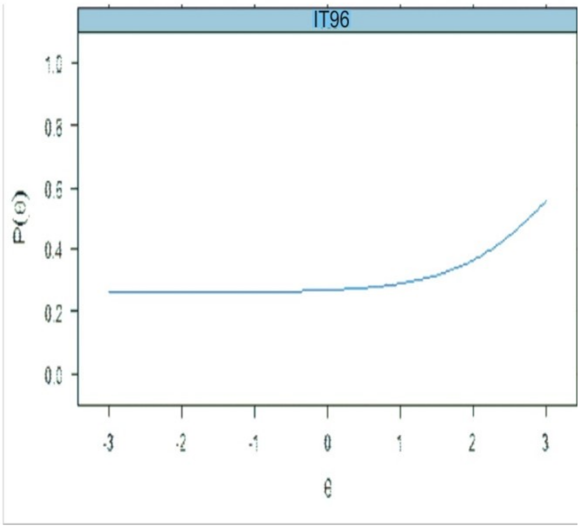

IT91
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

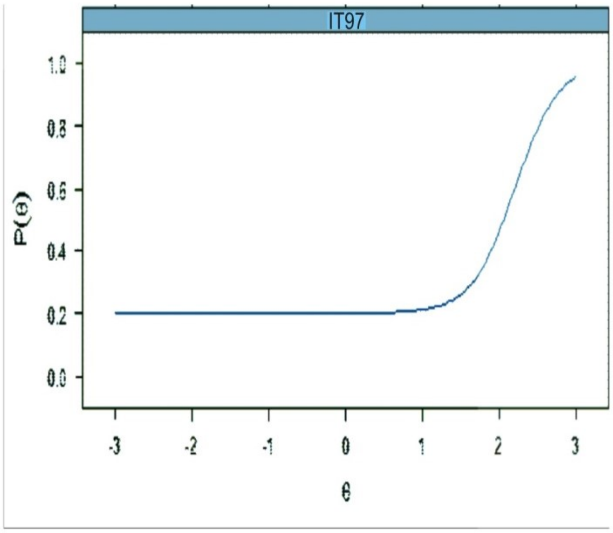

IT92
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

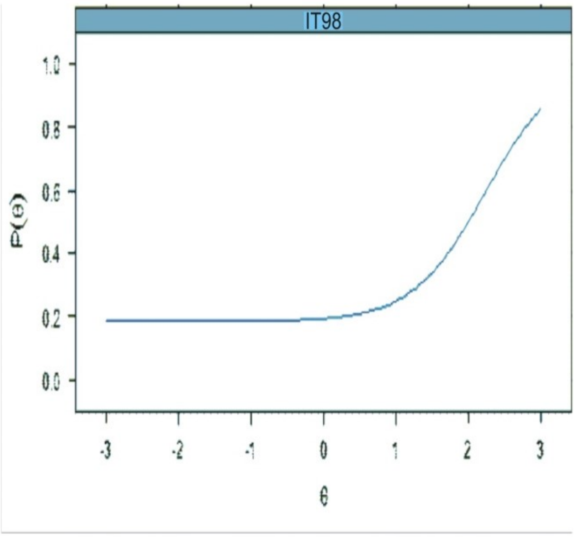

IT93
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

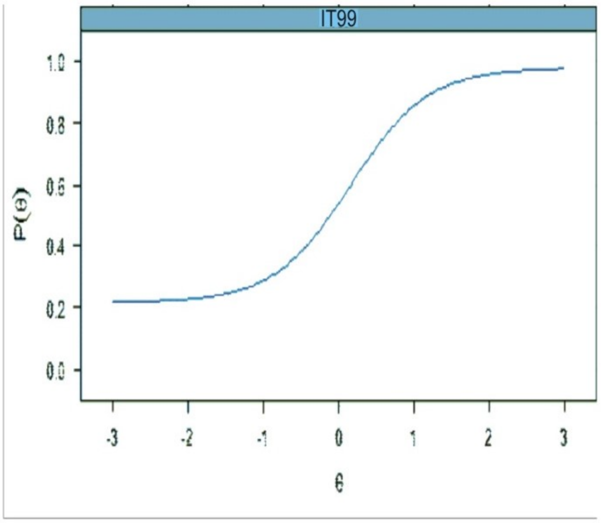

IT94
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

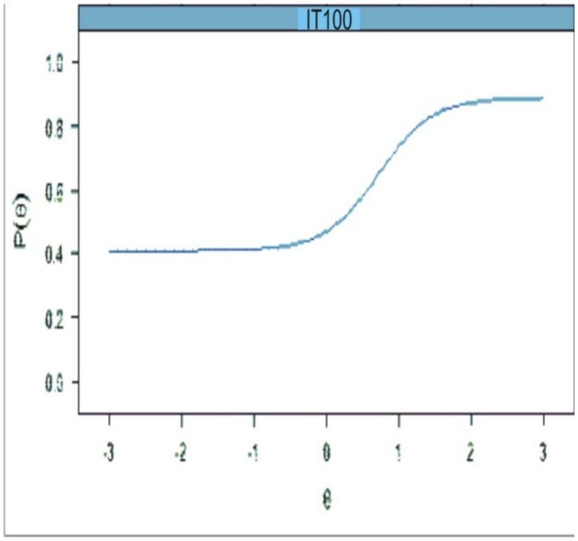

IT95
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

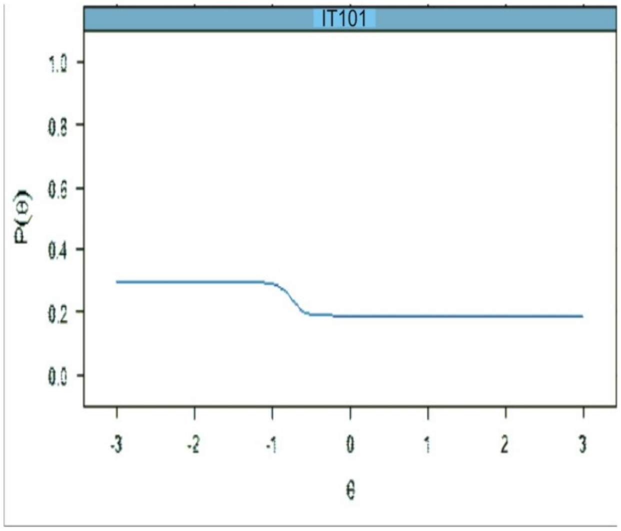

IT96
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

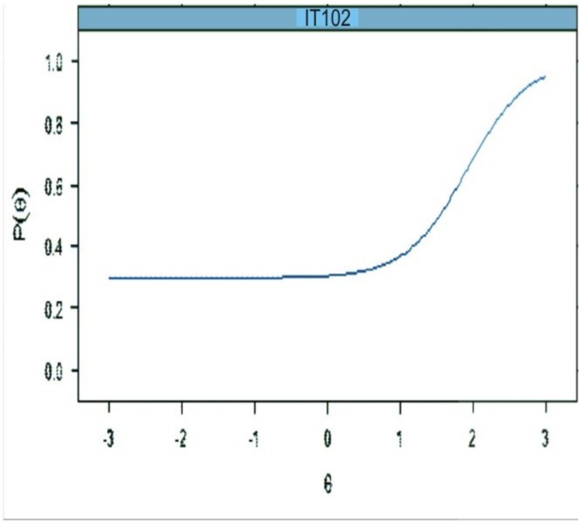

IT97
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

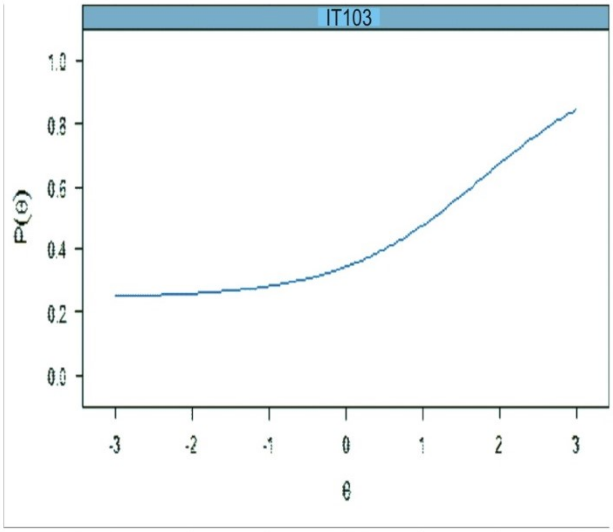

IT98
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

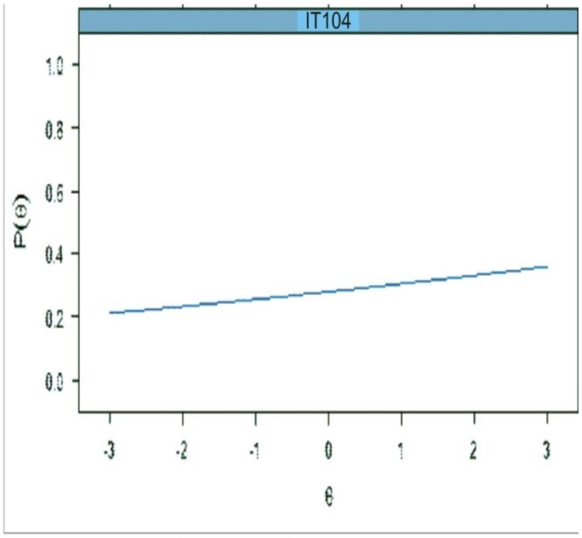

IT99
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

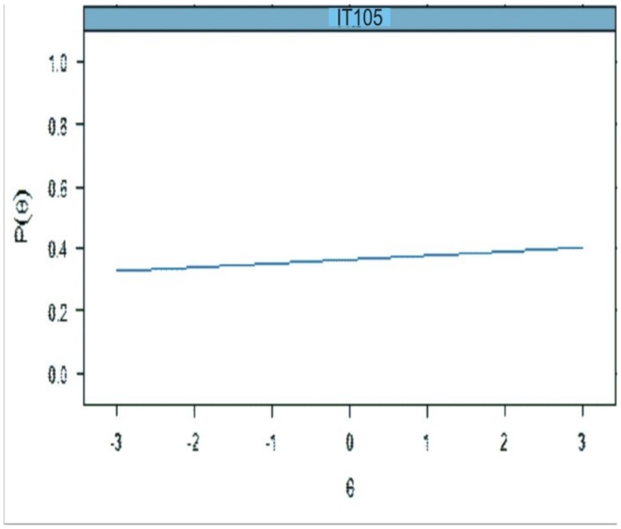

IT100
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

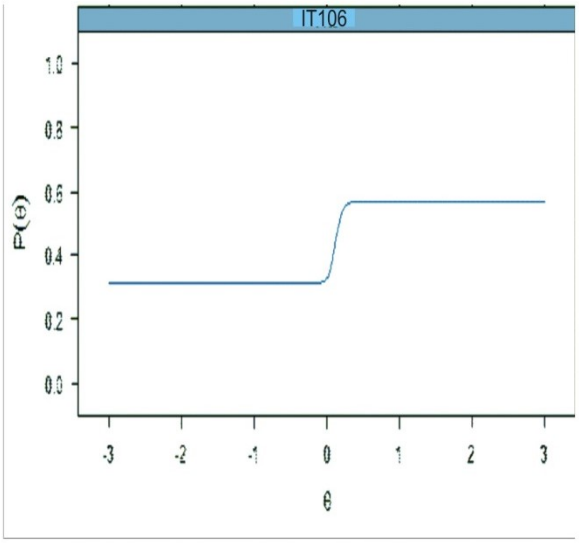

IT101
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

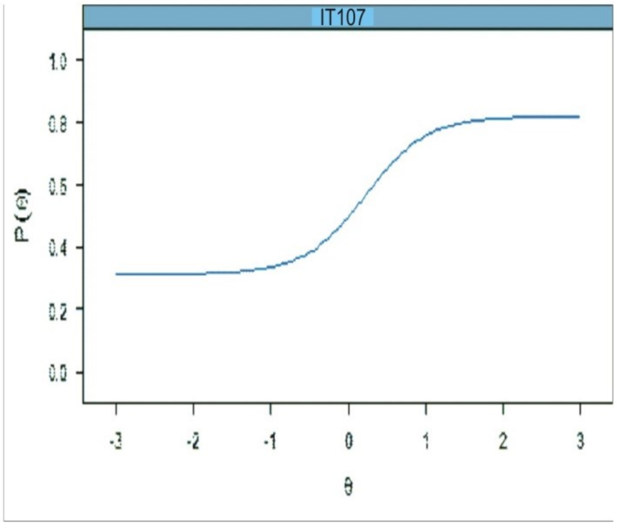

IT102
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

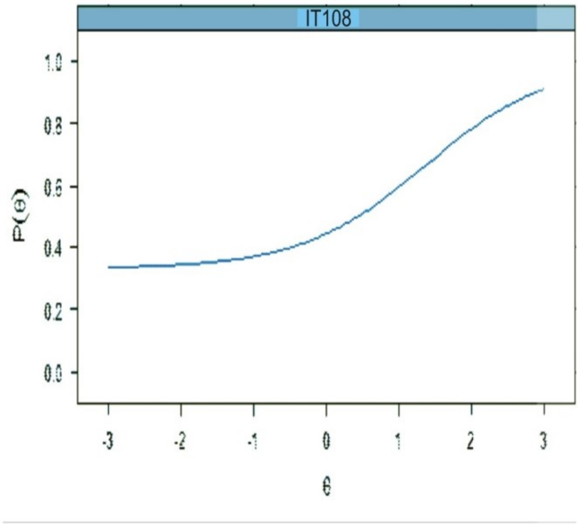

IT103
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

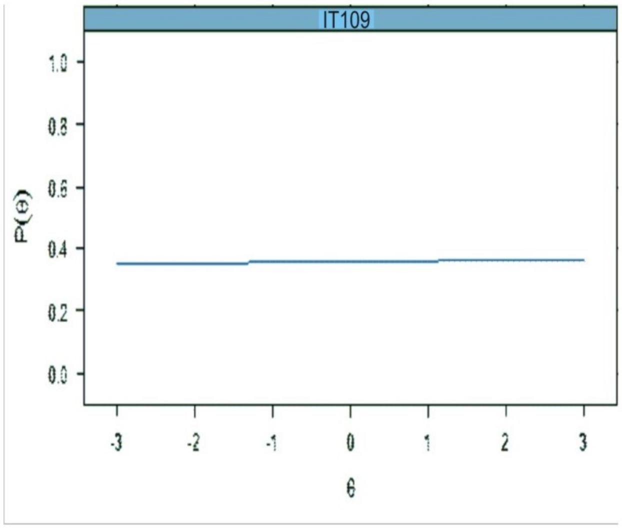

IT104
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

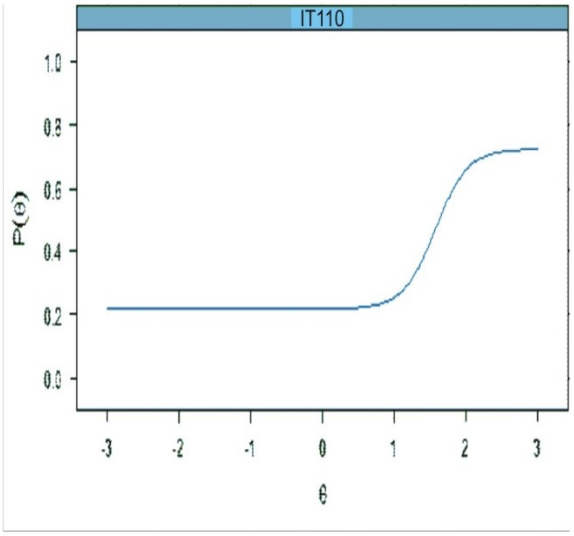

IT105
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

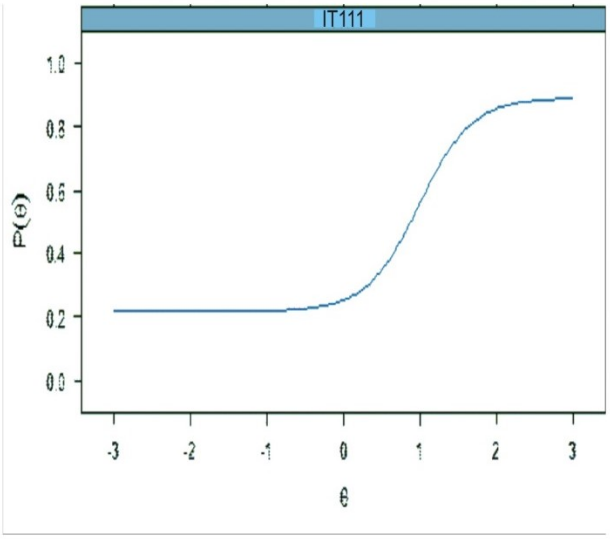

IT106
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

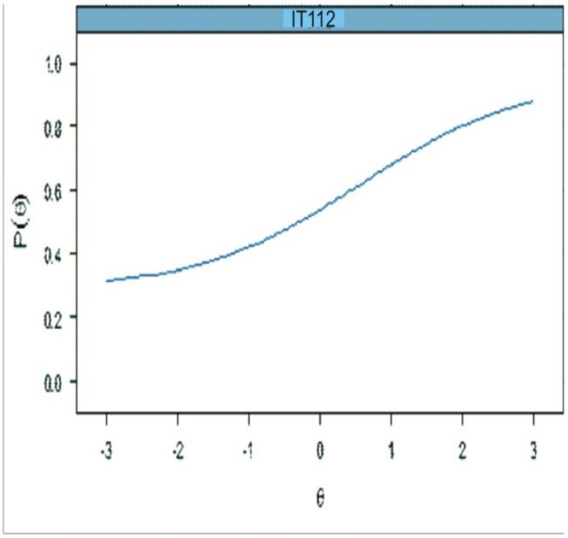

IT107
P(θ)
1.0
0.8
0.5
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

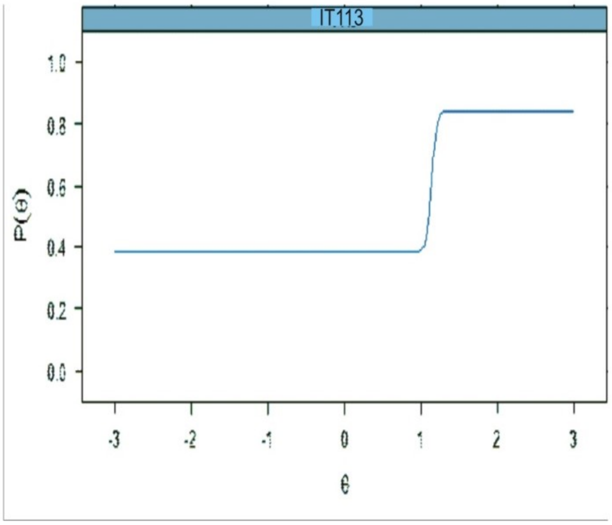

IT108
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

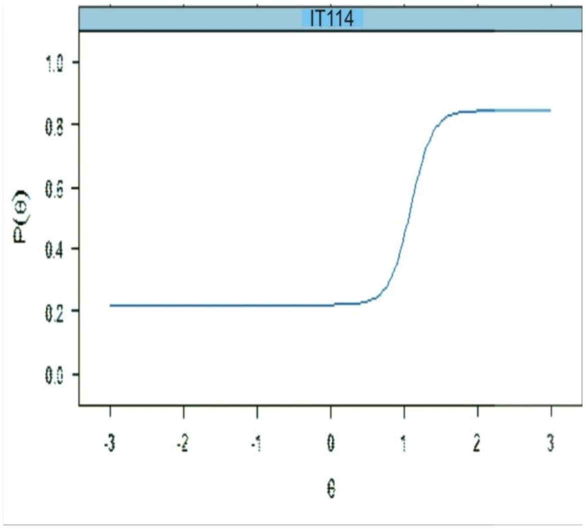

IT109
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ


IT110
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

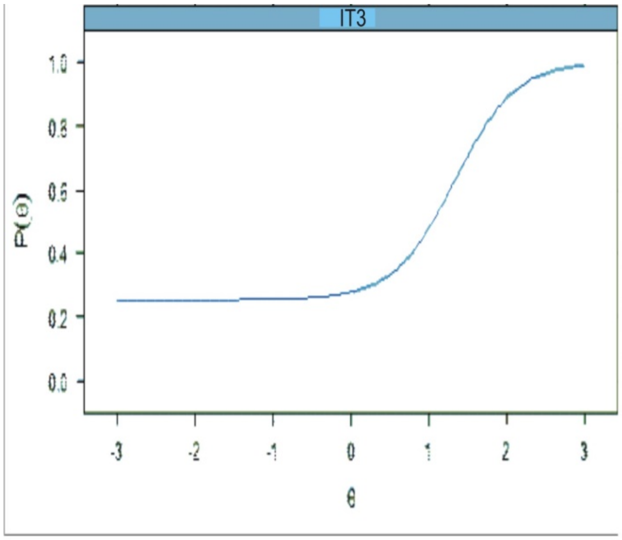

IT111
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

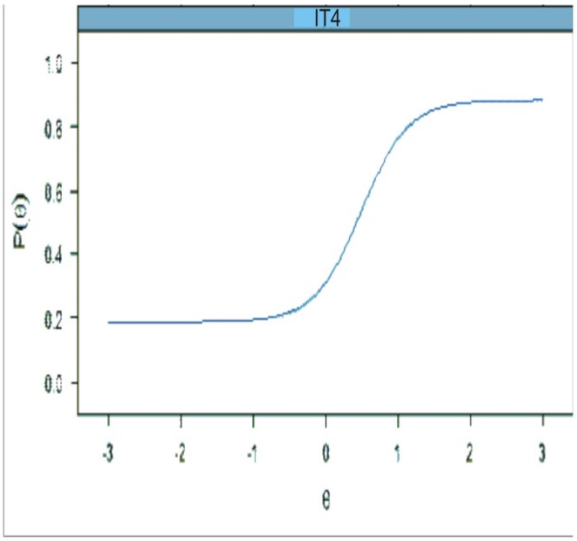

IT112
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

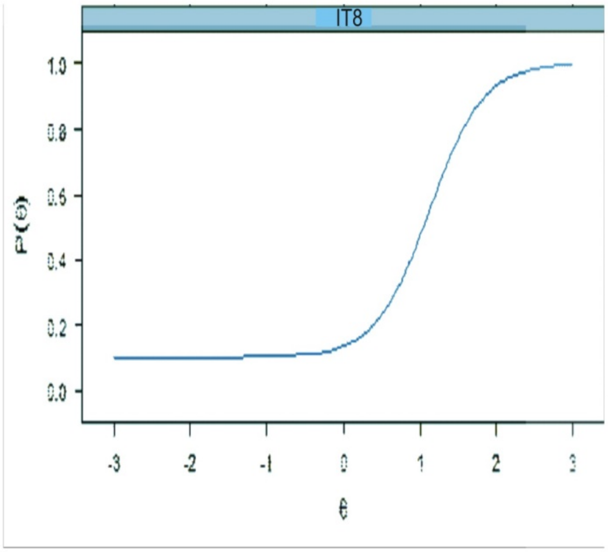

IT113
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

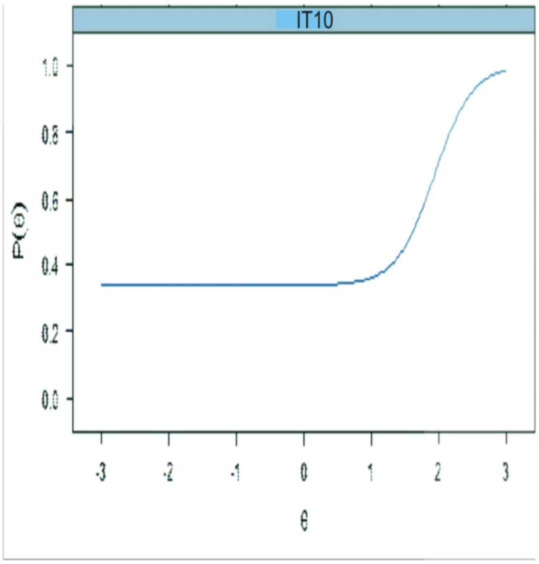

IT114
1.0
0.8
0.6
0.4
0.2
0.0
P(θ)
-3
-2
-1
0
1
2
3
θ

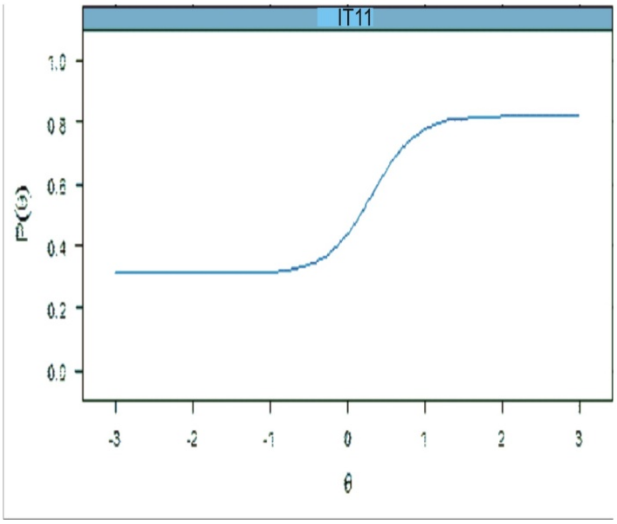
APPENDIX X (ICCs for the final CBMAT items)

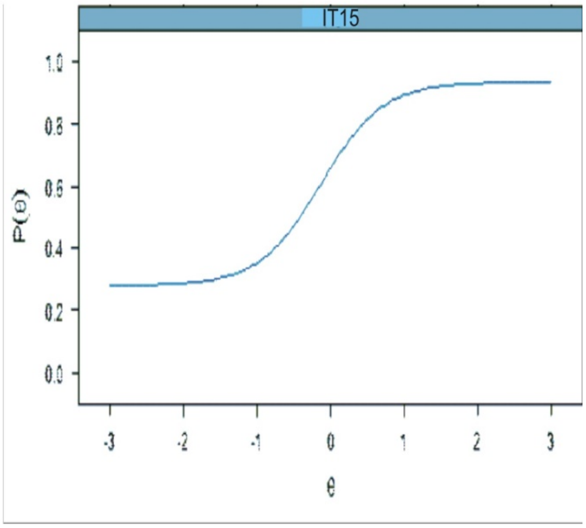


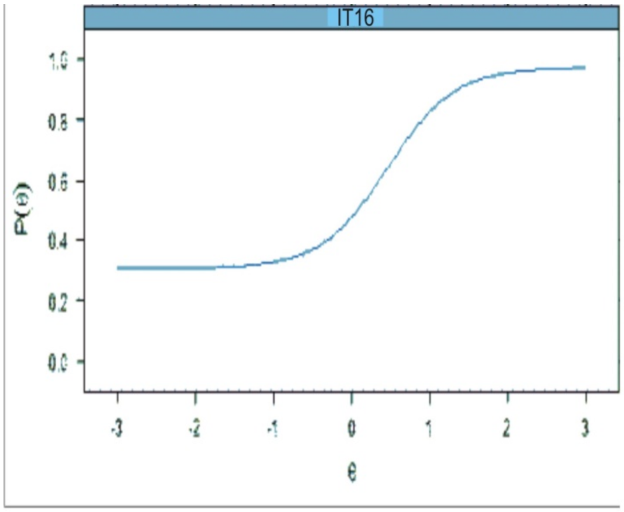


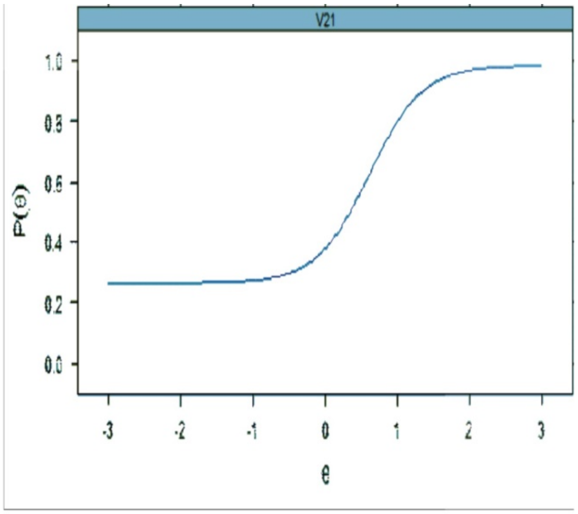


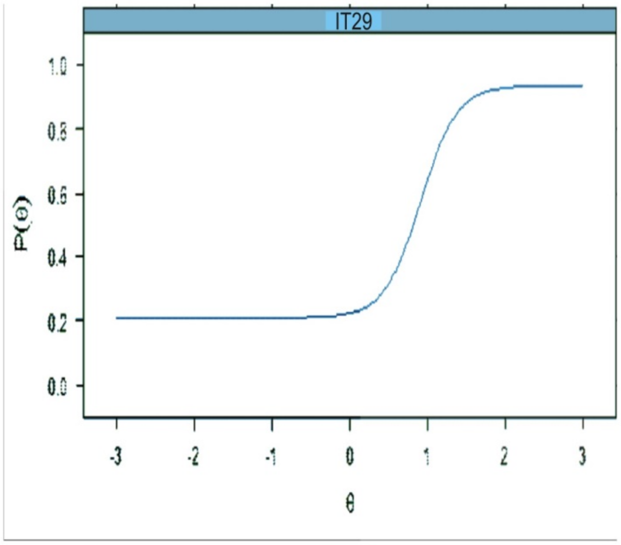


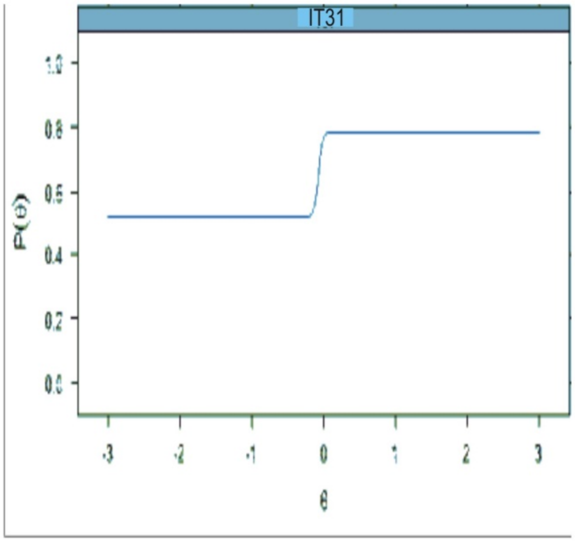


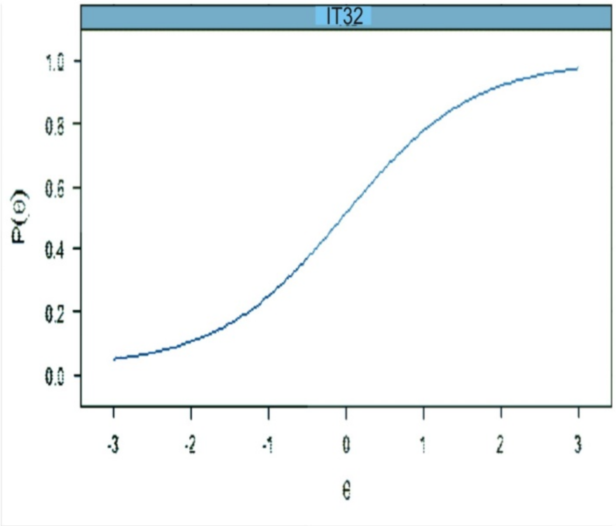


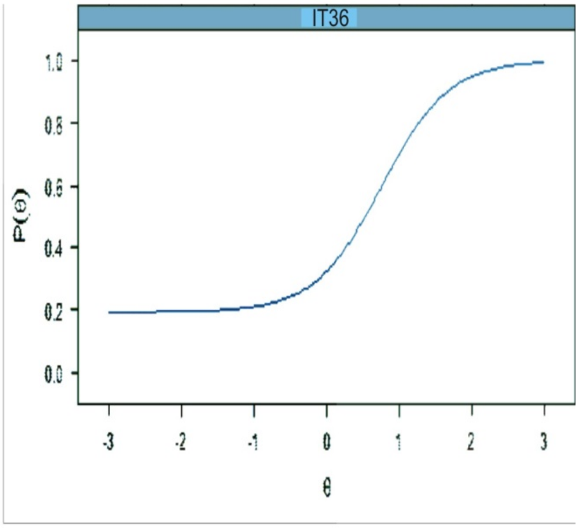

IT16
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

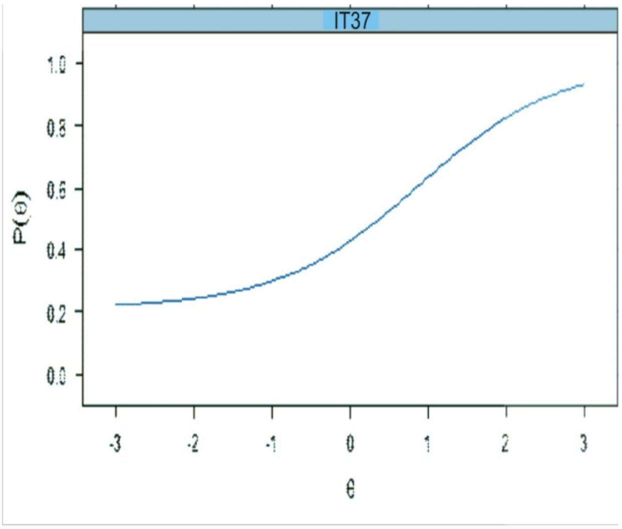

V21
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

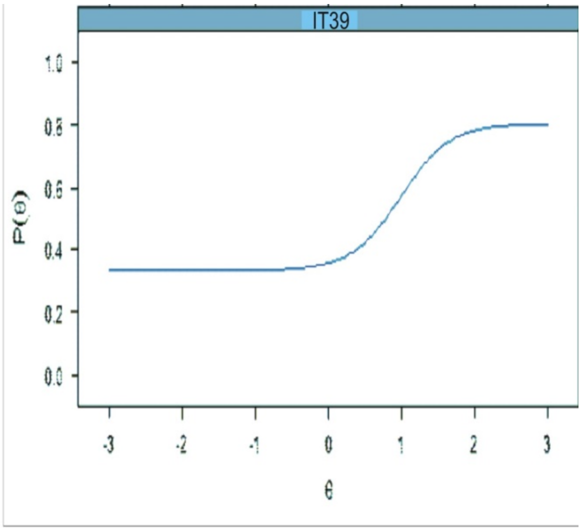

IT29
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

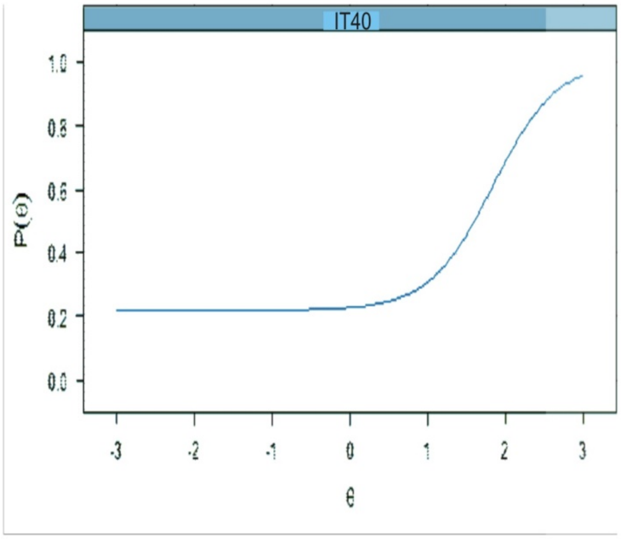

IT31
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

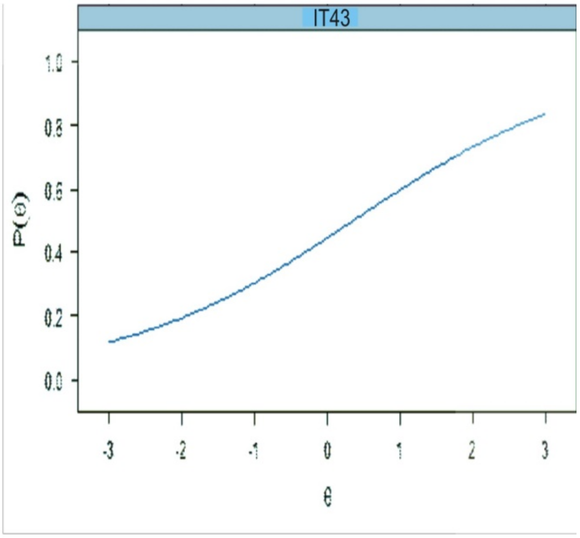

IT32
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

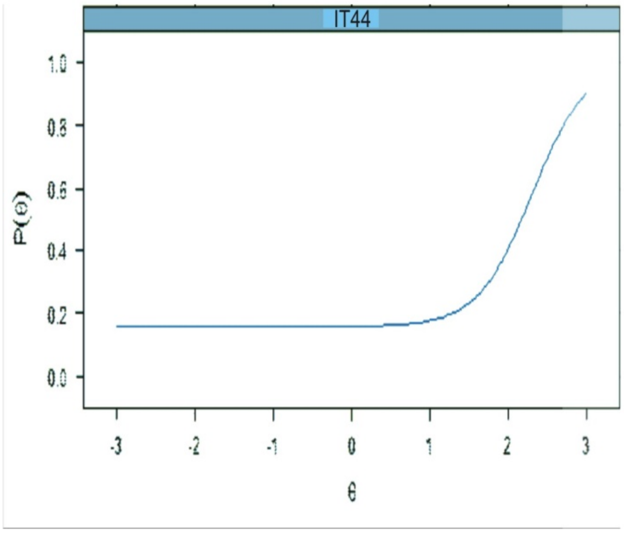

IT36
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

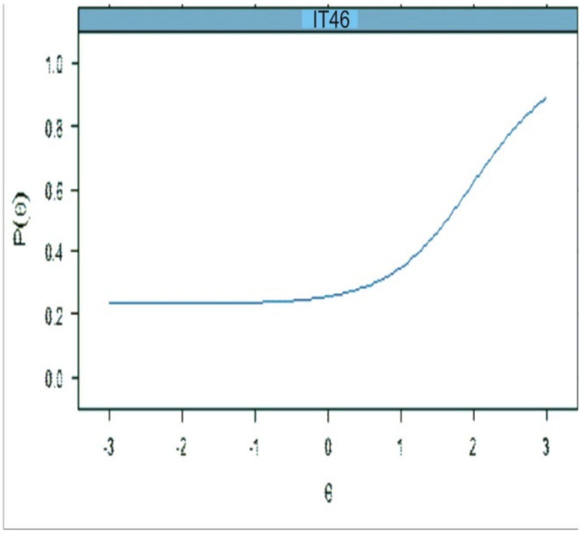

IT37
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

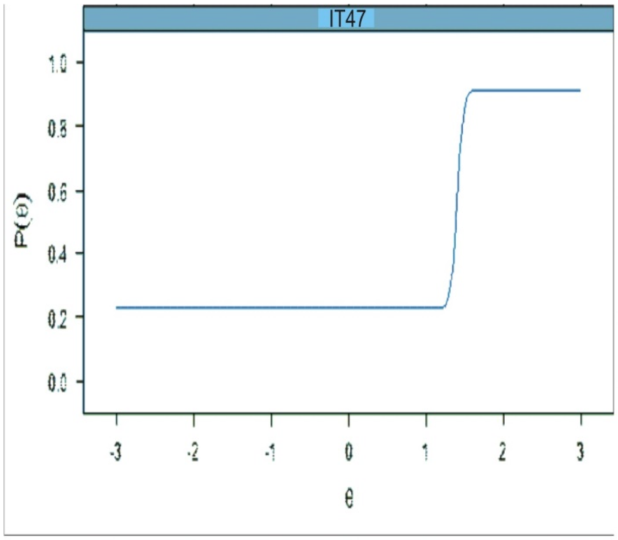

IT39
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

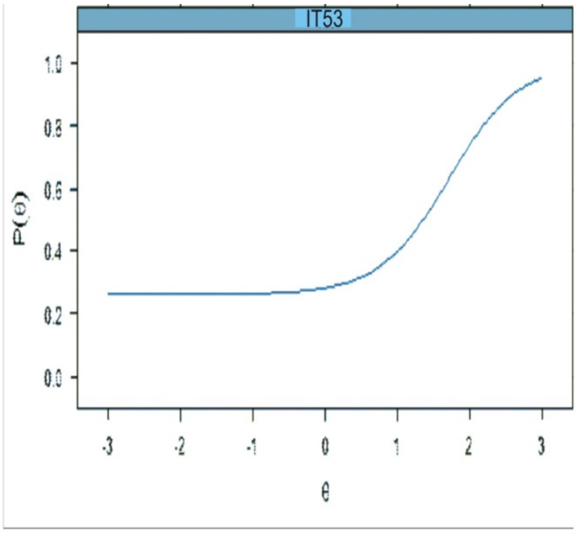

IT40
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

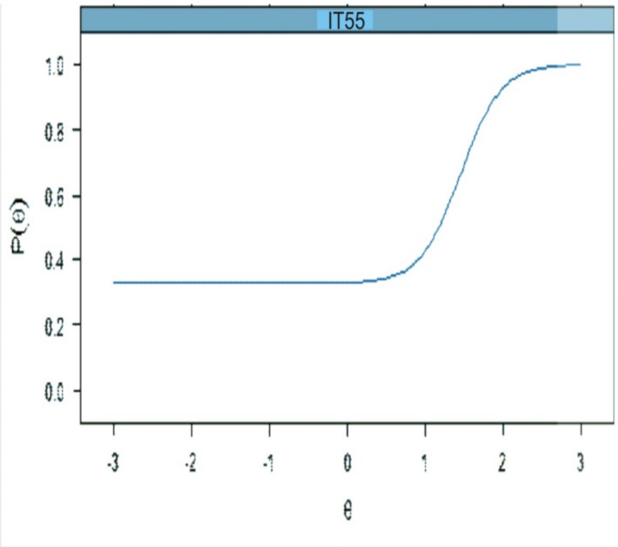

IT43
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

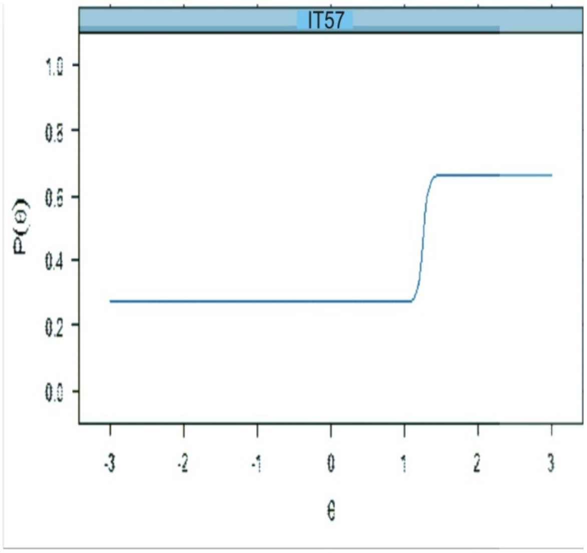

IT44
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

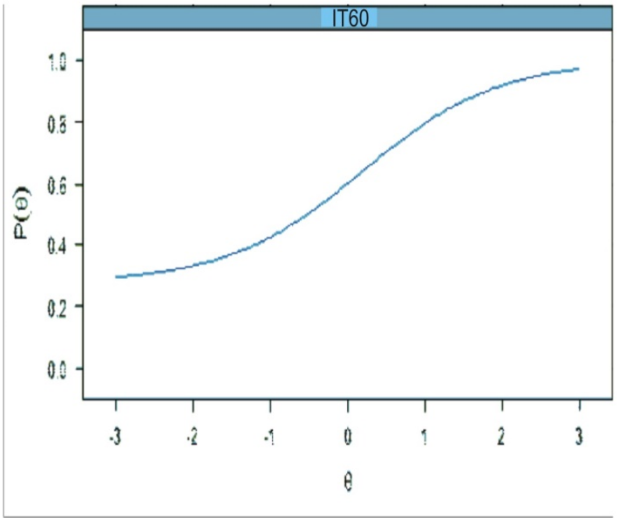

IT46
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

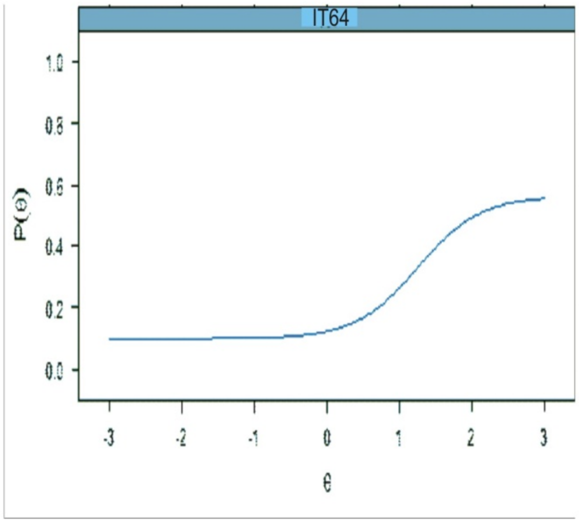

IT47
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

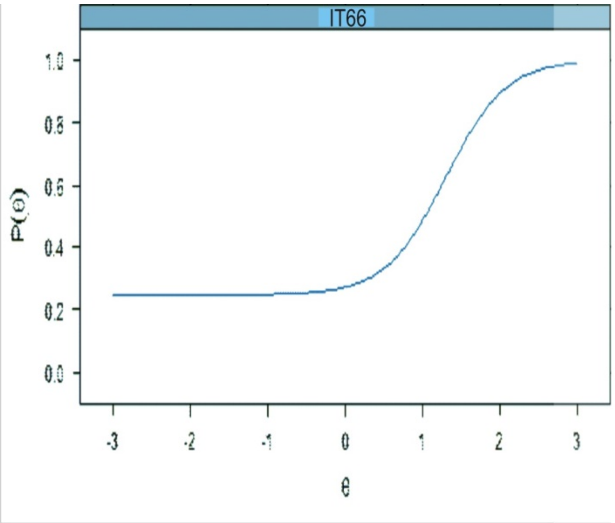

IT53
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

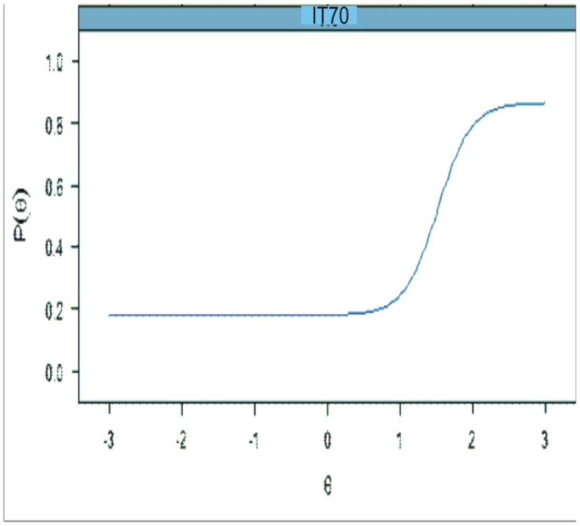

IT55
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

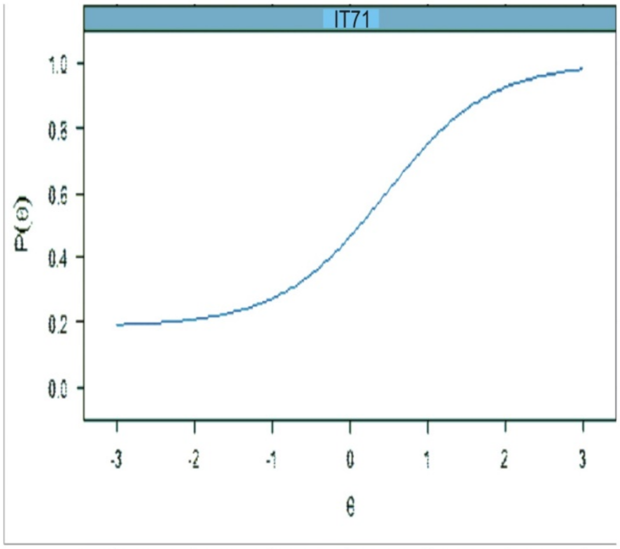

IT57
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

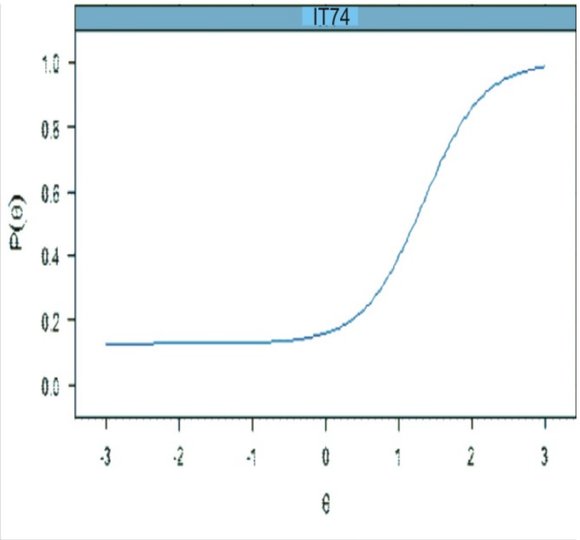

IT60
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

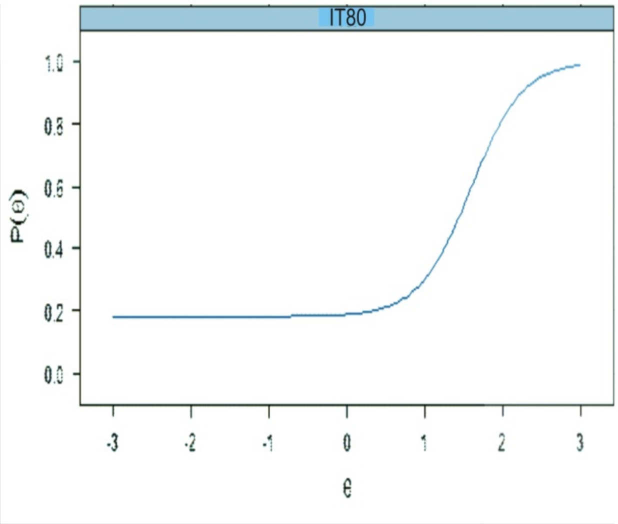

IT64
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

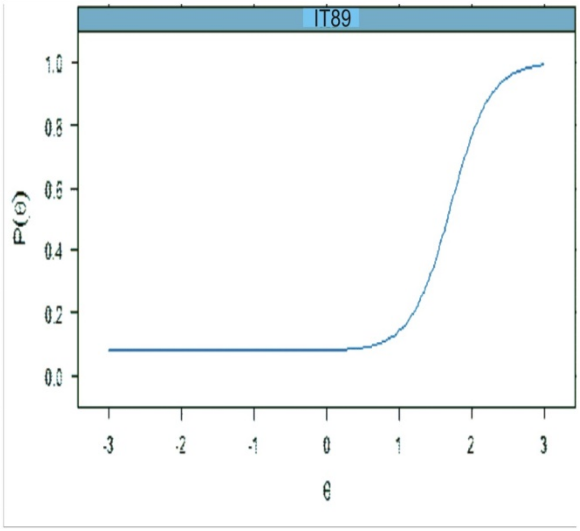

IT66
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

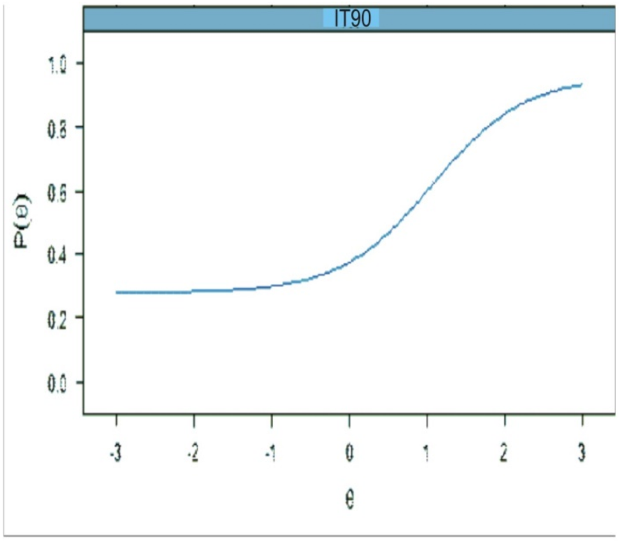

IT70
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

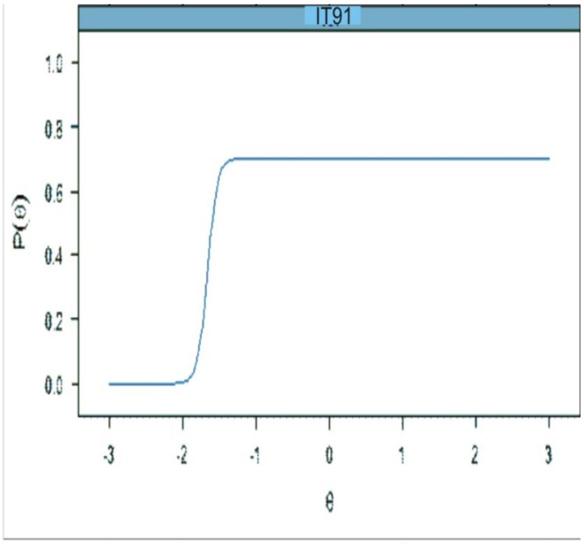

IT71
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

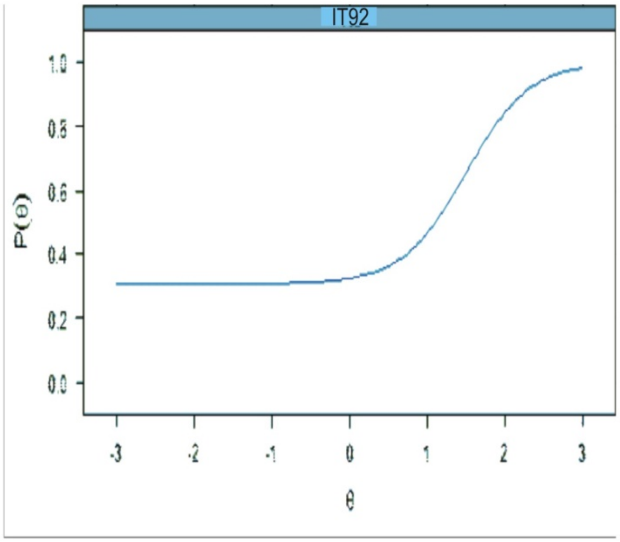

IT74
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

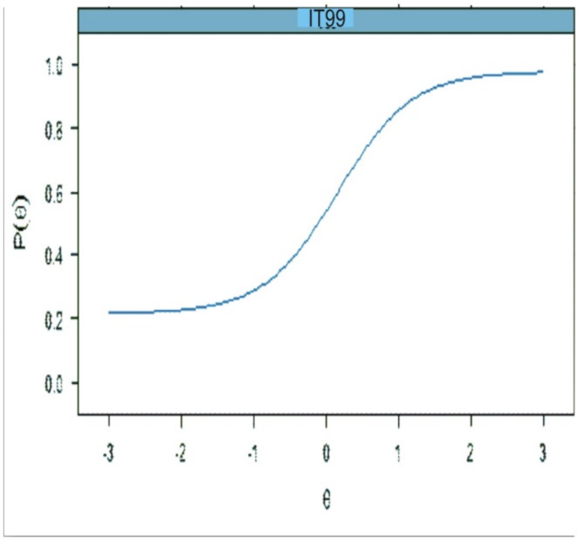

IT80
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

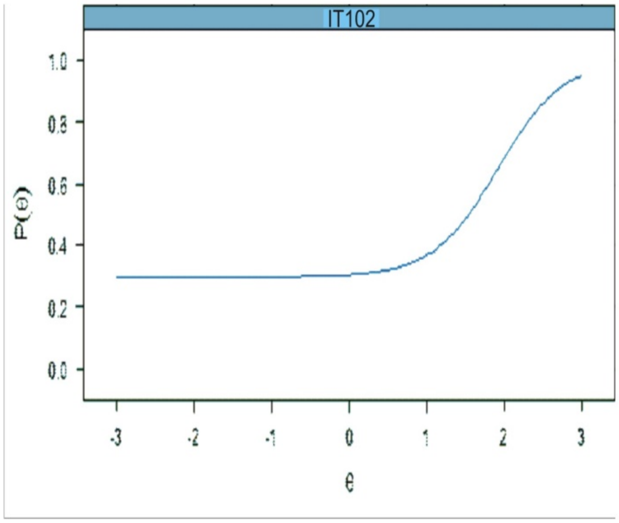

IT89
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
θ

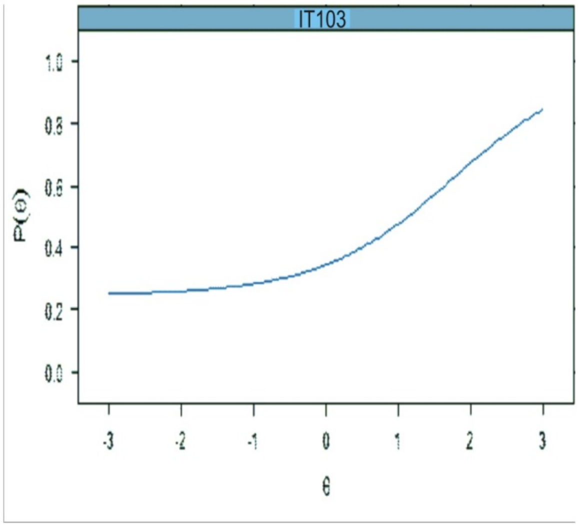

IT90
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

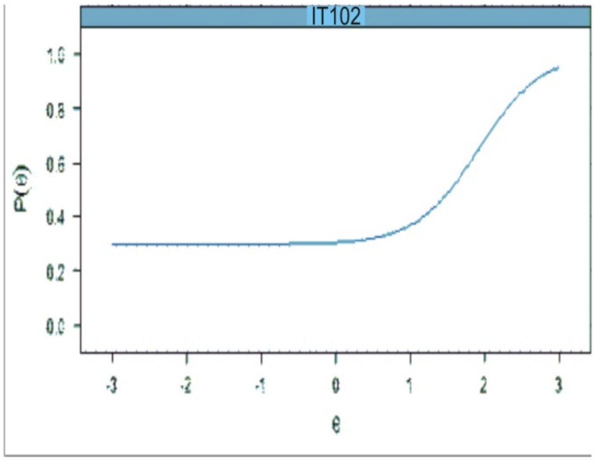

IT91
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

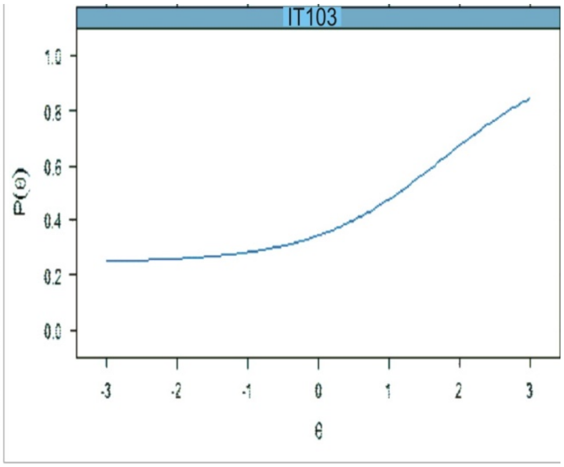

IT92
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

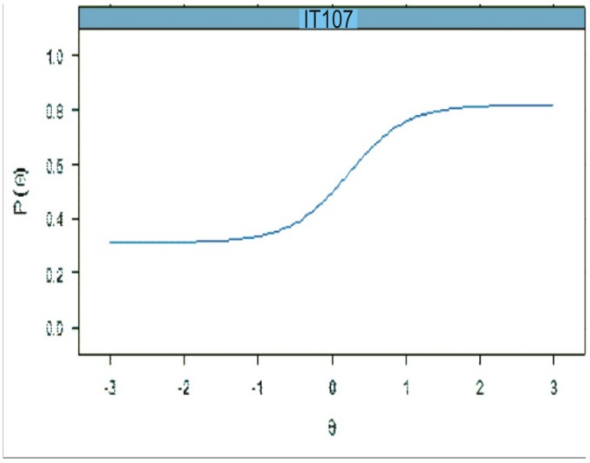

IT99
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

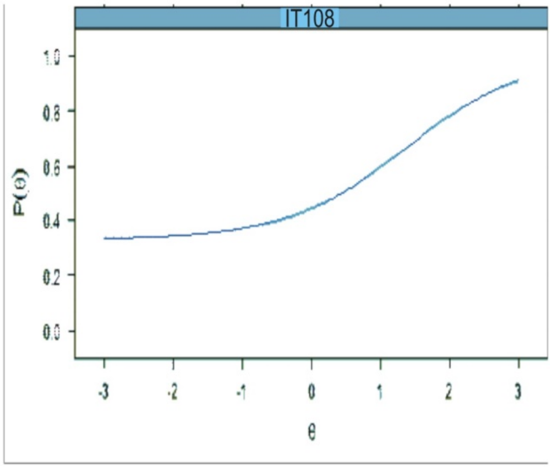

IT102
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

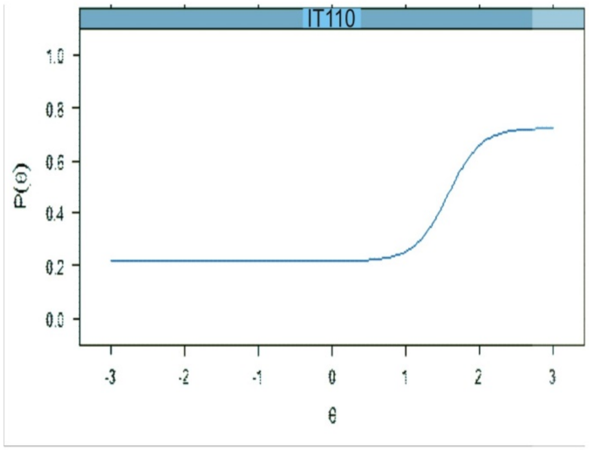

IT103
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

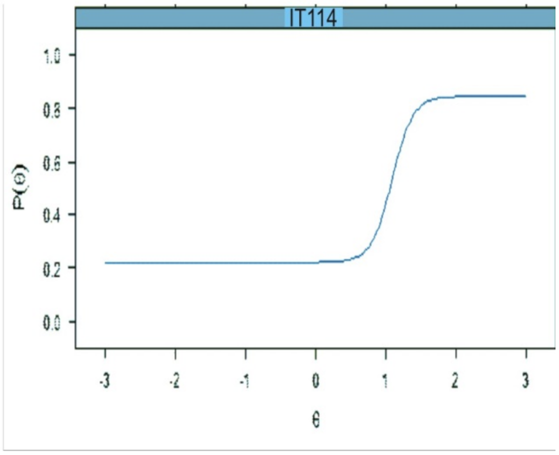

IT102
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ


IT103
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ


IT107
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ


IT108
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ


IT110
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ


IT114
P(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
θ

APPENDIX XI

21st September, 2018

TG/PS Education District 1,
Dairy Farm, Agege.
Lagos State.

Dear Sir/Ma,

LETTER OF INTRODUCTION - LAWAL R. Omonike (MATRIC NO: 125511)

I hereby wish to introduce the bearer who is a Post-Graduate Student (Ph.D) in the Institute of Education, University of Ibadan. Her research work is based on Assessment/Testing in Education which is targeted towards improving students' present level of performance both at school-based and external examinations levels.

She has been authorised to collect data on the research topic **"Application of 4-Parameter Logistic and Response-Time IRT Models in the Calibration of Senior Secondary School Computer-Based Mathematics Test in Southwest Nigeria.**

Due to the nature of her research work that involves the usage of Computer-Based Testing, the researcher will like to access Senior Secondary Schools with Computer Laboratories.

In view of this, I hereby solicit your kind support to ensure that approval is granted from your honourable office to access all the schools she intends collecting data for the study. I confirm that the data so collected will be treated with utmost confidentiality and used for research purpose only.

Kindly accord her the necessary assistance.

Yours Faithfully,

Prof. J. G. Adewale
Head of Unit, ICEE
Institute of Education.
08033263534

APPENDIX XII

17th October, 2018

Office of the Head of Service,
The Secretariat,
Alausa Ikeja,
Lagos State.

Dear Sir/Ma,

LETTER OF INTRODUCTION - LAWAL R. Omonike (MATRIC NO: 125511)

I hereby wish to introduce the bearer who is a Post-Graduate Student (Ph.D) in the Institute of Education, University of Ibadan. Her research work is based on Assessment/Testing in Education which is targeted towards improving students' present level of performance both at school-based and external examinations levels.

She has been authorised to collect data on the research topic **"Application of 4-Parameter Logistic and Response-Time IRT Models in the Calibration of Senior Secondary School Computer-Based Mathematics Test in Southwest Nigeria.**

Due to the nature of her research work that involves the usage of Computer-Based Testing, the researcher will like to access Senior Secondary Schools with Computer Laboratories.

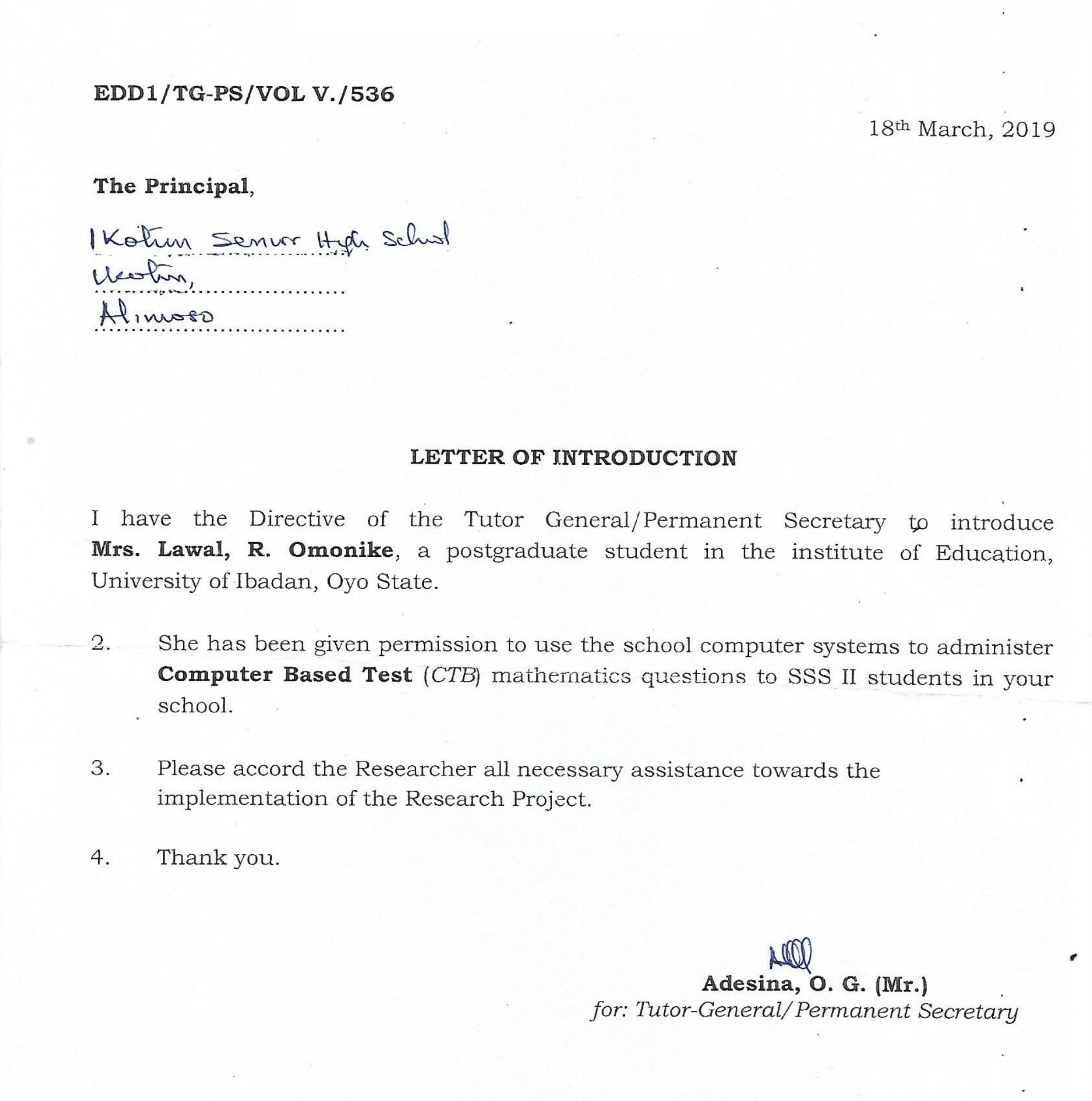
In view of this, I hereby solicit your kind support to ensure that approval is granted from your honourable office to access all the schools she intends collecting data for the study. I confirm that the data so collected will be treated with utmost confidentiality and used for research purpose only.

Kindly accord her the necessary assistance.

Yours Faithfully,

Prof. J. G. Adewale
Head of Unit, ICEE,
Institute of Education.
08033263534

APPENDIX XIII



EDD1/TG-PS/VOL V./536

18th March, 2019

The Principal,

Ikotun Senior High School
Ikotun,
Alimoso

LETTER OF INTRODUCTION

I have the Directive of the Tutor General/Permanent Secretary to introduce **Mrs. Lawal, R. Omonike**, a postgraduate student in the institute of Education, University of Ibadan, Oyo State.

2. She has been given permission to use the school computer systems to administer **Computer Based Test** (*CTB*) mathematics questions to SSS II students in your school.

3. Please accord the Researcher all necessary assistance towards the implementation of the Research Project.

4. Thank you.

Adesina, O. G. (Mr.)
for: Tutor-General/Permanent Secretary

APPENDIX XIV

The Institute of Education
University of Ibadan

17th May, 2019.

The Principal,
State Senior High School.
Oyewole.

Dear Sir,

LETTER OF APPRECIATION

On behalf of the researcher (Mrs Lawal Omonike) a PhD student of the Institute of Education, University of Ibadan, the Supervisor (Prof J. G. Adewale) and the entire University, we wish to acknowledge the permission granted by the school to make use of the ICT laboratory and the SSII students in carrying out data collection for this research work.

The researcher is also attesting to the support given by the Principal, Vice-Principals and the entire staff specifically the teacher in-charge of the ICT centre (Mr Oyebola) who out of his busy schedules attended to every need of the researcher in terms of putting all facilities in place and making sure that the Mathematics Computer-Based Test was successful.

The students' maximum cooperation and participation were indeed worth mentioning and they had in one way or the other benefitted from the training which will in-turn aid their performances positively either in school-based or external examinations in the nearest future. Their performances in the CBT exams were great as well. Thanks so much for making this aspect of the research work a reality. God bless you sir.

Yours faithfully.

Lawal R.Omonike.

APPENDIX XV

DIMTEST SUMMARY OUTPUT FOR THE POOLED CBMAT INSTRUMENT

```
  -------------------------------------------------
Number of Items Used:                114
Number of Examinees Used to
  Calculate DIMTEST Statistic:       731
  Minimum Cell Size for
Calculating DIMTEST Statistic:       2
  Number of Examinees After
  Deleting Sparse Cells:             719
  Proportion of Examinees Used to
  Calculate DIMTEST Statistic:       0.9836
  Number of Simulations Used
to Calculate TGbar:                  100
  Randomization Seed:                99991
  Estimate of Examinee
  Guessing on Test:                  0.2500
  -------------------------------------------------

     AT List                  PT List
  -------------------------------------------------
 2 3 4 5 7 9 10 11 12 13 15 16 18    1    6    8   14  17  21
 19 20 25 26 27 29 31 41 45 48 49   22   23   24   28  30  32
 50 51 58 59 60 61 65 66 68 71 72   33   34   35   36  37  38
 75 78 82 83 86 87 92 93 95 98 99   39   40   42   43  44  46
 100 101 102 105 106 107 109 112   47   52   53   54  55  56
 114                        57   62   63   64  67  69
                 70   73   74   76  77  79
                  80   81   84   85  88  89
                  90   91   94   96  97 103
                 104 108 110 111 113


TL=sum(TL,k)/sqrt(sum(S2,k)) {using original data}
TG=sum(TL,k)/sqrt(sum(S2,k)) {using simulated data}
TGbar = mean of ** TGs
T=(TL-TGbar)/sqrt(1+1/**)

          DIMTEST STATISTIC
  -------------------------------------------------
    TL      TGbar       T       p-value
  -------------------------------------------------
   7.4255    6.3919    1.0285    0.1518
```

APPENDIX XVI

Representative of the Item Local Independence Assessment of the pooled CBMAT

	1	2	3	4	5	6	7	8	9	10	11
1		0.051	-0.027	0.028	0.028	-0.025	-0.037	0.047	0.069	-0.013	-0.031
2	0.051		0.051	0.040	-0.002	-0.076	0.017	0.050	0.054	-0.030	-0.055
3	-0.027	0.051		0.039	0.020	-0.035	0.003	0.025	0.014	-0.081	0.042
4	0.028	0.040	0.039		-0.009	-0.097	-0.005	-0.053	-0.065	-0.017	0.031
5	0.028	-0.002	0.020	-0.009		0.016	0.002	-0.041	0.001	-0.068	0.029
6	-0.025	-0.076	-0.035	-0.097	0.016		-0.038	-0.090	0.030	0.058	0.046
7	-0.037	0.017	0.003	-0.005	0.002	-0.038		-0.013	0.051	-0.045	-0.025
8	0.047	0.050	0.025	-0.053	-0.041	-0.090	-0.013		0.030	0.029	-0.023
9	0.069	0.054	0.014	-0.065	0.001	0.030	0.051	0.030		-0.057	0.010
10	-0.013	-0.030	-0.081	-0.017	-0.068	0.058	-0.045	0.029	-0.057		0.053
11	-0.031	-0.055	0.042	0.031	0.029	0.046	-0.025	-0.023	0.010	0.053	
12	0.006	-0.004	-0.011	0.038	-0.004	-0.011	0.312	0.021	0.066	-0.012	0.032
13	0.010	0.010	0.041	-0.012	0.000	-0.050	-0.105	0.073	-0.030	0.037	0.035
14	0.026	0.050	-0.029	-0.026	-0.100	0.037	-0.001	0.006	-0.064	0.012	-0.031
15	-0.081	0.000	0.089	-0.087	0.000	0.046	-0.086	-0.091	-0.070	0.123	0.018
16	-0.031	0.005	-0.008	-0.069	0.021	0.028	0.025	0.051	0.014	0.036	-0.044
17	0.011	0.018	-0.109	0.014	-0.025	-0.033	-0.016	0.044	-0.054	-0.010	-0.026
18	0.060	0.017	0.013	0.035	-0.005	0.037	0.003	0.032	-0.041	0.029	0.067
19	0.057	0.066	0.019	0.038	-0.018	-0.011	-0.012	-0.056	0.045	-0.053	0.023
20	0.003	-0.034	-0.001	-0.056	0.013	-0.006	-0.028	0.063	0.087	0.092	-0.064
21	-0.006	0.026	0.060	-0.010	0.046	-0.014	-0.017	0.054	0.035	-0.101	0.068
22	0.056	-0.026	0.028	0.024	-0.040	-0.063	0.049	-0.070	0.010	0.024	-0.015
23	0.046	-0.058	-0.043	0.026	0.051	0.043	0.027	-0.033	0.014	-0.030	0.053
V53	-0.004	-0.016	0.015	-0.023	-0.028	-0.002	0.011	-0.023	0.005	0.037	-0.112
V54	-0.051	-0.088	-0.008	0.009	0.028	-0.012	-0.070	0.061	0.035	-0.044	0.024
V55	-0.005	-0.015	-0.083	-0.066	-0.032	-0.048	-0.047	-0.010	-0.039	0.044	-0.052
V56	-0.103	-0.003	-0.019	-0.003	-0.004	-0.015	-0.021	0.019	-0.043	0.022	-0.077
V57	0.008	-0.038	0.057	0.043	0.005	-0.072	0.069	-0.003	0.016	-0.011	0.005
V58	0.038	0.068	0.047	0.002	0.009	0.038	-0.017	0.024	0.040	0.033	0.047
V59	-0.051	-0.007	-0.109	0.033	0.006	0.015	-0.082	-0.030	-0.090	-0.004	-0.088
V107	-0.091	0.003	-0.030	0.010	0.044	-0.056	-0.112	0.012	0.018	-0.013	-0.066
V108	0.013	-0.034	0.047	-0.037	-0.052	-0.047	-0.015	-0.021	-0.060	0.056	-0.016
V109	-0.047	0.008	0.072	-0.087	0.006	-0.024	-0.009	0.010	-0.001	-0.014	-0.048
V110	-0.069	-0.025	-0.061	-0.088	0.023	0.010	-0.045	0.039	0.046	0.057	-0.014
V111	0.018	-0.082	0.016	0.019	-0.040	-0.026	-0.092	0.001	-0.077	0.075	0.053
V112	-0.042	-0.018	0.106	0.022	-0.024	0.000	-0.027	0.052	-0.015	0.028	-0.081
V113	-0.016	0.015	0.065	0.078	-0.006	-0.007	-0.029	0.046	0.030	0.039	-0.017
V114	-0.036	-0.026	0.017	-0.049	0.004	-0.039	0.017	-0.012	-0.009	-0.032	-0.049

APPENDIX XVII

Log-normal response time IRT modelling

```
LNIRT(RT, Y, data, XG = 1000, guess = FALSE, par1 = FALSE, residual = FALSE, td
=TRUE,WL = FALSE, alpha, beta, XPA = NULL, XPT = NULL, XIA = NULL, XIT =
NULL)
```

Arguments

RT a Person-x-Item matrix of log-response times (time spent on solving an item).

Y a Person-x-Item matrix of responses.

Data a list containing the response time and response matrices and optionally the predictors of both parameters of item and person.

XG the number of MCMC iterations to perform (default: 1000).

Guess include guessing parameters in the IRT model (default: false).

par1 use alternative parameterization (default: false).

residual compute residuals, requires > 1000 iterations (default: false).

Td estimate the time-discrimination parameter(default: true).

WL define the time-discrimination parameter as measurement error variance parameter (default: false).

Alpha an optional vector of pre-defined item-discrimination parameters.

Beta an optional vector of pre-defined item-difficulty parameters.

XPA an optional matrix of predictors for the person ability parameters.

XPT an optional matrix of predictors for the person speed parameters.

XIA an optional matrix of predictors of difficulty parameter estimates.

XIT an optional matrix for predictors for the time intensity parameters estimates.

APPENDIX XVIII
Examinees Parameter Estimates of the LNIRT Model

S/N	ABILITY	SPEED	S/N	ABILITY	SPEED	S/N	ABILITY	SPEED
1	0.9865277	0.134701	43	-0.2632291	0.882487	85	-0.6421883	-0.08051
2	-0.0748075	0.366629	44	0.5037541	0.093399	86	0.229878	0.640442
3	-0.2315218	0.044995	45	-0.2682245	-0.04233	87	0.1867756	0.018854
4	0.1879129	-0.12266	46	-0.0620407	1.131648	88	-0.0585331	0.10412
5	0.0877586	0.94866	47	0.0970531	-0.21198	89	-0.2352021	-0.33137
6	-0.3177967	0.344508	48	0.6554183	0.609657	90	-0.3970409	1.110571
7	0.2969874	-0.06416	49	0.0246653	0.285198	91	-0.1195436	0.911381
8	0.3231764	-0.22727	50	-0.1326875	0.044053	92	-0.4180508	0.295525
9	0.0600412	-0.23011	51	-0.1612834	0.033547	93	-0.4862121	0.869651
10	0.1646352	-0.15438	52	-0.4399214	0.493902	94	-0.4458569	0.252846
11	-0.05195	-0.23237	53	0.2873551	0.539993	95	-0.3264921	0.432988
12	0.1318013	0.068249	54	0.6779593	0.201341	96	-0.5150022	0.093089
13	-0.1645632	0.43556	55	0.0741149	0.255362	97	0.0042795	-0.03851
14	0.511433	-0.20526	56	-0.0980587	-0.25372	98	-0.2595366	-0.04404
15	0.112572	0.447084	57	-0.3439426	1.081828	99	-0.2395392	0.882038
16	-0.1075312	0.058743	58	-0.067057	0.108452	100	0.5121216	0.095601
17	-0.438224	-0.06404	59	0.3192013	-0.0903	101	0.0350833	0.322036
18	0.2372346	-0.03712	60	0.0990555	0.046632	102	-0.1587592	-0.06222
19	-0.1021678	1.123906	61	-0.36628	-0.07781	103	-0.4713605	0.050592
20	-0.2990273	-0.12051	62	-0.1735545	0.675285	104	-0.473591	0.390341
21	0.5959114	0.077508	63	0.1796657	-0.13566	105	-0.3504424	0.178735
22	0.4035387	-0.01091	64	-0.3762472	-0.37767	106	-0.3696559	0.119371
23	-0.0536223	0.115044	65	0.1206186	0.064922	107	-0.3170329	-0.08976
24	-0.0534776	-0.37618	66	0.0285618	-0.26418	108	0.3090082	0.50732
25	-0.1874876	-0.42214	67	-0.1794155	-0.19778	109	0.8020337	1.177284
26	0.1795302	0.017133	68	-0.0602109	0.104781	110	-0.0332548	0.447298
27	-0.245952	-0.334	69	-0.1964955	0.724952	111	0.0048486	0.808078
28	-0.1271619	-0.38683	70	-0.2393122	0.317631	112	-0.6279146	0.268513
29	-0.128137	-0.13763	71	0.3476796	-0.18794	113	-0.3089856	0.167046
30	0.338999	-0.33469	72	-0.4177038	-0.01168	114	-0.0930576	0.842946
31	0.1613478	1.056513	73	0.1397598	-0.01058	115	-0.226034	0.156783
32	-0.2762943	0.672673	74	-0.5030512	0.867606	116	-0.6084707	0.014665
33	0.2445177	0.108972	75	-0.0148358	0.60745	117	-0.5267725	0.995829
34	0.473111	-0.1729	76	-0.3924837	1.111732	118	-0.6572169	0.102052
35	0.3949361	-0.04859	77	-0.1765889	-0.1997	119	0.2304089	0.623196
36	-0.3081077	0.822784	78	-0.0929274	0.774858	120	-0.7237195	0.094507
37	0.4466921	0.301263	79	-0.2106533	0.507927	121	-0.0364329	-0.18448
38	-0.0964428	0.774455	80	-0.3857072	0.686598	122	-0.2298458	0.460435
39	0.6080198	0.076661	81	-0.0412466	-0.00397	123	-0.0491764	-0.21489
40	0.4041266	-0.01076	82	0.3421069	-0.18437	124	0.1786252	0.262954
41	0.2983343	0.929108	83	-0.3281349	0.059283	125	-0.0170824	-0.02995
42	-0.2039556	0.055267	84	-0.3466558	0.351776	126	-0.1814796	0.143088

S/N	ABILITY	SPEED
127	-0.1870443	0.253025
128	-0.6269449	0.03129
129	-0.4550238	0.110605
130	-0.2234491	0.007902
131	-0.4208738	-0.03196
132	-0.2066529	-0.25187
133	-0.0819993	0.104525
134	-0.2405581	0.059824
135	-0.0974165	0.138915
136	-0.1370387	0.20815
137	-0.3327174	-0.35741
138	-0.275483	0.315623
139	0.4911626	0.931267
140	0.0165138	-0.06841
141	-0.3841235	-0.21607
142	-0.2543586	-0.11895
143	-0.1445772	-0.10602
144	-0.1993885	-0.03125
145	0.1079884	-0.2535
146	0.0358467	0.184113
147	0.5281208	-0.04322
148	0.6594544	0.16847
149	-0.1103416	-0.04855
150	-0.3769883	-0.13096
151	-0.4925156	0.212282
152	-0.1463171	1.138607
153	0.6858848	1.174236
154	-0.0387753	-0.30249
155	-0.194394	-0.22203
156	-0.4210837	-0.25114
157	0.0701267	-0.07635
158	-0.6620708	-0.27205
159	-0.1278488	0.004715
160	-0.2462081	0.003749
161	0.0732396	-0.33742
162	0.0999674	-0.25458
163	-0.2760686	0.10956
164	-0.2954893	0.006441
165	0.3344589	0.098563
166	0.0984187	-0.09341
167	-0.3168495	-0.10751
168	0.1129349	-0.06737
169	-0.4477677	-0.04377
170	0.029382	-0.27823
171	-0.3685171	0.258468
172	0.1175584	-0.14605
173	-0.342848	0.25854
174	-0.358627	-0.35309
175	-0.3125724	-0.09458
176	0.3266848	0.304631
177	0.1101872	-0.10857
178	-0.0579908	-0.16176
179	0.3380062	-0.31061
180	-0.1446347	-0.30282
181	0.2517274	-0.04333
182	-0.3792345	-0.09231
183	0.5225344	0.14922
184	0.1266127	-0.16561
185	0.038585	-0.22183
186	0.204779	-0.25386
187	-0.469337	-0.1599
188	-0.1244595	-0.16931
189	-0.3503908	-0.06885
190	-0.1948503	-0.27847
191	0.1158244	-0.06764
192	-0.3249066	-0.01951
193	-0.0363329	-0.08775
194	0.7600975	-0.23716
195	-0.1937982	-0.11809
196	0.5818456	0.019203
197	0.310202	-0.28888
198	0.2686889	-0.05437
199	-0.4978858	-0.15965
200	0.4464261	0.099615
201	0.614261	-0.09608
202	0.0550755	0.413037
203	0.3101226	0.022475
204	0.01901	-0.03638
205	0.8029792	0.146774
206	0.0864233	-0.00673
207	0.1365195	-0.07573
208	1.093988	-0.01345
209	-0.5881457	-0.16162
210	-0.2901101	-0.28434
211	0.7680734	0.19135
212	-0.5914906	-0.17983
213	0.5345016	-0.09786
214	0.037119	-0.21306
215	-0.395672	-0.15548
216	0.223858	-0.34997
217	-0.0863933	-0.34449
218	-0.0861644	-0.42962
219	-0.4842999	-0.53387
220	-0.3175624	-0.34856
221	0.0418534	-0.27441
222	-0.1586659	-0.22431
223	-0.1719693	-0.11988
224	-0.3197939	1.063939
225	0.0473438	0.955379
226	-0.3706475	-0.09149
227	-0.1049933	-0.3036
228	-0.3801047	-0.07354
229	0.1365706	-0.09171
230	-0.5715757	-0.10481
231	-0.5222824	-0.17798
232	0.1449673	-0.30195
233	0.1118395	-0.31854
234	0.4357768	-0.16078
235	0.1780539	-0.29016
236	-0.1770654	0.436211
237	-0.2960154	-0.30913
238	-0.1116694	-0.19364
239	-0.3369063	-0.41531
240	-0.2249498	0.291535
241	-0.1866309	-0.25384
242	0.1163936	0.88804
243	-0.0548828	0.883032
244	0.1147496	-0.04889
245	-0.2198557	1.139461
246	0.1325899	-0.2015
247	0.5667496	-0.17657
248	0.68407	0.045118
249	-0.1957068	-0.25193
250	-0.1450939	-0.22511
251	-0.375653	1.105188
252	1.420279	0.005325

S/N	ABILITY	SPEED
253	-0.1225414	-0.07942
254	0.6010941	-0.17184
255	0.2532784	-0.03743
256	0.6351344	0.188019
257	-0.0908399	-0.05432
258	1.232302	-0.15093
259	0.1880013	-0.11692
260	0.8549335	1.013259
261	0.4393436	-0.06407
262	-0.3831465	0.448771
263	1.157551	-0.21394
264	1.183894	0.376841
265	0.1387431	-0.18735
266	0.4495934	-0.13941
267	0.1732636	-0.1445
268	0.7786796	0.015744
269	-0.3129866	0.057793
270	0.0071523	0.002821
271	0.9418782	-0.10249
272	0.200566	-0.3183
273	0.4893661	-0.128
274	-0.2048153	-0.36611
275	-0.1702463	-0.32696
276	0.0547749	-0.26908
277	-0.5180115	-0.41457
278	-0.4799271	-0.08415
279	0.2946906	-0.23576
280	-0.0612542	0.53017
281	-0.0605731	-0.37679
282	0.8891717	-0.12203
283	0.6002738	-0.1873
284	0.1278998	-0.04505
285	-0.059799	-0.12977
286	-0.4827694	-0.19981
287	-0.2930617	-0.13137
288	-0.0370279	-0.02001
289	0.932018	-0.12194
290	-0.2705738	0.258678
291	-0.075243	-0.24406
292	0.296944	-0.09065
293	-0.2742426	0.3691
294	0.2390795	-0.09943
295	-0.5966012	0.063654
296	0.2919534	-0.12095
297	-0.364225	-0.10618
298	1.213258	0.062732
299	-0.7832281	-0.02728
300	-0.2056432	0.066351
301	0.2966038	-0.35953
302	-0.0177915	-0.4612
303	1.494983	0.575904
304	-0.1983799	0.065889
305	0.4156176	-0.00868
306	-0.0450565	-0.00986
307	-0.3922005	-0.02597
308	-0.5857238	-0.2871
309	0.0529311	0.348041
310	0.0669034	-0.32455
311	0.5313607	0.033848
312	-0.4979716	-0.13021
313	-0.2805241	0.081863
314	-0.1878847	-0.152
315	0.0889885	-0.07578
316	-0.2232561	-0.09653
317	-0.4968958	-0.09536
318	0.2453806	-0.01062
319	0.1845593	0.060572
320	-0.1771749	-0.09665
321	0.5873566	-0.28495
322	-0.1599822	-0.36934
323	0.5555324	-0.09326
324	0.1770456	-0.18892
325	0.2639562	-0.22086
326	1.362485	0.225217
327	0.6470574	-0.00302
328	0.5950057	0.0002
329	0.7336404	-0.22218
330	0.5838946	0.021878
331	0.453063	-0.09708
332	-0.093833	0.127717
333	-0.2852209	0.061776
334	-0.2750342	0.029028
335	-0.2579293	-0.29053
336	0.5111878	0.072151
337	0.3300425	-0.14622
338	0.9319353	0.01491
339	-0.2477368	0.451749
340	-0.0076095	-0.22533
341	-0.5024199	-0.1253
342	-0.3156859	-0.109
343	0.1075707	-0.06725
344	0.0821603	0.112066
345	1.57922	0.048975
346	0.03877	-0.38432
347	0.178047	-0.20214
348	1.299936	0.036704
349	-0.0413072	-0.28295
350	0.1840999	0.027883
351	0.3692361	0.091526
352	-0.2191287	-0.33599
353	-0.1990695	0.182311
354	0.2854087	0.023473
355	0.1973352	0.076571
356	0.2012337	0.599665
357	0.5419184	-0.07397
358	0.0360999	1.140815
359	-0.5805568	-0.06892
360	0.748055	-0.11711
361	0.1237101	0.049173
362	-0.0217051	-0.15552
363	0.6393555	0.01083
364	0.0611456	-0.01594
365	0.1903145	-0.10593
366	0.0424394	0.063644
367	0.0889045	-0.31422
368	-0.0627846	-0.29598
369	0.2830578	-0.16104
370	-0.3826689	-0.36563
371	0.4517313	-0.15634
372	-0.1926518	-0.17528
373	0.2642196	-0.38687
374	-0.0212605	-0.25699
375	0.8857288	-0.22794
376	0.2140026	-0.25091
377	-0.0960932	0.040983
378	0.2178529	-0.25056

S/N	ABILITY	SPEED	S/N	ABILITY	SPEED	S/N	ABILITY	SPEED
379	0.0915385	-0.08089	421	-0.2718131	-0.08223	463	0.3053831	0.096971
380	-0.2322019	0.047796	422	-0.0022118	-0.04708	464	-0.0872994	0.050611
381	0.0292118	-0.30687	423	-0.0545713	-0.19169	465	-0.1255235	0.161975
382	-0.2348062	0.071672	424	0.2739011	-0.31586	466	-0.1434812	0.072972
383	0.5006123	-0.01158	425	0.0387376	-0.16873	467	-0.2445244	0.109243
384	-0.0452712	-0.26632	426	-0.2531896	0.590184	468	-0.203453	-0.11315
385	0.3259247	0.112695	427	-0.2201474	0.014057	469	-0.4304397	0.071259
386	-0.1267705	0.132055	428	0.0103667	-0.16704	470	0.220547	0.235782
387	0.3355681	-0.15567	429	-0.5270929	0.439429	471	-0.1437188	-0.14282
388	-0.1258046	0.081823	430	0.4471444	-0.25646	472	0.0446433	-0.03331
389	-0.2578026	0.033685	431	0.1561348	-0.01504	473	-0.0135675	0.081444
390	-0.4062277	-0.09704	432	-0.176453	-0.02835	474	0.4916407	0.214313
391	0.9961199	-0.07732	433	0.1856341	-0.37969	475	0.9566717	-0.12844
392	0.1439309	-0.27532	434	1.476816	0.037113	476	-0.2841164	-0.03194
393	0.3546178	0.105691	435	0.1285226	0.05295	477	-0.4197868	0.028021
394	0.2662154	-0.02603	436	0.5248399	-0.08007	478	0.1879551	-0.09033
395	-0.0725163	-0.00768	437	0.6939419	-0.23332	479	0.0559119	-0.21984
396	0.3416395	-0.37105	438	-0.2076234	0.003402	480	-0.4616864	-0.14089
397	0.6506665	-0.00156	439	-0.0485098	0.954073	481	-0.2548513	-0.18952
398	0.4852795	-0.09363	440	0.4296982	-0.34342	482	0.4838006	-0.0976
399	0.2712184	-0.0718	441	0.3220763	0.129236	483	-0.2435807	-0.29236
400	0.3976026	-0.24102	442	0.231738	-0.05003	484	0.6878939	0.166266
401	-0.1749666	0.189036	443	0.3113967	-0.28146	485	0.2539537	-0.03959
402	-0.1700162	0.114341	444	0.6195596	-0.19444	486	0.1978635	-0.13447
403	0.2816474	-0.31733	445	0.6479342	-0.13138	487	-0.5131893	-0.26449
404	-0.3041458	-0.17802	446	-0.3250461	0.077967	488	-0.468807	-0.18183
405	-0.2672589	-0.03687	447	-0.60808	-0.29138	489	-0.397361	0.245246
406	0.9939279	-0.10182	448	-0.0567305	-0.08867	490	-0.0352362	0.032381
407	0.0684977	-0.20774	449	0.4735808	-0.07044	491	0.4404778	-0.11209
408	0.8068632	-0.15138	450	-0.2852948	0.287539	492	-0.198287	-0.04105
409	0.8115243	-0.0995	451	0.7560071	-0.1381	493	0.2586183	0.246428
410	-0.8097406	-0.03901	452	0.9189155	-0.26662	494	-0.0775332	0.018305
411	-0.174552	0.100943	453	-0.0067535	0.124651	495	-0.4304272	-0.09291
412	-0.5094985	-0.27722	454	0.0278258	-0.16099	496	-0.5404821	-0.1266
413	-0.0557265	-0.2239	455	0.2713054	-0.20165	497	0.0975841	0.099436
414	-0.5001544	-0.15325	456	0.8500673	0.229309	498	-0.017551	-0.0228
415	-0.0052131	-0.07475	457	-0.0378583	-0.06929	499	0.0901782	-0.13329
416	0.8135044	-0.13523	458	0.318053	-0.05782	500	0.4157316	-0.24524
417	0.0161615	-0.11652	459	0.3354112	-0.05582	501	-0.2470551	-0.03859
418	0.06726	-0.0397	460	0.4543123	-0.065	502	0.9947566	-0.10397
419	0.5889087	-0.0821	461	-0.533342	-0.01023	503	0.0615184	-0.20247
420	0.7331544	-0.21119	462	0.1029135	-0.16588	504	9.12E-06	-0.25602

S/N	ABILITY	SPEED
505	0.6148834	-0.12359
506	0.3901293	-0.24896
507	-0.0284832	-0.25478
508	-0.1317611	-0.02072
509	-0.230911	0.19112
510	-0.2680148	-0.03105
511	0.2644378	-0.20639
512	0.2694399	-0.19103
513	-0.3084895	-0.14827
514	-0.3007624	-0.32993
515	-0.4795957	-0.12104
516	0.0704445	-0.31884
517	0.6084305	0.422523
518	0.3836965	-0.11017
519	0.0758206	0.404257
520	0.5940163	-0.03386
521	0.153687	-0.19469
522	0.0828525	-0.04948
523	-0.3745401	0.352357
524	-0.4614077	0.392086
525	0.8335116	-0.19907
526	0.1178455	0.083533
527	-0.5448363	0.730627
528	-0.4450564	0.394217
529	0.6064027	-0.01544
530	0.506241	0.019805
531	0.5107657	-0.2107
532	0.7470387	-0.00104
533	-0.1484998	-0.07789
534	0.660006	0.154084
535	-0.2605005	-0.1838
536	1.217023	-0.08677
537	0.0598878	0.018682
538	-0.2196022	-0.09305
539	-0.0818286	-0.29065
540	0.1277098	0.046786
541	-0.1401877	-0.22782
542	-0.3233872	-0.4455
543	0.1327677	-0.11479
544	0.7638324	-0.3054
545	0.9077615	-0.1629
546	-0.5764651	-0.05667
547	-0.4389984	0.170416
548	0.1562355	-0.06748
549	0.7966511	-0.12311
550	-0.4528621	0.037447
551	0.5442017	-0.24658
552	0.1324481	-0.18761
553	0.159477	-0.25082
554	0.2581825	-0.38909
555	-0.1405278	0.072543
556	-0.2590177	0.552448
557	0.0994072	-0.21081
558	-0.2899375	-0.14211
559	0.3426829	0.102298
560	0.3637613	0.085912
561	-0.0693173	0.492996
562	0.9789212	-0.20809
563	1.064282	-0.00433
564	-0.3263518	-0.14237
565	-0.0457913	-0.07888
566	0.729483	0.169863
567	0.6516006	-0.18476
568	-0.3346758	0.21888
569	0.3572883	-0.15275
570	0.6286271	-0.36542
571	1.032857	0.011265
572	-0.2394381	-0.0825
573	-0.4109241	-0.12609
574	0.909187	0.050924
575	0.374061	-0.21619
576	0.0193227	-0.21607
577	-0.0738072	-0.11984
578	0.2343855	0.038559
579	0.2305572	-0.28586
580	0.3016547	-0.34198
581	-0.4957139	-0.12264
582	0.9084496	-0.25867
583	-0.1522268	-0.07819
584	0.6688661	0.154336
585	-0.261217	-0.18441
586	1.211546	-0.09044
587	-0.0579678	0.138166
588	0.2645024	-0.15394
589	0.0761258	0.110608
590	-0.12203	-0.17734
591	0.6326297	-0.36837
592	1.039604	0.009677
593	-0.2309346	-0.08366
594	0.1411394	-0.13441
595	0.5781022	0.23466
596	0.2075352	-0.20839
597	-0.5434747	-0.27629
598	-0.4904794	0.061074
599	-0.4254485	-0.50722
600	-0.2629368	0.166261
601	0.0578271	0.021221
602	-0.2220007	-0.0924
603	-0.0903642	-0.28937
604	0.1286455	0.046681
605	0.0643177	-0.02805
606	0.4436465	-0.16696
607	-0.4748365	-0.18216
608	-0.2042583	-0.07701
609	-0.1041159	-0.20037
610	0.046444	-0.26183
611	-0.2144476	-0.17768
612	-0.0673286	-0.21258
613	0.2985775	-0.17283
614	-0.0141066	0.212545
615	0.2142324	-0.37701
616	-0.0212096	-0.41668
617	0.0477392	0.047296
618	0.7516795	0.023854
619	0.061295	-0.21744
620	-0.6713455	-0.49145
621	0.2431385	0.010269
622	0.2770558	-0.02281
623	-0.6037086	-0.16438
624	-0.0897007	-0.07089
625	0.0105733	-0.1178
626	-0.4332288	-0.22262
627	-0.3780948	0.011266
628	-0.0295313	-0.07463
629	-0.2734417	1.221126
630	1.10256	-0.01111

S/N	ABILITY	SPEED
631	0.5948786	0.136705
632	-0.2135856	-0.02206
633	-0.2113855	-0.0213
634	0.1770808	-0.08894
635	-0.3481714	0.000593
636	-0.4805782	-0.04533
637	-0.2827758	-0.27337
638	0.1345372	0.215446
639	0.0130026	-0.40395
640	-0.1594672	-0.1471
641	-0.2754134	0.284526
642	-0.4377731	-0.30371
643	-0.1578655	0.040308
644	-0.4143273	0.063178
645	-0.3050272	0.166829
646	-0.322042	0.33839
647	1.378064	-0.01444
648	0.1072678	0.078452
649	0.2197594	-0.2713
650	-0.1063268	0.007719
651	0.3527628	0.198815
652	0.3944889	-0.15191
653	-0.4703433	-0.04966
654	0.132926	-0.38953
655	0.0062657	0.080789
656	0.8946259	0.149401
657	-0.2072086	0.14567
658	-0.0137835	-0.39934
659	0.0598318	-0.03101
660	0.4413348	0.008702
661	-0.1296163	0.083638
662	-0.4162009	-0.24712
663	-0.0743566	-0.05469
664	-0.4513243	-0.35159
665	-0.3109582	-0.00182
666	0.1150803	-0.24019
667	-0.1391375	-0.18739
668	0.2813832	-0.03946
669	-0.0037753	-0.07044
670	0.0023563	-0.06017
671	-0.2415348	-0.14187
672	-0.0912403	0.391571
673	-0.3505552	-0.17442
674	-0.1083125	-0.37196
675	0.4559986	-0.17061
676	-0.319513	-0.02592
677	-0.3443615	-0.17307
678	0.2946184	-0.16833
679	-0.5895741	0.926785
680	0.5854543	-0.21667
681	-0.3948071	0.064533
682	-0.4316634	0.039332
683	-0.287051	0.614525
684	-0.2812235	-0.21445
685	0.0611722	0.548936
686	-0.4326045	0.224959
687	-0.2641621	0.016379
688	-0.3152623	0.231073
689	-0.37185	-0.29592
690	-0.3570792	-0.51093
691	-0.1240686	0.057588
692	-0.1714454	0.230175
693	0.1137923	-0.08091
694	0.0271253	-0.01696
695	0.6520237	-0.25515
696	-0.3115816	0.015933
697	-0.4038826	0.006033
698	-0.0355982	-0.16602
699	0.3162554	-0.25958
700	0.0852602	0.06527
701	-0.1824477	0.025215
702	-0.086636	-0.11025
703	-0.3750087	0.048539
704	-0.0765525	0.248071
705	-0.5578535	-0.08303
706	-0.0783685	0.546424
707	-0.4159195	0.004213
708	-0.0889696	0.053733
709	-0.1217211	-0.25713
710	-0.0291967	0.692407
711	-0.1875672	-0.42286
712	-0.2399714	0.128654
713	-0.2167909	-0.1609
714	-0.2379344	0.21
715	-0.155383	-0.28206
716	-0.4942012	-0.27386
717	-0.2062699	-0.20326
718	-0.1821098	0.045349
719	-0.1800479	-0.35132
720	0.1461783	0.073655
721	-0.3818606	0.326742
722	-0.5164863	0.204165
723	-0.0707863	-0.07827
724	-0.4971329	0.100821
725	-0.1238362	-0.25985
726	0.0003272	-0.03382
727	-0.4008623	-0.12904
728	-0.4323692	0.250129
729	-0.5111966	-0.02414
730	-0.193509	-0.09105
731	-0.2286228	-0.23814
732	-0.5410435	-0.08378
733	-0.4851935	-0.02553
734	-0.013255	-0.14967
735	-0.5110066	0.177581
736	-0.4630961	0.048249
737	0.0597582	-0.39475
738	0.0582046	-0.04176
739	-0.5621066	-0.11933
740	-0.4199724	-0.39676
741	0.1619008	-0.03668
742	-0.30747	-0.07085
743	-0.0958741	-0.26182
744	-0.229454	0.098544
745	-0.3858385	-0.07219
746	-0.3810136	0.120687
747	-0.3013009	-0.1985
748	-0.3406479	-0.25006
749	-0.496453	-0.0334
750	-1.108126	-0.0725
751	-0.2117993	-0.18074
752	0.0639847	0.12352
753	-0.5030479	-0.67622
754	-0.0257801	-0.08361
755	-0.0784755	-0.30237
756	-0.3751986	0.119761

S/N	ABILITY	SPEED
757	-0.3109855	-0.02279
758	-0.0184862	-0.12724
759	-0.0336321	-0.10786
760	-0.6280277	-0.23007
761	-0.7289122	0.067494
762	0.4942165	-0.35145
763	0.074747	-0.30297
764	-0.6530263	-0.22585
765	-0.1446049	-0.14934
766	0.8335332	-0.04613
767	0.2394223	-0.20928
768	-0.6145036	-0.34266
769	-0.1606463	0.092683
770	0.1237597	-0.3382
771	-0.2740328	0.001986
772	-0.1403068	-0.24835
773	0.1295701	0.103575
774	0.0398715	0.238633
775	-0.4678027	-0.35878
776	0.0955458	0.277846
777	1.070518	-0.26965
778	-0.1844642	-0.34882
779	0.1634148	-0.20059
780	0.0684764	0.069026
781	-0.2671329	-0.25327
782	-0.5503118	0.380602
783	0.1526925	-0.11469
784	0.0777323	-0.00256
785	0.3643239	-0.10253
786	0.7613538	-0.18821
787	-0.3220594	-0.06536
788	-0.1447299	-0.32853
789	-0.0031259	0.221706
790	0.1899578	0.204655
791	0.2065086	-0.08702
792	-0.4171732	-0.35917
793	-0.524149	0.069069
794	-0.5798879	0.014182
795	-0.4629419	0.376419
796	-0.4331093	0.028365
797	0.4421162	-0.1522
798	-0.2308001	-0.03491
799	-0.2084375	-0.30399
800	-0.3981283	2.095571
801	-0.2543637	1.058598
802	0.537631	-0.117
803	-0.4659775	-0.24785
804	0.3463918	0.737728
805	-0.4549872	-0.06147
806	0.1241394	0.142917
807	-0.4494025	-0.11972
808	-0.511352	-0.14649
809	-0.0083248	-0.25472
810	-0.1305202	0.058751
811	-0.2703001	-0.22029
812	-0.1813245	-0.53343
813	-0.2181102	-0.24699
814	-0.4168764	0.24208
815	-0.6337759	0.046312
816	-0.2534779	0.636395
817	-0.429635	-0.26064
818	-0.6163205	0.324846
819	-0.5613752	0.015354
820	-0.4570088	0.377658
821	-0.4172251	0.030051
822	0.442847	-0.1504
823	-0.2273199	-0.03075
824	-0.2575744	0.050742
825	-0.6370277	0.025284
826	-0.2076225	0.30926
827	-0.1874065	0.508522
828	-0.4217645	-0.03561
829	-0.0348577	0.119571
830	-0.4191914	-0.33297
831	-0.4069475	0.351759
832	0.0080748	-0.11679
833	-0.4275212	0.381649
834	0.5988663	-0.36653
835	0.099015	-0.22666
836	-0.2966173	-0.30391
837	0.7830119	-0.09279
838	-0.0716715	-0.12034
839	-0.4812183	0.135041
840	-0.4612924	-0.13267
841	-0.3836505	0.118037
842	-0.3077512	-0.02221
843	-0.020339	-0.12677
844	-0.0345872	-0.10712
845	-0.6518856	-0.23137
846	-0.4335602	-0.23791
847	-0.2041616	-0.2538
848	-0.2930038	-0.21027
849	-0.5634815	-0.14485
850	-0.3249277	-0.22123
851	-0.27106	0.507094
852	-0.4638158	-0.39812
853	-0.5370444	-0.05105
854	0.1491251	-0.22426
855	-0.4666274	0.072188
856	-0.1547891	0.135329
857	-0.223149	-0.34125
858	-0.6795247	0.027919
859	-0.2083842	1.173432
860	0.2722415	-0.23746
861	-0.2274772	-0.0285
862	-0.4975951	-0.47672
863	0.6102766	-0.2803
864	-0.1100748	0.419288
865	-0.0350924	-0.18576
866	-0.4894606	0.361983
867	-0.4467788	0.275314
868	-0.4676011	-0.27705
869	-0.1080641	-0.22242
870	-0.5373834	0.419381
871	-0.764577	0.786713
872	-0.4200359	-0.07868
873	-0.0963811	-0.36218
874	-0.7231757	-0.02073

APPENDIX XIX

```
        Log-Normal RT-IRT Modeling
        Summary of Results

        *Individual Appropriateness Assessment (Log-Normal Speed)*

         Proportion of Outliers of Examinees at 0.05 level
lZ
         14.53 %
         95% Posterior Probability:  13.04 %

         *Assessment of Question Appropriateness*
           Abnormal Questions at 0.05 level
           Any question is not found

           *Residual Analysis *
            Percentage Extreme Residuals (.95 Posterior Probability)
            0.1173 % (general average across persons and items)

                Extreme Residuals
                Person    Item RT
                  29       16    748.0000
                  51        5      3.0000
                 109       32     65.1000
                 119       39    193.1000
                 139        1    ***.0000
                 151       25      2.0000
                 153       27    378.0000
                 180       14      3.1000
                 205       36    830.0000
                 207       23    688.1000
                 214       20    518.1000
                 221       11    492.1000
                 224       32    257.0000
                 225       40    589.0000
                 240       39    268.0000
                 252       12    430.1000
                 260       13     66.1000
                 290        1    ***.0000
                 309       32    200.1000
                 312       21      3.0000
                 321       39      2.0000
                 372       13      3.1000
                 388       21      3.0000
                 395        1    ***.1000
                 405       23      3.0000
```

411	39	550.1000
439	1	***.1000
445	8	3.0000
474	4	2.0000
501	23	3.0000
580	37	3.0000
580	38	2.0000
626	38	1.0000
679	37	325.1000
683	1	***.1000
686	39	489.1000
722	39	515.1000
742	32	282.0000
779	38	3.0000
800	1	***.0000
859	1	550.1000

Test of Kolmogorov Smirnov at 0.50 significant level
An asymmetric distribution of the underlying residuals shows 52.5% for items

Question	P-value
1	0.004
4	0.000
7	0.000
8	0.000
10	0.000
11	0.011
16	0.014
18	0.001
19	0.000
20	0.023
21	0.001
22	0.044
23	0.044
28	0.003
29	0.041
30	0.000
31	0.034
32	0.003
37	0.022
38	0.017
39	0.001

Individual Appropriateness Assessment (IRT Model for Ability)

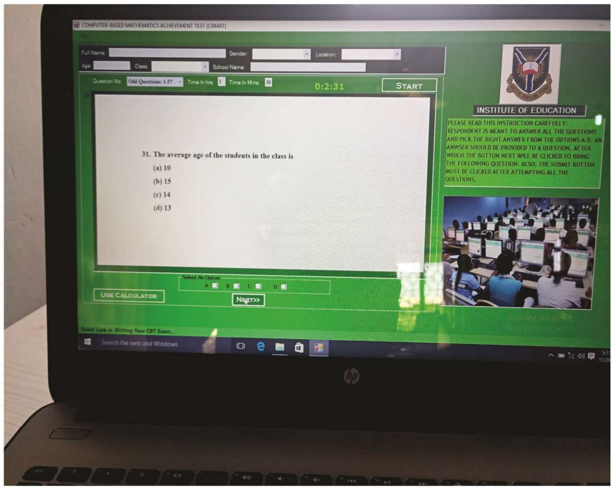
```
 Proportion Outliers Individuals at 0.05% level
  Log-likelihood Estimate
  0.11 %
  95% Posterior Probability:  0 %
  95% Posterior Probability (Ability and Speed):  0 %

*Assessment of Question Appropriateness*
 Abnormal Questions at 0.05 level
 Any question is not found

*Residual Assessment*
  Proportion of Extreme Residuals (95% of Subsequent Chance)
  0.0229 % (broad mean across individuals and responses)
Extreme Residuals
 Individual Question Reaction EAP Theta
     85         24          1        -0.6307
    230         24          1        -0.5738
    598         24          1        -0.5130
    724         24          1        -0.4897
    760         24          1        -0.6590
    782         24          1        -0.5596
    845         24          1        -0.6315
    871         24          1        -0.7791
Test of Kolmogorov Smirnov at 0.50 significant level
An asymmetric distribution of the underlying residuals is shown by 70% of the items
Question P-value
      1      0.000        28       0.000
      4      0.001        29       0.000
      5      0.000        30       0.000
      6      0.000        32       0.000
      7      0.001        33       0.007
      9      0.001        34       0.000
     12      0.000        36       0.000
     13      0.000        37       0.000
     14      0.000        38       0.000
     15      0.001        40       0.000
     16      0.000
     17      0.000
     18      0.000
     22      0.005
     23      0.000
     24      0.000
     26      0.000
     27      0.001
```

APPENDIX XX

Installation at Islamic High School, Basorun.



COMPUTER-BASED MATHEMATICS ACHIEVEMENT TEST (CBMAT)

Question No: Odd Questions: 1-57 Time In hrs: 1 Time In Mins: 30

0:2:31

31. The average age of the students in the class is

(a) 10

(b) 15

(c) 14

(d) 13

INSTITUTE OF EDUCATION

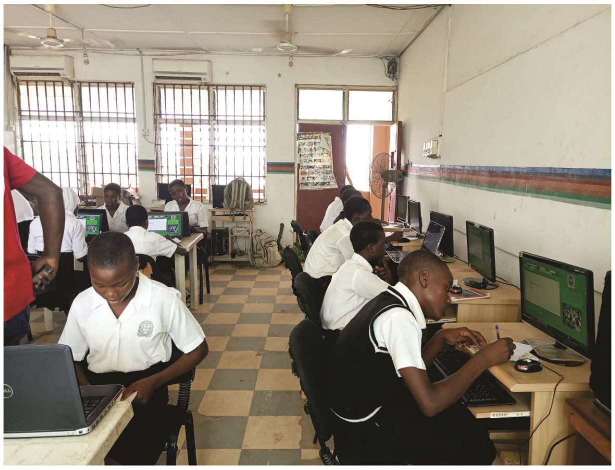
PLEASE READ THIS INSTRUCTION CAREFULLY: RESPONDENT IS MEANT TO ANSWER ALL THE QUESTIONS AND PICK THE RIGHT ANSWER FROM THE OPTIONS A-D. AN ANWSER SHOULD BE PROVIDED TO A QUESTION, AFTER WHICH THE BUTTON NEXT WILL BE CLICKED TO BRING THE FOLLOWING QUESTION. ALSO, THE SUBMIT BUTTON MUST BE CLICKED AFTER ATTEMPTING ALL THE QUESTIONS.

Good Luck in Writing Your CBT Exam...

CBMAT work environment

Testing at Abadina College

Testing



Testing at Abadina College

Testing at Anglican Grammar School

Olivet Baptist High School, Oyo.

Testing at Olivet Baptist High School, Oyo.

Testing at Olivet Baptist High School, Oyo.

The Researcher & Research Assistants with the School Vice-Principal and Computer Teacher

Testing in Progress

Testing at Loyola College

Ibadan Grammar School ICT Centre, Molete.

CBMAT in progress at Ibadan Grammar School.

Government College, Ibadan.

Testing at Government College, Ibadan.

Girls Senior High School, Agege.

Testing in progress

Respondents awaiting CBMAT

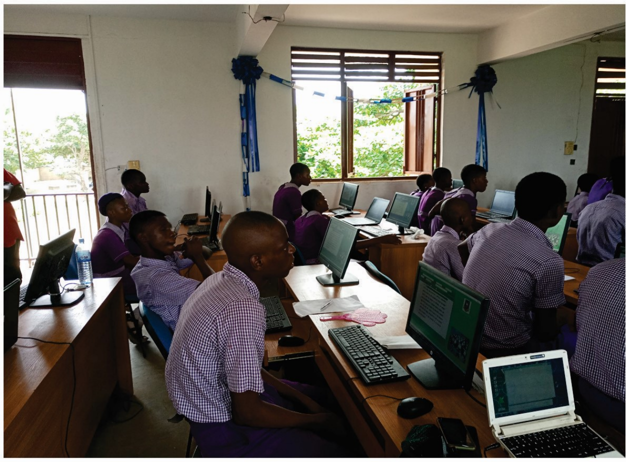
Testing at Government Senior College, Agege

CBMAT in progress, St. Anne's School, Molete

Sonmori Senior Comprehensive High School, Ifako

VETLAND JUNIOR/SENIOR GRAMMAR SCHOOL IFAKO-IJAYE
Powered by
LAGOS SOLAR
LAGOS STATE GOVERNMENT
Lagos Solar is a Lagos State Government Energy Initiative that will provide sustainable solar power to all public, rural, riverine and boarding secondary schools and healthcare facilities

Vetland Grammar School, Ifako-Ijaye



Instruction mood

Keke Senior High School, Ifako-Ijaye

CBMAT in progress

CBMAT in progress at Bishop Philip's Academy, Iwo-Road

Abiodun Atiba Memorial Academy, Awe

CBMAT in progress at Islamic High School, Basorun

CBMAT in progress at Abadina College, University of Ibadan