COMPARISON OF TECHNIQUES FOR ESTIMATING MODEL-FIT OF ITEM RESPONSE THEORY USING NBTCE 2018 MATHEMATICS MULTIPLE CHOICE TEST ITEMS.

Eucharia Ekene OKOYE

FEBRUARY 2022

COMPARISON OF TECHNIQUES FOR ESTIMATING MODEL-FIT OF ITEM RESPONSE THEORY USING NBTCE 2018 MATHEMATICS MULTIPLE CHOICE TEST ITEMS.

Eucharia Ekene OKOYE
PG/EDU8902708
B.Sc.Ed, M.Ed (Benin)

A THESIS WRITTEN IN THE DEPARTMENT OF EDUCATIONAL EVALUATION AND COUNSELLING PSYCHOLOGY AND SUBMITTED TO THE SCHOOL OF POST GRADUATE STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN MEASUREMENT AND EVALUATION OF THE UNIVERSITY OF BENIN, BENIN CITY NIGERIA.

FEBRUARY 2022

CERTIFICATION

We, the undersigned, certify that this thesis was carried out by Eucharia Ekene OKOYE in the Department of Educational Evaluation and Counselling Psychology, Faculty of Education University of Benin, Benin city Nigeria.

_____ _____

Prof. O.K. Omorogiuwa **Date**
Chief Supervisor

_____ _____

Prof. A.U. Osunde **Date**
Co-supervisor

_____ _____

Dr. (Mrs.) M.U Orheruata **Date**
Head of Department

DEDICATION

This thesis is dedicated to Almighty God for His infinite mercy upon me and my family and granting me His grace throughout the course of this programme. To Him, be all praise, honour and adoration forever. Amen.

ACKNOWLEDGEMENTS

The researcher is very grateful to God Almighty for enabling her to complete this thesis. She expresses her profound gratitude to her mentors; Professor O.K Omorogiuwa (Chief supervisor and Dean of the faculty) and Professor A. U. Osunde, (Co-Supervisor) both of who found time out of their tight schedules to scrutinize the work stage by stage in order to ensure the successful completion of the work.

She is also grateful to Dr. (Mrs.) M.U Orheruata the Head of Department, Professors E.O. Egbochuku, A.N.G. Alutu, G.I Osa-Edoh, V.E.I Audu and I.H. Alika, Doctors; U.C. Ataha, Rev. (Fr.) A.A Adubale, M.N Igbineweka, A. V. Uyigue, I. O Adeleke, B.I Ohanaka (late), O.N. Aihie, J.H. Osarunwense, , F.T Adeyemi, P.K. Adeosun, Y.O. Osunde and V. O. Idusogie for all the constructive criticisms and corrections during seminars and outside the seminars which helped to shape this thesis.

She appreciates other lecturers in the faculty. Professors; E.O.S. Iyamu, I. Owie, and F.E.O. Omoruyi, C.N. Omoifo, and B.O. Ogonor, for their analytical lectures and numerous advices during the course work. Her gratitude also goes to the non-academic staff in the Department of Educational Evaluation and Counselling Psychology, who were always ready to give prompt assistance when needed.

The researcher is also grateful to Mrs. F. Otabor and O. G. Omorodion, her principals, Mrs. F.N. Chukwudiebube the vice principal and other Emotan Junior College staff for their moral support and encouragement. A huge debt is owed to Dr. K.E. Aituriagbon and Mrs. J. Ezomo and other students for their concern, encouragement and support at all times. She is most grateful to the National Business and Technical Examination Board (NABTEB) staff especially the Registrar and Chief Executive of the Board Professor (Mrs.) I. M. Isiugo-Abanihe and Dr. (Mrs.) P. E. Iro-Aghedo for making

it possible for the researcher to obtain the data for the study. The researcher is ever thankful to Dr. U.I Ezenwani for providing the softwares used in data analysis.

Finally her sincere appreciation goes to her husband Chief J.O Okoye and her children Sally, Ogechukwu, Emmanuel-Anthony and Sylvester-Anthony for their encouragement, moral and financial supports.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

ABSTRACT

The purpose of this study was to examine the performance of five model-fitting estimation techniques; Likelihood Ratio Test (LRT), Akaike Information Criterion (AIC), Bayesian information criterion (BIC), Deviance Information Criterion (DIC) and Cross-Validation Log-Likelihood (CVLL) techniques that effectively selected an IRT dichotomous model which fitted NABTCE 2018 May/June Mathematics multiple choice test. This was carried out by comparing the performances of the five techniques used based on relative fit. Four research questions guided the study. No hypothesis was formulated and tested, due to the fact that the techniques used in this study were non-significant statistics.

The research design employed was the descriptive survey of the ex-post facto method. The population of the study consisted of 49,581 candidates who sat for the National Business and Technical Certificate Examinations in 2018 in the six Geo-Political Zones in Nigeria. The sample size comprised 4,948 candidates and a statistical sample of 50 items. The Multistage simple random sampling technique was employed for randomly selecting the sample for the study. The instrument used to collect data was 50-item Mathematics multiple choice test from NBTCE May/June 2018. The instrument was a standardized instrument and as such it was valid and reliable. Item parameters were estimated from the examinees' responses to the items using the computer programme BILOG-MG3. For the five estimation techniques BILOG-MG3 was used for LR, AIC and BIC. WinBUGS 1.4 was used for DIC, while MATLAB was used for CVLL which answered the research questions.

The findings showed that NBTCE May/ June 2018 Mathematics multiple choice test items met the assumptions of unidimensionality and local independence of IRT. The result generated item characteristic curves all of which were in form of cumulative logistic

curve thereby satisfying the item characteristics curve assumption of IRT. The item parameter *b* (difficulty) ranged from -3.305 to 18.265, parameter *a* (discrimination) ranged from 0.035 to 3.858. Parameter *c*, probability of guessing correctly in the test ranged from 0.028 to 0.500. The relative indices (LR, AIC, BIC, DIC and CVLL) used in the study revealed that LR, AIC, BIC and CVLL selected 3-parameter logistic model (3-PLM) as the best IRT dichotomous model that fitted the NABTCE 2018 Mathematics Multiple Choice Test Items. The result also showed the CVLL as the best of the five estimation techniques. On the basis of the findings it was therefore recommended that, in test construction and standardisation, examination bodies such as National Business and Technical Examination Board (NABTEB), National Examination Council (NECO), West African Examination Council (WAEC), Joint Admission and Matriculation Board (JAMB), National Teachers Institute (NTI) and other institutions should employ relative fit in model-fitting to real data.

CHAPTER ONE

INTRODUCTION

Background to the Study

In education, one of the evaluation tools to find out if persons who have attained proper education have been trained with the necessary skills that will bring about national development and social changes is examination. Examination is an essential part of education; it is a pivotal force for promotion, placement, certification without which the learning process is incomplete. Examination has a significant dynamic role in the growth of students (Rind & Mari, 2019). It is a compulsory criterion to determine whether or not a student is qualified to move on to the next level. Examination gives quantitative data of the extent of learning or skill acquired over a period of time. It builds a platform for assessing if educational aims and objectives have been met by all the stakeholders.

Public examinations setting are extremely skillful and expected to be done by experts. This comprises not only the construction of the questions and making sure that the objectives of the curriculum is achieved and that the test items should also be standardised. An examination is said to be standardized when it is given and marked in a dependable way to guarantee that it is capable of being defended validly. These examinations are mostly carried out in professional certification, psychology, education and so on. They are usually the same when it comes to difficulty, format and scope. Educational institutions and examination bodies typically conduct these examinations on certain dates.

In Nigeria, examinations and the award of ordinary level certificates are mainly done by examination bodies one of which is NABTEB. NABTEB is an examination body saddled with the responsibility of public educational assessment and certification in Nigeria. The certificate conferred by NABTEB empowers the candidates for future career studies as well as to further their academics in any professional course in any tertiary

institution. The assessment consists of two series within a year for the ordinary level certificates. The ordinary level (O'L) has four constituents made up of General Education ordinary level subjects that are compulsory, Trade Group subjects, Trade-Related subjects and Elective education subjects. In each of these four constituents English and Mathematics are core subjects. These are for National Business Certificates (NBC) and National Technical Certificate (NTC) May/June and November/December examinations.

In the senior secondary school certificate examinations, Mathematics is one of the compulsory subjects written by students. It is very significance and useful when it comes to science, research in technology and teaching. Mathematics is an essential subject in science which is the basis for getting the knowledge to understand difficulties in other areas. It is truly known as the science language and it is one of the main subjects in the primary and secondary school curriculum in Nigeria (Federal republic of Nigeria, 2014). It is also a compulsory subject to all students who are aspiring to be admitted to study science or science related courses in institutions of higher learning in Nigeria, in which a credit pass is required.

The importance of Mathematics in generating useful and ingenious graduates who will mostly be useful in economic development of any country cannot be over stressed. Musa and Dauda (2014) stressed that it is required for the comprehension of the majority of the areas of knowledge. The authors also said that order than Mathematics, there is no other subject that is very strong in the science world. The implication of this is that Mathematics as a subject in the school curriculum is very important for technical, human and scientific advancement. It serves as an instrument in choosing a career and also for the useful existence of the individual.

The Mathematics test items set in NABTEB consists of objective items and essay items. The focus of this study was on the objective items (multiple choices). The objective

item is among the test devices for the assessment of the performance of the students academically during and after teaching and learning. In the multiple choice test students are expected to respond to a set of questions by selecting the best probable answer among the choices provided as the options for each question or test item. The multiple choice consist of two parts, the stem which is the problem part and a list of optional solutions or answers. The examinees can select from among the options. The right answer to the question is called the key while the options that are not the answers are called the distractors.

The quality and standard of the Mathematics multiple choice test items set by the National Business and Technical Examination Board (NABTEB) depends on the procedures for determining the indices or parameters of the test items which in turn also depends on the measurement theory used.

Psychometricians are now interested in models of assessments, concepts and advancement such as item response theory (IRT). Banghael (2008) said that, this type of theory is also known as hidden or latent trait theory. Adebule (2013) indicated that for a very long time a significant discovery has been made that examinees' perceived marks and the actual marks are different from the scores of their abilities. The scholar also mentioned that the scores of the ability prove to be more necessary since they are not dependent on the test. The assumption of item response theory (IRT) is that, there exists a Mathematical function which defines the association between the likelihood that a student may get an item right and the student's aptitude. In item response theory, there are dichotomously and polytomously scored models. The dichotomously scored models include the Rasch or one-parameter (1p), two-parameter (2p) and three-parameter (3p) logistic models. While the polytomously scored models are the Rating scale model, Nominal model, Partial Credit model and Graded Response model.

This study was centred on the dichotomously scored models of the item response theory. In the light of the dichotomous models the likelihood of getting a right answer to an item is mathematically modelled by the use of normal Ogive model or logistic model. The association can be graphically denoted using the item characteristic curve (ICC). Unidimensionality and local independent are other assumptions of IRT. The meaning of unidimensionality is that the questions are assessing one construct (variable). Local independence on the other hand means that examinees' answers to the questions are independent of one another assuming ability is held constant (Ani, 2014). Each examinee is assumed to possess a reasonable quantity of the construct or trait that is being assessed. The variety of traits being assessed may be Mathematical reasoning, Mathematical manipulative skill, motor skill, spatial memory or verbal proficiency (Adebule 2013). In this study the underlying latent trait is Mathematical proficiency which includes Mathematical reasoning, Mathematical manipulative skills and others.

The item characteristic function can be verified using the 1-parameter, 2-parameter and 3-parameter logistic models. Using these models, the item statistics (or parameters), the item difficulty (*b*-parameter), item discrimination which is *a*-parameter and guessing (*c*-parameter) can be verified for items that are dichotomously scored. The Rasch model or one-parameter can verify parameter-*b* which is the item difficulty, the two-parameter model or Birnbaums model can verify *b* and *a* parameters which are the item difficulty and item discrimination, while the three-parameter model or Lords models can verify *b*, *a* and *c* parameters.

The logistic model which dictates difficulty of item parameter (*b*), the item discrimination parameter (*a*), and pseudo guessing parameter (*c*) is also known as the 3-parameter logistic model. The difficulty of an item gives a clue as to the level of how hard the items are and mainly prescribes the location on the item characteristic curve with

regards to the scale of ability (θ). When the item difficulty parameter is big it shows more difficult items and move the item characteristic curve up the scale with regards to the scale of ability. In the case of item discrimination, this gives the clear view of how well the items differentiate among the students who have the same level of ability. This gives the degree of the gradient (slope) of the item characteristic curve. When the parameter of discrimination is big this results in a bigger discrimination index (power) and also produces an item characteristic curve which the slope is steeper. Lastly the parameter for pseudo guessing gives the clue of how the students with very low ability could achieve on an item taking the chance to guess the right answer.

There had been several model selection and model-fit statistics for dichotomous item response theory (IRT) models that were proposed in evaluating the suitability of choosing item response theory models and calibration procedure in terms of the model-data fit for, one-parameter, two-parameter and three-parameter logistic models. Model-fit to data must precisely indicate the true association between students' achievement and ability on the item. Correct use of models is centred on the principle on which certain amount of item response theory postulations (assumptions) are made based on the type of data to ensure that the model perfectly represents the data. When these postulations are not achieved the conclusion with respect to the type of questions and tests may be faulty and the purpose of model-data fit analysis carries the risk of drawing incorrect conclusion. According to Essen, Idaka and Metibemu (2017), the measure of model-fit to the data could be on the basis of three kinds of proofs. Firstly, the validity of the assumption of the particular model for the set of data such as; test not being speeded, unidimensionality, minimal guessing for 1-parameter and 2-parameter logistic models. That for 1-parameter logistic model all the items should be of the same (equal) discrimination. Secondly, that the expected properties are obtained to reflect the estimates of the ability parameter and

item invariance. Lastly model prediction accuracies are assessed through the analysis of item residuals.

Model parameters estimation included in the latent theory is actually significant when it comes to the field of measurement as being a basic step for designing tests and measurements, analysing examination items, creating item-banks, and constructing computerized adaptive tests. Therefore, accuracy should exist when taking a decision on the methods of estimating model parameter(s). The aim of using model selection method is to select a model which gives comprehensive data fit and also have the capacity to be used universally in predicting future data that are not the same. There are different methods of estimating model-fit. This study compared the Likelihood ratio test (LR), Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC) and Cross Validation Log-Likelihood (CVLL) estimation techniques.

Likelihood ratio test which is a G^2 statistics is basically a χ^2 (chi-square) statistic which is computed as the variance amid two deviances from the two models that are being compared. When the difference is sought for the two models in question, this is distributed in form of a χ^2 (chi-square) and thus can be caused to undergo significant test in order to determine the model that fits better. The one-parameter, two-parameter and three-parameter logistic models are embedded sequential models and the Likelihood ratio test technique can be used to run them. When the models are embedded, the models that have many parameters will continually fit along with the models with lesser parameters. Likelihood ratio test has the capacity to decide if the model-fit will reduce significantly when some parameters are removed (Brown, Templin, & Cohen, 2014). For instance, in using absolute model-fit and the result specify that the 3-parameter logistic model is appropriate; a Likelihood ratio test may be employed to decide if reducing to 2-parameter logistic model is suitable. When the reduction to 2-parameter logistic model is true then

another Likelihood ratio test may be applied to decide if further reducing it to 1-parameter logistic model can allowed. When it comes to item level selection Likelihood ratio tests may also be used for model selection.

Akaike information criterion (AIC) is a valuable statistics for statistical model selection and evaluation. The Akaike information criterion advantage lies in its simplicity which does not require the use of table. It is used when models are non-nested. Once the estimation of the Maximum likelihood parameters of a model is determined it is easy to calculate Akaike information criterion. The model with a minimum value of Akaike information criterion is chosen as the best fitting model.

Bayesian information criterion (BIC) on the other hand, is an information criterion that is consistent, which means that as the size of the sample increases, this criterion will select a true model of finite dimension. This technique is also used when models are not nested. The best model is the model which has the least value of Bayesian Information Criterion.

Deviance information criterion (DIC) is built on the log-likelihood posterior distribution or the deviation. When the likelihood function is obtainable in closed form the models' posterior distributions are gotten by the simulation of Markov Chain Monte Carlo (MCMC).

Cross-validation Log-likelihood (CVLL) is a procedure of model validation for evaluating the generalization of a given set of data by the result of the statistical analysis. The CVLL is mostly applied in situations which aim is for prediction. The intention is to estimate the accuracy of the predictive model in terms of practical performance.

This study specifically compared the methods of model-fitting estimation techniques that selected the 1-paramter, 2-parameter and the 3-parameter logistic models of item response theory (IRT) for model-data fit.

Statement of the Problem

The idea of an examination is to assess the students' achievement, ability or some latent construct of interest. National Business and Technical Certificate Examination Mathematics paper 1 May/June 2018 is a multiple choice examination comprising 50 items with 4 options (A-D). These test items are dichotomously scored. Item response theory dichotomous models are the one-parameter (1p), two-parameter (2p) and three-parameter (3p) logistic models.

The main objective of model-fitting is to find a useful approximating model which may not only fit well and has parameters that are easily interpreted but also has the capability to generalise to different data or make predictions of future data. There is no standardized procedure of estimating fit of a model to a particular data set (Sinharay, 2005). Most of these estimations were carried out using small samples and limited test items. Some are used with simulated data. In addition there is no standard range of sample size and no specific range for test item length.

The basis for estimating goodness-of-fit in item response theory is usually by the use of Chi-square (χ^2) statistics when it comes to traditional methods. This is applied to examine model-fit, for instance Pearson's chi-square, Yen's (1981) $Q1$ statistics, Bock's Chi-Square statistics and many others. However, there are several drawbacks in using χ^2 statistics in assessing model-fit. The most common critique about the χ^2-based statistics according to Hambleton and Swaminathan (1985) is in their being sensitive to sample size. The Chi-square statistical test discards or rejects every model if the sample size is big. For instance when estimating model-fit using Chi-Square statistics through item fit, out of 40 items, 3 items might fit 1-parameter logistic model, 5 items might fit 2-parameter logistic model and 10 items might fit 3-parameter logistic model, what happens to the rest of the items? This study therefore, sought to compare the five estimation techniques (LRT, AIC,

BIC, DIC AND CVLL) to ascertain which of them would select the best IRT dichotomous model among one- parameter, two-parameter and three-parameter logistic models which are nested models, that fitted National Business and Technical Certificate Examination Mathematics paper 1 May/June 2018 using relative fit and real data.

Research Questions

In this study these research questions guided the study.

1. What are the difficulty indexes of the item parameter estimates of NBTCE May/June 2018 Mathematics Multiple Choice Test Items?
2. What are the discrimination indexes of the item parameter estimates of NBTCE May/June 2018 Mathematics Multiple Choice Test Items?
3. What are the guessing indexes of the item parameter estimates of NBTCE May/June 2018 Mathematics Multiple Choice Test Items?
4. How do the model selection methods for data from NBTCE May/June 2018 Mathematics Multiple Choice Test Items compare based on the one-parameter, two-parameter and three-parameter logistic models using the five (LRT, AIC, BIC, DIC and CVLL) techniques?

Hypothesis

No hypothesis was formulated and tested. The estimation methods are non-significant statistics. Therefore the comparisons were done based on the fact that a lower index expresses a better model-fit in terms of both model-data fit as well as parsimony. That is the model with the least value is chosen as being the best model that fits the data.

Purpose of the Study

The aim of the study was to compare five techniques for estimating Model-fit of item response theory dichotomous models on NBTCE May/June 2018 Mathematics multiple choice test items using different estimation techniques. Precisely, the study;

- determined the difficulty indexes of the items of NBTCE 2018 May/June Mathematics multiple choice test items.
- determined the discrimination indexes of the items of NBTCE 2018 May/June Mathematics multiple choice test items.
- determined the guessing indexes of the items of NBTCE May/June 2018 Mathematics multiple choice test items.
- determined which of these estimation techniques, likelihood ratio test, Akaike information criterion, Bayesian information criterion, Deviance information criterion and Cross-Validation Log-Likelihood selected the item response theory dichotomous models which fitted NBTCE May/June 2018 Mathematics multiple choice test items.

Significance of the Study

The findings from this study will be of great advantage to Psychometricians, test developers, examination bodies and teachers.

To the Psychometricians it will add to the literature on educational and psychological testing. The findings will serve as guide to measurement experts in selecting the most suitable model-fitting technique(s) that can help in identify the item response theory model that can fit dichotomously scored items.

Furthermore, the finding will encourage test developers to identify the best estimation technique(s) to use in model-fitting to data during test construction and standardization.

To the examination bodies, the findings of this study will aid in establishing the worth of examination carried out by National Business and Technical Examination Board (NABTEB), National Examination Council (NECO), West African Examination Council (WAEC), Joint Admission and Matriculation Board (JAMB), National Teachers Institute (NTI) and other institutions. It will involve measures that can help strengthen the standard of the examinations set and conducted by NABTEB, NECO, WAEC, JAMB, NTI and

other institutions. This will go a long way in helping to promote public trust and acceptability of results from these examinations conducted by examination bodies. For the purpose of more objective and complete judgement on the students' achievement by these examination bodies, the model-fitting of the data need to be determined using a more precise model-fitting estimation technique(s) of test theory, the latent theory. It will enable them to further improve test construction practices, standardisation, administration and analysis. For these examination bodies, this study will also heighten clearer understanding of their adoption and acceptance of IRT framework in model-fitting of multiple choice test items of dichotomous models.

The findings of this study will benefit teachers as it will help to give accurate and exact report on the achievement of the students by offering ideal and significant explanation of the student's performance via the latent trait model. Hence, it will aid in boosting their performance as it would be harmonized with their underlying latent trait (ability).

Scope and Delimitation of the Study

Basically, this study was restricted to the use of five model-fitting estimation techniques; Likelihood ratio test, Akaike information criterion, Bayesian information criterion, Deviance information criterion and Cross-validation log-likelihood in comparison of how these techniques selected the item response theory dichotomous model that fitted in NBTCE 2018 Mathematics multiple choice test items. The study was delimited to candidates who sat for NBTCE 2018 Mathematics multiple choice test items in the South-East and the South-South Geo-political zones of Nigeria.

Definition of Terms

The following terms are operationally defined for the purpose of this study:

IRT dichotomous Models: A dichotomous item response theory model is applied when a test comprises a question with two types of answers, 1for correct response and 0 for wrong answer.

Item Characteristic Curve (ICC): The Item Characteristic Curve represents the association between a student's likelihood for answering an item type and the level of the latent trait which is being assessed by the scale.

Item Difficulty, **or Threshold, Parameter b**: This is the position on the scale of ability θ where a person has a 50% chance of responding positively to the item on the scale.

Item Discrimination, or Slope, Parameter a: This defines the strong point of discrimination of the item amid examinees with levels of ability (θ) below and above the item difficulty.

Item Response Theory (IRT): This is also known as the hidden trait theory. It is made use of in psychological and educational measurement (achievement tests, rating scales, and inventories) that investigates a Mathematical association between individuals' abilities (cognitive construct) and response to items.

Local Independence: This means that the answer to an item is not dependent on answers to other items in a scale assuming that the latent construct is held constant.

Model-Fit: This is the extent to which the assumptions (expectations) of the model are met by the item response data.

Model selection: This refers to the choice of the statistical model that describes the data best among several contending models.

Test Characteristic Curve (TCC): Test characteristic curve describes the expected number of scale items endorsed as a function of the underlying latent variable.

Theta (θ): Theta is the unobservable hidden trait being assessed by the items.

Unidimensionality: This assumes that a set of questions are assessing one continuous underlying latent construct.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

This chapter provides the relevant literature related to the study which is organized under the following sub-headings;

- Theoretical Framework
- History and Functions of National Business and Technical Examination Board
- Concept of Test and Objective Test
- Item Response Theory
- Model-Fit Estimation Techniques
- Summary of Reviewed Literature

Theoretical Framework

This study is mainly anchored on Item Response Theory (IRT). The theoretical basis of IRT was put down by Thurstone (1925) after which Lord (1952) presented the notion of hidden construct or talent trait and distinguished this trait from observed marks for the test. Item response theory (IRT) models, also known as latent trait models, play an important role in educational testing and psychological measurement as well as numerous areas of behavioural and cognitive measurement. IRT analyses each item and each examinee's response on an item separately estimating item level parameters, as well as examinee's ability levels. The Item response theory is a set of hidden construct technique specifically designed to model the relationship between leaner's ability and the probability of getting a correct response to an item (Chalmers, 2012). This provides unique estimates of the difficulty, discrimination and guessing of each item for examinees at different levels of knowledge. The focus is on the pattern of responses on each item rather than total scores.

IRT is built on the idea that the likelihood of a correct answer to an item is a function of person and item parameter. The person parameter is constructed as a single

latent trait. The traits are not directly measured since they are not observable therefore, they are said to be latent or hidden traits. In this respect a response to an item is considered as an indicator of an underlying ability. The assumption is that whatever the latent trait, it can be measured on a scale having a midpoint of zero, a unit of measurement which range from negative infinity to positive infinity ($-\infty$ to $+\infty$). Since there is a unit of measurement and an arbitrary zero point, with such a scale the fundamental idea is that if one can physically ascertain the ability of a person, the measurement scale can be used to tell how much ability a particular individual has and the ability of several individuals can also be related.

Estimates of examinee's ability (θ) are based not only on the responses they provide but also on the characteristics. Each item is characterised by one or more model parameters such as, the item difficulty (b parameter), which is a point on the ability scale (θ) where an examinee has 50% chance of answering positively to an item. The item discrimination (a parameter) describes the strength of an item's discrimination between examinees with ability levels (θ) below and above the difficulty (b parameter). The pseudo-guessing (c parameter) explains why low achievers can respond positively to an item (Mislevy, 2011).

History and Functions of National Business and Technical Examination Board (NABTEB)

NABTEB exists as a specialised public educational assessment certification body that was established by an act of parliament under Decree 70 now Act 70, August 1993 in Nigeria. It was set up to control the ordinary and advanced levels of vocational education as a critical element of educational system in Nigeria. It was also mandated to take over the technical and vocational examinations previously offered in Nigeria but conducted by the foreign examination bodies, such as City and Guilds, Institute of London, Pitman,

Royal Society of Arts (RSA) and the West African Examination council (WAEC) Technical and Business. The mandate and objectives of NABTEB was to;

(a) Carry out examinations which lead to giving the award of

National Business Certificate (NBC);

National Technical Certificate (NTC);

Advanced National Business Certificate (ANBC);

Advanced National Technical Certificate (ANTC).

(b) Take over the conduct of Technical and Business examinations previously conducted by West African Examination Council (WAEC)

(c) Conduct entrance examination for admission into Technical Colleges;

(d) Monitor, collect and keep records of continuous assessment in Technical Colleges and associated institutes toward the award of certificate in National Technical and Business Examinations;

(e) Carry out research, circulate the statistics to the public and further information in order to develope suitable examinations, tests and syllabi in Technical and Business studies,

(f) Give results and certificates and make awards in examinations conducted by the board.

Concept of Test and Objective Test

An assessment of an examinee's intellectual ability, talent, aptitude, or classification in many other areas is carried out through test or examination. The examination may be given orally, written on paper, it might be computerised, or in a prearranged area that needs the individual to show or accomplish some of his skills (For instance, athletics or auditioning). According to Ughamadu, Onwuegbu and Osunde (1991), test represents series of questions to be responded to by the learners. To them it can also be

referred to as fixed items, tasks or problems intended to assess the extent of knowledge, aptitude, intelligent and other cognitive traits possessed by learners. In the opinion of Omorogiuwa (2010), a test is a set of items presented to an individual or a group of individuals to which they are required to give answer(s) to under specific conditions with the intention of determining the extent to which such a trait is absent or present in the respondents. He also mentioned that tests are vital in determining the extent to which learning has occurred.

School examination is generally supposed to be merely for the intention of assessing learners in addition to conveying scores or ratings to them. Roediger, Putnam and Smith (2011), outlined ten things that can be profited from testing; firstly they said that, retrieval of knowledge helps remembering in future, secondly testing finds out the information the learners do not possess. Others are that giving test make learners to acquire more knowledge and understanding from the subsequent material for learning, it yields improved knowledge organisation, it increases the chance to transmit what one had learned in one situation to different settings, giving test enable recovery of material that was not being tested for, that test increases observation of intellectual ability, test inhibits interruption from previous information when fresh material are being learnt, it offers advice to teachers in form of how much knowledge the student has and finally giving test regularly inspires learners to be studious.

Types of Test

Test in education according to Omorogiuwa (2010), can generally be classified into two types, Psychological and Achievement tests. Achievement tests according to Ughamadu et al (1991) are tests intended to assess a student's present height of knowledge, performance or motor skill. Ali (2006), regarded test of attainment as a device given to a

learner or group of leaners in form of questions. This may prompt a positive chosen or anticipated answer which signifies his or her talent or capability.

Achievement test may be classified into teacher made test and standardized test. Teacher made test are teacher's own test (Onunkwo, 2002). They are tests constructed by individual teachers in their schools for assessing their students. In the opinion of Ifeakor (2011), standardized test is the one that has norms. Norms are a set of descriptive data which make it possible to determine the standing of a candidate in relation to a specified reference group. Standardized tests provide a uniform set of questions, instructions and method of administration. Standardized tests as opined by Ughamadu et al (1991) are tests designed by test experts and administered, scored and interpreted under standard conditions. Agreeing with this is Popham (2008), who described standardized test as test that are constructed, given to examinees and following a proper procedure of interpretation using standard measures.

Tests for measuring the achievement of objectives in the cognitive domain fall mainly into two categories: the essay test and objective test. Onunkwo (2002) defined essay test as a test in which students are required to provide answers to questions and offers students the opportunity to organize thought and express their ideas in writing. The objective tests are of four types, true/false items, completion items, matching items, and multiple choice items. A matching item is an item that offers a definite term which the examinee needs in order to recognise the features that enables him or her to match the right term. The completion item also known as fill-in-the-blank item offers an examinee the ability to recognise the features and entails him or her to remember the right word. True/False questions present examinees with two options (that is two different choices); a statement may either be true or false. The multiple choice item has two parts namely; the stem and the alternatives (that is the answer options). The stem is the direct questions or

incomplete question while the alternatives are the options from which the examinees are instructed to pick only one which is most correct (Ani, 2014).

Item Response Theory (IRT)

Item response theory or the latent trait theory offers an ample statistical mechanism for the analysis of test in institutions of learning and mental assessment scale. According to Bielinski and Davison (2001), in educational testing latent trait theory has been generally used. It is a model for expressing the association concerning an examinee's answer to a question (item parameter) and the basic underlying construct (known as ability or trait) which is assessed through a device. It is usually used to generate a response curve (likelihood of an examinee with a specific aptitude to give an accurate response to the question) with regards to every question and also to generating a scaled mark for the entire examination on the basis of the knowledge about every question (Wendy & Carl, 2010).

Item response theory to Osterlind (2012) is a method used in contemporary academic and mental assessment which speculates a specific concept mostly mental, in addition generates complex statistics to assess mental developments. Its objective is to reliably calibrate examinees and items or questions for the test on a standard measure that is taken to show the person's talent or aptitude and definite features of the questions for the test. Item response theory can be applied to numerous real-world examinations, such as generalisation of results from test, countless item analyses, examining test bias and differential item functioning, equating test forms, estimating construct parameters, domain scoring, and adaptive testing. To Reckase (2009), item response models explain the relationship of examinees and examination questions. Therefore, item response theory according to DeMars (2010) is a universal structure for identifying the Mathematical functions that describe the association between one's aptitude or attribute as quantified using a device and the individual's answers to the different questions in the instrument.

Nering and Ostini (2010), sees item response theory as strong true score theory, modern mental test theory, or latent trait theory, a model for the analysis, design, and giving marks to tests, questionnaire and similar instruments assessing aptitudes, attitudes and various attributes. According to Palmieri (2012), it is a model-based form of test principle that a Mathematical function is applied to explain the relationship between an examinee's position on a hidden attribute and how he or she respond to items. When the right model is carefully chosen, the possibility that a person may answer the question in the right way is a function of the individual's stand point on the fundamental trait. In addition the difficulty of the item and item discrimination modelled as a function of person's achievement level of the attribute that is measured and the features of the items that were responded to. To Embretson and Reise, (2000) item response theory is likewise a Mathematical model which defines the way individuals relate the questions in a test. In the opinion of Bichi and Talib (2018), item responses can either be discrete or continuous and can be dichotomously or polychotomously scored. Item score types can be sequential or disordered; there can be one or numerous latent traits basic to the achievement on the test and there exists various methods in which the association between item responses and the underlying ability or abilities can be specified. In the overall, item response theory context, several models have been articulated and applied to actual data for test. Schumacker (2010) opined that, item response theory centred on hidden trait theory, includes assessment postulations as regards to the individual, question and achievement on the test and in what way achievement associates with learning as per being assessed via the questions on the examination. In item response theory, individuals and questions are positioned on a similar scale. A great number of item response theory models are of the assumption that the underlying construct is defined using a unidimensional continuum (scale). In addition, for a value to be placed on an item, that particular item ought to be capable to discriminate

amongst individuals positioned at various locations along the scale. If an item is able to discriminate amongst individuals it diminishes the likelihood as regards to their positions. This ability to discriminate amongst individuals with dissimilar locations might be kept steady or permitted to differ through the questions that are being used. Hence, persons are categorized in relation to their minimum locations on the underlying construct. Questions are identified with respect to their locations and ability to differentiate amongst individuals (Ani, 2014).

The fundamental hidden construct in Mathematics could be a latent trait that can be assessed like, cognitive reasoning, quantitative aptitude, Mathematics skills or Mathematics achievement. The aim of item response theory is to suggest models that allow the connection of the hidden construct to certain features of the student that can be observed, especially the individual's capabilities to properly answer to a number of items that make up an assessment (Magis 2007; Bichi, Embong, Mamat & Maiwada, 2015). Item response theory, parameters of the items consist of difficulty of the items (location), discrimination of the item (slope), and guessing (which is the lower asymptote). These are assessed precisely by means of logistic models rather than proportions. Cappelleri, Lundy and Hays (2014) said that numerous item response theory models differ in the amount of parameters and they can carry out only dichotomous or polytomous items in general. One-parameter (1-p or Rasch model), two- parameter (2-p) and three- parameter (3-p) logistic models are the three frequently used item response theory dichotomous models. The features of item response theory models are summarized by Hambleton and Swaminathan (1985), as follows, at first, an item response theory model needs to identify the association among the perceived answer and basic trait that is hidden. Furthermore, the model requires making available a method to assess total marks on the talent. In addition, the student's marks determine the base for the assessment of the basic trait. Lastly, an item response

theory model is of the assumption that how the individual performs in the test can absolutely be foreseen or accounted for using single or additional traits.

Item Response Theory Assumptions

The three item response theory assumptions are Unidimensionality, Local independent and item characteristic curve (Item Response Function). To Ojerinde (2013) these assumptions are very important and ought to be realized regardless of the hidden construct model that is being used. This suggests that, data from a test may be used for the assessment of models for the underlying construct if and only if these required assumptions are attained.

Unidimensionality

According to Reeve (2002), model of item response theory is built on the theory that the items are assessing only one constant hidden construct (unobservable construct) which ranges from negative infinity to positive infinity ($-\infty$ to ∞). Item response theory according to Ojerinde (2013), tries to model the latent trait of an individual and the likelihood of getting the right answer to a question on the test on the basis of the design of answers to the items that make up a test. In addition Ojerinde (2013) clarified that, the model of hidden variable presumes that certain constructs bring about performance of the test. The ability of the examinee in a set of unidimensional hidden (or dormant) variable can be symbolized by a vector of scores of ability such as (i.e., $\theta 1$, $\theta 2$, $\theta 3$….. θn). The latent trait theory models that undertake a single hidden trait is said to be unidimensional. Unidimensionality denotes that examinees and the questions can be determined by one construct (Yen & Fitzpatrick, 2006; Kyung, 2013). This points out that the achievement of each of the student is presumed to be ruled by a variable, which is said to be knowledge. This supposition is met when all of the questions consisting of a test assess the same latent

trait and candidates use this construct or knowledge to reply to the test. Failure to uphold unidimensionality would lead to deceptive result that is severe (Ani, 2014).

Ani (2014), also said that, since persons' mental and peculiar attributes effect test achievement and cannot frequently be checked, it is not always possible to meet the assumption of unidimensionality. The unidimensionality of a test may be spoken about only when there is just one prevailing hidden trait in the test. The unidimensionality assumption is powerful in the sense that certain questions may entail the use of various construct to get a right answer (Topczewski, 2013).

To satisfy this assumption according to Ojerinde and Ifewulu (2012), one can employ one of the following eleven techniques for unidimensionality assessment; Factor analysis, Random baseline test, Biserial test, Eigenvalue test, Cronbach analysis test, Factor loading test, Congruence test, Part/Whole test, Communality test, Vector frequency test and Confirmatory factor analysis (CFA) and Structural equation modelling (SEM) test, by the use of Statistical package for social sciences. Ojerinde (2013), further said that the advocate for unidimensionality of the items on the scale will be on the condition that the model fits the data very well and that no remarkable residual correlations exists (that is, there will not be any correlations ≥ 0.20). Failure to uphold this assumption of unidimensionality the result would be insufficiency of the model to describe the data, thus untrustworthy assessment of the ability of the examinee. Consequently, the precise description of the amount of measurements of latent variable is connected to the construct validity of the assessment (Rijn, Sinharay, Haberman & Johnson, 2016).

Local Independence

Answers to the questions are not dependent on one another's certain capability, as soon as an individual's degree of ability is recognised; the student answers to questions are not dependent of one another. In the opinion of Ojerinde (2013), the assumption of local

independence means that the likelihood of an individual answering the question rightly is not affected by the response to other questions in the assessment. He further said that local independence does not suggest that items in the test do not associate with one another, but that the achievement on dissimilar questions is not dependent; however it is determined by the ability of the examinee. It suggests that the likelihood that an examinee will respond accurately to two questions requires the result of the likelihood that the student will rightly respond to every question separately. The latent trait, that has effect on answers to any set of questions in an assessment, is stable at a specific period of assessment. Hence, the association amid two questions ought not to be different from zero significantly, or otherwise, it can be opinioned that the answers to the questions are as a result of the effect of some diverse variables that are not relevant, other than whatever the device is designated to assess (Ojerinde, 2013). The assumption is a unique one in latent trait theory. When the independence is conditional it offers the likelihood of statistical independent for the questions. In the opinion of Reeve (2002), assumption of local independence states that the answer of an examinee to a question is independent of the individual's response to a different question as long as the latent trait being assessed by the scale is held constant. The assumption of local independent is interrelated to that of unidimensionality, if success is specified by a single underlying construct on every single question, likewise achievement on the question is systematically influenced by the individual's latent trait exclusively (Ani, 2014).

The assumption of Local Independence may not be kept in diverse means. First and foremost devices for the test are frequently constructed using contents of the item, every single item assessing their particular feature of the hidden trait and most distinct hidden construct only may not be able to give explanation for the relationships amongst items in a similar test. This sort of local dependence can be understood as a failure to uphold the

assumption of unidimensionality. In addition, the assumption of local independence can fail to hold if the answer used on one item gives clue to the answer given to a different question. This may occur owing to likenesses in content of the item or in the way the items are responded to in academic examination, if the right answer to the first question contains a hint or gives indication to the right answer to the next question.

Item Characteristic Curve

An item characteristic curve (ICC) also known as item response function (IRF) is a Mathematical function which associates the likelihood of achievement on a given question (individual test question for a dichotomous answer) for persons who have certain trait level. This likelihood is not dependent on the ordering of the wellbeing of the individuals (Eleje, Onah & Abanobi, 2018). As the likelihood is not dependent on the number of different individuals that are positioned at the similar location on the latent trait continuum in the population of the examinees, the numerous item response theory models, which are different versions of logistic models. These are merely diverse Mathematical functions for defining item characteristic curves as the association of an item's characteristics and individual's level on the ability (for example, difficulty of the items, discrimination of the items and guessing) with the likelihood of an exact answer on that item assessing the identical ability (Ani, 2014; Cappelleri et al, 2014).

The item characteristics curve offers a clear difference among diverse item response theory models. There exist technical attributes which can describe item characteristic curve. Paramount is the item difficulty. In the latent trait theory, the item difficulty designates the point where the item functions along the ability scale. For instance, a question that is easy functions among the students with low-ability and item that is difficult functions among the students with high-ability. Therefore, difficulty is an index for location (position). The second attribute is item discrimination; this refers to how well

an item can distinguish amongst students who have abilities that are under the item location and students who have abilities directly above the item location. This feature principally shows the gradient, that is how the item characteristic curve inclines in its midsection. The third attribute is the lower asymptote which is guessing. The item characteristic curve is monotonic and takes the form of a normal ogive. How many parameters necessary to establish an item characteristic curve to Ojerinde (2013) is influenced by the specific logistic model used.

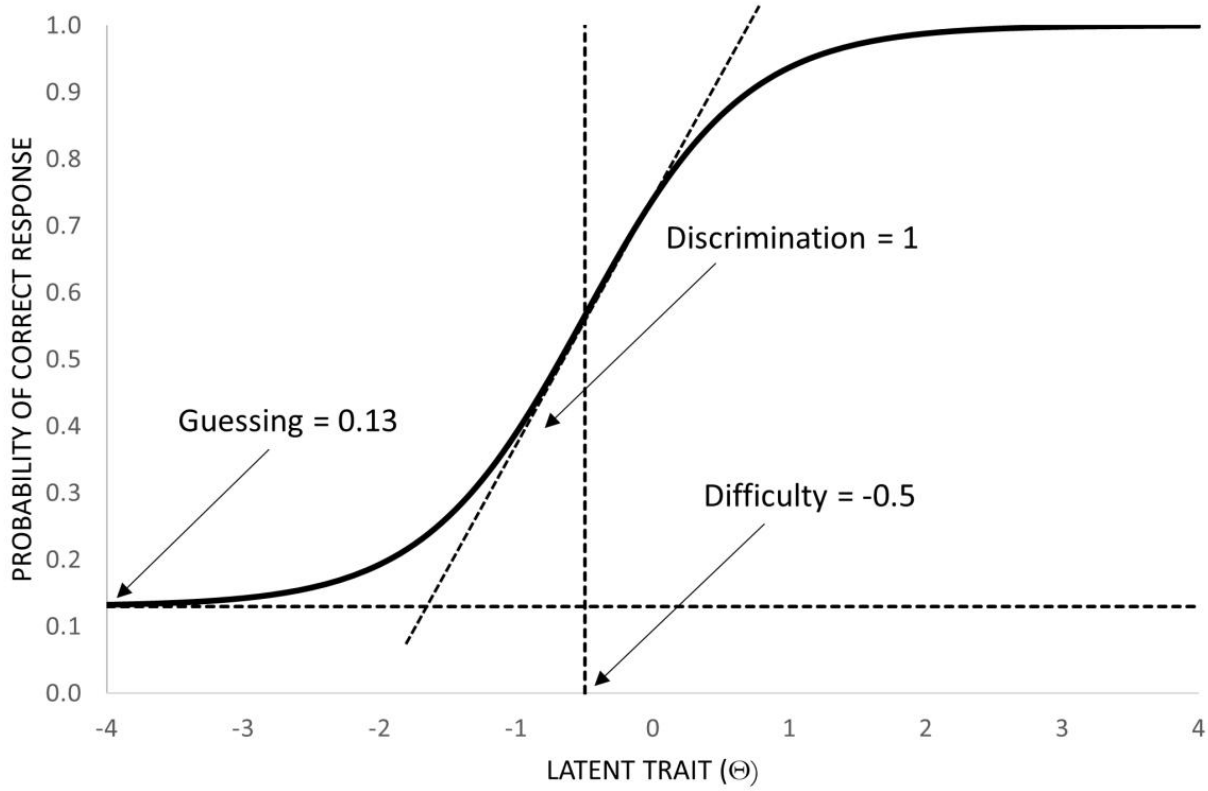


Figure 1. A sample of a typical item characteristic curve
Source: Bulut (2015)

The likelihood of a precise answer is near zero at the lowermost levels of the underlying latent variable. It escalates till the maximum points of ability, and the likelihood of accurate answer moves closer to 1. The S-shaped curve defines the association amid the likelihood of the right answer to a question and the scale of ability.

Test Characteristic Curve

The latent trait theory and procedures can also be applied to the test level as well as item level. The item response theory concept of test characteristic curve (TCC) is an offshoot of the ability of item response theory (Hambleton & Slater, 1997). Test characteristic curves are test level equivalents of item characteristic curves that characterises a non-linear regression of total assessment score on latent trait. Simply put, a test characteristic curve is created by the summation of the entire item characteristic curves through the length of the ability scale. The test characteristic curve is a beneficial instrument for appraising the extent of measurement and the amount of discrimination at various points of the scale of ability. De Ayala (2009) adds that, the extent to which the test characteristic curve is linear offers a suggestion to the degree to which the measure gives interval scale or linear measurement.

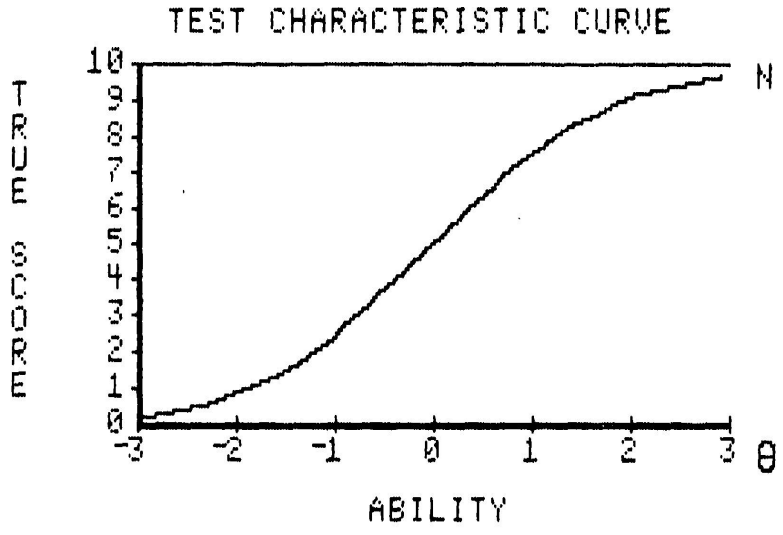


Figure 2: A typical Test Characteristic Curve
Source: Baker (2001)

Looking at Figure 2 above it can be seen that the estimation of the ability is plotted on the x-axis as for an item characteristic curve and the exact total marks on the y-

axis. A test characteristic curve signifies the association amid the true scores and the latent trait (ability) scales. Like the item characteristic curve it can be explained approximately in the same manner. By whatever means the value of true score is influenced by the variations in ability it has effect on the gradient of the curve. Nonlinear curve is used to explain majority of the assessments. Test characteristic curves do not possess a specific procedure which can be helpful in their estimations. Therefore, the best way to explain the curve is in words after scrutinizing it visually.

Item Information Function

Another essential feature of item response theory models is the item information function. This is a guide signifying the array of level of ability θ above which a question or examination is greatly valuable for differentiating amongst examinees. Every question in an assessment offers several facts concerning the latent trait of the individual, however the extent of this facts is subject to exactly how meticulously the item difficulty equals the individual's latent trait. This implies that the information function describes the accuracy of assessment for the individuals at separate levels of fundamental hidden trait with greater facts, indicating added accuracy. The graph of information function positions a person's level of ability on the horizontal axis and volume of facts on vertical axes.

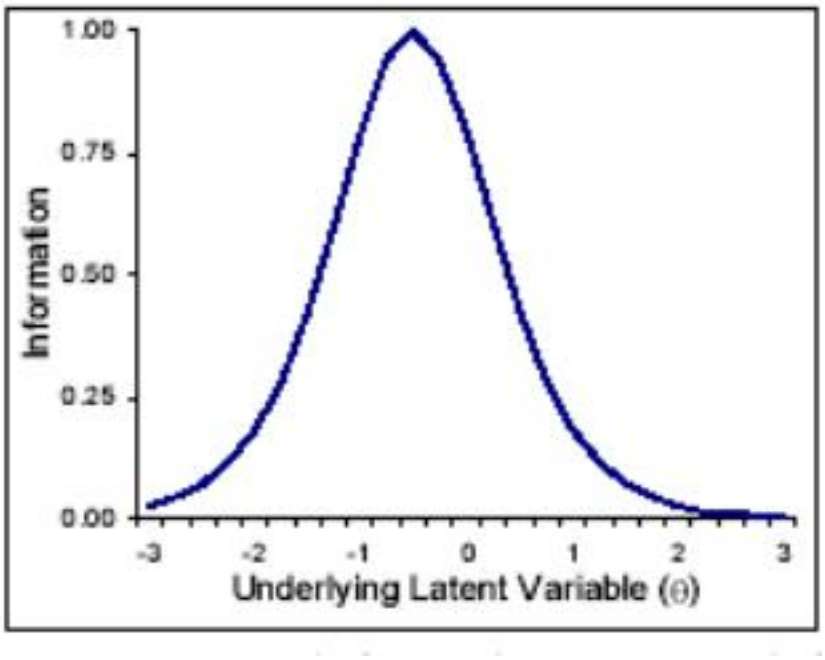


Figure 3: An example of Item information function
Source: Adonu (2014)

The item information function (IIf) shape depends actually on the parameters of the item. High discrimination of item implies additional information function that is pointed at the top. Higher discrimination parameters offer more facts about persons whose ability level θ are situated near the item's threshold (difficulty) value. The difficulty of the item

parameter defines the position that the item information function is situated (Flannery, Reise & Widaman 1995). Using the local independence assumption, the values of the item information function can be added throughout all the items in the scale in order to make up the test information curve (Lord 1980).

Test Information Function

Test information function (TIF) is an enormously valuable attribute of item response theory. It shows the extent to which the test is performing in assessing latent trait over the entire array of scores of the latent trait. As an examination assesses the ability of an examinee, the volume of factual knowledge generated through the examination at every level of trait can likewise be achieved. An examination is a set of questions hence, the test information at a specified level of trait remains basically the summation of the factual knowledge of the item at that particular level. As the test factual knowledge is obtained through adding the factual knowledge of the item at a specified level of trait, the quantity of facts is determined at the item level. Therefore, the test information function is defined as presented below:

$$I(\theta) = \sum_{i}^{N} Ii(\theta)$$

where I (θ) is the aggregate of the test information at every level of ability θ, I*i*(θ) is the aggregate of information for item I at every level of ability θ. N is the amount of questions in a test.

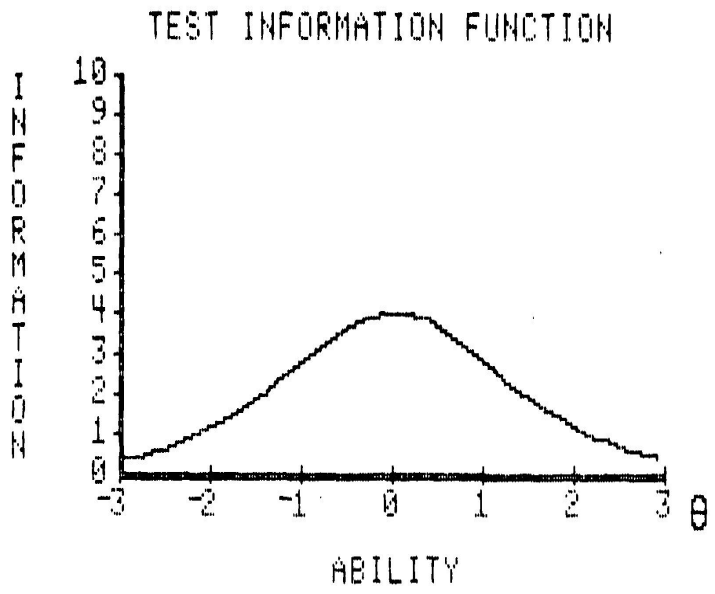


Figure 4: Test information function
Source: Natarajan (2009)

The highest worth of the test information function in Figure 4 is moderate as the quantity of information reduces quite gradually as the level of ability varies from that coinciding with the highest. Consequently, the latent trait is assessed through certain precision close to the midpoint of the latent trait scale. Nevertheless, as the latent trait level moves closer to the most extreme point of the scale, the volume of test factual knowledge decreases considerably.

Item Response Theory Parameters

Item response theory parameters according to Reeve (2002) are independent on the group of examinees employed to determine the parameters, and are presumed to be uniform through different distribution in a population enquiry and across populations.

The *b* Parameter

The *b* parameter is likewise called the difficulty of the item (threshold). This is a position indicator alongside the *x*-axis that states exactly how simple or how hard a question may be. The indicator of a position of the item is the location on the *x*-axis where the curve crosses the 0.5 value of probability on the *y*-axis. Difficulty is defined as the

probability of a right answer, not in relations to the presumed difficulty or extent of effort that is necessary. Negative item difficulty estimates point out that the questions are very simple, whereas positive item difficulty estimations show that the questions are difficult. A question that is easy functions among the individuals that have low-ability, while a question that is hard functions among the individuals that have high-ability (Thorpe & Favia, 2012). That is *b* values (item difficulty) that are greater than 1 show a very hard item and items with low *b* (item difficulty) values below -1 specify easy items. When the *b* values are between -0.5 and 0.5, it means the test questions that have that range of difficulty indexes possess average difficulty level.

According to Baker (2001), hypothetically difficulty values can range from - ∞ to + ∞, but practically, difficulty values generally are in the range of - 3 to + 3. Ceniza and Cereno (2012), on the other hand gave the explanation for values of item threshold or difficulty (b-values) as follows: Very Simple = Less than -2, Simple = -0.50 to -2.00, Average = -0.49 to 0.49, Hard = 0.50 to 2.00 and Very Hard = Greater than 2.00.

The *a* Parameter

The *a* parameter is a value that can be communicated graphically by the gradient of the item characteristic curve (ICC). The *a* parameter similarly acknowledged as discrimination of the item (slope) indicates how well an item can discriminate amongst examinees with latent traits who are to the left of the item location from those with latent traits who are to the right of the item location (Thorpe & Favia, 2012). This item parameter expresses the way fastness of the likelihood of success changes with latent trait when it is close to the difficulty of the item. If *a* parameter (discrimination) has a positive value, it simply means that the examinees' with greater latent trait possess a great likelihood of responding to a question accurately and examinees' with lesser latent trait possess less likelihood to respond to a question accurately. When the *a* parameter

(discrimination of item) has a negative value , it means that examinees' who have great latent trait possess a minute likelihood of responding to a question accurately whereas examinees' who have less latent trait possess a greater likelihood of responding to a question accurately. An item with a large discrimination value has a great association amid the ability and the likelihood of achievement on that particular question. Furthermore, a question which has big discrimination value can differentiate better amid low and high ability levels of the latent trait.

The discrimination values (parameter a-values) of question that are good according to Adedoyin and Mokobi (2013), range from 0.5 to 2. If the rate of the discrimination of the item is above 1, this is usually a required value for a good examination question. Baker (2001) classifies the range and analysis of values for discrimination of the item in the following manner: very low; 0.01 to 0.34, low; 0.35 to 0.64, moderate; 0.65 to 1.34, high; 1.35 to 1.69 and very high; 1.70 and above.

The *c* Parameter

The *c* parameter is likewise called pseudo-guessing parameter. According to Thorpe and Favia (2012), it is the probability that a student who has very low ability may be able to respond rightly to a question by guessing and consequently has a greater-than-zero likelihood of getting the correct answers. A student who chooses answers to questions with four choices at random can answer these questions accurately approximately 1 out of 4 times, which means that the likelihood of guessing accurately is about 0.25. Eleje and Esomonu (2018), said that the questions which has 0.30 or greater *c* parameter (guessing) values are considered not very good questions, instead *c* parameter values of 0.20 or lesser are suitable. Akindele (2003) noted that there are no questions that have perfect *c* parameter values since students do not guess at random the minute they do not have the knowledge of the right answer.

Item Response Theory Dichotomous Models

The parameters of latent trait theory are independent of the testers used to create the parameters. In addition they are presumed to be invariable throughout the different subgroups in a population that is being researched on. According to Schumacker (2010), item response models are different depending on whether the association amid performance of the item and what has been learned is considered a one, two or three parameter logistic function. He further said that different item response theory models adjust for different item properties of item leading to diverse latent trait estimation. 1-parameter item response theory model adjusts for difficulty of the item, 2-parameter item response theory model explains the difficulty of an item and discrimination of an item and 3-parameter item response theory model takes into account the effect of difficulty of an item, discrimination of an item and guessing.

The one- Parameter Logistic Model or the Rasch Model (Item Difficulty)

The Rasch model was made known openly by Danish Mathematician named Georg Rasch in the 1960s. He advanced the assessment of statistics for test from a likelihood concept point of view. He employed a logistic function to deduce an item characteristic curve in place of the normal ogive function (however for a period he described his model in a different way), and his classic added to making the normal ogive model to be simple and the difficulty involved in the calculation. In the Rasch model, the discrimination parameter of the one-parameter logistic model remains permanent at a value of a = 1.0 for all the questions, the only parameter that varies is the difficulty parameter (b). For this reason, the Rasch model is frequently talked about as the one- parameter logistic model. The following is the equation for Rasch model,

$$Pi(\theta) = \frac{1}{1 + e^{-(\theta - bi)}}$$

The following is the equation for 1-parameter (1-pl) model,

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

Where P_i (θ) = Probability of getting the correct answer to item *i* of a person with ability θ, b_i = the difficulty of item *i*, e as a sign in the equation denotes the basis for natural logarithm, that has an unchanging value of 2.71828 while the value of D= 1.7 is the scaling factor.

The two-Parameter Logistic Model (Item Difficulty and Item Discrimination)

The two-Parameter logistic model permits the discrimination of the item parameter (*a*) to differ throughout the items rather than being compelled to be equal such as in the one-parameter logistic or Rasch model. That is two-parameter model makes use of the *b* parameter (item difficulty) and an additional element that indicates how well an item separates students into different ability levels, this parameter is called item discrimination (*a*). The item discrimination (*a*) parameter used in the two- parameter logistic model is equal to the slope of the item characteristics curve, while it is at its steepest. The 2-parameter logistic model trace line for the likelihood of an answer that is positive to item *i* for an individual with ability level θ is

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

Where P_i (θ) = Likelihood of obtaining the right response to item *i* of an individual with latent trait θ, b_i = item difficulty of item *i*, a_i = item discrimination of item *i*, *e* as a representation in the equation means the base for natural logarithm, that has a stable value of 2.71828, while *D*= 1.7 is the scaling factor. Discrimination parameter (slope), contain information about how an actual question is related to the latent variable that is being assessed. When the slope is higher the item responses attribute to dissimilarities in the latent variable becomes more inconsistent (Edwards, 2009)

The three –Parameter Logistic Model (Difficulty, Discrimination, and Guessing)

The three-parameter logistic model has additional parameter, a third parameter called pseudo-guessing for a question which is assigned the letter '*c*'. This is given by the intercept of the likelihood axis that indicates the likelihood of getting an answer correctly by guessing. The guessing parameter is distinctive to the question and is not dependent of examinees' latent trait. The three-parameter logistic model trace line for the likelihood of answering positively to item *i* for an individual with ability level θ is

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

where P_i (θ) is the likelihood of getting the right answer to item i of a person with latent trait θ, b_i is item difficulty of item i, a_i is item discrimination of item i, D= 1.7 is the scaling factor, while *c* represents the likelihood that individuals at very low levels of the ability response to item *i* correctly.

Item Parameter Estimation and Ability

The likelihood of a right answer in the latent theory models is contingent on the individual's latent trait and the parameters that characterize these questions. Since the real estimates of the item parameters are unknown, when a test is analysed under item response theory, one of the jobs done is to assess these parameters. The item parameter values that are obtained then offer facts as to the procedural characteristics of the questions in the test. This technique is called maximum likelihood estimation (Bock & Aitken, 1981). In item parameter assessment the process of performing the Mathematical operation is very demanding and has to be performed with the use of computer programmes specially intended for similar undertaking. The first software programme concentrated on Maximum likelihood assessment as a device for assessing the parameters of the item**.** These programmes in the long run had to incorporate various spur of the moment compelling instruments in order to evade several complications related to "pure" Maximum likelihood

assessment of item response theory item parameters. Earlier item parameter assessment procedures needed tests that were rather lengthy and samples that were very large (numerous thousands of students) so as to get the exact latent theory item parameter assessments.

Through the application of the Maximum likelihood procedure, logical assessments of item response theory item parameters can be deduced from short tests (as short as 30 items) and samples of students that are small (for instance 1000 students or fewer than that). Item response theory incorporates plain models for the likelihood of every probable answer to an examination and therefore its substitute term, 'probabilistic test theory'. Every endeavour at testing is heralded by a calibration study, that is, the questions are administered to students that are adequate. The answers obtained are then used to assess the parameters of the item.

According to Mallikarjuna and Natarajan (2012), assessment of the parameters of the item is an important task to be done when an examination is analysed under latent trait theory (Since real values are not known). These estimates provide facts about the procedural characteristics of the examination questions. If the parameters of the item are identified there and then the latent trait can be assessed, or on the other hand the parameters of the item can be estimated if the real latent traits of the examinees are identified.

Estimation techniques for item parameters in item response theory consist of (i) correlation method (ii) regression method (iii) approximation (PROX) method (iv) maximum likelihood procedure. Correlation method was used by Lord (1968) for estimation of discrimination of item indices for all the items in the test at the same time using factor analysis of the matrix of inter item correlation. The parameter estimating the item difficulty is evaluated by the normal deviate that matched the ratio of the subjects in

the total cluster of students that responded rightly to the item. Also discrimination index was estimated by the item loading in the factor analysis.

The regression method according to Adonu (2014), consist of the regression of an item on the ability based on the item characteristics curve. In this, the discrimination index of an item is the slope of the item characteristic curve while the difficulty of the item is the value of the ability θ at which the likelihood of right answers to an item is 0.5 or 50%.

Izard and White (1980) as cited in Spearitt,(Ed.), had an approximation (PROX) procedure for item analysis of the latent trait models. In PROX analysis, the answers given by examinees are listed in students' by item matrix. The analyses are based on marginal totals of correct and incorrect responses (frequency counts) for each item and subjects. The techniques estimate any examinee with perfect or zero scores. Items not attempted by any examinee are deleted from the matrix before calculation. This enables validation of the whole items. If N is the number of individuals that attempted an item, S is the total marks on a question that is total number of correct responses for a question**.**

$$b = Ln\frac{N-S}{S}$$

Likewise if Y be the number of right respondents to an item, N is the number of examinees that attended to the item. Izard and White (1980) established that the ability θ estimate is

$$\theta = Ln\frac{Y}{N-Y}$$

Lord (1968) also employed estimation procedure in statistics known as Maximum likelihood estimation. In this procedure there is conditional and unconditional Maximum likelihood. For conditional likelihood the data for item parameter are estimated on the basis of examinee's ability scores and the item difficulty indices. But for unconditional likelihood the student's ability score is removed from the estimation equation and so the item parameters are estimated with no respect to student's latent ability. On the whole,

Maximum likelihood estimation is a statistical method that determines the maximum likelihood function produced from the effect of a population distribution with the individual trace curve related with every item's correct or incorrect response. The pattern of the subjects score, 1 if correct, 0 if wrong is the basic data for the item analysis in this method. Among other technical steps, the indices of item parameters and estimate of individual latent ability scores are obtained.

Examinees Ability

The main purpose of administering a test under item response theory is to locate the examinee on the scale of ability by way of evaluating how much underlying ability he or she has. Based on this, the examinee can be assigned grades, award scholarships and so on. Procedure to estimate the score of ability (that is the parameter) for an examinee is that the test that is used to assess an ability that is not known comprises N items, every one of which assesses certain aspect of the latent trait. To estimate an examinee's parameter of the latent trait that is not known, it is presumed that the statistical values of the parameters of the item are known. The immediate significance of this assumption is that the metric of the scale of latent trait will be similar to the metric of the known parameters of the item. As soon as the examination has been done with, an examinee answers to every of the N questions in the examination, and the answers will be scored dichotomously (Score of 1 or 0 for each item). This set of 1's and 0's for the N items is termed the examinee's item response vector. The item response vector thus obtained and the parameters of the item that is known will then be used to estimate the examinee's parameter of the latent trait that is not known.

In view of the techniques described above that are used in latent trait theory to estimate parameters, the Maximum likelihood technique is frequently used and more

preferred. This is because the Maximum Likelihood technique produces comprehensively the item parameters and ability estimates for all examinees (Baker, 1977).

Analysis of Fit

The analysis of statistical fit according to Obinne (2013) is an examination of the internal validity. Inside the ability test model, the internal validity of a test is evaluated in relations to the statistical fit of every single item to the model. Adonu (2014) opined that if the statistical fit of an item is adequate, hence the item is valid. The item response theory has three models that are dichotomous models: 1-parameter model, 2-parameter model and 3-parameter model. When a given set of questions fits the model, this is the indication that the questions relates to unidimensional ability. Model-fit, furthermore denotes that discriminations of items are identical and significant, also that there are no mistakes during scoring of the item. According to Adonu (2014) when the fit statistics is positive and big it shows that there is no fitting whereas when the statistics is low closer to one, this shows a better fit.

Model-fit analysis refers to an examination of whether the statistical model used in an application sufficiently describes the significant characteristic of the set of data that is available. Model selection denotes the choice of the statistical model that defines the data best amongst numerous models that are contending. In order to evaluate which item response theory model ought to be used, three conditions summarised in Hambleton and Swaminathan (1985) are normally used. The first is to verify the model assumptions. Secondly, the anticipated model characteristics should be examined then; thirdly model predictions of actual test results should be examined. The three conditions are a summary of the conditions that test evaluators or constructors can possibly put to use. These conditions can further be classified into subgroup.

Validating the model Assumptions

Unidimensionality

The assumption of unidimensionality states that a test ought to assess only one underlying latent trait in a test. This prerequisite applies to most item response theory models. Wiberg (2004) proposes that assumption of unidimensionality can be examined by using the eigenvalues in a factor analysis. Unidimensionality is established if the eigenvalues is being plotted (from the biggest to the smallest) of the inter-item correlation matrix there is one first factor that is predominant. Alternative possible way to determine unidimensionality is to compute the ratio of the first and second eigenvalues. If the ratio is high, that is, above a critical value the test is unidimensional.

Equal Discrimination

Verification of equal discrimination can be carried out by determining the correlation between items *i* and the total score on the test score, that is, the point biserial correlation or with the biserial correlation. The standard deviation should be small if there is equal discrimination. In the opinion of Hambleton and Swaminathan (1985) whereby the items are not discriminating equally then it is better to make use of the 2PL model or 3PL model rather than the 1PL model.

Possibility of Guessing the Correct Answer

One and only method of investigating if guessing occurs is to investigate how the test takers with low abilities respond to the most difficult items in the test. Guessing can be disregarded from the model if the examinees with low ability answer the most difficult items wrongly. If the examinees with low ability answer the most difficult items correctly to some extent a guessing parameter for guessing ought to be incorporated in the model. That is the 3-Parameter Logistic model is more suitable than the 1-Parameter Logistic model or the 2-Parameter Logistic model.

Expected Model Structures

The next condition, expected model structures is of importance irrespective of the model that is being used. First, it is necessary to scrutinise the invariance of the ability parameter assessments. This means that the estimations of the abilities of the examinees θ, should not depend on whether or not the items in the test are easy or difficult (Hambleton, Swaminathan & Rogers, 1991). Secondly, it is also necessary to scrutinise the invariance of the item parameter assessments. This means that it must not matter if the parameters of the item are estimated with separate groups in the sample that is groups with low or high abilities. Shepard, Camilli, and Williams, (1984) said that there ought to be a linear correlation between these estimates and that it is most simply scrutinised by means of scatter plots.

Model Predictions of Real Test Results

The third condition which is model prediction of real results of the test can be scrutinised by comparing the item characteristic curves (ICC) for every item with each other. This condition can also be scrutinized by means of plots of observed score and anticipated score distributions or chi-square tests (Wiberg, 2004). According to Hambleton et al (1991) Chi-squared test can be used to investigate degree to which the models predict observed data.

A number of model-fit statistics for item-fit index, for dichotomous latent trait theory models had been proposed (Orlando & Thissen, 2003) to evaluate the suitability of the chosen latent trait theory models and calibration procedure in terms of the model-data fit for, 1-parameter logistic model, 2-parameter logistic model and 3-parameter logistic model. Wells, Wollack, and Serlin (2005) was of the opinion that, fit of model to the data must accurately demonstrate the true association between latent trait and achievement on the item. Orlando and Thissen (2003) were of the opinion that the proper use of models is

predicated on the principle that a number of item response theory suppositions are made about the property of the data, to ensure the model perfectly represents the data. When these assumptions are not complied with, conclusions concerning the property of the items and tests can be inaccurate, and the possible benefits of using item response theory are not earned. Failure to ensure the suitability of model-data fit analysis carries the risk of making incorrect decisions (Sinharay 2005).

Essen et al. (2017), noted that the assessment of model-data fit ought to be on the basis of three types of evidence. Firstly, the validity of the assumption of the data model set such as: (a) unidimensionality, (b) un-speeded test, (c) negligible guessing for 1-plm and 2-plm, (d) equal discrimination of all the items for 1-plm. Secondly, that the expected properties are obtained to reflect invariance of item and ability parameter estimates. Finally, the accuracy of the model prediction should be assessed through the analysis of item residuals. In addition, Sijtsma and Hemker (2000) and Sheng (2005) suggested that the basic latent theory assumptions of unidimensionality, local independence, and item characteristic curve should be properly assessed as standard measures to investigate model-data fit analysis.

Zhoa (2008) suggested that the model-fit to the test data ought to be on the basis of four stages. The first is to choose a software and initial analysis. The next step is to check the unidimensionality assumption and local independence assumption. Then fit of the model-data is assessed and lastly parameter invariance of the model, parameter invariance of the item and parameter invariance of the ability are checked.

Model-Fit Estimation Techniques

In order to estimate the parameters of the model for any item response theory model, consideration of the test takers and the items in the study are essential. Fundamental assessment methods for latent trait models presume that the persons taking

part in the study are not dependent on one another and that, items behave in the same way for all persons.

Approaches to Item Response Theory Model-Fit Assessment

Assessment of item response theory fit of model-data includes the statistical technique and the graphical method. A general method for estimating model-fit at the item level involves comparing observed data with model predicted expectations (Hambleton & Han, 2005). Both statistical significance tests and graphical analyses based on the residuals can be used for this purpose. A goodness-of-fit study can generally be defined as the assessment of the relationship between the result that has been observed and anticipated (predicted) results. In the framework of item response theory, this usually entails first and foremost, assessing the parameters of an item response theory model, secondly making use of those parameter assessments to predict, by the method of the item response theory model response formats of the test taker, and thirdly comparing the anticipated response formats to authentic perceived response formats of the test taker.

Traditional Method: Chi-Square (χ^2) Based Fit Statistics

Many of the traditional χ^2 based techniques for estimating model-fit are on the basis of the following steps:

1. Estimate item and ability parameters.
2. Place examinees into groups along the ability scale using proficiency estimates.
3. Each of the latent trait observed score response distribution has to be constructed by cross-classifying test takers making use of their latent trait assessments and responses for the score.
4. Develope a predicted score response distribution across score classifications making use of item parameter evaluations and latent trait assessments which represents each latent trait group.

5. Observed and anticipated score response distributions are compared making use of a χ^2 fit statistic or examination of residuals. These are:

Pearson's Chi-Squared Tests

Chi-square (χ^2) test is centred on the dissimilarity amid the observed and the anticipated estimates for every single group. χ^2 statistic can be defined as

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where *Oi* is the perceived total number of samples in group *i*, and E_i is the anticipated total number of samples in group *i*. This Chi-square statistic can be gotten by computing the dissimilarity between the observed number of cases and the anticipated number of cases in every group. This dissimilarity is squared and divided by the anticipated number of cases in that group. These estimates are then added for all the groups, and the total is referred to as the Chi-squared estimate.

Yen's Q1. Yen (1981)

Yen's Q1 index that is suggested for use with dichotomous items is computed in the following way:

$$Q1 = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})}{E_{ij}(1 - E_{ij})}$$

Where N_j is the total of test takers in latent trait interval j; O_{ij} is the observed proportion of test takers in interval j whose response item i is correct; and E_{ij} is the likelihood centred on the model in interval j responding to item i rightly. The number of intervals is stable and it is 10 (though, if the expected frequency in an interval is less than five, then the groups may be combined so that the expected frequency is greater than 5). Yen's Q1 under the null hypothesis, is distributed as a χ^2 with degrees of freedom equal 10 minus the number of parameters actually assessed.

Likelihood-Ratio Test

The likelihood ratio test model-fit statistics (McKinley and Mills, 1985), designated G^2, is described by two frequently used item response theory software packages that are commercial, BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 2003) and PARSCALE (Muraki & Bock, 2003). It can be used to compare the goodness- of-fit of two statistical models, a null model against an alternative model. This is based on the likelihood ratio that states exactly countless times most probably the data are under one model than the other. This likelihood ratio, or its equivalent logarithm, can be used to calculate a *p*-value, or compared to a critical value to conclude if the null model can be rejected or not.

Once the logarithm of the likelihood ratio is used, the statistics is recognized as a log-likelihood ratio statistics, and the likelihood distribution of this test statistics, in assumption that the null model is true, can be estimated. In the case of differentiating between two models, every one of them has no unknown parameters, it can be reasonable to make use of the likelihood ratio test which determines that such a test has the highest statistical power amongst all competitors (Casella & Berger, 2001). On the other hand the likelihood ratio test statistics, G^2, is a Chi-square-based statistics and can be computed as the discrepancy of two deviances from two models that are being compared. The discrepancy for two models on its own is distributed as a χ^2 and this can make it to be subjected to significance tests to decide which model fits better (Baker & Kim, 2004).

G^2 is computed for item *i* in the following way:

$$G^2 = 2\sum_{h=i}^{n_g}\left[r_{hi} \log \frac{r_{hi}}{N_h p_i\left(\overline{\theta_h}\right)} + \left(N_h - r_{hi}\right)\log \frac{N_h - r_{hi}}{N\left[1 - p_i\left(\overline{\theta_h}\right)\right]}\right]$$

where n_g specifies the amount of intervals; r_{hi} signifies the observed occurrence of right answers for item *i* in interval *h*; N_h is the total of persons in interval *h*; and $P_i\left(\overline{\theta_h}\right)$ denotes

the model-predicted proportion right for item *i* at $\overline{\theta_h}$. $\overline{\theta_h}$ is the mean of theta assessments in interval *h*. G^2 is distributed under the null hypothesis approximately as a χ^2 with degrees of freedom equal to the number of intervals. Fundamentally G^2 is a log-likelihood based statistics where the test takers are convened into latent trait intervals based on their latent trait assessments, and at that point the experimental amount right against the model-based amount right within that interval is assessed, increasing across intervals.

Orlando and Thissen's *S-X²*.

Orlando and Thissen (2000) redressed the weakness of Yen's Q1 statistics by grouping test takers on the basis of their number right 10 (NC) scores in its place of their model-based latent trait assessments. The fit statistics S-X^2, based on Yen's (1981) Q1 statistics, is approximately distributed as a χ^2 with *G-m* degrees of freedom, with m describing the number of item parameters assessed and G is the total number of score groups (*G* = n-1 when no groups have been collapsed, where n is total number of score points). The general formula for testing the fit of dichotomous items is as follows:

$$S - X^2 = \sum_{t=t_{\min}}^{t_{\max}} \sum_{c_i=1}^{m_i} \frac{\left(O_{tci} - E_{tci}\right)^2}{E_{tci}}$$

Where *t* is summation of score, m is total number of groups, O is observed number of occurrence and E is expected number of occurrence, *i* is item. Cells are collapsed before computing *S-X²*, in order to achieve anticipated cell number of occurrences over a given number so as to circumvent scanty anticipated number of occurrences.

Lagrange Multiplier (LM) Test

Glas (1998), proposed a Lagrange Multiplier (LM) procedure to evaluate item and latent trait parameter assessments for 2PLM and 3PLM. Rationale behind LM tests is to test a restricted model in comparison to a more universal alternative, where the restricted

model is the derivative of the universal model by exacting restrictions on one or more parameters. The Lagrange Multiplier test is built on the assessment of the first-order partial derivatives of the log-likelihood function of the universal model, assessed by making use of the maximum likelihood assessments of the model that has been constrained. One of the benefits of making use of Lagrange Multiplier method for the assessment of fit to the item is that the asymptotic distribution of the statistics that is involved is compliance from asymptotic concept (Glas & Falcón, 2003)

Non Traditional Estimation Techniques

Likelihood Based Estimation Method (Likelihood Algorithm)

According to Palmieri (2012), the latent trait theory models link the likelihood of a correct answer to the person's and items' parameter. Assuming for each examinee the answers to all the items in the test are independent given his or her ability and that the examinee's responses are independent from each other, the likelihood function is thus:

$$L(\theta,\beta \mid y)=\prod_{i=1}^{n}\prod_{j=1}^{k}P_{ij}^{yij}\left(1-P_{ij}\right)^{(1-yij)}$$

Maximum Likelihood Technique

The procedure of Maximum likelihood (ML) is a technique to assess the parameters of a model and test hypotheses about those parameters. Two vital components exist in maximum likelihood. These are data set (which is essential for all investigations) and a Mathematical model that defines the distribution of the variables in the set of data. The model that describes the distribution of the variables will have certain unknown quantities in it. These are called parameters. The purpose of Maximum likelihood is to find the parameters of the model that best explain the data with the intention of producing the biggest likelihood of explaining the data.

Akaike Information Criterion (AIC)

Akaike information criterion (Akaike, 1974) can be said to be an estimation of the relative quality of statistical models for a known data set. When a known group of models for the data has been specified, Akaike information criterion assesses the quality of each model, in relation to every single of the other models. Thus, Akaike information criterion offers a means for selection of model. Once the maximum likelihood assessment of the parameter of a model is determined it is easy to calculate Akaike information criterion. The AIC can be defined as;

$$AIC = d + 2k$$

Where $d = 2 \times \log$ (maximum likelihood) attainable by the model and k is the number of model parameters (Akaike 1974). The model that has the minimum AIC value is the best model, and the models are not required to be nested (Liddle, 2007).

Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) (Schwarz, 1978) is a standard for the selection of model amongst a limited set of models. The model that has the least BIC is selected. It is based, in part, on the likelihood function. When fitting models, it is possible to increase the likelihood by adding parameters, on the other hand over fitting could be as a result of doing that. The BIC is defined as,

$$BIC = d + k \log N$$

Where $d = 2 \times \log$ (maximum likelihood), k is the total number of the parameters and N is the sample size that is used in the fit.

Bayesian Based Estimation Method (Bayesian Algorithm)

The Bayesian technique for assessment of item response theory models is the same as that of the marginal likelihood procedure. Nevertheless, additionally in presuming a distribution that is mixed for the tendencies, Bayesian analysis fixes a prior distribution on

every parameters of the model. Assessment of posterior quantities can equally be carried out concurrently for both the items and the test takers in the set of data. Bayesian analysis of an item response theory model has some deficiencies, one of which is that the posterior distributions have to be approximated using numerical integration procedures (Johnson, 2007).

The Bayesian estimation techniques perform better than the Maximum likelihood methods when the test size is short and with small sample size. (Ojerinde, Ojo & Popoola, 2012).

Posterior Predictive Model Checking

The posterior predictive model checking (PPMC) technique (Rubin, 1984) is a well-known Bayesian model checking instrument since it is not sophisticated, very strong theoretical foundation, in addition is connected closely to the classical goodness-of-fit, (Sinharay, 2016). The technique mainly involves comparing the data that has been observed with the simulated data, those anticipated by the model, making use of procedures that are not consistent. Basically numerous simulated sets of data are created from the predictive distribution of duplicated (simulated) data subject to the observed data (known as the posterior predictive distribution). Deficit on the part of the model to explain the data is indicated by the systematic discrepancy between the observed sets of data and duplicated sets of data.

Bayes Factor

In the opinion of (Lee, 2012) Bayes factor is a proportion of the probability of two contending hypotheses, typically a null and an alternative hypotheses. The use of Bayes factors is established in the situation where inference rather than decision-making are not certain. This is simply to ascertain which of the hypothesis is true, instead of really taking a decision based of this information. Mathematically, the idea is operationalized using

Bayes Theorem or Bayes formula,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

If the likelihood of the event A is P(A), then the formula states that, after observing event B, the uncertainty about A can be updated to P(A\B), based on the information that B provides, according to the Bayes formula.

Cross-Validation Log-Likelihood (CVLL)

One of the most commonly used techniques of evaluating predictive performances of a model is Cross-validation log-likelihood (CVLL). It is based on the theory of the pseudo-Bayes factor (PsBF; Bolt, Cohen, & Wollack, 2001), which is given a priori or established by a modelling procedure. Fundamentally, based on data splitting, fragment of the data is used for fitting every single contending model and the rest of the data is used to assess the predictive performances of the models by the validation errors, and the model that performed best is selected. That is two different samples are drawn, a calibration sample, Y_{cal} in which test takers will be sampled randomly from the test takers that take the test and a cross-validation sample, Y_{cv} in which a second sample will be drawn randomly from the test takers that are remaining. The sample that is used for calibration will be used to update prior distributions of parameters of the model to posterior distributions. In the opinion of Bolt and Lall (2003), the likelihood of the Y_{cv} for a model is then computed with the use of the updated posterior distribution as a prior:

$$P(Y_{cv} \mid \mathrm{mod}\, el) = \int P(Y_{cv} \mid \theta_i, Y_{cal,}\, \mathrm{mod}\, el) f_\theta(\theta_i \mid Y_{cal,}\, \mathrm{mod}\, el) d\theta_i$$

Where, $P(Y_{cv} \mid \theta_i, Y_{cal,}\, \mathrm{model})$ represents the conditional likelihood, and $f_\theta(\theta_i \mid Y_{cal,}\, \mathrm{model})$ is the conditional posterior distribution.

Deviance Information Criterion (DIC)

Spiegelhalter, Best, Carlin, and Van der Linde (2002) developed Deviance information criterion (DIC) in order to be able to deal with Bayesian posterior assessments of parameters of the model. Deviance information criterion comprises a Bayesian measure of fit known as the posterior mean deviance $\overline{D(\theta)}$, and a penalty for model complexity, P_D the number of free parameters in the model.

$$DIC = \overline{D(\theta)} + P_D = D(\bar{\theta}) + 2 \times P_D$$

where $\overline{D(\theta)}$ is a Bayesian measure of fit, $D(\bar{\theta})$ is the deviance of the posterior model (that is, the deviance at the posterior evaluations of the parameters of interest), and $P_D = \overline{D(\theta)} - D(\bar{\theta})$

Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) techniques are more universal in that many more statistical models can be used to fit by them. Usually they comprise numerous separate stages which makes it simple so that the algorithms will be used on more structures that are complex. These processes are on the basis of simulation so that rather than simply producing point estimates the methods are run for many iterations, and at each iteration an estimate for each unknown parameter is produced. At every iteration the estimations will not be independent; the estimations from the previous iteration are used to yield assessments that are new. The objective of the method at that point is to produce a sample of values from the posterior distribution of the unidentified parameters. This shows that the procedures are valuable for generating precise interval estimations. The Markov Chain Monte Carlo (MCMC) can take a lot of time to compute for large sets of data and utmost care needs to be taken to make sure that the resultant assessments to the posterior distribution are useable.

Least Squares

The technique of least squares is a typical method in regression analysis to assess the result of over determined procedure, in order words, arrays of equations comprising additional equations than unknown variables. Least squares mean that the general result reduces the sum of the squares of the residuals created in the outcomes of every single equation. The utmost significant operation is in fitting of data. The best fit when it comes to least-squares reduces the sum of squared residuals (a residual is the disparity amongst the value that is observed, and the value supplied by a model that is fitted). When there are considerable uncertainties of problem in the independent variable (the *x* variable), then simple regression and least-squares approaches have hitches; in such circumstances, the procedure necessary to fit errors-in-variables models might be taken into consideration rather than that for least squares.

There are two types of Least-squares complications; linear or ordinary least squares and nonlinear least squares, subject to if or not the residuals are linear in all unknowns. The problem of the linear least-squares take place in statistical regression analysis, it has a closed-form solution. The problem of nonlinear is typically solved by iterative improvement. At every iteration the method is estimated by a linear one, therefore the main computation is the same in two illustrations.

One of the common assumptions fundamental to most process of modelling approaches, linear and nonlinear least squares regression inclusive, is that, every point of data offers exact facts concerning the deterministic part of the total disparity in the procedure. In addition, it is expected that the standard deviation of the error term is steady above all values of the predictor or descriptive variables. This assumption obviously may not be true, tantamount to approximating in each modelling exercise.

The normal linear regression model $y = a + bx$ is of the assumption that the entire random error components are the same and not dependently distributed with variance that is steady. The failure of the assumption to hold, ordinary least squares estimator of regression coefficient loses its property of least variance in the category of evaluators that are linear and unbiased.

A number of studies were reviewed. Rijn, Sinharay, Haberman and Johnson (2016), assessed the fit of latent theory models that was used in large-scale educational survey investigation, using log-likelihood. They used two types of approaches for the estimation of absolute fit of the NAEP IRT model; item-fit analysis making use of residuals and generalized residual analysis. They used ETS mirt software (Haberman 2013) to calculate residuals for fit to the item and to perform the computations for the generalized residuals for the NAEP data sets. The data used was NAEP 2004 and 2008 long term trend mathematics assessment which was for the nine year olds, NAEP 2002 and 2005 reading which was for grade twelve. The assessment contained a total of 145 multiple choice items and constructed-response items which were distributed over 38 booklets. The data set comprised 26,800 test takers with 14,700 students from the 2002 sample and 12,100 students from the 2005 sample. Their result showed that significant misfit (in the form of significant residuals) was established statistically for all the sets of data.

Kose (2014) assessed the model-fit of unidimensional models using computer-generated data of 1000 examinees and found that the computer-generated data were fit for the 2pl model estimated using -2log likelihood ratio. Bulut (2015) investigated the fit of unidimensional item response theory models to the Entrance Examination for Graduate Studies which is a high-stake large-scale estimation in Turkey required for the application to graduate programme in Turkish universities. Findings suggested that the 3-parameter item response theory model revealed the best model-data fit for the Entrance Examination

for Graduate Studies when he compared the overall fit of 1-parameter logistic, 2-parameter logistic, and 3-parameter logistic models on the basis of Likelihood ratio test.

Hishamuddin and Siti (2016) explored the best suitable model that could be employed in the analysis of dichotomous items of the Anatomy and Physiology course. The study comprised 971 student nurses studying in the Ministry of Health Malaysia training colleges. Exploratory factor analysis was executed on the data retrieved from final examination paper comprising 40 multiple-choice items. The results of the analysis revealed that the assumptions of unidimensionality and local independence were met. The data calibration was executed by the means of an IRT-based software, X-calibre which is based on the negative twice the log-likelihood statistic (-2LL). The results revealed that the 3PL model is the best suitable model for analysing the data of the study. The study came to the conclusion that the 3PL model ought to be given precedence in analysing the items that are scored dichotomously which contain guessing components.

Galli, Chiesi and Primi (2010) made use of latent theory in the assessment of Mathematics competence in introductory statistics courses. The research involved 600 psychological students (84% females) of the University of Florence. The age range were from 19 to 58, with a mean age of 21.19 (SD = 4.01) years. They registered for two year programme in introductory statistics courses (2007 and 2008). The disparities of -2loglikelihood for models that were nested were calculated using Multilog software. The results revealed that the 2PL model was the best appropriate model analysing the PMP.

Maydeu-Olivares and Montaño (2013), investigated the performance of three statistics, $R1$, $R2$ and $M2$ to evaluate the inclusive fit of a one-parameter logistic model assessed by (marginal) maximum likelihood (ML). R_1 and R_2 were explicitly intended to target precise assumptions of Rasch models, while M_2 was intended for an overall test statistics. The three statistics were found to have great power than Pearson's X^2 compared

to two and three-parameter logistic that can be alternatives (2PL and 3PL), and also compared to multidimensional 1PL models. The outcomes imply that there is no perfect benefit with the use of goodness-of-fit statistics precisely intended for Rasch-type models to assess these models when marginal maximum likelihood assessment is used.

Kang and Cohen (2007) looked into the comparison of the several methods of relative model fit and how they performed. Akaike information criterion (AIC), Bayesian information criterion (BIC), the Likelihood ratio (LR), Deviance information criterion (DIC) and Cross- validation log likelihood (CVLL) test in order to select a latent theory dichotomous model using answers of Grade 8 test takers to the 1996 State NAEP Mathematics test. The DIC, CVLL, LR and AIC selected the 3PM as the model that is the best. The Cross-validation log-likelihood (CVLL) seemed to be the unsurpassed of the five models for the circumstances assumed in the study.

Kang, Cohen, and Sung (2009) also looked into the comparison of LR test, AIC, BIC, DIC, and CVLL in their performance in the selection of the right model amongst the graded response model (GRM), the partial credit model (PCM), the generalized partial credit model (GPCM), and the rating scale model (RSM). They random sampled 3,000 examinees for the calibration sample from the 2000 state NAEP Mathematics test data. The estimates for each one of the five techniques, LR test, AIC, BIC, DIC and CVLL were computed. They found that in view of the fact that the computer-generated situation affected the performance of every one of the technique, CVLL performed very well on the average.

Li, Cohen, Kim and Cho (2009), examined the techniques for model selection used to assess dichotomous mixture item response theory (IRT) models. The sample comprised 1,200 test takers who were drawn randomly from the total statewide sample of 172,887 pupils in Grade 3. They considered five indices; Pseudo-Bayes factor (PsBF), and

Posterior predictive model checks (PPMC), Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC). The five techniques offer rather endorsements that are not the same for a set of data that are real. The results from a computer-generated study showed that BIC selected the right model fit under maximum conditions that were replicated and for the three of the dichotomous mixture item response theory (IRT) models that were considered, they found that PsBF was also equally efficient. AIC and PPMC have a tendency to select the model that is more complex under certain situations. DIC was the least in terms of effectiveness. They concluded that model selection indices are not in agreement every time.

Wang and Liu (2006) compared Akaike information criterion (AIC) and Bayesian information criterion (BIC) methods, in choice of stock–recruitment associations to fit fish stock–recruitment (SR) data to a SR statistical model. They used the maximum likelihood technique to fit the six statistical SR models on six sets of computer-generated SR data. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) techniques were used to choose the associations that were the best respectively. These have benefit of testing the significance of the difference between the functions of model specifications that are not the same. The exercises were similarly administered on eight sets of fisheries SR data that were real. The outcomes revealed that both AIC and BIC are valid in choosing the SR association that were highly appropriate. As regards to the models that are nested they found out that, BIC is better than AIC.

Kim and Bolt (2007) estimated latent theory model using Markov Chain Monte Carlo method and found the 2pl model was fit to the item response data from a university Mathematics placement using WINBUG 1.4 software. Farnsworth, Kang and Ragan (2016) compared the 1pl, 2pl and 3pl models to determine the best suitable model for the analysis of Standard Assessment of Concussion (SAC) using a sample of one hundred and forty-

five high school athletes. The model fit was assessed by χ^2 and Bayesian Inference Criterion (BIC). He found that the 1pl model was appropriate for SAC using χ^2 and BIC.

Lou and Al-Harbi (2017) compared how the Likelihood ratio test (LRT), Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC), Widely available information criterion (WAIC) and Leave-one-out cross-validation (LOO) performed with a population of 36886 student using 52 items and different sample sizes of 500, 1000 and 2000 . They found out that the performance of WAIC and LOO were better than that of LRT, AIC, BIC and DIC when data were generated using 3PLM.

Adedoyin and Mokobi (2013), used MULTILOG to assess item level and model fit statistics in a 3-parameter logistic model with 2010 Botswana Junior Certificate Examination Mathematics paper one. The Mathematics paper 1 comprised forty (40) multiple choice test items which were generated by means of the three year JC Mathematics curriculum. The total population for the study was all the 36,940 learners who sat for the JC Mathematics examination in 2010, out of which a sample of 10,000 learners were randomly selected by the use of SPSS computer software. They made use of a χ^2 goodness-of-fit statistics in assessing item fit to 1PL, 2PL and 3PL models. The outcomes revealed that 10 items fitted the 1PL, 11 items fitted the 2PL model and 24 items fitted the 3PL models. Therefore, the 3PL model was used for the analysis.

Likewise, Essen (2015), examined model-data fit of 50 item dichotomously scored JAMB Mathematics items data with Chi-square goodness-of-fit statistics using BILOG-MG, 3.0 software programme. No item fitted the 1-parameter model, 26 items fitted 2- parameter model, while 3-paramater model displayed some irregularities. Therefore, the 2-paramter logistic model was best for the data.

Essen et al. (2017), examined item level diagnostic statistics and model-fit to the data with 1pl and 2pl models by means of IRTPROV3.0 and BILOG- MG V3.0. The design of ex-post facto was used for the study. The population for the study comprised 11,538 students' answers who participated in Type L 2014 Unified Tertiary Matriculation Examination (UTME) Mathematics paper in Akwa Ibom State. Stratified sampling technique was used to select the samples of 5,192 (45%) students randomly. BILOG-MG V3.0 and IRTPROV3.0 computer software were used to calibrate the students' answers. Pearson's χ^2 and S - χ^2 statistics as an item fit index for dichotomous latent theory models were employed. The results showed that only 1 item fitted 1-parameter model in BILOG-MG V3.0 and IRTPRO V3.0. In addition, the results showed that 26 items fitted 2-parameter models when BILOG-MG V3.0 was used. In IRTPRO five items fitted 2-parameter models.

Also Dodeen (2004) used BILOG 3.11 software to fit the 3PL model to the created sets of data and for calculating the values of the χ^2 and G^2 statistic. The statistics S−χ^2 and S−G^2 were calculated by the use of GOODFIT programme. The ratio for the S−χ^2 and χ^2 were significantly low and near to the minimal level for all test conditions. The statistics χ^2 and G^2 were calculated using the IRTFIT, RESAMPLE programme. The item fit statistics mean, the ratio of item fit statistics were significant at 1percent (1%) level and the relationship between the item parameters were used for generation and the item fit statistics mean over the 100 replications under any test condition were calculated within the stipulations of every one of the nine test.

Yalcin (2018), compared the fit of data of mixture latent trait theory models and classical model on a sample of 5000 students using 19 multiple choice items in TEOG science and technology subtest. Item fit statistics S-G^2 were computed using Goodfit, all the items fit 2pl model.

In the study by Chernyshenko, Stark, Chan, Drasgow, and Williams (2001) fit of several IRT models were compared to two instruments used for personality evaluation. Data were collected from 13,059 persons who responded to the US-English version of the Fifth Edition of the Sixteen Personality Factor Questionnaire (16PF) and 1,770 persons who responded to Goldberg's 50 item Big Five Personality measure were analysed. Two of the parametric models designed for dichotomously scored items (that is, the 2pl and 3pl models) were investigated and a parametric model for polytomous items (Samejima's graded response model). χ^2 statistics was used for model-data fit estimation. The result showed that presenting a guessing parameter had slight influence on data fit to the model and as such the 2PL model was suitable for these measures.

Nworgu and Agah (2012) applied three parameter logistic model (3PLM) in the calibration of a Mathematics achievement test. The aim of the research was to use 3PLM of latent trait theory in the calibration of a Mathematics achievement test. The sample of the study was 1514 SS III (SS3) students from Rivers and Cross river states. The instrument for the study constructed by the researchers was a 40 multiple choice test items. The data analysis was done using BILOG-MG3, IRT computer software that estimated the item parameter and their corresponding standard error of measurement. The Chi-square goodness of fit was used to decide the goodness-of-fit of the items of the instrument to three-parameter logistic model. The study also generated item characteristics curve to ascertain if the items in the tests were adequate for the evaluation of the abilities of the students. A reliability coefficient of 0.79 was obtained. The indices for the item parameter obtained showed that the discrimination parameter (a) ranged from 0.29 to 2.05; item difficulty from -0.40 to 1.79; the likelihood of guessing correctly in the test ranged from 0.02 to 0.50 for all the ability levels.

Summary of Reviewed Literature

The reviewed literature in this study looked at the theoretical background of Item response theory (IRT). The explanations of rudimentary conceptions in the latent trait theory were made. These were on the basis of the latent trait theory dichotomous models, the 1-parameter (1PL), 2-parameter (2PL) and 3-parameter logistic (3PL) models. The theory guided the researcher in ascertaining and deciding on ideas that roused the mind of the researcher in selecting the suitable measurement framework in analysing the scores of the students. The theory is appropriate for this study in that IRT expresses the relationship between an examinee's given answer to an item and the ability (dormant or hidden concept) that is being measured by the instrument. Also in IRT items are scored 1 for correct answer and 0 for wrong answer. The instrument here was National Business and Technical Certificate Examination Mathematics paper 1 May/June 2018 which is also scored dichotomously.

From the empirical studies reviewed the studies carried out were item response theory dichotomous model-data fitting using one or two estimation technique and comparison of two or three techniques which were very scanty. Most of these studies were done in developed countries and the researcher is aware that issues of abilities may be influenced by environmental background. Most of the studies reviewed used simulated data. Simulated data are computer generated responses.

Most of the studies reviewed focused on certain aspect of model fit such as Kose (2014) carried out a study on assessing model fit of Unidimensional models using simulated data and employing the software BILOG for the assessment of model-data fit. In the same vain Adedoyin (2015) carried out a study on psychometric analysis in investigating the quality of Botswana 2010 Junior Certificate Mathematics multiple choice

examination in Botswana using IRT dichotomous models with the softwares MULTILOG 3.0 and BILOG MG3

During the review, no study could be assessed that was conducted in Nigeria that used more than one technique. What had been found was the use of Chi-square goodness of fit for model-data fitting. The Chi-square is mostly used for absolute model-data fit. This study stands out as it compared these five model-fitting estimation techniques (Likelihood ratio, Akaike information criterion, Bayesian information criterion, Deviance information criterion and Cross-validation log-likelihood) for model-fitting using relative fit and real data from the field (candidates' responses). This will go a long way in filling the gap which exists during test construction and standardization in Nigeria.

CHAPTER THREE

METHODOLOGY

In this chapter the procedures that were used to conduct the study are organized under the following sub-headings;

- Design of the Study
- Population of the Study
- Sample and Sampling Technique
- Research Instrument
- Validity of the Instrument
- Reliability of the Instrument
- Method of Data Collection
- Method of Data Analysis

Design of the Study

This study employed the descriptive survey research design using ex-post-facto method. The survey design is considered most appropriate because only a part of the population was studied and finding was used to generalise for the whole population. The Ex-Post Facto was considered since secondary data was used and the instrument was not re-administered by the researcher. The independent variable in this thesis was Technique which has five categories. The dependent variable was model-fit.

Population of the Study

The population of this study consisted of forty-nine thousand, five hundred and eighty one (49,581) candidates who enrolled and sat for the National Business and Technical Certificate Examination (NBTCE) 2018 May/June Mathematics multiple choice examination in the six Geo-political zones in Nigeria.

Table 1; Population distribution of the number of candidates in the Geo-Political Zones in Nigeria

Geo-political zones	Number of candidates
North-Central	18969
North-East	8069
North-West	8021
South-East	3640
South-South	3528
South-West	7354
Total	**49581**

Source: Information and Communication Technology Department, NABTEB (2018)

Sample and Sampling Technique

The sample size for this study was four thousand nine hundred and forty eight (4,948) candidates who sat for National Business and Technical Certificate Examination (NBTCE) 2018 May/June. The Multistage simple random sampling technique which involves sampling stages was adopted for the selection of the sample.

Stage 1: Simple random sampling procedure was used to select two zones (South- East and South-South) from the six Geo-political zones which are North-Central, North-East, North-West, South-East, South-South and South-West zones in Nigeria.

Stage 2: Simple random sampling was also used to select three states each (Anambra, Enugu and Imo) from the South-East zone while Cross-River, Delta and Edo states were randomly selected from the South-South zones. The random selections processes were carried out by balloting and all the candidates in the six selected states were purposely used for the study.

Statistical Sample

In view of the nature of the study which dealt with item analysis, there was also a statistical sample of 50 items.

Table 2; Sample Size Distribution of candidates from six states obtained from two Geo-political Zones

Geo-Political Zones/ States	Number of Candidates
South-East	
Anambra	1154
Enugu	1265
Imo	718
South-South	
Cross-River	896
Delta	477
Edo	438
Total	**4948**

Source: Information and Communication Technology Department, NABTEB (2018)

Research Instrument

The instrument that was used for this study was National Business and Technical Certificate Examination Mathematics Paper 1 May/June 2018. It was a dichotomously scored multiple choice examination consisting of 50 items with 4 options (A-D). One (1) mark was given for each correct answer and zero (0) for an incorrect answer based on the senior secondary school Mathematics curriculum as prescribed by the Federal Ministry of Education.

Validity of the Instrument

The instrument was validated and standardized by the Examination Development Department of National Business and Technical Examination Board (NABTEB) prior to its administration on the students.

Reliability of the Instrument

The instrument was a standardized test items developed by experts in National Business and Technical Examination Board (NABTEB) as such it was deemed reliable.

Method of Data Collection

The softcopy of the candidates' responses of the fifty Mathematics multiple choice test items in form of students' matrix score was collected by the researcher from the Information and Communication Technology (ICT) Department of NABTEB.

Method of Data Analysis

Principal component analysis (PCA) was carried out to test for Unidimensionality of the NBTCE 2018 Mathematics multiple choice items in order to determine whether the items were measuring a single hidden trait. Research questions 1-3 were answered using Maximum likelihood statistics of the computer programme BILOG-MG3, while research question 4 was answered using the five estimation techniques. BILOG-MG3 was used to estimate for Likelihood ratio test, Akaike information criterion (AIC), Bayesian information criterion (BIC), WINBUGS 1.4 was used for Deviance information criterion (DIC), while MATLAB was used for Cross-Validation Log-Likelihood (CVLL). In LRT the model that has the least deviance, is the best fitting model. Also with AIC, BIC and DIC the model that has the least value is said to be the model that fits the data (Kang & Cohen, 2007). In Cross-Validation Log-Likelihood the model with the largest CVLL fits the data best (Kang & Cohen 2007). This study has one dependent variable model-fit. In addition it has one independent variable (techniques) with five categories (Likelihood Ratio, Akaike information criterion, Bayesian information criterion, Deviance information criterion and Cross-validation log-likelihood). No hypothesis was formulated and tested since the statistics are non-significant.

CHAPTER FOUR

PRESENTATION OF RESULTS AND DISCUSSION OF FINDINGS

In this chapter results from the data collected are presented. The findings are also discussed.

Presentation of Results

Test for Unidimensionality

Principal component analysis (PCA) was used to assess Unidimensionality of the NBTCE Mathematics multiple choice items in order to determine whether the items were measuring a single hidden trait. This element would characterise the trait which indicates the Mathematics skills assessed by the examination. The Principal component analysis carried out on the 50 items of the National Business and Technical Certificate Examination (NBTCE) 2018 Mathematics multiple choice test items yielded twelve eigenvalues greater than one. The first eigenvalue was 8.328 greater than the next eleventh Eigen values (3.165, 2.434, 1.912, 1.655, 1.339, 1.246, 1.227, 1.152, 1.081, 1.041 and 1.020). The first factor explained 16.657% of the variance in the data set. The second factor explained 6.330% of the remaining variance. The third factor explained 4.868% of the remaining variance. The rest of the variance was explained by the remaining 47 factors with 2 factors each having a percentage of variance between 4 and 3, 7 factors each having a percentage of variance between 3 and 2 and 28 factors in which every one of them has a percentage of variance of between 1 and 2, 10 factors each having a percentage of variance of between 1 and 0. Hence, the first unrotated factors in the principal component analysis yielded 12 eigenvalue greater than one (> 1) accounting for 51.2% of the total variance for the NBTCE May/June 2018 Mathematics multiple choice tests. Going by these, the Reckase's (1979) least criterion of 20% required to guarantee that unidimensionality of data was met. The local independent assumption is equal to the assumption of Unidimensionality. If

Unidimensionality is met then local independence is also met since they are equivalent (Ani 2014). The Scree plot was plotted to guide in the determination of whether unidimensionality could be established. Unidimensionality was established because of the presence of a dominating factor which is a single latent trait.

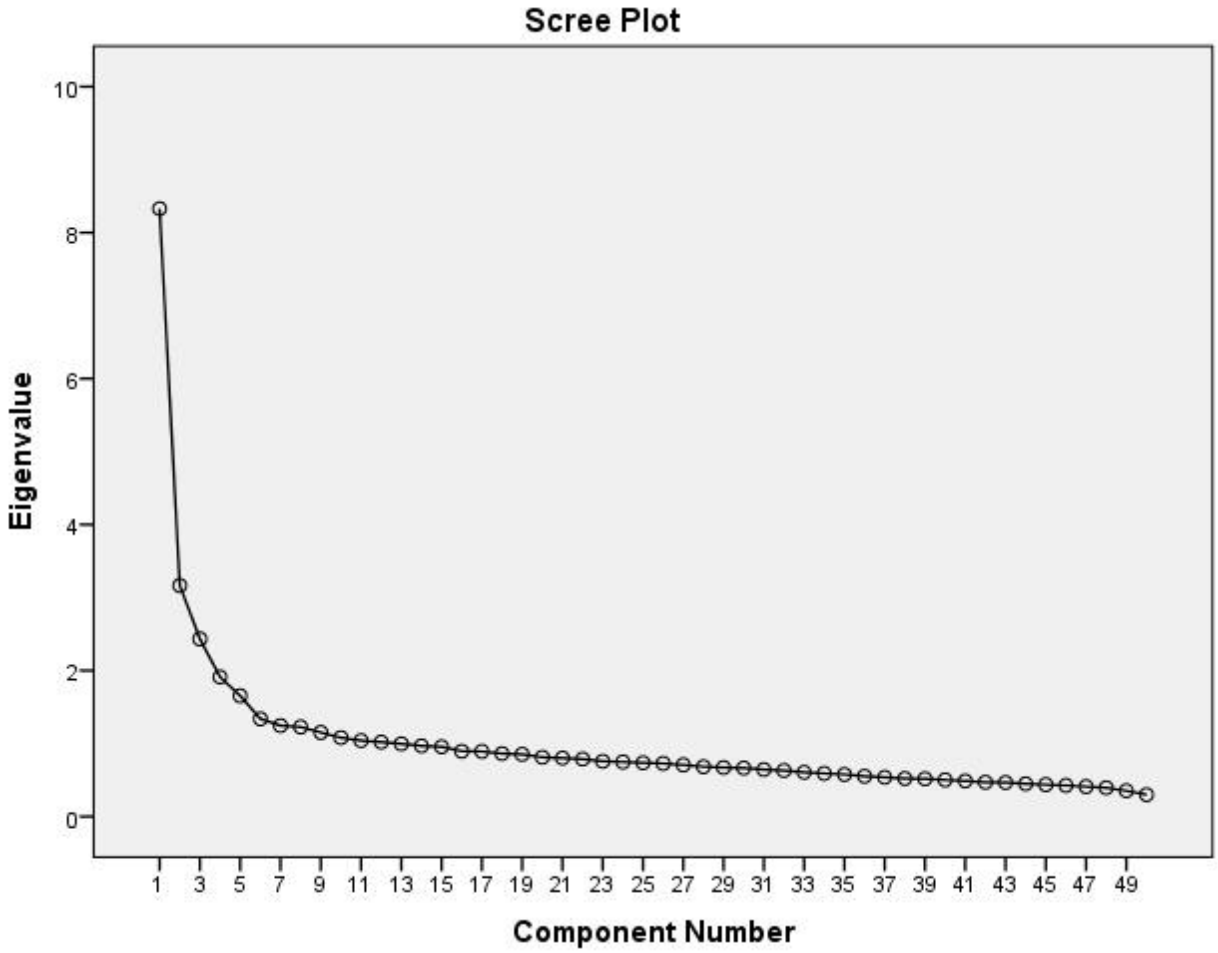


Figure 5: Scree plot of NBTCE May/June 2018 Mathematics Multiple Choice Test.

Research Question One: What are the difficulty indexes of item parameter estimates of NBTCE May/June 2018 Mathematics Multiple Choice Test Items?

Table 3: Item parameters (Difficulty) of the items of NBTCE May/June 2018 Mathematics multiple choice tests based on one, two and three parameter logistic models.

Difficulty Parameter Estimates (Threshold)				Difficulty Parameter Estimates (Threshold)			
ITEM	**1 PLM**	**2 PLM**	**3 PLM**	**ITEM**	**1 PLM**	**2 PLM**	**3 PLM**
1	-2.561	-2.198	-2.051	26	0.891	17.789	18.265
2	-3.578	-3.698	-3.305	27	-1.732	-1.560	-0.623
3	-2.315	-1.842	-1.546	28	-1.563	-1.322	-0.153
4	-3.112	-2.462	-2.251	29	-1.003	-0.674	-0.198
5	-1.243	-0.867	-0.304	30	-2.834	-2.484	-2.128
6	-1.557	-1.058	-0.445	31	0.618	INITIAL SLOPE LESS THAN -0.15	
7	-0.898	-2.138	0.797	32	-1.224	-0.780	-0.293
8	-2.217	-1.784	-1.525	33	-2.643	-2.411	-1.557
9	-2.280	-2.119	-1.803	34	-1.126	-1.725	0.322
10	-2.968	-1.924	-1.943	35	-0.950	-0.902	-0.300
11	-1.118	-0.833	-0.435	36	-2.598	-1.780	-1.529
12	-2.502	-2.371	-2.070	37	-0.238	-0.250	0.526
13	0.144	0.104	0.586	38	-1.071	-0.908	0.149
14	-2.795	-2.616	-2.227	39	-1.555	-1.035	-0.366
15	-2.638	-2.077	-1.922	40	-2.658	-1.907	-1.853
16	-1.104	-0.776	-0.265	41	-2.713	-2.280	-1.990
17	-2.11	-1.631	-1.352	42	0.881	INITIAL SLOPE LESS THAN -0.15	
18	-0.787	-0.757	0.004	43	-1.926	-1.710	-0.504
19	-2.239	-1.562	-0.998	44	-2.460	-1.747	-1.389
20	-2.237	-1.969	-1.711	45	-2.956	-2.349	-2.066
21	-2.513	-1.966	-1.818	46	-2.990	-2.189	-2.032
22	-2.575	-1.867	-1.672	47	-2.492	-2.036	-1.712
23	-1.578	-1.122	-0.818	48	-2.485	-2.146	-1.861
24	-2.032	-1.289	-1.000	49	-2.705	-1.699	-1.504
25	-1.984	-1.707	-0.798	50	-1.804	-1.975	-1.487

Note: Items 31 and 42 were not calibrated for 2plm and 3plm because their initial slopes were less than -0.15
Values < -1 are easy items. Values > 1 are difficult items.

Table 3 reveals difficulty parameter estimates of NBTCE May/June 2018 Mathematics Multiple Choice Test Items. The item difficulty, also known as the *b* parameter, is established if an examinee requires greater talent so as to respond to the item

accurately. For 1PLM the item difficulty ranges from -3.578 to 0.891, for 2PM from -3.698 to 17.789 and for 3PM -3.305 to 18.265. Mean difficulty was -1.843 (SD=1.037). The result revealed that for 1PLM 42 items were easy, none was hard and 8 items had optimum *b*-values. It also revealed that for 2PLM, 37 items were very easy, 1 item was hard and 10 items had optimum *b*-values while item 31 and 42 could not be calibrated. Finally for 3PLM 26 items were very easy, 1 was difficult, 21 had optimum *b*-values and item 31 and 42 could not be calibrated. The items that were very easy show that the students had mastery of the contents and were able to answer the items correctly. The items that were not calibrated are those that could not meet up with the initial slope values for the continuation of their calibration.

Research Question Two: What are the discrimination indexes of the item parameter estimates of NBTCE 2018 May/June Mathematics Multiple Choice Test Items?

Table 4: Item parameters (discrimination) of the test items of NBTCE 2018 May/June Mathematics multiple choice test based on two and three parameter logistic (PL) models.

Discrimination Parameter Estimates (Slope)			**Discrimination Parameter Estimates (Slope)**		
ITEM	2 PLM	3 PLM	ITEM	2 PLM	3 PLM
1.	1.102 *	1.030*	26	0.031	0.035
2.	0.854	0.807	27	1.050*	1.388*
3.	1.239*	1.260*	28	1.160*	2.827*
4.	1.225*	1.180*	29	1.900*	3.522*
5.	1.670*	2.808*	30	1.065*	1.041*
6.	1.701*	2.870*	31	INITIAL SLOPE LESS THAN -0.15	
7.	0.339	0.972	32	2.083*	3.858*
8.	1.221*	1.257*	33	1.011*	1.154*
9.	0.993	1.020*	34	0.547*	1.420*
10.	1.700*	1.541*	35	1.008*	1.256*
11.	1.483*	1.838*	36	1.565*	1.572*
12.	0.963	0.954	37	1.003*	3.647*
13.	1.029*	2.332*	38	1.186*	4.302*
14.	0.975	0.976	39	1.780*	3.064*
15.	1.247 *	1.212*	40	1.443*	1.336*
16.	1.672*	2.793*	41	1.132*	1.079*
17.	1.307*	1.360*	42	INITIAL SLOPE LESS THAN -0.15	
18.	0.998	1.546*	43	1.066*	1.629*
19.	1.540*	1.799*	44	1.479*	1.531*
20.	1.071*	1.053*	45	1.221*	1.171*
21.	1.264*	1.215*	46	1.383*	1.289 *
22.	1.423*	1.415*	47	1.185*	1.172*
23.	1.552*	1.694*	48	1.094*	1.086*
24.	1.882*	2.008*	49	1.828 *	1.764*
25.	1.114*	1.420*	50	0.808	0.809

* *a*-value greater than or equal to one. (a-value > or = 1) discriminated very well between high and low achievers

Table 4 reveals discrimination parameters based on item response theory. A good item must be able to discriminate very well. For an item to discriminate very well, its *a*-value must be greater than one or equal to one (>1 or =1). Using 2PLM, mean difficulty

was -1.263 (SD=2.893) and mean discrimination 1.241 (SD=0.393). The result showed that for 2PLM 40 items discriminates very well between high and low achievers. For 3PLM 42 items discriminated very well between high and low achievers. This explains how well an item can separate examinees that possess ability under the item location and those above the item location. This therefore is a measure of discriminating capabilities of the items.

Research Question Three: What are the guessing parameter estimates of NBTCE 2018 May/June Mathematics Multiple Choice Test Items?

Table 5: Item parameters (guessing) of the test items of NBTCE 2018 May/June Mathematics Multiple Choice Test based on three parameter logistic (3PL) model.

ITEM	3 PLM	ITEM	3 PLM
	Asymptote		**Asymptote**
1.	0.195*	26	0.028*
2.	0.352	27	0.400
3.	0.215*	28	0.500
4.	0.251	29	0.262
5.	0.298	30	0.288
6.	0.330	31	Not calibrated
7.	0.500	32	0.282
8.	0.177*	33	0.413
9.	0.190*	34	0.500
10.	0.132*	35	0.246*
11.	0.204*	36	0.232*
12.	0.217*	37	0.351
13.	0.238*	38	0.463
14.	0.271	39	0.375
15.	0.175*	40	0.149*
16.	0.267	41	0.272
17.	0.187*	42	Not calibrated
18.	0.300	43	0.494
19.	0.353	44	0.275
20.	0.197*	45	0.288
21.	0.173*	46	0.244*
22.	0.188*	47	0.255
23.	0.183*	48	0.221*
24.	0.215*	49	0.234*
25.	0.417	50	0.240*

*Test Items not susceptible to guessing

Table 5 shows the guessing (asymptote) estimates of the NBTCE Mathematics multiple choice test items on the basis of three parameter logistic (3pl) model. The guessing parameter, ***c,*** indicates the probability that a student who has low ability will be able to get the right answer to an item by guessing. In this study, the ***c-***value was between 0.00 to 0.25. Using the 3PLM the mean difficulty was –0.732 (SD=2.945), mean discrimination 1.673 (SD=0.893) and mean guessing was 0.276 (SD=0,105). The result revealed that 22 items had *c*-values between 0-0.25 while 26 items had *c*-values greater than 0.25. This implies that these 26 items were prone to guessing. This meant that the students that did not know the correct answer had to guess.

Research Question Four: How do the model selection methods for data from NBTCE May/June 2018 Mathematics Multiple Choice Test Items compare based on the one-parameter, two-parameter and three-parameter logistic models using the five (LR, AIC, BIC, DIC and CVLL) techniques?

Table 6: Comparisons of Model Selection Methods for data from NBTCE 2018 May/June Mathematics Multiple Choice Test Items based on 1PLM, 2PLM and 3PLM using LR, AIC, BIC, DIC and CVLL.

N	Model	LR Test		AIC	BIC	DIC	CVLL
		G2	LR				
4948	1PM	210889.4		210989.4	211314.7	184387.000	-126362.0
	2PM	190893.7	19995.7*	190989.0	191710.3	182865.000	-121265.6
	3PM	190163.0	730.7*	190355.0	191388.0	183206.000	-120356.6

Note. NBTCE = National Business and Technical Certificate Education; N= sample size; 1PM = one-parameter model; 2PM = two-parameter model; 3PM = three-parameter model; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion; DIC = Deviance information criterion; CVLL = cross-validation log-likelihood;
*p ($\chi^2_{df=50}$>67.51) <:05.

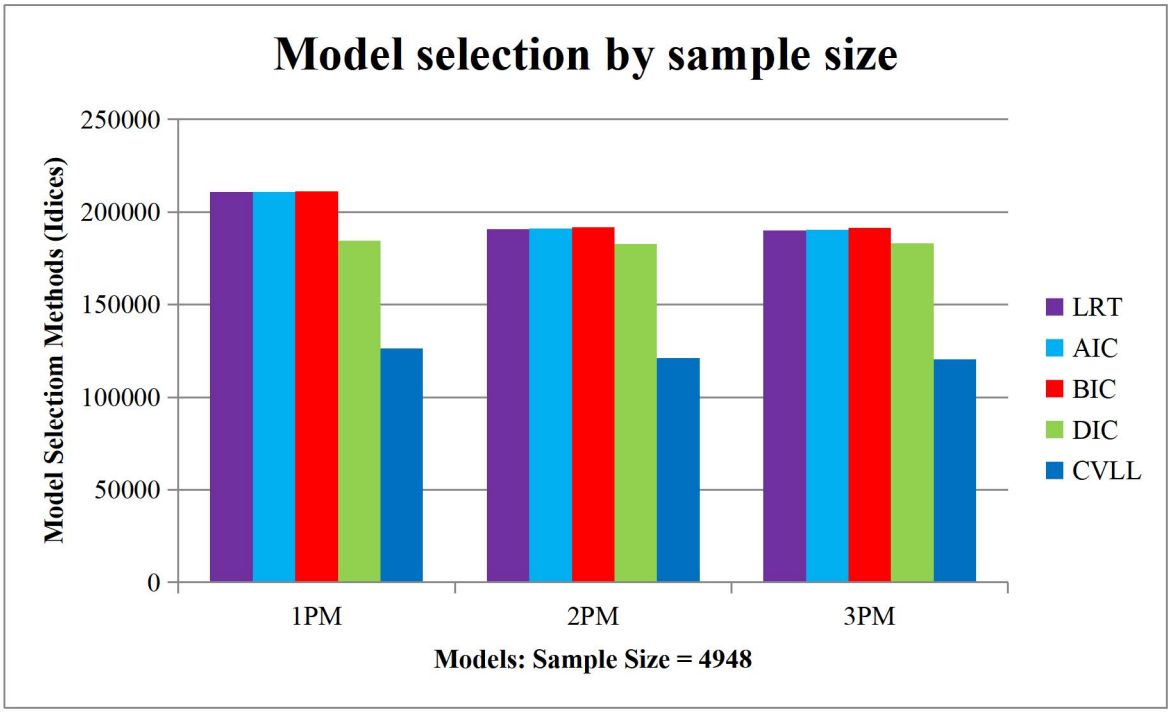
Table 6 revealed that four of the methods LR, AIC, BIC, and CVLL, provided steady results detected the 3PM as the best model that fitted NBTCE 2018 Mathematics Multiple Choice Test Items while DIC selected 2PM.

Comparing each model among the techniques, for 1PLM model CVLL had the least value followed by DIC, LR, AIC and BIC. For 2PLM CVLL had the least value, followed by

DIC, LR, AIC and BIC. For 3PLM CVLL also had the least value followed by DIC, LR, AIC and BIC. Since four out of five methods selected 3PLM, then 3PLM is the best model that fitted NBTCE May/June 2018 Mathematics multiple choice test items. Comparing the techniques CVLL which use Bayesian algorithm seemed to be best technique.

The graphical comparison of the five model selection methods is presented in figure 6 below.

Figure 6: Graphical Model Selection by Sample Size



Note. 1PM=one-parameter model; 2PM=two-parameter model; 3PM=three-parameter model; LR=likelihood ratio; AIC=Akaike information criterion; BIC=Bayesian information criterion,; DIC=deviance information criterion; CVLL=Cross-validation log-likelihood.

Discussion of Findings

Fit of model to the data is a serious issue in psychometric models specifically when the models are dichotomous. Comparison of relative fit of contending models has been a common exercise in the progress of fitting the model to the data. In this study real data were made use of to investigate how the five relative fit indices performed in selecting the item response theory dichotomous model that fitted the NBTCE 2018 Mathematics

multiple choice test items. LR, AIC and BIC methods use the Frequentist algorithm, while DIC and CVLL make use of the Bayesian algorithm. Selecting a suitable item response theory dichotomous model is very important to obtain the benefits of the item response theory application like test equating, item banking, test development, detection of differential item functioning and computer adaptive testing among others.

The finding revealed that items with high *b* values (*b* greater than 1) are hard items, and items that have low *b* values (*b* values below -1) show items that are easy. The optimum or best *b*-value for an item is between -1 and 1. In using 1PLM, 84% (42) of the items were easy, none of the items was difficult (1,2,3,4,5,6,8,9,19,11,12,14,15,16,17,19,20,21,22,23,24,25,27,28,29,30,32,33,34,36,38,39, 40,41,43, 44,45,46,47,48,49,50) and 16% (8) items had optimum b-values and were considered good items (7,13,18,26,31,35,37,42). Using 2PLM, 74% (37) of the items were easy (1,2,3,4,6,7,8,9,10,12,14,15,17,19,20,21,22,23,24,25,27,28,30,33,34,36,39,40,41,43,44,45, 46,47,48, 49 and 50), 2% (1) of the items was difficult (item 26), 20% (10) had optimum b-values (5,11,13,16,18,29,32,35,37,38) and 2(4%) were not calibrated because their initial slopes were less than -0.15 (31 and 42). Finally, when 3PLM was used to calibrate the test items, 52% (26) of the items were easy (1,2,3,4,8,9,10,12,14,15,17,20,21,22,30,33,36,40,41,44,45,46,47,48,49,50), 2% (1) of the items was very hard (item 26), 42% (21) of the items were with optimum b-values (5,6,7,11,13,16,18,19,23,24,25,27,28,29,32,34,35,37,38,39,43) and 4% (2) of the items were not calibrated because their initial slopes were less than -0.15 (31 and 42).

The finding also revealed that, making use of latent trait theory 2PL model to examine the quality of NBTCE 2018 Mathematics Multiple Choice Test, 40 test items representing 80% of the total items discriminated very well

(1,3,4,5,6,8,10,11,13,15,16,17,19,20,21,22,23,24,25,27,28,29,30,32,33,34,35,36,37,38,39, 40,41,43,44,45,46,47,48,49). In addition making use of the latent trait theory 3PL model to examine the quality of test items, 42 test items representing 84% of the total items discriminated very well (1,3,4,5,6,8,9,10,11,13,15,16,17,18,19,20,21,22,23,24,25,27,28,29,30,32,33,34,35,36,37,38,39,40,41,43,44,45,46,47,48,49).

It also revealed that test items with *c* values greater than 0.25 were reasoned to be very prone to guessing and were classified as poor test items. 22 (44%) items, that is, items 1, 2, 8, 9, 10, 11,12, 13, 15, 17, 20, 21, 22, 24, 26, 35, 36, 40, 46, 48, 49 and 50, were with the c-value range of 0.00 to 0.25 and showed that the likelihood of successfully answering rightly by guessing was low. While 26 items (52%) had c-values greater than 0.25, this implies that the items were susceptible to guessing, that is, the likelihood of getting an accurate response by guessing was very high. The result generated item characteristic curve to determine whether the items in the test are good enough for the assessment of the students. The curves are all in the form of cumulative logistic curve thereby satisfying the item characteristics curve assumption of IRT. The item parameter *b* (difficulty) ranged from -3.305 to 18.265, parameter *a* (discrimination) ranged from 0.035 to 3.858. Parameter *c*, probability of guessing correctly in the test ranged from 0.028 to 0.500.

The result after the estimation using the five techniques revealed that the LR test showed that the 3PM described the data better than the 2PM and that the 2PM described the data better than the 1PM.The smallest AIC was for the 3PM, the BIC of 191388.0 for the 3PM was smaller than the BICs for the 2PM and 1PM models and CVLL for the 3PM was the largest among the values of CVLL. DIC however, specified the 2PM as the better fitting model. However, because four of the five indices suggested the 3PM, hence 3PM

was considered as the best item response theory dichotomous model that fitted NBTCE 2018 Mathematics multiple choice test items with sample size 4948.

This finding of this study is in agreement with the study by kang and Cohen (2007) who reported that DIC, CVLL, LR, and AIC selected the 3PM as the best model. The Cross-validation log-likelihood (CVLL) appeared to be the best of the five techniques for the stipulations that they fixed. The only difference is that while BIC selected 3PM in this study it selected 2PM in the study by Kang and Cohen and DIC selected 3PM in their work it selected 2PM in this study.

The findings of the study is also in agreement with Kang et al. (2009) who compared the performances of AIC, BIC, DIC, and CVLL in selecting the correct model among the graded response model (GRM) the partial credit model (PCM), the generalized partial credit model (GPCM) and the rating scale model (RSM). They established that in view of the fact that the condition for simulation affected the manner in which each method performed well; CVLL on the basis of average had the best performance.

In addition it agrees with the finding of LOU and Al-Harbi (2017) who compared the performance of Likelihood ratio test (LRT), Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC), Widely available information criterion (WAIC) and Leave-one-out cross-validation (LOO) with a population of 36886 student using 52 items and different sample sizes of 500, 1000 and 2000 . They found out that the better performance was done by WAIC and LOO which use Bayesian algorithm than LRT, AIC, BIC and DIC when data were generated using 3PLM.

The finding is in disagreement with Farmsworth et al, (2016) study that compared the 1PL, 2PLand 3PL models to determine the best suitable model for the analysis of Standard Assessment of Concussion (SAC) using χ^2 and BIC. They found out that χ^2 and BIC selected 1PL model as the best fitting model.

The findings of the study also disagrees with Li et al (2009) who examined the techniques for model selection used to assess dichotomous mixture item response theory (IRT) models. They considered five indices; Pseudo-Bayes factor (PsBF), and Posterior predictive model checks (PPMC), Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC). The five techniques offer rather endorsements that were not the same for a set of data that were real. They concluded that model selection indices were not in agreement every time.

In this study none of these methods at any point selected 1PM. LRT, AIC, BIC and CVLL selected 3PM as the best fitting model and CVLL appeared to be the best in terms of performance out of the five estimation techniques. This was predictable since every item was a multiple choice item with four options, so it was expected that examinees would use a guessing approach when they did not know the correct answer. Therefore 3PM is the model that fitted NBTCE May/June 2018 Mathematics multiple choice test items. Examining the model selection graphically it can be observed that Cross-validation log-likelihood which uses Bayesian algorithm is the most efficient technique. This means that the Bayesian estimation method performed better than the Maximum likelihood as reported by Ojerinde et al (2012) though using short length and small sample. This study had established that Bayesian method also performed better with longer test length and large sample size.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

In this chapter, the summary, conclusion and recommendations of the study based on the findings are presented.

Summary

The study investigated the comparison of model-fitting techniques in choosing item response theory dichotomous model that fits NBTCE May/June 2018 Mathematics multiple choice test using five different estimation techniques. The study specifically examined how effective Likelihood Ratio test, Akaike information criterion, Bayesian information criterion, Deviance information criterion and Cross-validation log-likelihood in selecting a dichotomous model that fitted NBTCE May/June 2018 Mathematics multiple choice test items. This was carried out by comparing the performances of the five techniques used in the estimations.

Four research questions guided this study. No hypothesis was formulated and tested due to the fact that the techniques used in this study are non-significant statistics. The descriptive Survey research design was employed in the study using Ex-Post Facto approach. The population of this study comprised forty-nine thousand, five hundred and eighty one (49,581) candidates who enrolled and sat for the National Business and Technical Certificate Examination (NBTCE) May/June 2018 Mathematics multiple choice examination in the six Geo-political zones in Nigeria. The sample size of this study was four thousand, nine hundred and forty eight (4,948) candidates who sat for National Business and Technical Examination Board (NBTCE) May/June 2018 Mathematics multiple choice examination. The Multistage simple random sampling technique which involved sampling stages was employed for random and effective selection of the sample. At the first stage, two zones (South- East and South-South) were randomly selected from

the six Geo-political zones which are North-Central, North-East, North-West, South-East, South-South and South-West zones in Nigeria. At the second stage three states each (Anambra, Enugu and Imo) were randomly selected from the South-East zone while Cross-River, Delta and Edo states were randomly selected from the South-South zones. All the random selection processes were carried out through balloting and all the candidates in the six randomly selected states were purposely used for the study. In view of the nature of the study which dealt with item analysis, the number of items in the study was 50. So, the study also used a Statistical sample of 50 items.

Principal component analysis was used to test for Unidimensionality using SPSS version 20. Item parameters were estimated from the candidates' responses to the items using the BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2003). For the five estimation techniques BILOG-MG3 was used for Likelihood ratio test (LR), Akaike information criterion (AIC) and Bayesian information criterion (BIC). WINBUGS 1.4 was used for Deviance information criterion (DIC), while MATLAB 8.5 was used for Cross-Validation Log-Likelihood (CVLL).

The result of the analysis revealed that:

- Principal component analysis yielded 12 eigenvalue greater than one (>1) accounting for 51.2% of the total variance for the NBTCE 2018 May/June Mathematics multiple choice tests, thereby meeting Reckase's (1979) minimum criterion of 20% required to guarantee that unidimensionality of data was met.
- Using one parameter logistic model (1PLM) forty-two (42) items of the fifty (50) items in the (NBTCE) May/June 2018 Mathematics multiple choice test were easy, eight (8) items had the optimum *b*-value. In using the two parameter logistic model (2PLM), 37 items of the 50 items were easy, 1 item was difficult, 10 items had the best *b*-value and 2 items were not calibrated because their initial slopes were less

than -0.15 (item 31 and 42). When three parameter logistic model (3PLM) was used, 26 items of the 50 items were easy, 1 item was difficult and 21 items had optimum *b*-values. Two (2) items were not calibrated because their initial slopes were less than -0.15 (item 31 and 42).

- Forty (40) items of the fifty (50) items in the (NBTCE) May/June 2018 Mathematics multiple choice test discriminated very well when using 2PL model. While in using 3PL model 42 items discriminated very well.
- Regarding pseudo-guessing 22 items had *c*-value range of 0.00-0.25 which showed that the probability of getting correct answers to these items by guessing was very low, while 26 items were prone to guessing, that is the probability of getting these items correct by guessing was very high.
- The result generated item characteristic curve to determine whether the items in the test were good enough for the assessment of the students. The curves were all in form of cumulative logistic curve thereby satisfying the Item characteristics curve assumption of IRT. The item parameter *b* (difficulty) ranged from -3.305 to 18.265, parameter *a* (discrimination) ranged from 0.035 to 3.858. Parameter *c*, probability of guessing correctly in the test ranged from 0.028 to 0.500.
- The likelihood ratio test (LRT), Akaike information criterion (AIC), Bayesian information criterion (BIC), and Cross Validation Log-Likelihood (CVLL) selected three parameter logistic (3PL) model as the best fitting latent trait theory dichotomous model that explains the data from NBTCE May/June 2018 Mathematics multiple choice tests very well.
- Out of the five estimation techniques the Cross-validation log-likelihood appeared to be the most efficient technique. The graphical representation of model selection also points to Cross-validation log-likelihood as the best technique used in this study.

Conclusion

The study concluded that the 3-parameter logistic model (3PLM) fitted NBTCE May/June 2018 Mathematics multiple choice test items. The Cross-Validation Log Likelihood which uses Bayesian algorithm is the best of the five techniques. This means that the Bayesian algorithm performed better than the Maximum likelihood algorithm.

Recommendations

On the basis of the findings the following recommendations are made

- In test construction, examination bodies such as National Business and Technical Examination Board (NABTEB), National Examination Council (NECO), West African Examination Council (WAEC), Joint Admission and Matriculation Board (JAMB), National Teachers Institute (NTI) and other institutions should employ relative fit in model-data fitting.
- In using relative fit for model-fitting to the data Psychometricians should make use of at least five estimation techniques.

Contribution to Knowledge

The study has been able to establish Cross validation log-likelihood as the most efficient technique in model-fitting to real data using relative fit compared to the Likelihood ratio test, Akaike information criterion, Bayesian information criterion and Deviance information criterion. It also established Bayesian algorithm is better than Maximum Likelihood algorithm

Suggestions for Further Studies

- Further studies should be carried out using relative fit to model-fit item response theory Polytomous models such as Graded Response Model (GRM), partial credit model (PCM), nominal model (NM) and rating scale model (RSM) during test construction and standardisation.
- Study should also be carried out on multidimensional items and mixture IRT models.
- Mixed format items (multiple choice and essay items) should also be studied using these methods
- Model-fit of the items from two or more examination bodies could be compared using these estimation techniques.

REFERENCES

Adebule S. O. (2013) Item response theory as a basis for measuring latent trait of interest. *Greener Journal of Social Sciences*. 3 (7), 378-382

Adedoyin, O.O. & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, 3(4), 992-1011.

Adonu, I.I. (2014) Psychometric analysis of WAEC and NECO practical physics tests using partial credit model. *PhD thesis University of Nigeria Nsukka.* https://www.unn.edu.ng/publications/files/Adonu%20I.I.pdf

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans-actions on Automatic Control*, 19, (6), 716-723. Doi:10.1109/TAC1974.1100705

Akindele, B. P. (2003). *The development of an item bank for selection tests into Nigerian universities: an exploratory study.* Unpublished doctoral dissertation, University of Ibadan, Nigeria.

Ali, A. (2006). *Conducting research in education and the social sciences*. Enugu: Tashiwa Networks Ltd.

Ani, E.N., (2014). Application of item response theory in the development and validation of multiple choice test in Economics. (*Master's Thesis*). *University of Nigeria, Nsukka.*

Baghael, P. (2008). The Rasch Model as a construct validation tool: Transaction of the Rasch measurement. *American Educational Research Association*. 22(1): 1145—1146.

Baker, F. B (2001). *The basics of item response theory*. Eric clearing house on assessments and evaluation. University of Maryland College. Park M.D.

Baker, F.B. (1977). Advances in item analysis. *Review of Educational Research* 47: 151-178

Baker. F. B. & Kim, S.H. (2004). *Item response theory: parameter estimation techniques* (2nd Ed.). New York: Marcel Dekker.

Bichi .A.A & Talib. R (2018), Item response theory: an introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education (IJERE)* 7 (2) 142-151

Bichi .A.A., Embong R., Mamat M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: a review of empirical studies. *Australian Journal of Basic and Applied Sciences, 9(7), 549-556.*

Bielinski, J. & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple choice Mathematics items administered to national probability samples. *Journal of Educational Measurement*, *38*, 51-77.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and non-compensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 27, 395-414.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics,* 26(4), 381-409.

Brown C, Templin J. & Cohen A. (2014). Comparing the two- and three-parameter logistic models via likelihood ratio tests. *Journal of Applied Psychological Measurement.* 39(5): 335–348.

Bulut O. (2015) Applying item response theory models to entrance examination for graduate studies: practical issues and insights. *Journal of Measurement and Evaluation in Education and Psychology* 6(2); 313-330

Cappelleri, J.C., Lundy, J.J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clinical Therapeutics ;36(5):648-662.*

Casella, G & Berger, R. L. (2001). *Statistical Inference (Second ed.). ISBN 0-534-24312-6*

Chalmers, R. P (2012). Mirt: A multidimensional item response theory package for R environment. *Journal of Statistical Software*, 48 (6), 1-29

Ceniza, J.C. & Cereno, D. C (2012). Development of Mathematics diagnostic test for dorshs**.** https://www.doscst.edu.ph/index.ph[/academics/graduateschool/publication/category/5-volum-1-issue-1-2012?

Chernyshenko O, Stark.S, Chan K, Drasgow F, & Williams B.(2001) Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research* 36(4):523-562

De Ayala R .J. (2009) *Theory and practice of item response theory*: Guilford Publications.

De Mars C. (2010), *Item response theory. Understanding statistics measurement*, Oxford University Press.

Dodeen, H., 2004. The relationship between item parameters and item fit. *Journal of Educational Measurement,* 41(3), 261-270.

Edwards .M.C (2009). An introduction to item response theory using the need for cognition scale *social and personality psychology Compass 3(*4), *507–529.*

Eleje L.I & Esomonu N. P. M (2018). Test of achievement in quantitative economics for secondary schools: construction and validation using item response theory. *Asian Journal of Education and Training,* 4(1): 18-28.

Eleje, L. I.; Onah, F. E. & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational and Social Sciences, 3 (1),* 71 - 89.

Essen C. B, Idaka E. I & Metibemu M.A (2017) Item level diagnostics and model - data fit in item response theory (irt) using BILOG - MG V3.0 and IRTPRO V3.0 Programmes. *Global Journal of Educational Research* 16, 87-94.

Essen, C. B., (2015). Differential item functioning of 2014 unified tertiary matriculation examination Mathematics of candidates in Akwa Ibom State, Nigeria. Unpublished Doctoral Dissertation. University of Calabar, Nigeria.

Farnsworth J.L; Kang, M & Ragan, B. G (2016) Comparison of item response theory models for analyzing standard assessment of concussion data. Conference Paper. https://researchgate.net/publication/309351237

Federal Republic of Nigeria (2014). *National policy on Education*. Nigeria, National Educational Research Council NERC Press.

Fischer, G. (1968). Psychologische test theorie [Psychological test theory]. Bern: Huber.

Flannery, W.P., Reise, S.P. & Widaman, K.F. (1995). An item response theory analysis of the general academic scales of self-description questionnaire II. *Journal of Research in Personality*, 29: 168-188.

Galli. S, Chiesi F. & Primi C. (2010) Assessing mathematics competence in introductory statistics courses: an application of the item response theory. *International Association of Statistical Education (IASE)*. https://iase-web.org/documents/papers/icots8/ICOTS8_C114_GALLI.pdf

Glas, C. A. W, & Falcon, J. C. (2003). A comparison of item fit statistics for the three-parameter logistic model *Journal of Applied Psychological Measurement* 27(2), 87-106,

Glas, C.A.W (1998). Detection of differential item functioning using Lagrange multiplier tests. *Stat Sinica* 8:647–667

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm (ETS Research Report RR-13-32)*. Princeton: ETS.

Hambleton R. K. & Slater S. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment*. 13 (1) 21–28

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications,* 57–78. Washington, DC: Degnon Associates.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory. Principles and application.* Boston, MA: Kluwer Academic publishers

Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Heidenheimer, A.J., Heclo, H. & Adams.C.T (1983). *Comparative Public Policy: The Politics of Social Choice in Europe and America.* 2ded. New York: St. Martin's Press.

Hishamuddin A. & Siti E. M. (2016), Is 3PL Item Response Theory an Appropriate Model for Dichotomous Item Analysis of The Anatomy & Physiology Final Examination? *Journal Pendidikan Sains & Matematik Malaysia 6 (1). 13-23*

Ifeakor, A. C. (2011). *Psychological measurement & evaluation in Education: Issues and application*. Onitsha: Folmech Printing and Publishing Co.Ltd

Izard, J.F. & White, V.D. (1980). *The use of Latent Trait Model in the Development and Analysis of Classroom tests.* In D. Spearitt (Ed.). The Improvement of Measurements in Education and Psychology: Contribution of Latent Theories. Australian Council for Education Research ACER.

Johnson S.M (2007) Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software* 20 (10), 1-24

Kang T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*(7), 499-518.

Kang, T. & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31,* 331-358.

Kim, J.S. & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26,* 38-51.

Kose, I. A., (2014). Assessing model data fit of unidimensional item response theory in simulated data. Educational Research and Reviews, 9 (17): 642-649.

Kyung, T. H. (2013).*Windows Software that Generates IRT parameters and Item Responses: Research and Evaluation Program Methods (REMP)*. University of Massachusetts Amherst.

Lee, P. M. (2012). *Bayesian statistics: An introduction* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Li. F, Cohen A.S, Kim S-H & Cho S-J (2009). Model Selection Methods for Mixture Dichotomous IRT Models. *Applied Psychological Measurement Volume 33: 5*

Liddle, A.R., (2007) "Information criteria for astrophysical model selection". *https:/ / arxiv. org/ PS_cache/ astro-ph/ pdf/0701/ 0701113v2.*

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Richmond: Psychometric Corporation.

Lord, F.M (1968). An analysis of the verbal scholastic aptitude test using three-parameter logistic model, *Educational and Psychological Measurement 28, 989-1020.*

Lord, F.M. (1980) *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Luo Y. & Al-Harbi K. (2017) Performances of LOO and WAIC as IRT Model Selection Methods. *Psychological Test and Assessment Modeling,* 59, (2), 183-205

Magis, D. (2007). *Influence, information and item response theory in discrete data analysis. http://bictel.ulg.ac.be/ETDdb/collection/available/ULgetd-06122007-100147/.*

Mallikarjuna, G. & Natarajan, V. (2014). Investigation into Invariance Properties of Item Response Theory (Irt) By Two and Three Parameter Models. *International Journal of Information Technology and Business Management.* 1 (1)

Maydeu-Olivares. A & Montaño. R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis *Psychometrika* 78 (1) 116–133.

McKinley, R. & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-5.Merit Trac Services (P) Ltd

Mislevy, R. J (2011) Evidence centered designed for simulated-based assessment (CRESST Report 800). Los Angeles CA**: University of Califonia, National Center for Research on Evaluation, Standard and Student Testing (CRESST)**

Muraki, E. & Bock, R.D. (2003).*PARSCAL4: Item analysis and test scoring for rating scale data.* (Computer program).Chicago, IL; Scientific Software. *https://www.umass.edu.remp>simcata*

Musa. M & Dauda. E. S (2014). Trends analyses of students' mathematics performance in West African senior secondary certificate examination from 2004 to 2013: Implication for Nigeria's Vision 20:2020 *British Journal of Education* 2(7), 50-64.

Natarajan (2009). *Basic Principles of IRT and Application to Practical Testing and Assessment*.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge.

Nworgu B.G. and Agah J.J. (2012). Application of three parameter logistic model in the Calibration of a mathematics Achievement Test. *Journal of Educational Assessment in Africa 29 (7) 162 – 172.*

Obinne, A.D.E. (2013). Test item validity: item response theory (irt) perspective for Nigeria. *Research Journal in Organizational Psychology & Educational Studies* 2(1) 1-6

Ojerinde, D. & Ifewulu, B. C. (2012). Item unidimensionality using 2010 unified tertiary matriculation examination mathematics pre-test. A paper presented at the 2012 international conference of IAEA, Kazastan.

Ojerinde, D. (2013). *Classical test theory (CTT) vs. item response theory (irt): an evaluation of comparability of item analysis results*. Lecture Presentation at the Institute of Education, University of Ibadan.

Ojerinde, D., Ojo. F.R & Popoola, O.O (2012). *Introduction to item response theory parameter, models, estimation & application.* Goshen Print Media Limited, Lagos Nigeria.

Omorogiuwa. O. K (2010*). Introduction to Educational Measurement and Evaluation.* Perfect Touch Prints: Benin City, Edo state.

Onunkwo, G.I.N. (2002). *Fundamentals of educational measurement and evaluation*. Cape Publishers International: Owerri, Imo State.

Orlando, M. (1997). Item fit in the context of item response theory (Doctoral dissertation, University of North Carolina. *Dissertation Abstracts International, 58/04-B*, 2175.

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement,* 27(4), 289-298.

Osterlind, S. J. (2012). Item response theory. *Journals of home school and academic learning*. http://www.education.com>Home>School and Academics> classroom learning.

Palmieri, P.A. (2012). *Item response theory method and application gaining support as assessment instrument.* https://www.istss.org/publications.

Popham, W. J. (2008). Timed tests for tykes? *Educational Leadership, 65*(8), 86-87.

Rasch, G. (1960*). Probabilistic models for some intelligent and attainment tests*: Chicago: MESA press ltd.

Reckase, M. D.(1979) Unifactor latent trait model applied to multifactor test. Result and implications. *Journal of Educational and Behavioral Statistics* 4 (3) 207-230

Reckase, M.D. (2009). *Multidimensional item response theory.* New York: Springer-Verlag.

Reeve, B. B. (2002). *An Introduction to Modern Measurement Theory*. Bethesda, Maryland: National cancer institution. https//appliedresearch.a.a.ncer.gov/areas/cognitive/immt. Pd.

Rijn, R.W, Sinharay, S., Haberman, S.J., & Johnson, M.S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education,* 4(1): 1-23.

Rind I.A & Mari M.A (2019) Analysing the impact of external examination on teaching and learning of English at the secondary level education. Cogent Education 6(1):1574947. https://doi.org/10.1080/2331186X.2019.1574947

Roediger H.L, Putnam A. L & Smith M. A (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, 55, 1-36

Rubin D.B. (1984) Bayesian justifiable and relevant frequency calculations for applied statistics. *Annals of statistics* 12, 1151-1172.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monographs,* 34(4pt 2), 17.

Schumacker, R. E. (2010) Item Response Theory. http://appliedmeasurementassociates.com/ama/assets/File/ITEM_RESPONSE_THEORY.pd

Schwarz, G.E (1978). Estimating the dimension of a model. *The Annals of Statistics,* 6 (2), 461-464. doi:101214/aos/1176344136

Sheng, Y. (2005). Bayesian analysis of hierarchical IRT models: comparing and combining the unidimensional and multidimensional IRT models. Unpublished *Doctoral Dissertation*. University of Missouri-Columbia.

Shephard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93-128.

Sijtsma, K & Hemker, B. T., (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal Educational Behavioral Statistics*, 25 (4),391-415.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement,* 42, 375–394.

Sinharay, S. (2016). Bayesian model fit and model comparison. In van der Linden W. (Ed.), *Handbook of item response theory: Statistical tools 2,* 379-394. Boca Raton, FL: Chapman & Hall/CRC Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series* B—Statistical Methodology, 64, 583-616.

Thorpe, G.L. & Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship*: 1-34.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology* 16, 433–451.

Topczewski, A. M. (2013) "Effect of violating unidimensional item response theory vertical scaling assumptions on developmental score scales." PhD (Doctor of Philosophy) thesis, University of Iowa, 2013. https://doi.org/10.17077/etd.

Ughamadu K.A, Onwuegbu O.C & Osunde A.U (1991). *Measurement and Evaluation in Education.* World of Books Publishers: Benin City, Edo state.

Wang Y. & Liu Q. (2006). Comparison of Akaike information criteria (AIC) and Bayesian information criteria (BIC) in selection of stock recruitment relationships. *Journal of Fiseries Research*. 77: 220-225. Doi:10.1016/j.fishres2005.08.011

Wells, C. S., Wollack, J. A & Serlin, R. C. (2005). An equivalency test for model fit. *Paper presented at the annual meeting of the National Council on Measurement in Education,* Montreal, Canada.

Wendy, K. A. & Carl, E. W. (2010). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*. *http://www.informaworld.com/smpp/title~content=t713737283*

Wiberg, M. (2004). *Classical Test Theory vs. Item Response Theory: An evaluation of the theory test in the Swedish driving-license test.* Working paper: EM No 50, UMEA University.

Yalçın .S. (2018) Data fit comparison of mixture item response theory models and traditional models. *International Journal of Assessment Tools in Education* 5(2), 301–313.

Yen .W.M (1981). Using simulated result to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262

Yen, W. M. & Fitzpatrick, A. R. (2006). Item Response Theory (pp.111-153). In R.L.Brennan (ED.), *Educational measurement* (4th Edition). Westport, CT: American Council of Education/Praeger.

Zhao, Y. (2008). Approaches for addressing the fit of item response theory model to educational test data. *Unpublished Doctoral Dissertation*. University of Massachusetts Amberst.

Zimowski, M.F, Muraki, E. Mislevy, R.J. & Bock, R.D. (2003). BILOG MG 3 (computer program).Chicago, IL; Scientific Software international.

APPENDICES

APPENDIX A

MATHEMATICS
PAPER CODE: **002-1**
Monday
21st May, 2018
(9.00 a.m. - 10.30 a.m.)

NATIONAL BUSINESS AND TECHNICAL EXAMINATIONS BOARD
NATIONAL BUSINESS AND TECHNICAL CERTIFICATE EXAMINATIONS

MATHEMATICS
PAPER I (OBJECTIVE)
(50 marks)

PAPER CODE:
002-1

Time: 1 Hour 30 Minutes.

GENERAL INSTRUCTIONS:

DO NOT OPEN THIS QUESTION PAPER UNTIL YOU ARE TOLD TO DO SO.
While you are waiting, read the following instructions carefully.

This paper consists of only **one section. All** candidates are to answer **all** questions in **1 hour 30 minutes**. Mathematical Tables and Calculator may be used in any question. Do all rough work on this question paper.

1. Before answering questions in this paper, read the instructions on the Objective Answer Sheet carefully. Ensure that your **Centre Code** and **Candidate Number** are **boldly** and correctly written and **shaded** on your Objective Answer Sheet.

2. Each question is followed by four options lettered A to D. Find out the correct option for each question and shade in **pencil** on your Objective Answer Sheet, the answer which bears the same letter as your chosen option. Give only **one** answer to each question.

3. If you change your mind on your option, **erase** the shading completely and shade another option.

4. The **code** for this paper is **0 0 2 - 1.**

   Write it in the space provided on your Objective Answer Sheet and shade accordingly.

1. Express 3.195 correct to 2 decimal places.
   A. 3.10
   B. 3.190
   C. 3.2
   D. 3.20

2. Given $37^2 - 33^2 = 8x$, find the value of x.
   A. 33
   B. 35
   C. 37
   D. 70

3. Find the seventh term of the arithmetic progression $\frac{1}{6}, \frac{1}{2}, \frac{5}{6}, \frac{7}{6}, \ldots$
   A. $\frac{5}{6}$
   B. $\frac{9}{6}$
   C. $\frac{5}{6}$
   D. $\frac{13}{6}$

4. What is the difference in longitude between A (40°S, 60°E) and B (40°S, 160°E)?
   A. 220°
   B. 200°
   C. 100°
   D. 80°

The direction of X from Y is 30°E. Use this information to answer questions 5- 6

5. What is the bearing of X from Y?
   A. 330°
   B. 210°
   C. 060°
   D. 030°

6. What is the bearing of Y from X?
   A. 210°
   B. 180°
   C. 60°
   D. 30°

7. If $\cos\theta = \frac{4}{5}$, find $\sin\theta$ for $0 \leq \theta \leq 90$

A. $\frac{4}{5}$

B. $\frac{3}{5}$

C. $\frac{5}{8}$

D. $\frac{1}{2}$

8. If the angles of a quadrilateral are 5x, 4x, 3x and 6x, what is the value of x?

A. 18°
B. 20°
C. 90°
D. 180°

9. The reciprocal of 0.05 is

A. 0.1
B. 5
C. 20
D. 25

10. Find the value of x that satisfies the inequalities $2x + 4 \leq 16 - 2x$

A. $x \leq 4$
B. $x < 3$
C. $x \leq -3$
D. $x \leq 3$

11. The expression 'a is at least 5' means

A. $a = 5$
B. $a < 5$
C. $a \leq 5$
D. $a \geq 5$

12. Find the value of x if $\sqrt{2^x} = 8$

A. 2
B. 3
C. 6
D. 9

13. A man with 120 cattle bought food for them for 35 days. But immediately sold 15 of them. How many days will same amount of food last the remaining cattle at the same rate of food consumption? (Approximate your answer to the nearest days)

A. 10 days
B. 20 days
C. 30 days
D. 40 days

14. Find the cost of 520 shares at ₦1.25 each

A. ₦520.00
B. ₦650.00
C. ₦675.00
D. ₦700.00

15. Which of the following is not a quadratic expression?

A. $y = 2x^2 - 5x$
B. $y = x(x - 5)$
C. $y = x^2 - 5$
D. $y = 5(x - 1)$

16. A trader sold an article for ₦4600.00 there by making a profit of 15%. Calculate the cost price of the article

A. ₦3,450.00
B. ₦3,910.00
C. ₦4,000.00
D. ₦5,290.00

A box contains 6 red balls, 10 blue balls and 4 green balls. If two balls are selected at random one after the other without replacement, use this information to answer questions 17 and 18.

17. What is the probability that the two balls are red?

A. $\frac{5}{33}$
B. $\frac{15}{190}$
C. $\frac{103}{132}$
D. $\frac{3}{38}$

18. What is the probability that one is green and the other is blue?

A. $\frac{4}{19}$

B. $\frac{5}{12}$

C. $\frac{2}{3}$

D. $\frac{2}{19}$

19. Simplify $4321_5 - 2424_5$

A. 1342_5

B. 2241_5

C. 3140_5

D. 4320_5

20. Find the LCM of $35x^2y$, $420x^3$ and $245xy^3$

A. $2840xy$

B. $2940x^3y^3$

C. $3940x^2y^2$

D. $4390x^5y^5$

21. Which of the following is an irrational number?

A. $\sqrt{121}$

B. $\sqrt{19}$

C. 25

D. $\frac{49}{625}$

22. The square root of $42\frac{1}{4}$ is

A. $3\frac{1}{4}$

B. $6\frac{1}{4}$

C. $6\frac{1}{2}$

D. $7\frac{1}{2}$

23. Calculate the volume of a cylinder of diameter 1m and height $1\frac{3}{4}$ m. $\pi = \frac{22}{7}$.

A. $2\frac{5}{8}m^3$

B. $2\frac{3}{8}m^3$

C. $1\frac{1}{2}m^3$

D. $1\frac{3}{8}m^3$

24. If 5 is multiplied by its multiplication inverse, the result is

A. 1

B. $\frac{1}{5}$

C. -5

D. 25

25. What progression is a + b, 3a + 2b, 5a +3b.....?

A. Arithmetic progression

B. Geometric progression

C. Harmonic progression

D. Arithmetic sequence

26. The numerical coefficient of x in $3x^2 - 8x + 11$ is

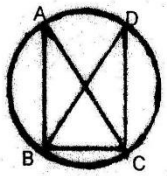
A. 11

B. 8

C. 2

D. -8

27. In the diagram below, $\angle$ BDC = 25° and $\angle$ ABC = 95°. Calculate $\angle$ ACB



A. 25°

B. 45°

C. 60°

D. 70°

28. Calculate the total surface area of a solid cube with sides 4cm each.

A. $46cm^3$
B. $64cm^3$
C. $96cm^3$
D. $108cm^3$

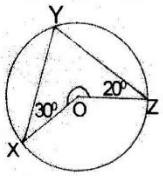
29. A man at the top of a cliff 120m high observe a boat at an angle of depression of 69°.What is the distance of the boat from the bottom of the cliff.

A 120 sin 69°
B. 120 sin 21°
C. 120 tan 21°
D. 120 tan 69°

30. Find the value of x in the equation $9^x = 729$

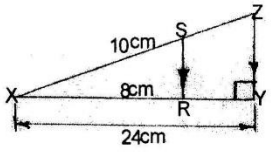
A. 27
B. 18
C. 9
D. 3

31. In the diagram below, O is the centre of the circle.if $\angle YXO = 30°$ and $\angle YZO = 20°$, what is the reflex angle XOZ?



A. 330°
B. 300°
C. 270°
D. 260°

32. XYZ is a right angled triangle in which YZ is parallel to RS. If XS =10cm, XR = 8cm and XY = 24cm, what is the lenght of YZ?



A. 18cm
B. 12cm
C. 6cm
D. 2cm

33 The number of cars sold by a dealer during the first six months of 2012 was as follows:

Jan	Feb	Mar	April	May	June
16	18	28	47	51	50

What was the monthly average?

A. 18
B. 35
C. 62
D. 93

34. Which of the following are characteristics of a bar chart?

(i) Equal width
(ii) Separated bars
(iii) Unequal spacing between bars

A. I and II only
B. II and III only
C. II only
D. III only

35. What percentage of observation lies outside inter-quartile range of any distribution?

A. 75%
B. 50%
C. 25%
D. $12\frac{1}{2}\%$

36. Ogive is a graph made with

A. index number
B. frequency polygon
C. cummulative frequency
D. time / distance

37. Which of the following figures has one line of symmetry only?

(i) Isosceles triangle
(ii) Rhombus
(iii) Kite

A. I and II only
B. I and III only
C. II and III only
D. I, II and III only

38. A trader makes a loss of 15% when selling an article. Find the ratio of selling price : cost price

A. 3 : 20
B. 3 : 17
C. 17 : 20
D. 20 : 23

39. $Cos\theta$ is negative and $\sin\theta$ is negative which of the following is true of θ?

A. $0° < \theta < 90°$
B. $90° < \theta < 180°$
C. $180° < \theta < 270°$
D. $270° < \theta < 360°$

40. Simplify $\left[\frac{3}{4}+\frac{1}{3}\right] \times 4\frac{1}{3} \div 3\frac{1}{4}$

A. $\frac{12}{13}$
B. $\frac{10}{9}$
C. $\frac{17}{12}$
D. $\frac{13}{9}$

41. While doing her chemistry practical, Mary recorded a reading as 1.2mg instead of 1.21mg. Calculate her percentage error.

A. 0.83%
B. 0.63%
C. 0.44%
D. 0.03%

42. The ratio of the number of men to the number of women in a 20 member committee is 3: 1. How many women must be added to the 20 member committee so as to make ratio of men to women 3 : 2

A. 2
B. 5
C. 7
D. 9

43. The compound interest on N500.00 for 2 years at 6% per annum is

A. N31.00
B. N61.80
C. N91.00
D. N92.80

44. Solve the equation $\frac{3}{m}-1=\frac{2}{3}$

A. -9

B. $1\frac{4}{5}$

C. $4\frac{1}{2}$

D. $5\frac{1}{5}$

45. Make g the subject of the formula $k=\pi r\sqrt{\frac{h^2}{g}}$

A. $\left[\frac{hr\pi}{k}\right]^2$

B. $\left[\frac{h}{\pi rk}\right]^2$

C. $\frac{h^2r^2\pi}{k}$

D. $\frac{h^2r^2}{\pi k}$

46. Solve the equation simultaneously $2m + n = 7$, $m - n = 2$

A. 3, 1

B. 1, 3

C. -1, 3

D. -3, 1

47. If $\sin 30^\circ = 0.5$, find $\tan 30^\circ$

A. $\sqrt{3}$

B. $\frac{1}{\sqrt{3}}$

C. $\frac{2}{\sqrt{3}}$

D. $\frac{\sqrt{3}}{2}$

48. What is the sum of the following vectors a = 2i +2j +3k, b = 3i + 4j + 5k?

A. 5i +4j +3k

B. 4i +6j +8k

C. 8i +6j +4k

D. 5i +6j +8k

49. What is the number of element in the set A = $\{m : 3 < m \leq 15\}$?

A. 12

B. 15

C. 18

D. 20

50. Find the circumference of latitude $\alpha^{o}N$ where Rkm is the radius of the earth.

A. $\pi R\cos\alpha$

B. $\pi R^2 \sin\alpha$

C. $2\pi R^2 \cos\alpha$

D. $2\pi R\cos\alpha$

APPENDIX B

Total Variance Explained by the result of the factor Analysis

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.328	16.657	16.657	8.328	16.657	16.657
2	3.165	6.330	22.986	3.165	6.330	22.986
3	2.434	4.868	27.855	2.434	4.868	27.855
4	1.912	3.824	31.679	1.912	3.824	31.679
5	1.655	3.310	34.989	1.655	3.310	34.989
6	1.339	2.679	37.667	1.339	2.679	37.667
7	1.246	2.492	40.159	1.246	2.492	40.159
8	1.227	2.455	42.614	1.227	2.455	42.614
9	1.152	2.305	44.918	1.152	2.305	44.918
10	1.081	2.162	47.080	1.081	2.162	47.080
11	1.041	2.081	49.161	1.041	2.081	49.161
12	1.020	2.040	51.201	1.020	2.040	51.201
13	.997	1.995	53.196			
14	.968	1.937	55.133			
15	.955	1.911	57.043			
16	.895	1.790	58.833			
17	.893	1.785	60.619			
18	.863	1.726	62.344			
19	.853	1.705	64.050			
20	.812	1.625	65.675			
21	.802	1.604	67.279			
22	.787	1.574	68.853			
23	.757	1.513	70.366			
24	.748	1.496	71.862			
25	.737	1.473	73.336			
26	.728	1.456	74.791			
27	.708	1.415	76.206			
28	.683	1.366	77.572			
29	.671	1.342	78.914			
30	.666	1.332	80.246			
31	.643	1.287	81.532			
32	.630	1.260	82.793			
33	.607	1.213	84.006			
34	.591	1.183	85.188			

35	.577	1.154	86.343			
36	.553	1.106	87.449			
37	.538	1.077	88.525			
38	.522	1.043	89.568			
39	.518	1.035	90.603			
40	.503	1.006	91.609			
41	.488	.975	92.584			
42	.470	.941	93.525			
43	.468	.935	94.461			
44	.449	.899	95.359			
45	.437	.874	96.233			
46	.427	.854	97.087			
47	.410	.820	97.907			
48	.394	.788	98.695			
49	.353	.706	99.401			
50	.299	.599	100.000			

Extraction Method: Principal Component Analysis.

APPENDIX C

Item Parameter Estimation (BILOG MG3)

ONE PARAMETER LOGISTIC MODEL

```
>GLOBAL DFName = 'C:\Users\UCHE\Desktop\NABTEB 1PLM\NABTEBMT.DAT',
       NPArm = 1,
       LOGistic,
       SAVe;
>SAVE MASter = 'NABTEB1P.MAS',
     CALib = 'NABTEB1P.CAL',
     PARm = 'NABTEB1P.PAR',
     SCOre = 'NABTEB1P.SCO',
     COVariance = 'NABTEB1P.COV',
     TSTat = 'NABTEB1P.TST',
     EXPected = 'NABTEB1P.EXP',
     ISTat = 'NABTEB1P.IST';
>LENGTH NITems = (50);
>INPUT NTOtal = 50,
      NALt = 2,
      NIDchar = 6;
>ITEMS ;
>TEST1 TNAme = 'NABTEB1P',
      INUmber = (1(1)50);
(6A1, 50A1)
>CALIB ACCel = 1.0000;
>SCORE METhod = 1;
```

MODEL FIT IRT FOR 2 PLM

```
>GLOBAL DFName = 'C:\Users\UCHE\Desktop\NABTEB 2 PLM\NABTEBMT.DAT',
       NPArm = 2,
       LOGistic,
       SAVe;
>SAVE MASter = 'NABTEB2PL.MAS',
     CALib = 'NABTEB2PL.CAL',
     PARm = 'NABTEB2PL.PAR',
     SCOre = 'NABTEB2PL.SCO',
     COVariance = 'NABTEB2PL.COV',
     TSTat = 'NABTEB2PL.TST',
     EXPected = 'NABTEB2PL.EXP',
     ISTat = 'NABTEB2PL.IST';
>LENGTH NITems = (50);
>INPUT NTOtal = 50,
      NALt = 2,
      NIDchar = 6;
>ITEMS ;
>TEST1 TNAme = 'NABTEB2P',
      INUmber = (1(1)50);
(6A1, 50A1)
>CALIB ACCel = 1.0000,
```

```
      TPRior,
      NOGprior;
>SCORE METhod = 1;
COMPARISON OF TECHNIQUES FOR ESTIMATING MODEL FIT OF ITEM
RESPONSE THEORY
DICHOTOMOUS MODELS USING NABTEB 2018 MATHEMATICS MULTIPLE
CHOICE TEST ITEMS
>COMMENT
COMPARISON OF TECHNIQUES FOR ESTIMATING MODEL FIT OF ITEM
RESPONSE THEORY
DICHOTOMOUS MODELS USING NABTEB 2018 MATHEMATICS MULTIPLE
CHOICE TEST ITEMS

>GLOBAL DFName = 'C:\Users\UCHE\Desktop\NABTEB\NABTEBMT.DAT',
       NPArm = 3,
       LOGistic,
       SAVe;
>SAVE MASter = 'NABTEBMTH.MAS',
     CALib = 'NABTEBMTH.CAL',
     PARm = 'NABTEBMTH.PAR',
     SCOre = 'NABTEBMTH.SCO',
     TSTat = 'NABTEBMTH.TST',
     EXPected = 'NABTEBMTH.EXP',
     ISTat = 'NABTEBMTH.IST';
>LENGTH NITems = (50);
>INPUT NTOtal = 50,
NALt = 2,
      NIDchar = 6;
>ITEMS ;
>TEST1 TNAme = 'NABTEBMT',
      INUmber = (1(1)50);
(6A1, 50A1)
>CALIB ACCel = 1.0000,
      TPRior,
      GPRior;
>SCORE METhod = 1;
```

APPENDIX D

LIKELIHOOD RATIO TEST

BILOG-MG V3.0
REV 19990329.1300

BILOG-MG ITEM MAINTENANCE PROGRAM: LOGISTIC ITEM RESPONSE MODEL

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

*** PHASE 2 ***

MODEL-FIT IRT FOR 1 PLM

>CALIB ACCel = 1.0000;

CALIBRATION PARAMETERS
==========

MAXIMUM NUMBER OF EM CYCLES: 20

MAXIMUM NUMBER OF NEWTON CYCLES: 2
CONVERGENCE CRITERION: 0.0100
ACCELERATION CONSTANT: 1.0000

LATENT DISTRIBUTION: NORMAL PRIOR FOR EACH GROUP
PLOT EMPIRICAL VS. FITTED ICC'S: NO
DATA HANDLING: DATA ON SCRATCH FILE
CONSTRAINT DISTRIBUTION ON SLOPES: NO
CONSTRAINT DISTRIBUTION ON THRESHOLDS: NO
1

****

CALIBRATION OF MAINTEST

NABTEB1P

****
METHOD OF SOLUTION:

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

QUADRATURE POINTS AND PRIOR WEIGHTS:

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

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

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

[E-M CYCLES]

-2 LOG LIKELIHOOD = 211000.458

CYCLE 1; LARGEST CHANGE= 0.06303

-2 LOG LIKELIHOOD = 210952.606

CYCLE 2; LARGEST CHANGE= 0.02343

-2 LOG LIKELIHOOD = 210926.695

CYCLE 3; LARGEST CHANGE= 0.01797

-2 LOG LIKELIHOOD = 210910.997

CYCLE 4; LARGEST CHANGE= 0.03437

-2 LOG LIKELIHOOD = 210893.358

CYCLE 5; LARGEST CHANGE= 0.01477

-2 LOG LIKELIHOOD = 210890.143

CYCLE 6; LARGEST CHANGE= 0.00592

[NEWTON CYCLES]

-2 LOG LIKELIHOOD: 210889.3579

CYCLE 7; LARGEST CHANGE= 0.00434

MODEL-FIT IRT FOR 2 PLM

>CALIB ACCel = 1.0000,

TPRior,

NOGprior;

CALIBRATION PARAMETERS
==========

MAXIMUM NUMBER OF EM CYCLES: 20

MAXIMUM NUMBER OF NEWTON CYCLES: 2
CONVERGENCE CRITERION: 0.0100
ACCELERATION CONSTANT: 1.0000

LATENT DISTRIBUTION: NORMAL PRIOR FOR EACH GROUP
PLOT EMPIRICAL VS. FITTED ICC'S: NO
DATA HANDLING: DATA ON SCRATCH FILE
CONSTRAINT DISTRIBUTION ON SLOPES: YES
CONSTRAINT DISTRIBUTION ON THRESHOLDS: YES
SOURCE OF ITEM CONSTRAINT DISTIBUTION
MEANS AND STANDARD DEVIATIONS: PROGRAM DEFAULTS
1

****

CALIBRATION OF MAINTEST

NABTEB2P

****

METHOD OF SOLUTION:

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

QUADRATURE POINTS AND PRIOR WEIGHTS:

1 2 3 4 5

```
POINT    -0.4000E+01 -0.3429E+01 -0.2857E+01 -0.2286E+01 -0.1714E+01
WEIGHT    0.7648E-04  0.6387E-03  0.3848E-02  0.1673E-01  0.5245E-01

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

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

***** ITEM:  31 OMITTED FROM CALIBRATION. *****
INITIAL SLOPE LESS THAN -0.15.


***** ITEM:  42 OMITTED FROM CALIBRATION. *****
INITIAL SLOPE LESS THAN -0.15.

[E-M CYCLES]


-2 LOG LIKELIHOOD =     191781.641

CYCLE      1;   LARGEST CHANGE=   0.86238

-2 LOG LIKELIHOOD =     191013.136


CYCLE      2;   LARGEST CHANGE=   0.16392

-2 LOG LIKELIHOOD =     190932.844

CYCLE      3;   LARGEST CHANGE=   0.08308

-2 LOG LIKELIHOOD =     190913.590

CYCLE      4;   LARGEST CHANGE=   0.09276

-2 LOG LIKELIHOOD =     190901.094

CYCLE      5;   LARGEST CHANGE=   0.03090

-2 LOG LIKELIHOOD =     190897.355

CYCLE      6;   LARGEST CHANGE=   0.02699

-2 LOG LIKELIHOOD =     190895.764
```

CYCLE 7; LARGEST CHANGE= 0.02238

-2 LOG LIKELIHOOD = 190894.445

CYCLE 8; LARGEST CHANGE= 0.02035

-2 LOG LIKELIHOOD = 190894.350

CYCLE 9; LARGEST CHANGE= 0.04801

-2 LOG LIKELIHOOD = 190894.624

CYCLE 10; LARGEST CHANGE= 0.03039

-2 LOG LIKELIHOOD = 190893.718

CYCLE 11; LARGEST CHANGE= 0.00125

[NEWTON CYCLES]

-2 LOG LIKELIHOOD: 190893.6625

CYCLE 12; LARGEST CHANGE= 0.00134

MODEL-FIT IRT FOR 3 PLM

>CALIB ACCel = 1.0000,

TPRior,

GPRior;

CALIBRATION PARAMETERS

==========

MAXIMUM NUMBER OF EM CYCLES: 20

MAXIMUM NUMBER OF NEWTON CYCLES: 2
CONVERGENCE CRITERION: 0.0100
ACCELERATION CONSTANT: 1.0000

LATENT DISTRIBUTION: NORMAL PRIOR FOR EACH GROUP
PLOT EMPIRICAL VS. FITTED ICC'S: NO
DATA HANDLING: DATA ON SCRATCH FILE

CONSTRAINT DISTRIBUTION ON ASYMPTOTES: YES
CONSTRAINT DISTRIBUTION ON SLOPES: YES
CONSTRAINT DISTRIBUTION ON THRESHOLDS: YES
SOURCE OF ITEM CONSTRAINT DISTIBUTION
MEANS AND STANDARD DEVIATIONS: PROGRAM DEFAULTS
1

****

CALIBRATION OF MAINTEST

NABTEBMT

****

METHOD OF SOLUTION:

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

QUADRATURE POINTS AND PRIOR WEIGHTS:

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

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

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

***** ITEM: 31 OMITTED FROM CALIBRATION. *****
INITIAL SLOPE LESS THAN -0.15.

***** ITEM: 42 OMITTED FROM CALIBRATION. *****
INITIAL SLOPE LESS THAN -0.15

[E-M CYCLES]

-2 LOG LIKELIHOOD = 201438.249

CYCLE 1; LARGEST CHANGE= 6.10092

-2 LOG LIKELIHOOD = 192035.372

CYCLE 2; LARGEST CHANGE= 0.79285

-2 LOG LIKELIHOOD = 190796.719

CYCLE 3; LARGEST CHANGE= 2.05516

-2 LOG LIKELIHOOD = 190401.884

CYCLE 4; LARGEST CHANGE= 0.77015

-2 LOG LIKELIHOOD = 190257.893

CYCLE 5; LARGEST CHANGE= 0.29081

-2 LOG LIKELIHOOD = 190203.338

CYCLE 6; LARGEST CHANGE= 0.24765

-2 LOG LIKELIHOOD = 190205.973

CYCLE 7; LARGEST CHANGE= 0.30482

-2 LOG LIKELIHOOD = 190167.064

CYCLE 8; LARGEST CHANGE= 0.10981

-2 LOG LIKELIHOOD = 190164.952

CYCLE 9; LARGEST CHANGE= 0.16835

-2 LOG LIKELIHOOD = 190165.169

CYCLE 10; LARGEST CHANGE= 0.03663

-2 LOG LIKELIHOOD = 190164.038

CYCLE 11; LARGEST CHANGE= 0.04842

-2 LOG LIKELIHOOD = 190164.283

CYCLE 12; LARGEST CHANGE= 0.07000

-2 LOG LIKELIHOOD = 190164.610

CYCLE 13; LARGEST CHANGE= 0.04521

-2 LOG LIKELIHOOD = 190163.924

CYCLE 14; LARGEST CHANGE= 0.00346

[NEWTON CYCLES]

-2 LOG LIKELIHOOD: 190162.9949

CYCLE 15; LARGEST CHANGE= 0.00385

APPENDIX E
AIC AND BIC ANALYSES

Akaike Information Criterion (AIC)

1PM $AIC(\text{Model}) = d + 2p = 210889.4 + 2(1 \times 50) = 210989.4$

2PM $AIC(\text{Model}) = d + 2p = 190893.7 + 2(2 \times 48) = 191959.0$

3PM $AIC(\text{Model}) = d + 2p = 190163.0 + 2(3 \times 48) = 190355.0$

Bayesian Information Criterion (BIC)

1PM $BIC(\text{Model}) = d + p\,(\ln N) = 210889.4 + 50(1 \times \ln 4948) = 210889.4 + 425.3 = 211314.7$

2PM $BIC(\text{Model}) = d + p\,(\ln N) = 190893.7 + 48(2 \times \ln 4948) = 190893.7 + 816.6 = 191710.3$

3PM $BIC(\text{Model}) = d + p\,(\ln N) = 190163.0 + 48(3 \times \ln 4948) = 190163.0 + 1225.0 = 191388.0$

APPENDIX F

WinBUGS Code Used for One, Two and Three-Parameter Logistic Models

model

```
{
# m1_ : 1 Parameter logistic model.
# m3222 : data generated with 1PM,
# number of items = 50,
# sample size = 4948
# mean theta ~ N(0,1).
# These data are an initial calibration sample N=1000.
for (j in 1:N){
for (k in 1:T){
r[j,k]<-resp[j,k]
}}
# 1PL model
for (j in 1:N){
for (k in 1:T){
tt[j,k]<-exp(-(theta[j] - beta[k]))
p[j,k]<-(1)/(1 + tt[j,k])
r[j,k]~dbern(p[j,k])
}
theta[j] ~ dnorm(0.,1.)
}
# Priors
for (k in 1:T){
beta[k]~dnorm(0.,1.)
}
}
model
{
# m2_ : 2 Parameter logistic model.
# m4948: data generated with 2PM,
# number of items = 50,
# sample size = 4948
# mean theta ~ N(0,1).
# These data are an initial calibration sample N=1000.
for (j in 1:N){
for (k in 1:T){
r[j,k]<-resp[j,k]
}}
# 2PL model
for (j in 1:N){
for (k in 1:T){
tt[j,k]<-exp(-a[k]*(theta[j] - beta[k]))
p[j,k]<-(1)/(1 + tt[j,k])
r[j,k]~dbern(p[j,k])
}
theta[j] ~ dnorm(0.,1.)
}
# Priors
```

```
for (k in 1:T){
a[k]~dlnorm(0.,1.)
beta[k]~dnorm(0.,1.)
}
}
model
{
# m3_ : 3 Parameter logistic model.
# m3222 : data generated with 3PM,
# number of items = 50,
# sample size = 4948
# mean theta ~ N(0,1).
# These data are an initial calibration sample N=1000.
for (j in 1:N){
for (k in 1:T){
r[j,k]<-resp[j,k]
}}
# 3PL model
for (j in 1:N){
for (k in 1:T){
tt[j,k]<-exp(-a[k]*(theta[j] - beta[k]))
p[j,k]<-c[k]+(1-c[k])/(1 + tt[j,k])
r[j,k]~dbern(p[j,k])
}
theta[j] ~ dnorm(0.,1.)
}
# Priors
for (k in 1:T){
a[k]~dlnorm(0.,1.)
beta[k]~dnorm(0.,1.)
c[k]~dbeta(5,17)
}
```

APPENDIX G
MATLAB Code Used for Calculating Cross-Validation Log-Likelihood (CVLL)

```
%Condition m3222: data generated following 1PM; n=50; N=4948; N(0,1)
%estimated item parameters by 1PL: each column contains beta of each data set
load estpar1.dat
%loading CV data set (1000 by 40)
load cvdata.dat; cvloglik=zeros(10,3);
%CV log-likelihood of 1PL of each data set
for z=1:10; n=50; N=4948;
a=ones(n,1); beta=estpar1(:,z); c=zeros(n,1);
cv=zeros(N,1);
for j=1:N
resp=cvdata(j,:);
cvj=cv(j);
end
cvloglik(z,1)=sum(cv);
end
%21 quadrature points between -4 to 4
k=-4:.4:4; K=length(k); prob=zeros(1,K); L=zeros(1,K);
%to calculate likelihood at each node
for t=1:K
p=0; lik=1; tt=zeros(1,n);
for i=1:n
tt(i)=exp(-a(i)*(k(t)-beta(i)));
p=c(i)+(1-c(i))/(1+ tt(i));
lik = lik * p^resp(i) * (1-p)^(1-resp(i));
end;
L(t)=lik;
end
%to cmpute a posterior probability
for t=1:K
prob(t)=L(t)*normpdf(k(t),0,1);
end
prob=prob/sum(prob);
%to get Cross-Validation log likelihood(CVLL)
cvj=0; for t=1:K
cvj=cvj+prob(t)*log(L(t));
CVLL = cvj*N;
end


%estimated item parameters by 2PL: each column contain a, beta, and c
%of each data set
load estpar2.dat
%loading CV data set (4948 by 50)
load cvdata.dat; cvloglik=zeros(10,3);
%CV log-likelihood of 2PL of each data set
for z=1:10; n=50; N=4948;
a=estpar2(1:n,z); beta=estpar2(n+1:2*n,z); c=zeros(n,1);
cv=zeros(N,1);
```

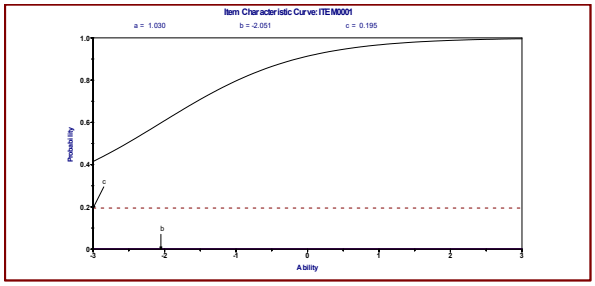
```
for j=1:N
resp=cvdata(j,:);
cvj=cv(j);
end
cvloglik(z,2)=sum(cv);
end
%21 quadrature points between -4 to 4
k=-4:.4:4; K=length(k); prob=zeros(1,K); L=zeros(1,K);
%to calculate likelihood at each node
for t=1:K
p=0; lik=1; tt=zeros(1,n);
for i=1:n
tt(i)=exp(-a(i)*(k(t)-beta(i)));
p=c(i)+(1-c(i))/(1+ tt(i));
lik = lik * p^resp(i) * (1-p)^(1-resp(i));
end;
L(t)=lik;
end
%to cmpute a posterior probability
for t=1:K
prob(t)=L(t)*normpdf(k(t),0,1);
end
prob=prob/sum(prob);
%to get CV log likelihood
cvj=-0; for t=1:K
cvj=cvj+prob(t)*log(L(t));
CVLL = cvj*N;
end


%estimated item parameters by 3PL: each column contain a, beta, and c
%of each data set
load estpar3.dat
%loading CV data set (4948 by 50)
load cvdata.dat; cvloglik=zeros(10,3);
%CV log-likelihood of 3PL of each data set
for z=1:10; n=50; N=4948;
a=estpar3(1:n,z); beta=estpar3(n+1:2*n,z); c=estpar3(2*n+1:3*n,z);
cv=zeros(N,1);
for j=1:N
resp=cvdata(j,:);
end
cvloglik(z,3)=sum(cv);
end
%21 quadrature points between -4 to 4
k=-4:.4:4; K=length(k); prob=zeros(1,K); L=zeros(1,K);
%to calculate likelihood at each node
for t=1:K
p=0; lik=1; tt=zeros(1,n);
for i=1:n
```

```
tt(i)=exp(-a(i)*(k(t)-beta(i)));
p=c(i)+(1-c(i))/(1+ tt(i));
lik = lik * p^resp(i) * (1-p)^(1-resp(i));
end;
L(t)=lik;
end
%to cmpute a posterior probability
for t=1:K
prob(t)=L(t)*normpdf(k(t),0,1);
end
prob=prob/sum(prob);
%to get CV log likelihood
cvj=0; for t=1:K
cvj=cvj+prob(t)*log(L(t));
CVLL = cvj*N;
end
```

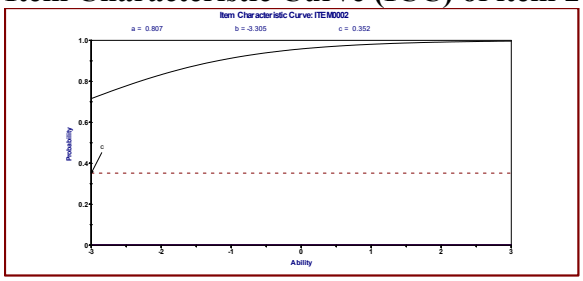
APPENDIX H

The Item Characteristics Curve of NABTEB 2018 mathematics multiple choice test for Individual Test Items

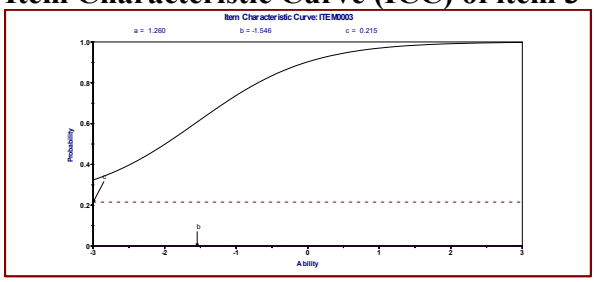
Item Characteristic Curve (ICC) of item 1



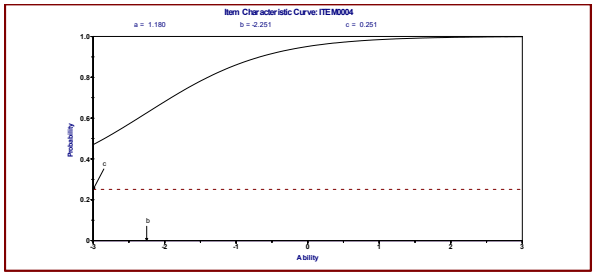
Item Characteristic Curve (ICC) of item 2



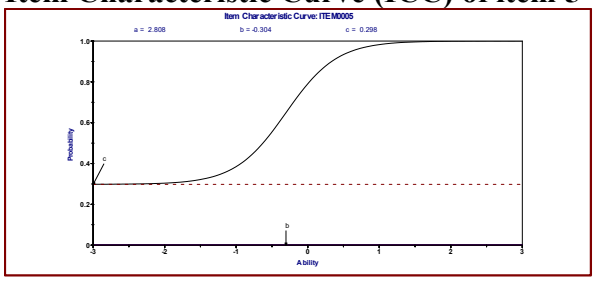
Item Characteristic Curve (ICC) of item 3



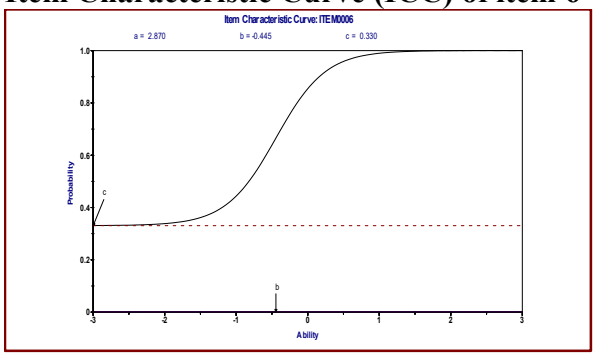
Item Characteristic Curve (ICC) of item 4



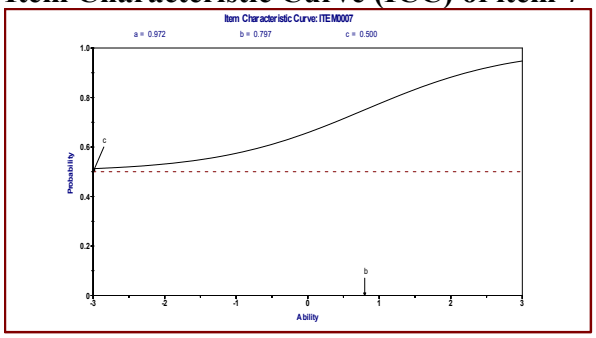
Item Characteristic Curve (ICC) of item 5



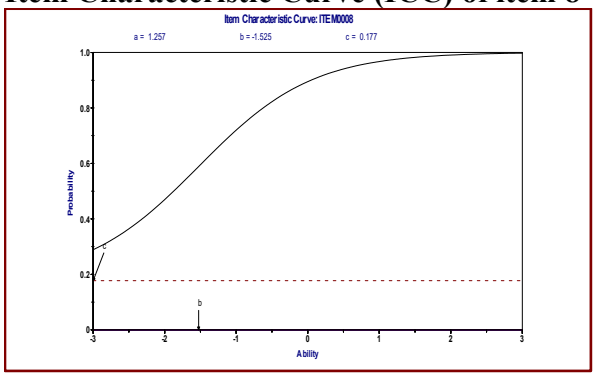
Item Characteristic Curve (ICC) of item 6



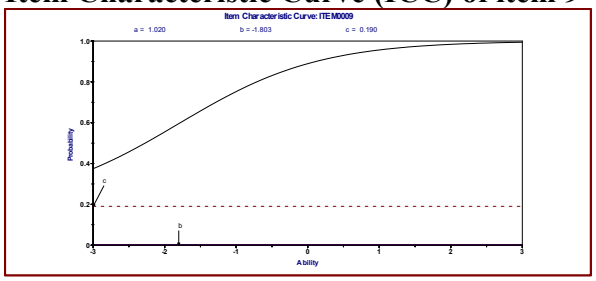
Item Characteristic Curve (ICC) of item 7



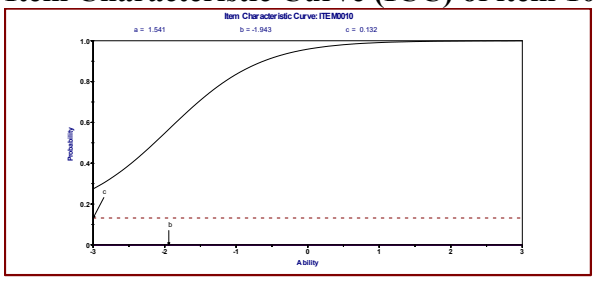
Item Characteristic Curve (ICC) of item 8



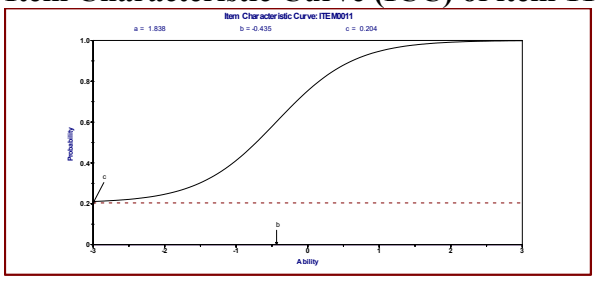
Item Characteristic Curve (ICC) of item 9



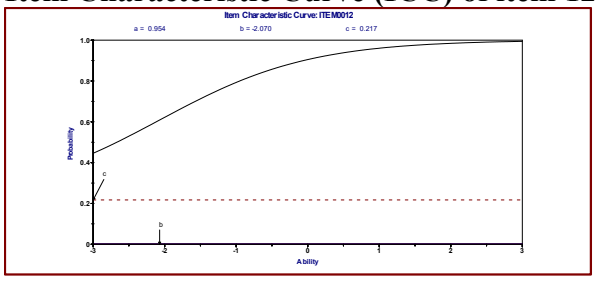
Item Characteristic Curve (ICC) of item 10



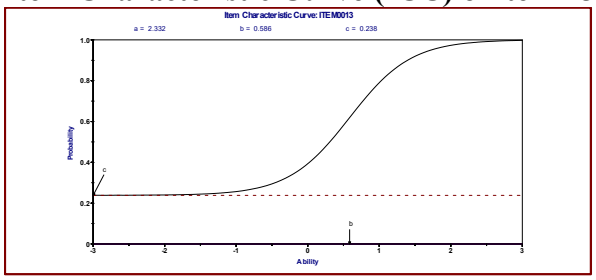
Item Characteristic Curve (ICC) of item 11



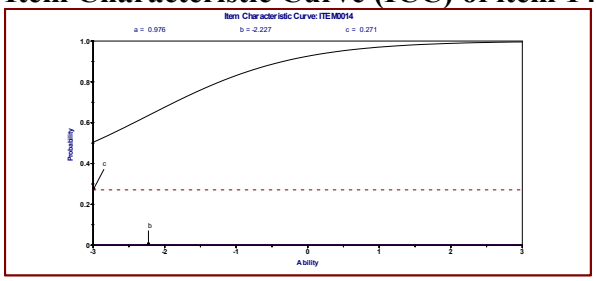
Item Characteristic Curve (ICC) of item 12



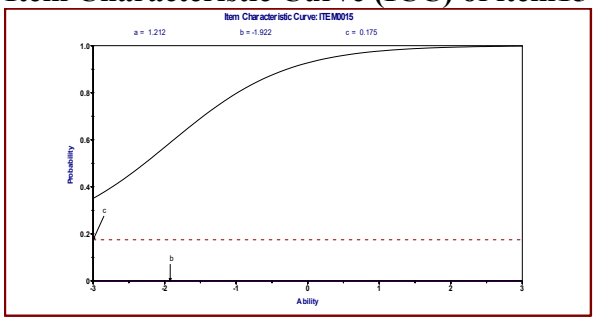
Item Characteristic Curve (ICC) of item 13



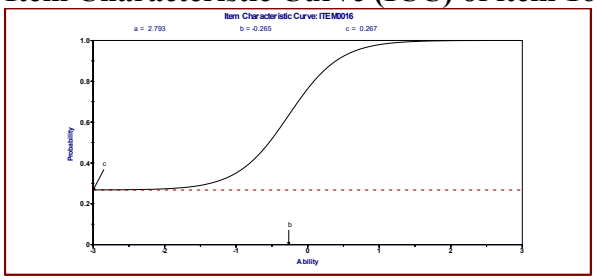
Item Characteristic Curve (ICC) of item 14



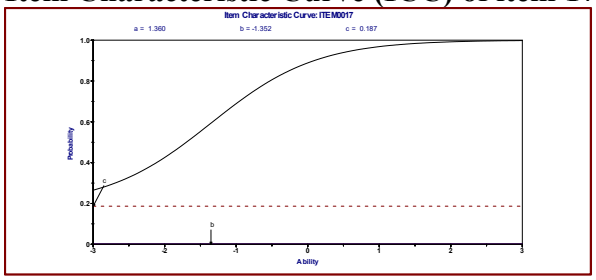
Item Characteristic Curve (ICC) of item15



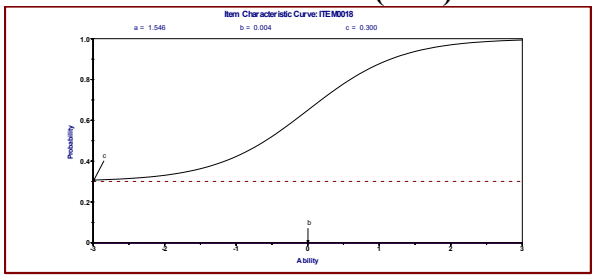
Item Characteristic Curve (ICC) of item 16



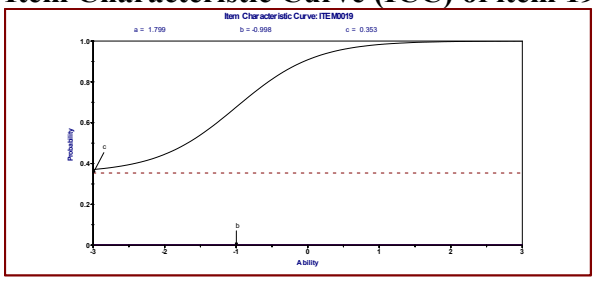
Item Characteristic Curve (ICC) of item 17



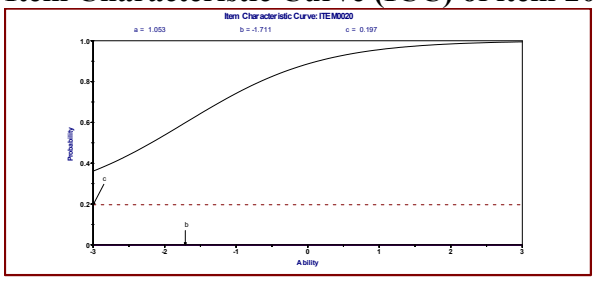
Item Characteristic Curve (ICC) of item 18



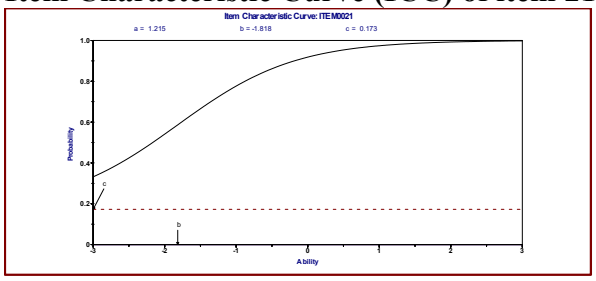
Item Characteristic Curve (ICC) of item 19



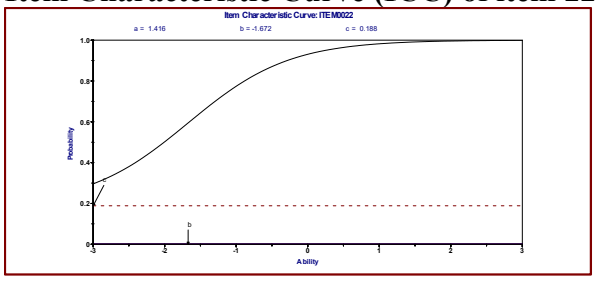
Item Characteristic Curve (ICC) of item 20



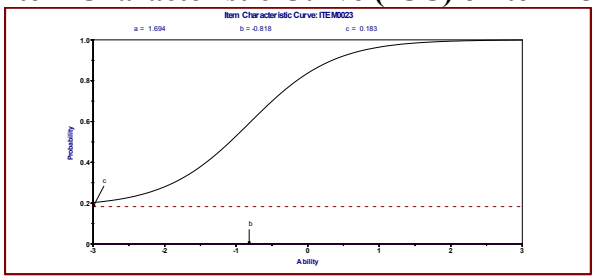
Item Characteristic Curve (ICC) of item 21



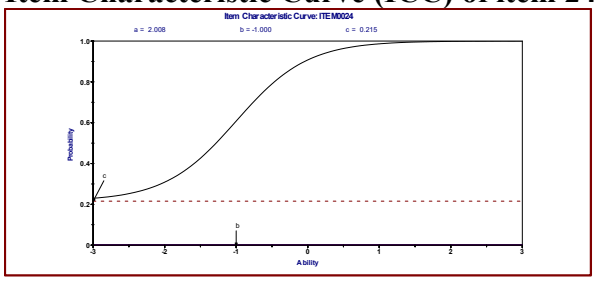
Item Characteristic Curve (ICC) of item 22



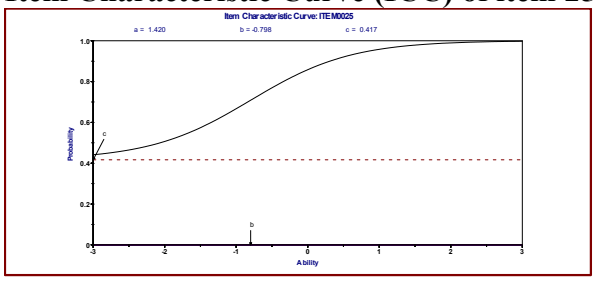
Item Characteristic Curve (ICC) of item 23



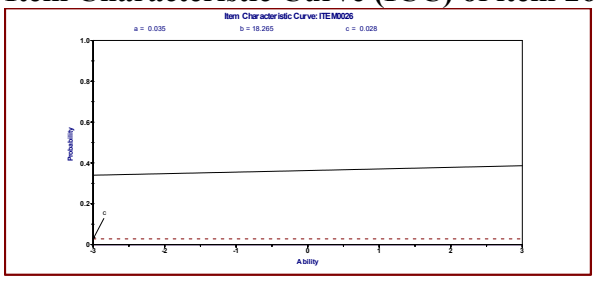
Item Characteristic Curve (ICC) of item 24



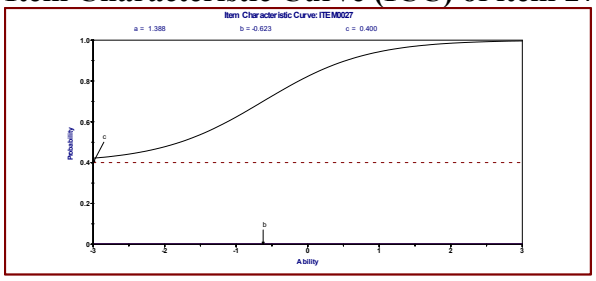
Item Characteristic Curve (ICC) of item 25



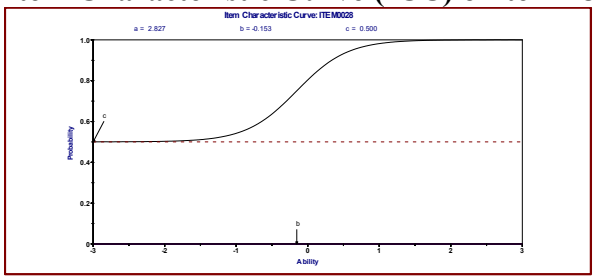
Item Characteristic Curve (ICC) of item 26



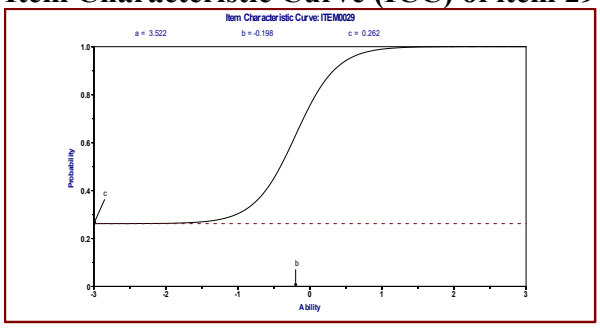
Item Characteristic Curve (ICC) of item 27



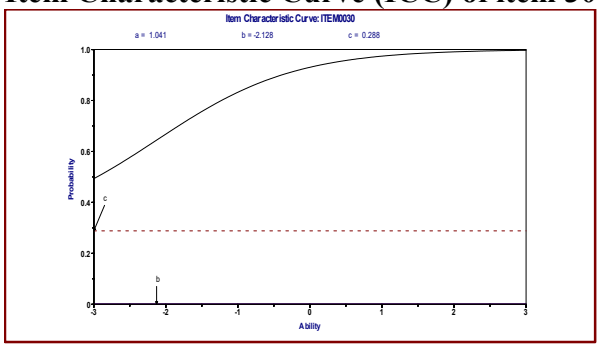
Item Characteristic Curve (ICC) of item 28



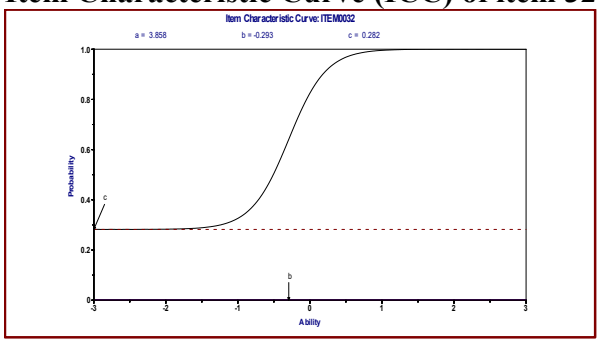
Item Characteristic Curve (ICC) of item 29



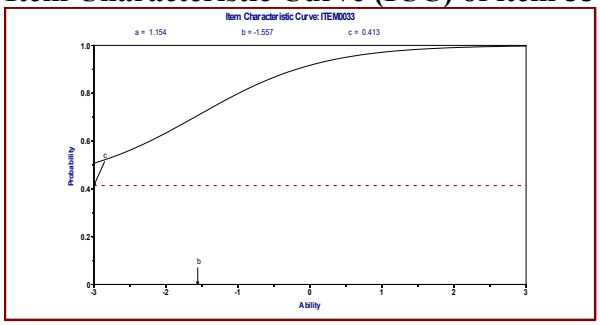
Item Characteristic Curve (ICC) of item 30



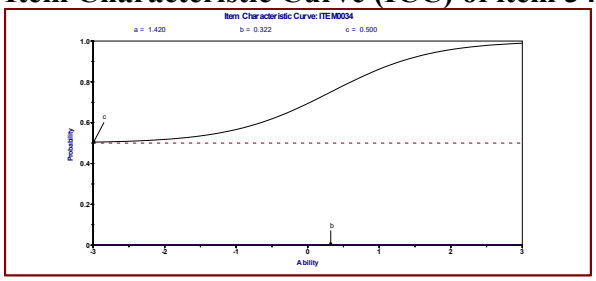
Item Characteristic Curve (ICC) of item 32



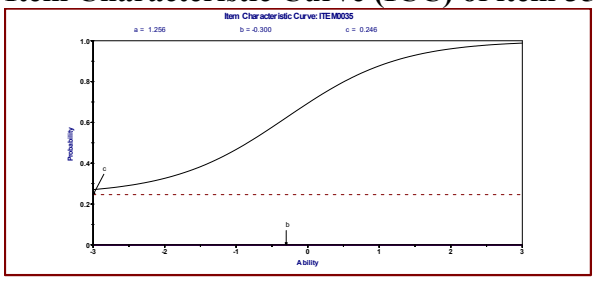
Item Characteristic Curve (ICC) of item 33



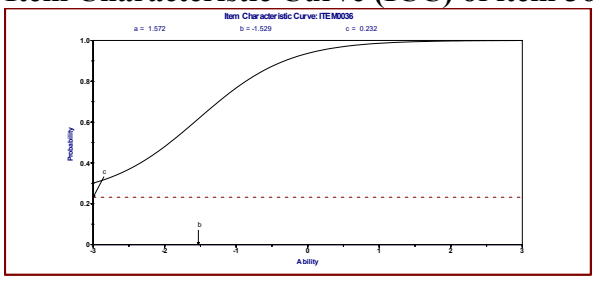
Item Characteristic Curve (ICC) of item 34



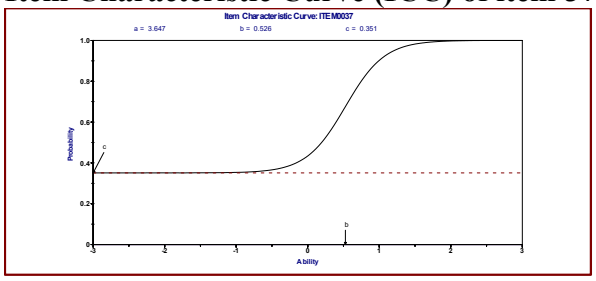
Item Characteristic Curve (ICC) of item 35



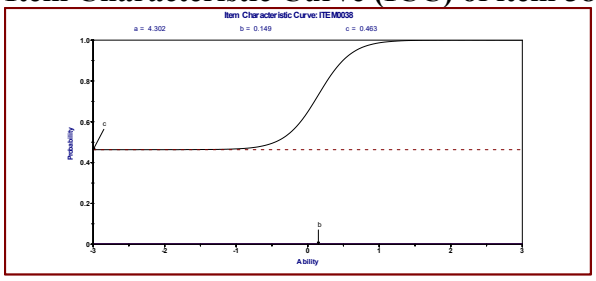
Item Characteristic Curve (ICC) of item 36



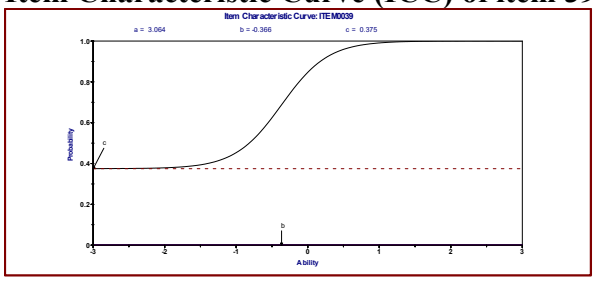
Item Characteristic Curve (ICC) of item 37



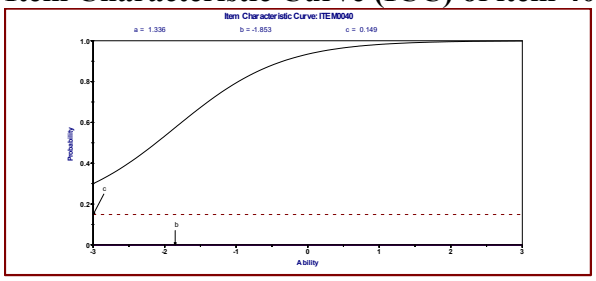
Item Characteristic Curve (ICC) of item 38



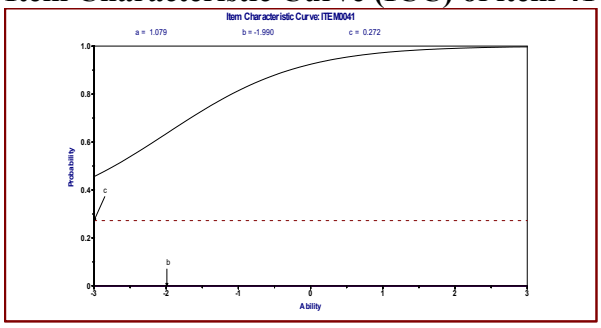
Item Characteristic Curve (ICC) of item 39



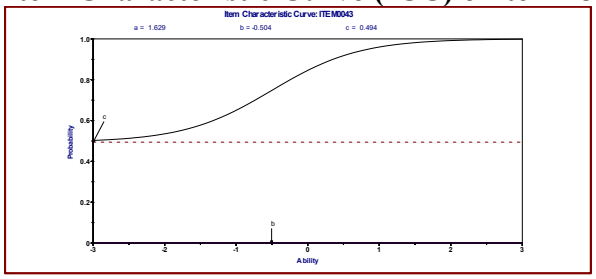
Item Characteristic Curve (ICC) of item 40



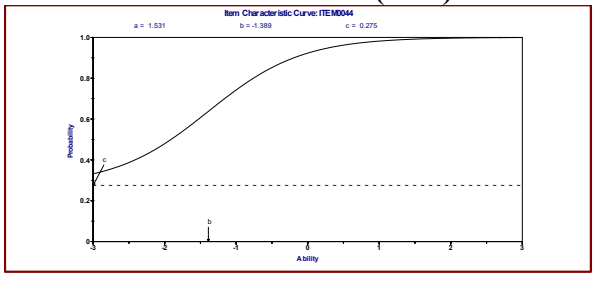
Item Characteristic Curve (ICC) of item 41



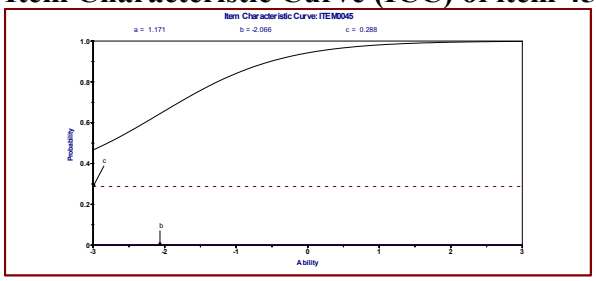
Item Characteristic Curve (ICC) of item 43



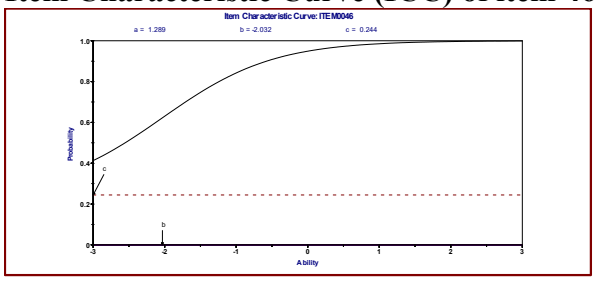
Item Characteristic Curve (ICC) of item 44



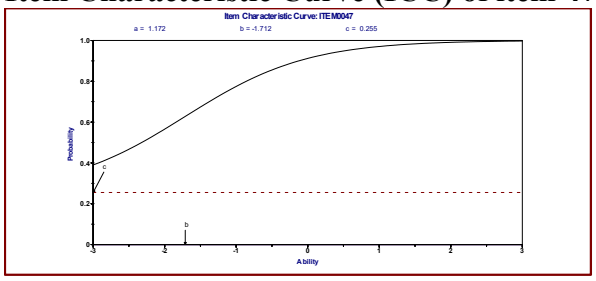
Item Characteristic Curve (ICC) of item 45



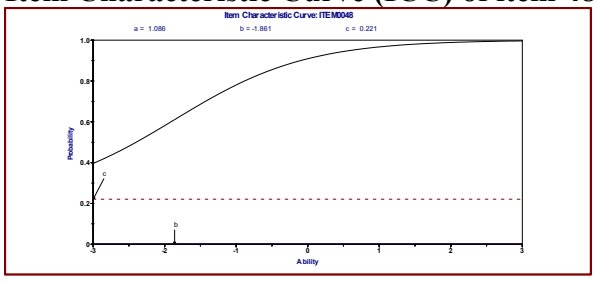
Item Characteristic Curve (ICC) of item 46



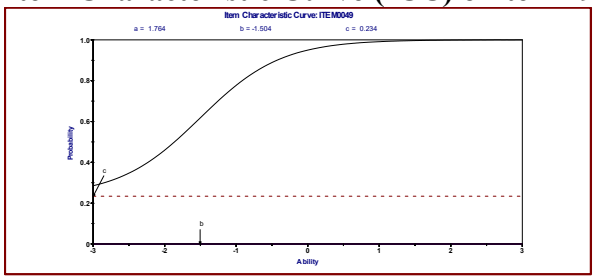
Item Characteristic Curve (ICC) of item 47



Item Characteristic Curve (ICC) of item 48



Item Characteristic Curve (ICC) of item 49



Item Characteristic Curve (ICC) of item 50