EXPLORING THE APPLICABILITY OF 4-PARAMETER LOGISTIC ITEM RESPONSE THEORY MODEL IN THE CALIBRATION OF COMPUTER-BASED MATHEMATICS ACHIEVEMENT TEST

Omonike R. Lawal & J. Gbenga Adewale

Abstract

This study explored the applicability of 4-parameter logistic Item Response Theory (IRT) model in the calibration of senior secondary school (SSS) Computer-Based Mathematics Achievement Test (CBMAT) in Lagos State, Nigeria. A causal-comparative design was adopted while Lagos Education District 1 was used with eight schools that had functional computer laboratories where 874 SSS 2 students were randomly selected. The CBMAT had IRT empirical reliability value of 0.89. Two research questions were raised and data was analysed using Stout's Test of Essential Dimensionality and 1-, 2-, 3- and 4PL models of Multidimensional IRT package of R programming. Results revealed that the CBMAT instrument was unidimensional and fitted the 4PL model. The popularly used models for calibration by researchers for so many years are the 1, 2 and 3PL models. This is due to the complexity that was attached to estimating parameters of the 4PL model. However, software has been made available to tackle this issue. It was therefore inferred from the findings that CBMAT revealed only a dominant mathematics ability trait. Based on the findings, it was recommended that suitable calibration procedures for the right choice of appropriate model should be imbibed by test developers, pyschometricians and researchers on any form of scale that is meant to measure students' learning outcomes.

Keywords: 4-parameter logistic IRT model, calibration, computer-based mathematics achievement test, unidimensional.

Introduction

The essence of being dynamic and the quest for new knowledge in every facet of human life have brought major multidimensional transformations into the society. Such transformations help in making life more meaningful and better. Meanwhile, whatever positive change that is observed in any society is traceable to the quality of education that is evident in that environment. It is on this note that practitioners/stakeholders in educational system, especially psychometricians are continually interested in making sure that measurement and assessment are done with utmost professionalism, which aligns with emerging trends. Troy-Gerard, (2004) and Adedoyin (2010) asserted that the field of educational measurement globally is undergoing a number of changes to meet the increasing demand for valid interpretation of examinee's score/performance. This effective explanation of scores must however be preceded by a carefully and objectively measured assessment/testing because of certain traits that are inherent in the examinee.

Nenty (2004) was of the opinion that every attempt of measuring such intrinsic features is susceptible to error. This is why operationalisable theories or models of measurement that will provide guiding principles in an attempt to measure appropriately is necessary. These theories, according to Ariyo (2015), provide a framework for considering issues and addressing technical problems in test design. They also specify the precise relationships among test items and ability scores so that careful test design work can be done to produce desired test score distributions and least error term that is tolerated. The most recent of these theories is the Item Response Theory (IRT), which is an improvement over the traditional approach (Classical Test Theory) for unbiased measurement. Adedoyin (2010) and Ojerinde (2013) described IRT framework as a renowned and acceptable, modern-day approach because of its impartial, more flexible and invariance property. Alordiah (2015) stated that in testing an individual's observed score (X_i), IRT is mathematically represented as:

$$X_i = \theta_i + \lambda_i + \varepsilon_i \text{..........................eqn.1}$$

Where θ_i is the true examinee ability, λ_i is the extraneous (systematic) error variable component of the score and ε_i is the random error. The acknowledgment of systematic error in IRT is a key development for over CTT approach. IRT is based on a ratio scale measurement, sample/item independent characteristics (invariance principle) and ability reported on both item and total instrument levels (Hambleton & Swaminathan, 1985; Adedoyin, 2010; Ojerinde, 2013). This approach reliably calibrates examinees and test item on a common scale that is understood to show individuals' ability and specified characteristics of the test item (Baker, 2001).

The values of the parameter estimate to be assessed solely depend on the kind of parameter model to be used. Ojerinde, Popoola, Ojo and Onyeneho (2012) classified IRT parameters into examinee and item parameters for a given test. Ateach ability level of the examinee, there is a certain probability, $P(\theta)$, that an examinee with that ability will give a correct response to the item. Basic assumptions that are essential for effective usage, appropriate interpretations of scores, precise and useful results of IRT methods are the trait dimensionality, item local independence as well as monotonicity of response function (Yen, 1993; Reeve, 2000).

Thus, applications of IRT models to investigating test items, the tests themselves and item score patterns are only valid if the IRT model holds (Cees & Rob, 2003). Although there are numerous models ranging from their different usage in terms of what the test is to measure and whether the response data is dichotomously or polytomously scored. Two basic types of IRT models that could be used for calibrating test are the unidimensional and multidimensional IRT models (Olonade, Metibemu & Adewale, 2017). Unidimensional models indicate that only a single dominant trait accounts for the difference observed in respondents' test performance while multidimensional models show that two or more traits account for the variation seen in performance.

The choice of a right model in the model-data fit analysis is subjected to the number of dimension(s) the test in question portrays. Even though IRT models were initially established for items that are dichotomously scored with unidimensional traits, its ideas and approaches were later extended to a wide range of multidimensional models (de Ayala, 2009; Reckase, 2009; de Mars, 2010). For models with dichotomous response format (correct/incorrect), there are usually three distinct models in the past that are predominantly used by researchers in fitting data set so as to estimate item and examinee's parameters. These are the 3-, 2- and 1-paramter logistic (PL) models. These models relate mathematical relationship in terms of the probability of correct response to the items. The Item Response Function (IRF) of a 3PL model is given as:

$$P_i(\theta_s) = \Pr\,(X_{is} = 1 \mid \theta_s, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1+e^{-1.7a_i(\theta s - b_i)}} \text{..........eqn.2}$$

where c_i = guessing parameter of item i; b_i = difficulty parameter of an item i; a_i = discrimination parameter of item i and θ_s = ability level of an examinees. However, If c = 0, equation 2 becomes IRF of a 2PL model and when c = 0 and a = 1, the equation becomes that of a 1PL model.

One Parameter Logistic (1PL) Model is the most common and simplest of the IRT model. It was formulated by a Danish mathematician, George Rasch (van der & Hambleton, 1997). An examinee's correct response to binary-optioned item is determined by his trait level and the difficulty of the item (b_i). *The strength of 1PL model is that* its solutions are monotonic with total score. However, it is limited in such a way that it cannot be used with large number of examinees and items only differ in how hard they are to be answered with assessing the latent trait of the examinee.

Two-Parameter Logistic (2PL) model was proposed by Birnbaum (1968). It was an extension of 1PL model with an additional estimate that is incorporated (discrimination parameter, *a*), which gives the 2PL model a better fit when assessed. The higher the discriminating value of an item when calibrated, the more such item contributes to a specific learner's ability estimate. The 2PL model is strictly applicable to items where guessing is very unlikely and items differ in how difficult they can be answered and how well they can differentiate each examinee on the ability continuum.

Three-Parameter Logistic (3PL) model was developed through the effort of Barton and Lord (1981) for test items where *n* alternatives are likely as options. A pseudo-guessing (c_i) parameter also known as lower asymptote was integrated to improve on the 2PL model. This model was formulated to suit where multiple choice test items are likely to cater for any form of correct guessing respondent might make on the key option available among the distracters. Ojerinde, Popoola, Ojo and Ariyo (2014) affirmed that an indisputable fact of life as far as testing is concerned is that respondents will sometimes guess some items aright. This means that a small probability due to guessing is included.

The 1-, 2- and 3PL models across the globe have been used in diverse studies for several decades as supported by Steinberg and Thissen (1995); Lanza, Foster, Taylor and Burns (2005) and Adegoke (2013). Other evidence could be seen in the works of Amarnani (2009), Ojerinde, Popoola, Ojo and Oyenecho (2012), Ojerinde (2014) as well as Enu (2015), Metibemu (2016), Fakayode (2017) and Olonade (2017).

In spite of the improvement in the 3PL model, where the probability of guessing was taken care of, the model seems to be more advantageous to low-ability learners, who could guess difficult items correctly in the cause of testing. Meanwhile, a high-ability examinee is left at a disadvantage to any easy item he/she might incorrectly respond to or misses, even if such examinee possesses the commensurate ability to accurately respond to the item. The reason behind a high-ability examinee that slightly misses an item that seems easy could be attributed to some other indicators, that were not considered in 3PL model, but later incorporated in the formulation of 4PL model. With this, 3PL model was seen to be fraught with some estimation errors that might flout the accurate measurability of learners' true ability, if such errors are not reduced to the barest minimum.

The *Four-Parameter Logistic (4PL) model* was developed to enhance the quality of what 3PL model can estimate in terms of addressing some of the flaws that were associated with the model. The 4th parameter, carelessness (that could be prompted by either mistake, stress, tiredness, inattention, anxiety, unfamiliarity with computer techniques, distraction by poor testing conditions or misreading of questions) was added to 3PL model. Its item response function is given as:

$$P\,\mathrm{I}\,(\theta_s) = \Pr(X_{is} = 1 \,|\, \theta_s,\, a_i,\, b_i,\, c_i,\, d_i) = c_i + (d_i - c_i)\,\frac{1}{1+e^{-1.7a_i(\theta s - b_i)}} \text{..........eqn.3}$$

The additional d_i represents the carelessness parameter of item *i;* all other entities remain the same as in the previous models.

Meanwhile, the application of IRT models to response data as seen in literature locally reviewed (Nigeria) has been limited to the application of 3-, 2- and 1PL models of the dichotomously scored response-type category up till the time of this research. Few studies in other climes such as the USA and Europe have explored the usefulness of 4PL model and found better fits with improved estimates of both item and examinee' parameters. The works of Chang and Yin (2008), Rulison and Loken (2009), Loken and Rulison (2010) and Liao, Ho, Yen and Cheng (2012) found enhanced parameter estimates of both item and examinee estimates with improved overall fit when 4PL was applied in place of 3-, 2- or 1PL models when Bayesian method was adopted. Reise and Waller (2003) utilized responses from the Minnesota Multiphasic Personality Inventory (MMPI) to calibrate 2- and 3PL models in a psychopathology research. It was discovered that the overall fits of the models were not really improved and suggested that a higher parameterized model is better used in modeling certain medical and personality scales for a better estimation of respondents' ability. Other studies that advocated the potential use of 4PL model in practice are the works of Osgood, McMorris and Potenza (2002), Tavares, de Andrade and Pereira (2004) and Waller and Reise (2009).

However, the 4PL model suffered some initial setbacks, which accounted for its limited usage. Some of these include:

- suggestions for its application have been rather isolated such that there is no clear consensus on the need for the utility of the model (Barton & Lord, 1981; Hambleton & Swaminathan, 1985).
- the model has been traditionally proved difficult to estimate using maximum-likelihood (ML) estimation methods (Waller & Reise, 2009) which made the fitting of the 4PL model to be considered difficult and this created a concern that estimates of *the upper asymptote (d_j)* would not be reliable given the problems often encountered when estimating c_j using ML (Embretson & Reise, 2000; Baker & Kim, 2004).
- the strong dominance of the 3PL model in the literature and the lack of consensus on the usefulness of 4PL model as pointed out by Linacre (2004) and Loken and Rulison (2010).

Based on the foregoing, 4PL model was later reconsidered and its computational power and resources were improved upon by developing new sophisticated and accurate statistical modeling software that could use Bayesian approach to estimate its parameters. This was what constituted a major breakthrough towards its wider consideration and usage for practical purposes. This is with the aim of reducing to the barest minimum any estimation errors that could ensue from aberrant responses of a highly-able examinee (Liao *et al*., 2012). It thus appears that little or no effort(s) has been made in Nigeria towards assessing the extent to which estimation errors can be significantly reduced in terms of measurability of examinees' true ability (performance) when estimating with 4PL model.

Moreover, the usage of Computer-Based Testing (CBT) as one of the modes of test administration became a necessity. Almost all assessment bodies (private or public), academic institutions and professional bodies in Nigeria are working towards engaging e-Examination for the conduct of either testing or online-registration for their candidates/examinees. The adoption of CBT became necessary when it was realized that IRT approach provides some modern psychometric bases that are very necessary to the implementation of CBT (Ojerinde, 2016). One of the current interests in the adoption of CBT modeis its affordability to measuring some variables that were originally termed difficult to measure in the PPT mode (Schnipe & Scrams, 1999).

However, the fourth parameter in the 4PL model that is tagged carelessness would have been hard to measure without the help of computer that affords the opportunity to automatically record some indicators while examinee responds to test items. Mathematics as one of the secondary school subjects was used as the focus subject for the study where examinees' mathematics ability is estimated from their performances (responses from the computer-based mathematic achievement test (CBMAT)). These responses served as the empirical data that was used to investigate more utility of the highly parameterized 4PL model. In view of this, this study explored the applicability of 4PL model in the calibration of senior secondary school CBMAT in Lagos state. Two research questions were addressed.

1. Does the CBMAT response-data fulfill the assumption of trait dimensionality of the IRT framework?
2. Which of the four IRT models for dichotomous test best fit the CBMAT response data?

Methods

The study adopted a causal comparative research design of non-experimental type. The population used for the study comprised all Senior Secondary School two (SSII) mathematics students in Lagos State, which is stratified into 6 educational districts. Purposive sampling was used to select Education District 1 that consisted of 3 Local Government Areas (LGAs). From these LGAs, the researchers purposively selected 8 schools that had functional computer laboratories (3 schools from each of Agege and Ifako/Ijaye LGAs while 2 schools were picked from Alimosho LGA). A sample of 874 students was randomly chosen from these schools and data collection was made with the CBMAT instrument that consisted of a 40 multiple-choice items with four response options. CBMAT was answered on the computer system and automatic recording of examinee responseswas programmed such that respondent's click on any of the response options provided was recorded. Provision for automatic scoring on response made availed the system to dichotomously score 1 for correct response and 0 for incorrect response. Face and content validity were done by giving the CBMAT scale to experts with atable showing the test-blueprint of Bloom's taxonomy of educational objectives in the cognitive domain that was developed. Trial testing was carried out with sample of the same characteristics in Oyo State to establish instrument usability and appropriateness in eliciting information from the respondents. Meanwhile, full information item factor analysis (FIFA) of multidimensional item response theory (MIRT) package was used to ascertain the psychometric properties of the test and items of the scale. IRT empirical reliability value of 0.89 was generated. Data was analyzed with Stout's test of essential dimensionality of DIMTEST 2.0 software to determine the number of scale dimension and multidimensional item response theory (MIRT) package of the R environment via R-studio with the 1-, 2-, 3- and 4-parameter logistic IRT models for model-data fit assessment.

Results

Research question 1: Does the CBMAT response data fulfill the assumption of trait dimensionality of the IRT framework?

To answer this research question, IRT assumption of trait dimensionality was assessed on the CBMAT response data to know the most appropriate model-category that should be used in calibrating, either the unidimensional or multidimensional category. The CBMAT test data was subjected to Stout's test of essential unidimensionality which was implemented in DIMTEST 2.0 software. A null hypothesis (Ho) that there is no significant difference between the partitioning subtest (PT) and Assessment subtest (AT) of the examinees' responses was tested. Failure to reject Ho means that the test-data is unidimensional. However, if Ho is rejected, multidimensionality is evident. To perform the test, the DIMTEST software divided the items into two subtests that are as dimensionally distinct. Items under AT were those that formed the secondary dimension (indication of measurement of two or more constructs) as a result of the clustering procedure in DIMTEST while those under PT measured the primary construct (mathematics ability).

Table 1: Essential Dimensionality of the CBMAT Instrument

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

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

The finding (as seen in Table 1) showed that the abilities measured by the AT were not significantly different from those measured by the PT (T = -0.0628, p> 0.05). An indication that the null hypothesis was not rejected. It could therefore be concluded that the CBMAT response data fulfils unidimensionality assumption. This means that only one dominant construct (mathematical ability) accounted for the variation observed in the examinees.

Research question 2: Which of the four IRT models for dichotomous test best fit the Computer-Based Mathematics Achievement Test (CBMAT) response data?

Having established that the instrument is unidimensional, model-data fit analysis was carried out to see which of the 1-, 2-, 3- and 4PL models of the dichotomously scored response format best explained the CBMAT response data. Analysis was done through the MIRT package of the R foundation for Statistical Computing Platform via R-Studio environment. Several iterations were run in the analysis and convergence at four stages was noted. At the end of the process, comparison of the different values of -2loglikelihood, Akaike and Bayesian Information Criteria (AIC & BIC) were made.

Table 2: Model-data fit Assessment of the CBMAT Instrument

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

Table 2 presents the model-data fit assessment showing the model that produced the best fit when calibration was done with the CBMAT scale. The table showed that when the fitness of 1PL and 2PL models to the data was compared, result indicated that 2PL model had values (-2Loglikelihhod = 53041.62, AIC = 53045.62 & BIC = 53049.44) that were lesser than those of 1PL model (-2Loglikelihhod = 53590.34, AIC = 53592.34 and BIC = 53594.25). In search of a better fit to the response data, statistic values of the result at convergence for 3PL were in turn compared. The same decision also applies to the comparison of 2PL and 3PL models as well as that of 3PL and 4PL models consecutively. Statistic values for 4PL model appeared the least in all. findings eventually revealed that 4PL model gave the best fit to the CBMAT response data.

Discussion

The outcome of this study as far as research question one is concerned tells that the CBMAT used in eliciting responses from examinees exhibited a single dominant trait. This means that when analysis was done to check the number of dimension underlying the variance observed in examinees' responses to the test, only mathematics ability was seen to be dominant. Trait dimensionality assumption as pointed out by Suh (2016) and Umobong and Tommy (2017) were of the view that if dimensionality is not appropriately checked, inferences resulting from the estimates generated may be erroneous, thereby jeopardizing the potential advantage of IRT. The result of this study is in support of the findings of Ayanwale (2018) and Okwilagwe and Ogunrinde (2017) when IRT assumptions were checked on the Draft Multiple Choice Mathematic Test and NECO Geography Achievement Test respectively. But, was contrary to the result of Olonade, Metibemu and Adewale (2017), where 2014 WASSCE mathematics test was significantly multidimensional.

In addition, this study also discovered that 4PL model offered the best fit to the CBMAT response data. This supports the claim of Magis (2013) that the core advantage of 4PL model is its capability of allowing a non-zero probability of responding to an item incorrectly by highly able examinees. The 4PL model is capable of handling guessing error that 3PL model could as well take care of and also help to leverage any mistakes (due to stress for instance) high-ability students might incur in the cause of responding to test items. Loken and Rulison (2010) in a computerized adaptive testing (CAT) situation showed how the impact of early mistakes made by brilliant students was strongly reduced with the application of 4PL model. Other study that is in support of the usefulness of 4PL model include the work of Liao, Ho, Yen and Cheng (2012), who saw the performance of 4PL model as a robust mechanism when it was when examined and compared with 3PL model. Their finding indicated that 4PL model gave a more effective and strong estimation method than 3PL model.

Conclusion and Recommendations

The campaign for innovations (new development) in the way assessment is done and carried out is a pointer to having more approaches to objectively measure students' learning outcomes so as to depict the true ability of examinees in terms of their performances. IRT methods as become one of the modern-day approaches of assessing impartially because of the many flexible models its principles are associated with. This has projected its methods to be a more robust tool to make-use of by test-developers, psychometricians and researchers in combatting the very many challenges the present-day world is facing in the place of measurement and assessment. IRT model has the capability of describing and predicting respondents' performance on item. However, if examinee performance is to be accurately predicted, it means that the right model that best explains the response data should be employed.

It is therefore concluded that the CBMAT test was unidimensional and 4PL model appeared to have the best fit in place of the 3-, 2- and 1PL model Hence, the authors recommended that model-fit assessment is essential to present the most appropriate model needed to predict examinees' true ability. This means that inappropriate choice of model in scale calibration could cause estimation error on student true ability. Also, 4PL model that has been found to lessen estimation error should be further explored to establish more of its utility.

References

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

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks: *Journal of Education and Practice,* 4(2).

Adegoke, B. A. (2014). The role of item analysis in detecting and improving faulty physics objective test items: *Journal of Education and Practice,* 5(21).

Alordiah, C. O. (2015). A progressive step in educational measurement: An application of the Rasch model on mathematics achievement test. *Nigerian Journal of Educational Research and Evaluation,* 14(3).

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

Ariyo, A. O. (2015). An overview of classical test theory and item response theory in test development. *Nigerian Journal of Educational Research and Evaluation,* 14(3).

Ayanwale, M. A. (2019). Efficacy of item response theory in score ranking and concurrent validity of dichotomous and polytomous response mathematics achievement test in Nigeria. An Unpublished Ph.D Thesis. International Centre for Educational Evaluation (ICEE), Institute of Education, University of Ibadan.

Baker, F. B. (2001). The basics of item response theory. College Park, MD: ERIC Clearing House on Assessment and Evaluation. Original work published in 1985. Retrieved from http://echo.edres.org:8080/irt/baker/

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

Barton, M. A. & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model.* Princeton, NJ Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores.* 397-479. Reading, MA: Addison-Wesley.

Cees A. Glas, W. & Meijer, R. R.(2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement,* 27(3), 217–233.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press, New York. NY 10012.

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

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Enu, V. O. (2015). The use of item response theory in the validation and calibration of mathematics and geography items of joint command schools promotion examination in Nigeria. A Ph.D thesis. International Centre for Educational Evaluation (ICEE), Institute of Education. University of Ibadan.

Fakayode, O. T. (2018). Relative Effectiveness of CTT and IRT in equating WAEC Mathematics test scores for June and November 2015. A Ph.D thesis, International Centre for Educational Evaluation (ICEE), Institute of Education.University of Ibadan.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory, Principles and Application*. Norwell, MA Kluwer Academic Publishers.

Lanza, S. T. Foster, M. Taylor, T. K. & Burns, L. (2005).Assessing the impact of measurement specificity in a behaviour problems checklist: An IRT analysis. Technical Report 05-75. University Park, PA: The Pennsylvania State University, the methodology centre.

Liao, W. W. Ho, R. G. Yen, Y. C. & Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behaviour and Personality,* 40, 1679-1694.

Linacre, J. M. (2004). Discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions*, *18*, 959-960.

Loken, E. & Rulison, K. L. (2010). Estimation of a 4-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology, 63(3),* 509-525.

Magis, D. (2013). A Note on the Item Information Function of the Four-parameter model. *Applied Psychological Measurement,* 37(4), 04-15.

Metibemu, M. A. (2016). Comparison of classical test theory and item response theory frameworks in the development and equation of physics achievement tests in Ondo State, Nigeria**.** A Ph.D thesis, Institute of Education, University of Ibadan.

Nenty, H. J. (2004). From classical test theory (CTT) to item response theory (IRT): An introduction to a desirable transition. In O. A. Afemikhe and J. G. Adewale (Eds.), *Issues in educational measurement and evaluation in Nigeria (in honour of*'Wole Falayajo) (Chapter 33, pp.371 – 384).Yaoundé, Cameroon: Educational Assessment and Research Network in Africa.

Ojerinde, D. Popoola, K, Ojo, F. & Onyeneho, P. (2012). *Introduction to item response theory: parameter models, estimation and application.* Goshen Print Media Ltd.

Ojerinde D. (2013). Implementing and sustaining ICT-based assessment and evaluation in the Nigerian education system. National conference on ICT in education, National University Commission (NUC) Auditorium, Maitama, Abuja, 19th -20th November.

Ojerinde, D. Popoola, K, Ojo, F. & Ariyo, A. (2014). *Practical applications of item response theory in large-scale assessment.* Nigeria: Marvelous Mike Press Limited.

Ojerinde, D. (2016). The preface of vital issues in the introduction of computer-based testing in large-scale assessment. A compilation of papers presented at Local and international conferences. Joint Admission and Matriculation Board (JAMB). ISBN: 978-978-953-759-4.

Olonade, P. O. (2017). Equating 2014 senior school certificate mathematics examinations of West African Examinations Council and National Examinations Council in Lagos state, Nigeria. A Ph.D Thesis, Institute of Education, University of Ibadan.

Okwilagwe, E. A. & Ogunrinde, M. A. (2017). Assessment of unidimensionality and local independence of WAEC and NECO 2013 geography achievement tests. *African Journal of Theory and practice of Educational Assessment.* 5, 31-45.

Olonade, P. O. Metibemu, M. A. & Adewale, J. G. (2017). Unidimensional item response theory versus multidimensional item response theory: Evaluating the similarity of item calibration results in mathematics test in Lagos State. *African Journal of Theory and practice of Educational Assessment,* 5, 73-86.

Osgood, D. W. McMorris, B. J. & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance: Item response theory scaling. *Journal of QuantitativeCriminology,*18, 267-296.

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

Reeve, B. B. (2000). *Item and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory*. Unpublished Doctoral Dissertation, University of North Carolina at Chapel Hill.

Reise, S. P. & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods,* 8(2), 164–184.

Rulison, K. L. & Loken, E. (2009). I've fallen and I can't get up: Can high ability students recover from early mistakes in computer adaptive testing? *Applied Psychological Measurement,* 33, 83– 101.

Steinberg, L. Thissen, D. (1995). Item response theory in personality research. In Shrout, P. E., Fiske S. T. *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* 161–181. Hillsdale, NJ Erlbaum.

Suh, H. (2016). A study of Bayesian estimation and comparison of response time models in item response theory. Unpublished Ph.d thesis. Department of Psychology and Research in Education. University of Kansa, USA.

Tavares, H. R. Andrade de, D. F. & Pereira, C. A. (2004). Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology,* 27 679–685.

Troy-Gerard, .C. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics.* Unpublished Doctoral Dissertation, University Texas.

Umobong, M. E. & Tommy, U. E. (2017). Dimensionality of national examinations' council's biology examinations: Assessing test quality in modern trend approach. *African Journal of Theory and practice of Educational Assessment,* 5, 14-30.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory.* Springer, New York.

Waller, N. G. & Reise, S. P.(2009). Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI. In Embretson, S., Roberts J. S. *New directions in psychological measurement with model-based approaches.* Washington, DC. American Psychological Association.

Yen, T. Y. (1993). A comparison of three statistical procedure to identify clusters of items with local dependency. Huynh University of Carolina.