Isiugo-Abanihe, F. M., **Ugwoke, M.E**. & Iluobe, I.O. (2018). An Empirical investigation of goodness-of-fit of 1, 2, and 3-parameter logistic models to achievement test data in English Language

AN EMPIRICAL INVESTIGATION OF GOODNESS-OF-FIT OF 1, 2, AND 3-PARAMET ER LOGISTIC MODELS TO ACHIEVEMENT TEST DATA IN ENGLISH LANGUAGE

Isiugo-Abanihe, Ifeoma. Mercy, Ugwoke, Mark Eze and Iluobe, Ovekairi Irene

National Business and Technical Examinations Board (NABTEB), Benin City, Nigeria

ABSTRACT

This study investigated the goodness-of-fit of 1, 2 and 3-Parameter Logistic Models of Item Response Theory (IRT) to Achievement Test Data in English Language. The increased use of achievement tests in selection, promotion and awards of certificates inevitably brings attention to the quality and fairness of achievement testing. To adequately address these issues require sophisticated mathematical methods. The traditional Classical Test Theory (CTT) approaches that evaluated psychological measures at the total test scores have been complemented by more recent IRT approaches that focus on item level data. The ex-post facto research design was adopted for the study. The sample consisted of 3,000 examinees' responses which were randomly selected from Edo and Delta States of Nigeria in English Language Multiple-Choice Test Items conducted by the National Business and Technical Examinations Board (NABTEB), in 2014, 2015 and 2016. The test is valid and reliable because NABTEB conducts standardized tests. Three research questions guided the study and three hypotheses were tested. Data analysis was carried out using eirt -Item Response Theory Excel Assistance Version 3.1 Software. Pearson Chi–Square was used to test the hypotheses at 0.05 significant level. Findings from the study revealed that there was no significant difference among the 1, 2 and 3 -Parameter Logistic Models fit in 2014 and 2015 NABTEB English Language Multiple-Choice Test Items while there was a significant difference among the 1, 2 and 3 -Parameter Logistic Models in 2016. Based on the findings it was concluded that the 1, 2 and 3 -Parameter Logistic Models fits the data across the three years under study, none was empirically superior to others. It was recommended among others that examining bodies should make sure that selected models fits the data to be confident of results generated from such data.

KEYWORDS: Item Response Theory, 1, 2, 3 -Parameter Logistic Models, Model fit, Achievement Data

Background

Educational tests are main sources of information about students' achievement in schools and in activities. The analysis of test data is essential in determining the quality of the test and the information the test generates. The worth of any educational assessment depends on the instruments and techniques used, if the instruments are poorly designed, the results could be misleading. In educational measurement there are two main theories by which a test and the items it contains can be analyzed. These theories are: the Classical Test Theory (CTT) and Item Response Theory (IRT). The CTT explains the link among the observed score, the true score and error score. Within this theoretical framework, models of various forms have been formulated. The most common model is known as "Classical test model". It is a simple linear model linking the observable test score (X) to the sum of two unobservable (or often called latent variables), that is true score (T) and error score (E).

$$X = T + E$$

There are two unknowns in the equation (X and E) and this makes it not easily solvable unless some simplifying assumptions are made. The assumptions in the classical test model are : (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on parallel test are uncorrelated (Adegoke, 2015). The focus of CTT is on the test level information. Its item statistics (difficulty and discrimination) are often denoted by P and D respectively. These item statistics are important parts of the model and are used in item analysis and item selection in the development of achievement tests.

CTT has been in use for many decades to solve testing problems however, some shortcomings such as weak theoretical assumptions, sample dependent, etc. To meet up with this challenge, Nenty (1996) pointed that a good test theory which is capable of addressing some of, if not all the testing problems should be used by examining bodies in order to ensure quality assurance in educational assessment and certification.

The shortcomings of CTT gave birth to the evolution of IRT which was first proposed in the field of psychometrics for the purpose of ability assessment. Its primary

concern is on the item – level information in contrast to the CTT's primary focus on test – level information. IRT is used for the design, analysis, scoring and comparison of tests and similar instruments whose purpose is to measure unobservable characteristics of the respondents (Stata Corp, 2016). It is concerned with accurate test scoring and development of test items. Test items are designed to measure various kinds of abilities, traits or behavioural characteristics. Responses to the items can be binary (such as correct or incorrect responses), ordinal (such as degree of agreement on Likert scales) and partial credit (such as essay test). IRT is widely used in education to calibrate and evaluate items in tests, questionnaires, and other instruments and to score subjects on their abilities, attitudes, or other latent traits. During the last decades, educational assessment has used more and more techniques to develop tests. Today all major educational tests, such as the Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE), are developed by using IRT, because the methodology can significantly improve measurement accuracy and reliability while providing potentially significant reductions in assessment time and effort, especially via computerized adaptive testing. In recent years, IRT-based models have also become increasingly popular in health outcome, quality- of-life research and clinical research (Hays, Morales, & Reise 2000; Edelen & Reeve 2007; Holman, Glas & Haan 2003 and Reise & Waller 2009).

Item Response Theory (IRT) is a measurement theory and its focus is on the item level rather than on the total test score. In IRT framework, parameters are classified into two basic components: first is related to the examinee's ability (latent trait), second to the task (test). The assumption is that, each examinee responding to a test item possesses some amount of underlying ability. Thus, one can consider each examinee to have a numerical value, a score which places him or her somewhere on the ability scale. This ability scale is called latent trait and it is denoted by theta the Greek letter (ø). At each ability level, there is a certain probability that an examinee with that ability, will give a correct answer to the item. Under IRT, P(ø) is used to represent this probability. The case of a typical test item, this probability will be low for examinees of low ability and high for examinees of high ability.

IRT is a modelling technique that tries to describe the relationship between an examinee's test performance and the latent trait underlying the performance (Hambleton & Jones, 1993).

IRT models describe the interactions of persons and test items (Reckarse, 2009). Hence, IRT is a general framework for specifying mathematical functions that characterize the relationship between a person responses to the separate items in the instrument (DeMars, 2010). The most widely used traditional IRT models are the One Parameter Logistic Model (1PLM), Two Parameter Logistic Model (2PLM), and Three Parameter Logistic Model (3PLM). The 1PLM utilizes a single item difficulty parameter. The 2PLM incorporates an item discrimination parameter as well as an item difficulty parameter and the 3PLM utilizes an item difficulty, item discrimination and pseudo-guessing parameter (Lord, & Kelkar cited in Chon, Lee & Ansley, 2007). The proposed Four - Parameter Logistic Model (4PLM) which incorporates response time and slowness parameter (Wang and Hanson, 2001) has not been formally incorporated into the traditional IRT models. Moreover, softwares for analysing it is yet readily available. (Hambleton and Swaminathan, cited in Chon, Lee and Ansley, 2007) suggest that model – data – fit improves with the inclusion of each additional model parameter.

Model - data - fit is regarded as a useful checking tool in model selection for a particular data set. When various models and calibration procedures are available the question that will arise is which one to choose? One way to assess the appropriateness of the chosen IRT model(s) and calibration procedure is to conduct an analysis of model – data – fit. Several studies have examined model- data – fit utilizing 1, 2 and 3PLM under different conditions with PARSCALE (Chon, Lee & Ansley, 2007).

External agencies (examining bodies) like the West African Examinations Council (WAEC), the National Examinations Council (NECO), and the National Business and Technical Examinations Board(NABTEB) were established to conduct examinations for both in- school candidates and out- of – school candidates and award certificates to successful candidates. The National Business and Technical Examinations Board (NABTEB) is one of the foremost examining body in Nigeria charged with the responsibility of

conducting valid and reliable examinations leading to the awards of certificates that are recognized locally and internationally. The National Technical Certificate (NTC) / National Business Certificate (NBC) examinations have three components:

- Trade related;
- Trade group;
- General education, where English Language is one of the general education subjects examined by NABTEB.

English language is a core subject offered at the post basic level (secondary level) in Nigeria. It is one of the general education subjects in which hundreds of candidates are tested by NABTEB during her May /June and November /December examination series. It is one of the compulsory general education subjects. A credit pass is required in English Language before a candidate is certified by NABTEB and it is also a prerequisite for admission into tertiary institutions.

Objectives

The objectives of this study were to:

- Fit the 1, 2 and 3 Parameter Logistic Models to 2014 NABTEB English Language Multiple – Choice Test Items.
- Fit the 1, 2 and 3 Parameter Logistic Models to 2015 NABTEB English Language Multiple – Choice Test Items.
- Fit the 1, 2 and 3 Parameter Logistic Models to 2016 NABTEB English Language Multiple – Choice Test Items.
- To establish which logistic model could be preferred empirically to others.

Statement of the problem

The need to ensure that scores are accurate reflection of students' knowledge and the content being measured has been an unending search in the field of psychological testing. Examining bodies in Nigeria have been doing this by conducting item analysis and establishing test psychometrics by using Classical Test Theory (CTT) approach. In recent years, attention of Psychometricians in these examining bodies has been focused on Item

Response Theory (IRT). Given the importance of IRT models and the emphasis placed on the good – model- data – fit in IRT application, it is logical to expect that misfit between an IRT model and empirical data may potentially threaten the ability- parameter estimates and invariant property of IRT model parameters. The advantages claimed for item response model can be realized if only the fit between the model and the test data set of interest is satisfactory. In IRT, there are three logistic models commonly used. These are 1, 2, and 3 PLM. The question that has to be answered is whether in the invent that examining bodies decide to adopt the IRT procedure for item analysis, then a problem may arise as to which of the logistic models should be used? A poorly fitting model could be misleading and cannot yield invariant item and ability parameter estimates. From literature, there appears to be a research vacuum in model – data – fit. Also, studies carried out on NABTEB English Language Multiple – Choice items are scarce. To this extent, there is no empirical evidence on the superiority of any model. This therefore, is the trust of this study. To guide this study, the following research questions were postulated:

1. Are there differences among the 1, 2 and 3 parameter logistic models fits in the scores for 2014, NABTEB English Language Multiple - Choice test?
2. Are there differences among the 1, 2 and 3 parameter logistic models fits in the scores for 2015, NABTEB English Language Multiple - Choice test?
3. Are there differences among the 1, 2 and 3 parameter logistic models fits in the scores for 2015, NABTEB English Language Multiple - Choice test?

Hypotheses

The following hypotheses were tested at 0.05 level of significance:

1. There is no significant difference among the 1, 2 and 3 parameter logistic models fits in the scores for 2014, NABTEB English Language Multiple - Choice test.
2. There is no significant difference among the 1, 2 and 3 parameter logistic models fits in the scores for 2015, NABTEB English Language Multiple - Choice test items.
3. There is no significant difference among the 1, 2 and 3 parameter logistic models fits in the scores for 2016, NABTEB English Language Multiple - Choice test items.

Review of Related Literature

This study is hinged on Item Response Theory (IRT). IRT is credited to Fredrick Lord. The theory models the relationship between the responses of each examinee of a given ability of each item in the test, (Lord, cited in Amasingha, 2015). The main idea of item response theory is that of the item response model that is, a mathematical function describing the probability of specified responses to an item, given some level of quantitative attributes of the respondent. This is explained by Item Characteristic Curve (ICC) which scales items and people onto a common metric, helps in standard setting serves as foundation of equating and makes meaning in terms of student ability.

ICC is illustrated by a line in a Cartesian system called Ogive which is defined by a logistic function shown below:

$$P_{ij}(1)\,|\theta,b = \frac{\text{Exp}(\theta_j - b_i)}{1+\exp(\theta_j - b_i)}$$

Where

b is the item parameter, and

Θis the person parameter

The equation represents the probability of responding correctly to item *i* given the ability of person *j* while figure 1 below represents ICC which shows the behaviour of a good item

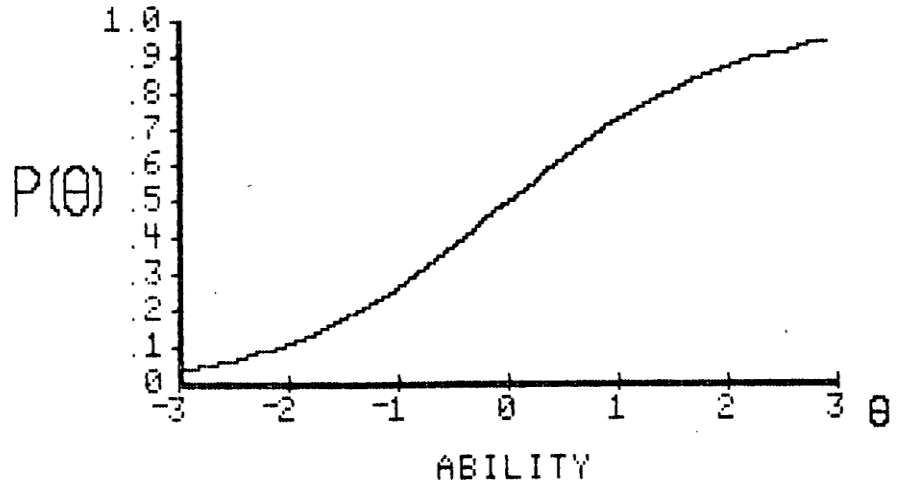

Figure 1: Item Characteristic Curve (ICC) (Source: Baker, 2001)

The item characteristic curve (ICC) is the basic building block of Item Response Theory; all the other constructs of the theory depend upon this curve (Baker, 2001). The main concept of IRT is the ICC. The ICC describes the probability that a person "succeeds" on a given item that is individual test question (Stata Corp, 2016). The vertical axis represents the probability (.0 to 1.0) of responding correctly to the item while the horizontal axis represents the latent trait/Ability (-3 to 3) of the respondents.

IRT is a set of models which, by relating the likelihood of a particular reaction by an individual with a given trait level to the characteristics of the item designed to elicit the level to which the individual possesses that trait, attempts to estimate the parameters involved, explains the process and predicts the results of such an encounter (Nenty,2004). Models in IRT mathematically define the probabilistic relationship between individuals' observed responses to a series of items and their location on the unobservable latent variable continua reflecting the constructs being measured (De Ayala, 2009; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Reckase, 2009). IRT has models for both dichotomously scored items (e.g., true/false), and polytomously scored questions (e.g., 5 category Likert-type scale). IRT item parameters are set to relate responses to the underlying trait (Embretson & Reise, 2000), thus, IRT can easily model the mixed item formats included in many surveys. Joshua (2014) states that IRT and the presentations of it in models predispose IRT to many applications in practical testing situations. Such applications include test construction (item selection), item +banking, test equating, adaptive testing, and study of item bias, to mention but a few.

IRT is used by researchers to analyze student's performance data from one testing situation, describe it succinctly, and are able to make predictions about item and test performance in other situations. IRT has three basic assumptions. These are monotonicity, local independence and unidimensionality. These three assumptions are very important and hold irrespective of the latent model used. According to Ojerinde, Popoola, Ojo and

Onyeneho (2012) a test data can only be useful for a latent trait model if these assumptions are met. The application of goodness of fit (GoF) methods in IRT framework, informs us of the discrepancy between the model and the data being fitted (Maydeu, 2010). The Goodness of Fit (GoF) is a statistical model that describes how well it fits into a set of observations. GoF indices summarize the discrepancy between the observed values and the values expected under a statistical model. Assessing absolute model fit (that is, the discrepancy between a model and the data) is critical in application, as inferences drawn on poorly fitting models may be misleading badly (Maydeu, 2010). Researchers are also interested in a relative model fit (that is, discrepancy between two models) when more than one substantive model is under consideration (Yuan & Bentler, 2004; Maydeu – Olivares & Cai, 2006).

Chom, Lee and Arisley (2007) Studied Assessing IRT model-data fit for mixed format tests. They examined various model combinations and calibration procedures for mixed format tests under different item response theory (IRT) models and calibration methods. The data used was data sets that consist of both dichotomous and polytomous items, nine possibly applicable IRT model mixtures and two calibration procedure were compared based on traditional and alternative goodness-of-fit statistics. Three dichotomous models and three polytomous models were combined to analyze mixed format test using both simultaneous and separate calibration methods. The PARSCALE's G^2 was used to assess goodness of fit. The findings revealed that the 1PLM had the largest misfit in items. Si (2002) carried out a study in ability estimation under different item parameterization and scoring models. The study employed a 7 x 4 x 3 factorial design. Seven models 1-2-3-PL dichotomous logistic model, the Generalized Partial Credit Model with item discrimination a_i set to a constant (GPCM-1), the Generalized Partial Credit Model (GPCM), the Multiple Choice Model (MCM), and the Nominal Categories Model (NCM), were compared. A set of polytomous item responses of 1,000 subjects to 30 items was simulated using a computer program (Monte Carlo Estimation). The result revealed that the 1-PL model (with only item difficulty) had the most accurate ability estimation, and the 3-PL model (with three types of

parameters) were less accurate in ability estimation among the three different types of item parameterization models.

Kose (2014) carried out a study in assessing model data fit of unidimensional item response theory models in simulated data. Responses of 1000 examinees to a dichotomously scoring 20 item test were simulated with 25 replications. Also, data were simulated to fit 2PL model. 4 – Step procedure was used for model – data fit and BILOG software was used. The result revealed that the 2PL model fits significantly better than the Rasch model. Similarly, the difference between the 3PLmodel and 2PLmodel was evaluated and the result showed that the 2PL model fits significantly better than the 3PL model.

Methodology

Ex - post- facto research design was adopted for this study. The population of this study consists of four thousand two hundred and fifty three (4,253) students' responses in 2014, 2015 and 2016 May/June NBC/NTC Examinations in English Language in Edo and Delta states, in Nigeria. The statistical population of items for this study is three hundred (300) items (100 each for 2014, 2015 and 2016) the three years under study.

Table 1: Population Distribution of Candidates in Edo and Delta that sat for NABTEB May/June NBC/NTC English Language in 2014, 2015 and 2016 Examinations

YEAR	STATE				TOTAL
	DELTA		EDO		
	Male	Female	Male	Female	
2014	340	100	800	200	1440
2015	398	200	758	200	1556
2016	253	100	704	200	1257
TOTAL	911	400	2262	600	4253

Multi – stage sampling technique was employed for this study.

Stage 1: the schools were stratified into private and public schools.

Stage 2: random sampling was employed to select ten (10) schools, five (5) private and five(5) public schools.

Stage 3: in each school one hundred (100) students were randomly selected in each year. The total number of participants in this study is three thousand (3,000) students made up of two

thousand (2000) males and one thousand (1000) females. The male students enrolls more than the female students in NABTEB examinations hence unequal sample between the male and female responses.

The instrument that was used for this study is the NBC/NTC English Language Multiple Choice Test Items question paper for 2014, 2015 and 2016 May/June Examinations conducted by National Business and Technical Examinations Board (NABTEB). The instrument consists of one hundred (100) items with four (4) options each lettered A-D. The candidates' were required to select from these options one correct answer. The responses to each item for all the students in the schools selected for the three years were obtained from the board. The validity and reliability of the instrument have been determined by the board because it is a standardized test. The researchers prepared a person by item matrix with the horizontal axis showing number of items and the vertical axis showing the number of persons and the cells indicating the responses of each examinee. The examinees responses were analyzed using IRT statistical software: eirt - Item Response Theory Assistant for Excel (Germain, Valois & Abdous, 2007), for the test items calibration to determine item parameters based on IRT framework. The output included: Item Parameter Estimates; ability estimates, test of fit, local independence and Item characteristics curves.

Findings

Research question one: Are there differences among the 1, 2 and 3 parameter logistic models fits in the scores for 2014, NABTEB English Language Multiple - Choice test?

Table 2: The Fit of Logistic Models in 2014, NABTEB English Language Multiple - Choice Test

PARAMETER LOGISTIC MODELS		ITEM FIT MISFIT ITEMS	FIT ITEMS	Total
1PLM	Count	34	66	100
	Expected Count	31.3	68.7	100.0
2PLM	Count	32	68	100
	Expected Count	31.3	68.7	100.0
3PLM	Count	28	72	100
	Expected Count	31.3	68.7	100.0
Total	Count	94	206	300
	Expected Count	94.0	206.0	300.0

Table 2 shows the result of the chi- square goodness of fit analysis for the NABTEB certificate examinations. It can be deduced that 34 items representing 34% misfit the One Parameter logistic Model (1PLM), while 66 items representing 66% fit the 1PLM. 32 items representing 32% misfit the Two Parameter Logistic Model (2PLM), while 68 items representing 68% fit the 2PLM. Also, 28 items representing 28% misfit the Three Logistic Model (3PLM), while 72 items representing 72% fits the 3PLM. From the result the 1PLM, 2PLM and 3PLM fitted 2014, NABTEB English Language Multiple - Choice test.

Research Question two: Which of the models 1, 2 and 3 parameter logistic models fits the scores for 2015, NABTEB English Language Multiple - Choice test?

Table 3: The Fit of Logistic Models in2015, NABTEB English Language Multiple - Choice Test

PARAMETER LOGISTIC MODELS		ITEM FITS		Total
		MISFIT ITEMS	FIT ITEMS	
1PLM	Count	31	69	100
	Expected Count	35.7	64.3	100.0
2PLM	Count	34	66	100
	Expected Count	35.7	64.3	100.0
3PLM	Count	42	58	100
	Expected Count	35.7	64.3	100.0
Total	Count	107	193	300
	Expected Count	107.0	193.0	300.0

Table 3 shows the result of the chi- square goodness of fit analysis for the NABTEB certificate examinations. It can be deduced that 31items representing 31% misfit the One Parameter logistic Model (1PLM), while 69 items representing 69% fit the 1PLM. 34 items representing 34% misfit the Two Parameter Logistic Model (2PLM), while 66 items representing 66% fit the 2PLM. Also, 42 items representing 42% misfit the Three Logistic Model (3PLM), while 58 items representing 58% fits the 3PLM. From the result, the 1PLM, 2PLM and 3PLM fitted the 2015, NABTEB English Language Multiple - Choice test.

Research Question three: Which of the models 1, 2 and 3 parameter logistic models fits the scores for 2016, NABTEB English Language Multiple - Choice Test?

Table 4: The fit of Logistic Models in 2016, NABTEB English Language Multiple -Choice Test

PARAMETER LOGISTIC MODELS		ITEM FITS		Total
		MISFIT ITEMS	FIT ITEMS	
1PLM	Count	20	80	100
	Expected Count	28.0	72.0	100.0
2PLM	Count	24	76	100
	Expected Count	28.0	72.0	100.0
3PLM	Count	40	60	100
	Expected Count	28.0	72.0	100.0
Total	Count	84	216	300
	Expected Count	84.0	216.0	300.0

Table 4 shows the result of the chi- square goodness of fit analysis for the NABTEB certificate examinations. It can be deduced that 20 items representing 20% misfit the One Parameter logistic Model (1PLM), while 80 items representing 80% fit the 1PLM. 24 items representing 24% misfit the Two Parameter Logistic Model (2PLM), while 76 items representing 76% fit the 2PLM. Also, 40 items representing 40% misfit the Three Logistic Model (3PLM), while 60 items representing 60% fit the 3PLM. From the result, the 1PLM, 2PLM and 3PLM fitted the 2016 NABTEB English Language Multiple - Choice test

Hypothesis one: There is no significant difference among the 1, 2 and 3 parameter logistic models fits in 2014, NABTEB English Language Multiple - Choice test

Table 5: Fit of 1, 2 and 3 Parameter Logistic Models in 2014 NABTEB English Language Multiple - Choice test

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.868[a]	2	.648
Likelihood Ratio	.873	2	.646
Linear-by-Linear Association	.834	1	.361
N of Valid Cases	300		

Significant at 0.05 level

Table 5 shows Chi-Square Tests carried out on Fit of 1, 2 and 3 Parameter Logistic Models in 2014 NABTEB English Language Multiple - Choice test. Pearson Chi-Square

depicts an F-ratio of .868 df 2 which is significant at p-value = .648. Comparing the p-value with the alpha level of .05, the p-value is greater than the alpha level of .05;therefore, the null hypothesis that says, “There is no significant difference among the 1, 2 and 3 parameter logistic models fits in 2014 NABTEB English Language Multiple - Choice test” is retained.

Hypothesis two: There is no significant difference among the 1, 2 and 3 parameter logistic models fits in the scores for 2015, NABTEB English Language Multiple - Choice test.

Table 6: Fit of 1, 2 and 3 Parameter Logistic Models in 2015 NABTEB English Language Multiple - Choice test

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.818[a]	2	.244
Likelihood Ratio	2.800	2	.247
Linear-by-Linear Association	2.628	1	.105
N of Valid Cases	300		

Significant at 0.05 level

Table 6 shows Chi-Square Tests carried out on Fit of 1, 2 and 3 Parameter Logistic Models in 2015 NABTEB English Language Multiple - Choice test. Pearson Chi-Square depicts an F-ratio of 2.818 df 2 which is significant at p-value = .244. Comparing the p-value with the alpha level of .05, the p-value is greater than the alpha level of .05; therefore, the null hypothesis that says, “There is no significant difference among the 1, 2 and 3 parameter logistic models fits in 2015 NABTEB English Language Multiple - Choice test” is retained.

Hypothesis three: There is no significant difference among the 1, 2 and 3 parameter logistic models fits in 2016 NABTEB English Language Multiple - Choice test

Table 7: Model Fit of 1, 2 and 3 Parameter Logistic Models in 2016 NABTEB English Language Multiple - Choice test

	Value	**Df**	**Asymp. Sig. (2-sided)**
Pearson Chi-Square	11.111[a]	2	.004
Likelihood Ratio	10.873	2	.004
Linear-by-Linear Association	9.888	1	.002
N of Valid Cases	300		

Significant at 0.05 level

Table 7 shows Chi-Square Tests carried out on Fit of 1, 2 and 3 Parameter Logistic Models in 2016 NABTEB English Language Multiple - Choice test. Pearson Chi-Square depicts an F-ratio of 11.111 df 2 which is significant at p-value = .004. Comparing the p-value with the alpha level of .05, the p-value is less than the alpha level of .05; therefore, the null hypothesis that says, "There is no significant difference among the 1, 2 and 3 parameter logistic models fits in 2016 NABTEB English Language Multiple - Choice test" is rejected. Thus, there exist a significant difference among 1, 2 and 3 parameter logistic models fits in 2016 NABTEB English Language Multiple - Choice test.

Discussion

Results of this study in tables 2, 3, and 4 revealed that the 1, 2, and 3 PLM fitted the 2014, 2015 and 2016 NABTEB English Language multiple – choice test. This result is not in agreement with Chom, Lee and Arisley (2007) who claimed that the 1PLM had the largest misfit in items.

Hypothesis one revealed that there is no significant difference among the 1, 2 and 3 parameter logistic models fits in the scores for 2014, NABTEB English Language Multiple - Choice test. This implies that 1, 2 and 3 PLM fitted the 2014 data. The findings are in disagreement with Chom, Lee and Arisley (2007) who claimed that the 1PLM had the largest misfit in items.

The second hypothesis revealed that there is no significant difference among the 1, 2 and 3 parameter logistic models fits in 2015 NABTEB English Language Multiple - Choice test," Hence the hypothesis was retained. This shows that the 1, 2, and 3 PLM fits the 2015 data. The finding is in disagreement with Chom, Lee and Arisley (2007), who claimed that the 1PLM had the largest misfit in items. Also, this finding is not in agreement with Si, (2002), Kose (2014) who claimed that the 1PLM and 2PLM respectively is superior to others.

The third hypothesis was tested with Pearson Chi –Square test. The likelihood ratio value fit evidence obtained was significant hence the hypothesis that says "there is no significant difference between the 1, 2 and 3 parameter logistic models fits in 2016,

NABTEB English Language Multiple - Choice test," was rejected. This shows that there exists a significant difference among 1, 2, and 3PLM fit in 2016 data. That is, though they all fit the data, the 1PLM and 2PLM has few misfit items compared to the 3PLM. The result is in concord with the findings of Si, (2002), Kose (2014) who claimed that the 1PLM and 2PLM respectively is superior to others.

Conclusion

Based on the findings, the researchers therefore concluded that the 1, 2 and 3 Parameter Logistic Models fitted the 2014, 2015 and 2016 NABTEB English Language Multiple - Choice test, therefore, none is empirically superior to others.

Recommendations

Based on the findings and the conclusion reached from this study, the researchers therefore recommends as follows:

1. The examining bodies should make sure that models used for selection of items fits the data. Otherwise, such results may be spurious and misleading.
2. Examining bodies should embrace IRT in item generation, assessment of candidates and analysis of results. Since it is the globally preferred method of test construction and analysis of results.
3. Examining bodies should engage the services of measurement experts who are proficient in IRT since IRT is informative.
4. There is the need for examining bodies to attend national workshops regularly to keep them abreast of advantages of IRT over CTT and to key into its use.

References

Adegoke, B.A. (2015). Comparison of psychometric properties and examinees' scores in two forms of physics objective tests of West African examination council. Paper presented at 17th annual national conference of the association of educational researchers and evaluation of Nigerian

Amasingha, K. J. (2015). Item response theory: The way forward to objectivity in educational measurement in the school system. *Nigerian Journal of Educational Research and Evaluation*. Vol. 14, 2, 17.

Baker, F. B. (2001).*The basics of item response theory* (2nd ed). ERIC Clearinghouse on Assessment and Evaluation.

Chon, K., Lee, W. & Ansley, T, (2007). Assessing IRT Model – Data Fit for Mixed Format Tests. Center for Advanced Studies in Measurement and Assessment (CASMA) Research Report Number 26.

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

De Mars, C. (2010). *Item Response Theory*. Understanding Statistics Measurement City: Oxford University Press.

Edelen, M.O. & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation and refinement. Quality of life Research, 6, 5-18.

Embretson, S.E., & Reise, S.P. (2000).Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Germain, S., Valois, P., & Abdous, B. (2007). eirt - Item Response Theory Assistant for Excel*(Freeware)*. Available online at: http://libirt.sf.net

Hambleton, R. K. & Jones, R.W.(1993). Comparison of classical test theory and item response theory and their applications to test development. An NCME Instructional Module16, Fall 1993. Retrieved on 08/07/04 fromhttp://www.ncme.org/ pubs/items/24.pdf.

Hambleton, R.K., Swaminthan, H., & Rogers, H.J. (1991). Fundamental of item response theory. Newbury Park, CA: Sage Publications, Inc.

Hays, R.D., Morals, L. S. & Reise, S.P. (2000). Item response theory and health outcomes measurement in the 21st century.

Holman, R., Glas, C. and Haan, R.J. (2003).Power analysis in randomized clinical trials. CONTROL CLIN TRIALS, 24(4): 390-410 (PubMed).

Joshua, M. T. (2014). Applications of Item Response Theory (IRT) In Testing &Associated Challenges (In Africa). Paper presented at IAEA Scholarship Training on Educational and Evaluation. Abuja, Nigeria

Kose, I.A. (2014). Assessing Model data fit of unidimensional item response theory model in simulated data academic journals 9(17), pp 642-649 retrieved from http://www.academicjournals. org/ERR on 5/11/2016.

Maydeu-Olivares, A. (2010). Goodness of fit Testing: International Encyclopedia of Education, 7:190-196

Meadeu-Olivares, A & Gai, L. (2006). A cautionary note on using G^2 (d_1f) to assess relative model fit in categorical data analysis. Multivariate Behavioural Research 41, 55-64.

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) (33), 371 – 384. Ibadan, Nigeria: Institute of Education, University of Ibadan.

Nenty, H. J. 1991. Transformation of test scores, why and how? In H.J. Nenty (Ed.), Fundamentals of Continuous Assessment (Unit 8). Executive Publishers, Owerri.

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

Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York, U.S.A: Springer- Verlag.

Reise, S.P., & Waller, N.G. (2009). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, (14), 45-58.

Si, C.B. (2002). Ability estimation under different item parameterization and scoring models. Unpublished Dissertation University of North Texas.

StataCorp (2016). Stata: Release 14. Statistical software. College station, TX: StataCorp LP.

Wang, T., & Hanson, A. (2001). Development and an item response model that incorporates response time. A paper presented to the Annual meeting of the American Education Research Association in Settle, April.

Yuan, K.H and Bentler, P.M. (2004).On chi-square difference and Z tests in mean and covariance structure analysis when the base model is MI specified. *Educational and psychological measurement* (64) 737-757