ASSESSMENT OF MODEL FIT FOR 2016 AND 2017 BIOLOGY MULTIPLE CHOICE TEST ITEMS OF THE NATIONAL BUSINESS AND TECHNICAL EXAMINATION BOARD

Judith Hannah OSARUMWENSE (Ph.D)
Department of Educational Evaluation and Counselling Psychology
Faculty of Education
University of Benin
Benin City.

Chisom Perpetual DURU
Master's Student
Department of Educational Evaluation and Counselling Psychology
Faculty of Education
University of Benin
Benin City.

Abstract

This study was based on the assessment of model fit for 2016 and 2017 Biology multiple choice test items of the National Business and Technical Examination Board. It aimed at empirically investigating the model fit of the 1, 2, and 3 Parameter Logistic Models (PLM) of the examinations using Item Response Theory. Three research questions were raised with two hypotheses formulated and tested. The expo-facto research design was adopted for this study. The population for the study was 5,115 and 4600 candidates in public and private schools in south-south geo-political zone in Nigeria for 2016 and 2017 respectively. A total of 2000 students were sampled using Simple random sampling technique. The instruments for data collection was the NABTEB 2016 and 2017 Biology multiple choice question papers. The instruments are said to be valid and reliable as they were developed by a standard examination body. The responses from the instruments were used for data analysis. The results obtained from the study revealed that the 1, 2 and 3 PLM fit the 2017 and 2016 NABTEB May/June Biology multiple choice test items. However, the 1PLM provided a better fit to the data than other models. Based on the findings of the study, it was recommended among others that the examining bodies should make sure that model fit the data well before they are used to make inferences regarding the data.

Key words; Item Response Theory, Biology multiple choice test items, model fit and unidimensionality

Background to the Study

Researchers in the field of educational assessment are continually developing new approaches to improve efficiency of assessments. They are often concerned with methodologies that can extract the most useful and accurate information from students' responses to test items. Psychometric theory is the statistical framework for measurement in many fields of psychology and education. These measurements may concern abilities, personality traits, attitudes, opinions, and achievement.

Item Response Theory (IRT) models play a prominent role in psychometric theory. In these models, the properties of a measurement instrument are completely described in terms of the properties of the items, and the responses are modeled as functions of item and person parameters. While many of the technical challenges that arise when applying IRT models have been resolved (example, Model Parameter Estimation), the assessment of model fit remains a major hurdle for effective IRT model implementation (Hambleton & Han, 2005).

The assessment of the IRT model fit to item response data is one of the crucial steps before an IRT model can be applied with confidence to estimate proficiency or ability levels of examinee (Stone & Zhang, 2003). The assessment of fit of IRT models usually involves the collection of a wide variety of diagnostic evidences for model fit and then making an informed judgment about model fit and usefulness of a model with a particular set of data (Hambleton & Han, 2005). Model data misfit can be attributed to violation of model assumptions or the specific parameterization for the IRT model (number of parameters). That is, exclusion of relevant item or ability parameters may influence the appropriateness of IRT model.

The model should be able to explain aspects of the data that influences the inferences made using the IRT model otherwise, the conclusions obtained using the model might not be relevant. IRT models are based on explicit assumptions; therefore, the methods for evaluation of model fit focus on these assumptions. IRT models are based on a number of explicit assumptions, so the method for the evaluation of model fit focus on these assumptions. The most important assumptions underlying these models are the unidimensionality of the items, the form of the Item Characteristic Curve (ICC), local stochastic independence, and item score pattern. Researchers have proposed significant number of fit statistics for assessing fit of IRT models. These statistics were developed to be sensitive to specific model violations (Jansen & Glas, 2005; Glas &Suárez-Falcón, 2003; Maydeu-Olivares & Joe, 2005).

Once an IRT has been applied to a set of data, its appropriateness should be investigated with data-model fit analysis otherwise, the researcher is under the risk of drawing incorrect conclusion regarding the scientific problem of interest. Substantial lack of fit should result in the replacement or extension of the model if possible (Sinharay, 2005). Traditional methods are most widely used to assess model fit; especially, the likelihood ratio, chi square goodness of fit statistics and these are provided in the most popular current software packages, such as BILOG, BILOG-MG and PARSCALE (Zhao, 2008). The most common criticism about the chi-square statistics is that they are sensitive to sample size (Hambleton & Swaminathan, 1985). An essential feature of these statistics is that they are based on information that is aggregated over persons; therefore they will be referred to as aggregate test statistics. Hence, to measure the underlying traits or abilities of examinees such as intelligence, mathematical prowess or stochastic aptitudes, these characteristics cannot be quantified directly as one would measure height or eye colour.

Instruments in the form of examination or questionnaire are commonly used to assess the desired latent variables.

There have been arguments among researchers about the appropriateness of some models among the dichotomous models of IRT to be used in assessing multiple choice examinations conducted by examining bodies. However, IRT model fit studies have not received the attention they deserve among test practitioners. Possible reasons for this neglect are the complexity of assessing fit, the lack of understanding of the fit statistics and the absence of comprehensive model fit software (Zhao, 2008). An important part of any modeling process is assessing or checking the fit of a model before using the model to make inferences regarding the data. This is because, it is important to establish that the model fits the data well enough due to some criteria. Researches in assessment of model fit have been carried out across different countries. For example, Kose (2014) carried out the study on assessing model fit of unidimensional Item Response Theory model in simulated data ,using the IRT software (BILOG) for the assessment of the fit of the model through the analyses it was revealed that the 2PL IRT model fits significantly better than the 3PL IRT model. Adedoyin and Mokobi (2015) carried out a study using IRT Psychometric analysis in examining the quality of Botswana 2010 Junior Certificate Mathematics multiple choice examinations in Botswana, using dichotomous IRT models with IRT software – MULTILOG 3.0 and BILOG MG through the analyses, it was found that out of the twenty three (23) item that fitted the 3PL IRT model, twelve (12) items were classified as poor test items, ten (10) items were classified as fairly good test items which could be revised or improved on and one (1) item was considered to be good test item.

Kyong, Won, and Timothy (2007) carried out a study on the assessment of IRT model data fit for mixed item format test of the IOWA tests of basic skills of IOWA City, U.S.A. They examined the various model combination and calibration procedures for mixed item format test (multiple choice items and constructed response test items) under different IRT models (dichotomous and polytomous models), the analyses revealed that the 3PLM combined with the generalized partial credit model among various IRT model combinations led to the best fit to the given data sets. Psychometric Quality of the Common Educational Proficiency Assessment (CEPA) – English test in school of Pittsburgh in the United Arab Emirates was investigated by Salma (2009) with dichotomous models of IRT, using the IRT software BILOG-MG and MULTILOG, the analyses revealed that the CEPA – English test demonstrated good psychometric properties and the test developer may want to evaluate items that misfit the 3PL model.

Statement of the Problem

One of the main emphases of the Nigerian education policy is that citizens must acquire scientific and technological education. Biology is one of the science subjects and has links with other science subjects. It is the general field of knowledge concerned with the study of all aspects of living organisms. According to Parker (1992), Biology embraces those principles of widest application to the origin, growth and development, structure, function, evolution and distribution of plants and animals. It is also the bedrock upon which some science subjects derive their being (origin). The students' poor performance in Biology has drawn attention of researchers and curriculum planners towards Biology as a subject (Kareem, 2003). Inspite of popularity and importance of Biology among Nigerian Students, performance in Biology at

Senior Secondary School level has been poor (Ahmed, 2008). Specifically, Auwalu, Mohd and Muhammad (2014) noted that in Kano State, Nigeria, 25.5%, 32.6%, 29.6%, 23.4% and 24.1% were the percentage passes of students with credit and above grades in Biology WAEC certificate examinations from 2007 to 2011 respectively.

Sakiyo and Badau (2015) also gave the average academic performance of students in WAEC examinations from 2008-2012 in Science subjects and Mathematics and English Language. According to them, Physics recorded the best average academic performance with mean of 56.01%, followed by English Language with 52.52% then, Mathematics with 47.44%, Chemistry 46.30% and the least was in Biology with 37.27%. It was also found that Biology had the highest failure rate of 28.66% followed by Mathematics with 24.39%, Chemistry with 22.52%, English Language with 21.89% and the least rate was in Physics with 13.08%. The desire to know the causes of the poor performance in Biology has been the focus of researchers for some time now. Researchers attributed the poor performance in Biology to some factors like teachers' laxity, poor study habit, parents' poor attitude to their children education and so on without much consideration of the quality of the assessment tools used for assessing the subject.

National Business and Technical Examination Board (NABTEB) is a specialized assessment and certification body in Nigeria that combines professional and general education, specially designed its test methods to meet the twin needs of education and the requirements of the world of work and in line with international best practices. The body without doubt, employs modern psychometric method (IRT model) for the calibration of test items. Since there are various arguments among researchers according to Chernysheko, Stark, Chan and Williams (2001) about the best model among the dichotomous models of IRT for assessing multiple choice examination conducted by examining bodies, this study sought to assess the model that fit the 2016 and 2017 Biology multiple choice test items of the National Business and Technical Examination Board.

Research Questions

To successfully attain the set objectives of this study, the following research questions were raised to guide this study.

1. How do the NABTEB May/June NBC/NTC 2016 and 2017 Biology Multiple choice test items fit into the 1PL, 2PL, and 3PL (dichotomous) IRT models?
2. Is there a difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June NBC/NTC 2016 Biology objective test items?
3. Is there a difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June NBC/NTC 2017 Biology objective test items?

Hypotheses

Research questions 2 and 3 were hypothesized for the study:

1. There is no significant difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June NBC/NTC 2016 Biology objective test items.

2. There is no significant difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June NBC/NTC 2017 Biology objective test items.

Methods

The expo-facto research design was adopted for the study. The population of the study comprised all senior secondary school (SS3) Biology students who sat for NABTEB certificate examination in south-south geo political zone of Nigeria. According to the statistics from NABTEB statistical (2017) A total number of five thousand, one hundred and fifteen (5,115) and Four thousand six hundred (4600) students' responses in south-south geo-political zone in Nigeria for 2016 and 2017 NABTEB Biology examination were used for the study. A sample size of 2000 students which consisted of 1000 males and 1000 females were randomly selected for the study. The instrument for data collection is a 50-item multiple choice question paper of NABTEB for 2016 and 2017 examinations.

The scores of 2016 and 2017 May/June NBC/NTC Biology Examinations of NABTEB were obtained from the board. A Principal Component Factor Analysis was used to assess the most important assumption common to IRT models (unidimensionality) with the help of SPSS statistical software. The examinees responses were analyzed using IRT statistical software EIRT Item Response Theory Assistant for Excel (Germain,Valois & Abdous,2007) to determine item parameters based on IRT framework. EIRT estimates coincides satisfactorily with those from BILOG-MG (Galdin & Laurencelle, 2010). The output include: Item parameter estimates, test of fit, and Item characteristics curves. The result from this analysis was used to answer research questions 1 while Cochran's Chi-square test statistics was used to test hypotheses 1 and 2 at 0.05 significant level.

The unidimensionality of the instruments that were used for data collection in this study are presented below.

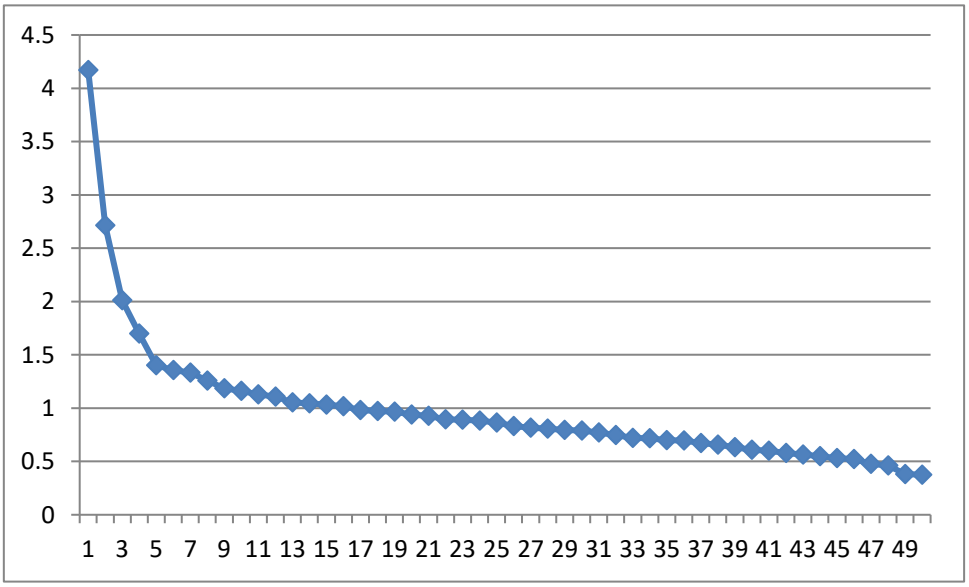

Fig. 1: The Eigen Value Scree Plots of the Component Factor Analysis for 2016 NABTEB May/June, 2016 Biology Multiple Choice Test Items.

The Figure 1 shows the principal component accounts for a maximum percentage of total variance and the variance is used as index of unidimensionality hence a knee point emerged after the first factors, this is a sign of unidimensionality. Also it could be seen that in 2016 that the first Eigen value 4.22 is greater than the next Eigen value 2.21, which showed that there is a dominant factor in the scree plot for the year 2016, thus the instruments satisfied one of the most important assumption necessary for IRT analysis.

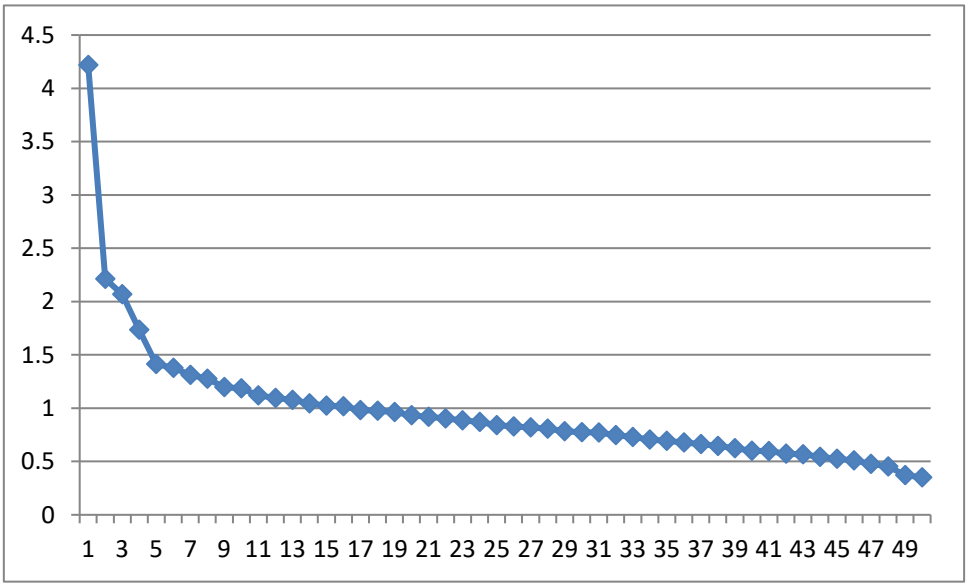

Fig. 2: The Eigen Value Scree Plot of the Component Factor Analysis for 2016 NABTEB May/June, 2017 Biology Multiple Choice Test Items.

The Figure 2 shows a knee point emerged after the first factors, therefore it could be seen that in 2017 that the first Eigen value 4.17 is greater than the next Eigen value 2.71, which showed that there is a dominant factor in the scree plot for the year 2017, thus the instruments satisfied one of the most important assumption necessary for IRT analysis. This concurred with the submission of Udom (2004); Wiberg (2004) who affirmed that a dominant factor is a confirmation of unidimensionality.

Research Question 1: How do NABTEB May/June NBC/NTC 2016 and 2017 Biology Multiple choice test items fit into the 1PL, 2PL, and 3PL(dichotomous) IRT models?

Table 1: NABTEB May/June 2016 and 2017 NBC/NTC Biology multiple-Choice Test Items Fit Analysis

IRT Models	No of Fitted Items (2016)	No of Non Fitted Items (2016)	No Of Fitted Items (2017)	No Of Non Fitted Items (2017)
1PL	44(88%)	6(12%)	23(46%)	27(54%)
2PL	29(58%)	2(4%)	29(58%)	21(42%)
3PL	35(70%)	15(30%)	32(64%)	18(36%)

The Table 1 shows that 44, 29 and 35 items representing 88%, 58% and 70% items that fitted 1PL, 2PL and 3PL IRT models for NABTEB 2016 Biology multiple choice items while 6, 2 and 15 items representing 12%, 4% and 30% did not fit into the IRT (dichotomous) IRT models used for its analysis, hence the table also showed that 1PL IRT model is the model that showed a better fit of the items for the NABTEB May/June,2017 NTC/NBC Biology multiple choice test items while 27, 21 and 18 items representing 54%, 42% and 36% items did not fit into the IRT dichotomous models. Hence most of the NABTEB May/June 2017 Biology multiple choice items fitted the dichotomous models used for its analysis without a significant difference.

Hypotheses Testing

Hypothesis 1: There is no significant difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June NBC/NTC 2016 Biology objective test.

Table 2: Cochran Test of 1PL, 2PL, 3PL Model Fit of 2016 NABTEB Biology Multiple Choice Test Items.

Logistic Model	Value		Cochran Q	Asymp. Significance
	0	1		
1PL	6	44		
2PL	21	29	11.793	0.003
3PL	15	35		

N=50,$\propto$ =0.05.

Table 2, shows the Cochran's Q of 11.793[a] with the P-Value of 0.003, tested at the Alpha level of 0.05, thus since the P-value is less than the Alpha level of 0.05, the null hypothesis which states that there is no significant difference in the fit of 1pl, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June 2016 Biology multiple choice test items is rejected. Therefore there is a significance difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June 2016 Biology multiple choice test items.

Table 3: Cochran Test of 2PL and 3PL Model Fit of 2016 NABTEB Biology Multiple Choice Test Items.

Logistic Model	Value		Cochran Q	Asymp. Significance
	0	1		
2PL	21	29	1.385[a]	0.239
3PL	15	35		

N=50, $\propto = \mathbf{0.05}$

As the result of the significance difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June 2016 Biology multiple choice test items, further comparison of the 2PL and 3PL IRT

model was done. Thus, Table 3 showed that there was no difference between the 2PL and the 3PL IRT models, hence the significant difference lies in the 1PL.This is because the number of fitted items in the 1PL IRT model is greater than that of the 2PL and 3PL IRT models.

Hypothesis 2: There is no significant difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June NBC/NTC 2017 Biology objective test items.

Table 4: Cochran Test of 1PL, 2PL, 3PL Model Fit of 2017 NABTEB Biology Multiple Choice Test Items.

Logistic Model	Value		Cochran Q	Asymp. Significance
	0	1		
1PL	27	23		
2PL	21	29	4.50	.105
3PL	18	32		

N=50,$\propto$ =0.05.

Table 4 shows the Cochran's Q OF 4.500[a]and P-level of 0.105tested at the alpha level of 0.05. Hence since the value of the P-value is greater than the Alpha level of 0.05, therefore the null hypothesis which states that there is no significant difference in the fit of 1PL, 2PL and 3PL (dichotomous) IRT models in NABTEB May/June 2016 Biology multiple choice test items is retained. It shows there is no disparity among any of the 3 models in the model fit.\

Discussion of Findings/Results

The result revealed that 44, 29 and 35 items representing 88%, 58%, and 70% items fitted 1PL, 2PL and 3PL IRT models for NABTEB Biology multiple choice items while 6, 2 and 15 items representing 12%, 4% and 30% did not fit into the IRT (dichotomous) IRT models used for its analysis. It was also found that 1PL IRT model is the model that showed a better fit of the items for the NABTEB May/June, 2016 NTC/NBC Biology multiple choice test items**.** Hence 23, 29 and 32 items representing 46%, 58% and 64% fitted the IRT dichotomous models while 27, 21 and 18 items representing 54%,42% and 36% items did not fit into the IRT dichotomous models for NABTEB **May/June,2016** NTC/NBC Biology multiple choice test items. Most of the NABTEB May/June 2017 Biology multiple choice items fitted the dichotomous models used for its analysis without a significant difference.

Findings revealed that there is a significant difference between the dichotomous IRT models in NABTEB May/June 2016 Biology test items, which was in favour of the 1PL (dichotomous) IRT model. The findings of this study is in line with that of Leeson and Fletcher (2003); Si, (2002), and Kose (2014) who claimed that the 1PLM and 2PLM respectively is superior to others. Contrary to the finding of this study, Chon, Lee and Arisley (2007) claimed that the 1PLM had the largest misfit in items.

It was also revealed that there is no significant difference among the 1, 2 and 3 parameter logistic model fits in NABTEB May/June 2017 Biology test items which shows that the 1, 2 and 3PLM fit the 2017

data. The implication of this empirical finding that supports the IRT theoretical claim is that, for data to be amenable to the IRT analysis there should be a fit of the model to the data set, in other words the data available must be, such that, allows the item to be modeled with an ICC that is derived from an Item Response Function otherwise, the conclusion obtained using the model might not be relevant. However, this finding is in disagreement with Chon, Lee, and Arisley (2007), who claimed that the 1PLM had the largest misfit in items.

Conclusion

It is concluded that the estimated ability in the 1, 2 and 3PL Models are not the same and that there is a significant difference between the 1, 2 and 3 parameter IRT models in NABTEB May/June 2016 Biology test items, which is in favour of 1PL(dichotomous) IRT models but no significant difference was found in NABTEB May/June 2017 Biology test items. Therefore, all the parameter logistic models fits the data for NABTEB May/June 2017 Biology multiple choice test items hence non is empirical superior to others.

Recommendations

The following recommendations are made based on the finding of this study;

1. The examining bodies should make sure that model fits the data well before they are used to make inferences regarding the data
2. Examination bodies should embrace IRT in item generation, assessment of candidates and analysis of results.
3. Examining bodies should engage the services of trained Psychometricians (measurement experts) who are proficient in Item Response Theory in order to ensure objectivity in assessment of examinees.
4. Items diagnosed misfit with any models should not be discarded instead it may be removed from that test form and rewritten or replaced in future test forms

REFERENCES

Adedoyin, O. O. & Mokobi, T. (2015). 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.

Ahmed, M. A. (2008). Influence of personality factor on biology Lecturers' Assessment of difficulty level of genetics concepts in Nigeria Colleges of Education (unpublished doctoral thesis). Ilorin. Nigeria, University of Ilorin.

Auwalu, R. A., Mohd, E. T. & Muhammad, B. G. (2014). Academic achievement in Biology with suggested solutions in selected secondary schools in Kano State, Nigeria. *International Journal of Education and Research*. 2(11) 215-224.

Chernysheko,O. S., Stark,S., Chan, K. Y., Drasgow, F. & Williams, B. (2001). Fitting Item Response Theory models to two personality inventories: Issues and Insights: *Multivariate Behavioural Resource* 36(4) 523-562.*doi:10.1207/S15327906MBR3604_03.*

Chon, K. H., Lee, W. & Ansley, T. N. (2007). Assessing IRT model-data fit for mixed format tests Iowa City, Iowa: Center for Advanced Studies in Measurement and Assessment, the University of Iowa. Retrieved on the 4/6/2018 from http://education.uiowa.edu/casma.

Galdin,M. & Laurencelle,L.(2010). Assessing parameter invariance in item response theory's logistic two parameter model: A Monte Carlo investigation: Tutorials in Quantitative Methods for Psychology 6(20) 39-51.

Germain,S.,Valois, P. & Abdous,B (2007). Eirt-Item response theory assistant for excel (Freeware). Retrieved on the 4/6/2018 from http://libirt.sf.net.

Glas, C. A. W. & Suarez-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.

Hambleton, R. K. & Han, N. (2005). Assessing the fit of IRT models: Some approaches & graphical displays. Paper presented at the annual meeting of the National Council on measurement in education, San Diego, California. Retrieved on the 5/4/2018 from https://www.scribd.com>document .

Hambleton, R. K. & Swaminthan, H.(1985). Item Response Theory: Principles and Application. Boston: Kluwer-Nijhof.

Jansen, M. G. H. & Glas, C. A. W. (2005). Checking the assumptions of Rasch's model for speed tests. *Psychometrika,*70(4) 671-684.Doi:10.1007/s11336-001-0929-2

Kareem, L. O. (2003). Effects of audio-graphics; self-instructional packages on senior-secondary school students' performance in Biology in Ilorin, Nigeria. Unpublished Ph.D. Thesis of the University of Ilorin, Ilorin.

Kose, I. A. (2014). Assessing model data fit of unidimensional item response theory model in simulated data. *Academic journals* 9(17) 642-649

Kyong, H. C., Won, C. L. & Timothy, N. A. (2007). Assessing IRT model-data fit for mixed format test; Centre for advanced studies in measurement and assessment. (CASMA).University of IOWA, IOWA City. U.S.A

Leeson, H. & Fletcher, R. (2003). An investigation of fit: Comparism of the 1-2-3 parameter IRT models to the project as Ttle data. School of Psychology, Massey University, New Zealand.

Maydeu – Olivares, A. & Joe, H. (2005). Limited and full estimation and testing in contingency tables .A unified framework. *Journal of the American statistical Association,* (100) 1009 – 1020.

Parker, I. (1992). Discourse dynamics: Critical analysis for social and individual Psychology Florence, K.Y, U.S: Taylor & Frances /Routledge.

Sakiyo, J. & Badau, K. M. (2015). Assessment of the trend of secondary school students' academic performance in the Sciences, Mathematics and English; Implications for the attainment of the milleneim development goals in Nigeria. *Advances in Social Sciences Research Journal*. 2(2) 31-38.

Salma, A., D. (2009). The eamination of the psychometricquality of the commoneducational proficiencyassessment (CEPA) English test. Unpublished Ph D thesis submitted to the Graduate Faculty of School of Education, University of Pittsburgh.

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

Sinharay S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational and Behavioural Statistics.* 42(4) 375 – 394

Stone, C. A. & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures*. Journal of Educational Measurement,* 40(4) 331–352.

Udom, S.U (2004). Item response theory dimensionality of west African Examination council (WAEC) Mathematics objective test and students performance. Unpublished M.Ed. Thesis. Faculty of Education, University of Calabar, Calabar, Nigeria.

Wiberg, M. (2004). Classical test theory versus item response theory: An evaluation of the theory test in Swedish driving license test. Sweden: Umea' University.

Zhao Y. (2008). Apparatus for addressing the fit of item response theory models to educational test data. Unpublished Doctoral dissertation, University of Massachusetts Amberst.