DOI: http://dx.doi.org/10.4314/gjedr.v16i2.2

GLOBAL JOURNAL OF EDUCATIONAL RESEARCH VOL 16, 2017: 87-94

 ISSN 1596-6224
www.globaljournalseries.com; Info@globaljournalseries.com

ITEM LEVEL DIAGNOSTICS AND MODEL - DATA FIT IN ITEM RESPONSE THEORY (IRT) USING BILOG - MG V3.0 AND IRTPRO V3.0 PROGRAMMES

CYRINUS B. ESSEN, IDAKA E. IDAKA AND MICHAEL A. METIBEMU

(Received 31, January 2017; Revision Accepted 13, April 2017)

ABSTRACT

Item response theory (IRT) is a framework for modeling and analyzing item response data. Item-level modeling gives IRT advantages over classical test theory. The fit of an item score pattern to an item response theory (IRT) models is a necessary condition that must be assessed for further use of item and models that best fit the data. The study investigated item level diagnostic statistics and model- data fit with one-and two- parameter models using IRTPROV3.0 and BILOG- MG V3.0. Ex-post facto design was adopted. The population for the study consisted of 11,538 candidates' responses who took Type L 2014 Unified Tertiary Matriculation Examination (UTME) Mathematics paper in Akwa Ibom State, Nigeria. The sample of 5,192(45%) responses was randomly selected through stratified sampling technique. BILOG-MG V3.0 and IRTPROV3.0 computer software was used to calibrate the candidates' responses. Two research questions were raised to guide the study. Pearson's χ^2 and S - χ^2 statistics as an item fit index for dichotomous item response theory models were used. The outputs from the two computer software were used to answer the questions. The findings revealed that only 1 item fitted 1-parameter model in BILOG- MG V3.0 and IRTPRO V3.0. Furthermore, the findings revealed that 26 items fitted 2-parameter models when using BILOG-MG V3.0. Five items fitted 2-parameter models in IRTPRO. It was recommended that the use of more than one IRT software programme offers more useful information for the choice of model that fit the data.

KEYWORDS: Item Level, Diagnostics, Statistics, Model - Data Fit, Item Response Theory (IRT).

INTRODUCTION

The crucial benefits of IRT models are realized to the degree that the data fit the different models, 1-, 2-, and 3 parameters. Model-data fit is a major concern when applying item response theory (IRT) models to real test data. Though, there is an argument that the evaluation of fit in IRT modeling has been challenging, the use of item response theory model checking and item fit statistics serve crucial factors to effective IRT use in psychometrics for information on items and model selections (Reise, 1990; Embretson & Reise, 2000).

Obtaining evidence of model-data- fit when an IRT model is used to make inferences from a data set is recommended as the standards for educational and psychological testing by the American Association of Educational Research, American Psychological Association, and National Council on Measurement in Education (2014). Failure to meet this requirement invalidates the application of IRT in real data set evaluation. Researches (Orlando and Thissen, 2000, 2003) indicated that model checking

Cyrinus B. Essen, Department of Educational Foundations, University of Calabar, Calabar, Nigeria.
Idaka E. Idaka, Department of Educational Foundations, University of Calabar, Calabar, Nigeria.
Michael A. Metibemu, Institute of Education, University of Ibadan, Ibadan, Nigeria.

remains a major hurdle to the effective implementation of item response theory in which, failure to assess item level and model-data- fit statistics in the applications of IRT models, according to Liu and Maydeu-Olivares (2014) before any inferences can be drawn from the fitted model, is capable of leading to any potentially misleading conclusions derived from poorly fitted models. The need to effectively assess model-data fit is imperative for correctly choosing the right model that adequately fits the data.

Studies have shown an extension beyond dichotomous IRT models to polytomous IRT models, including the generalized partial credit model and rating scale model on item fit statistics and model selection in recent times (Chon, Lee & Ansley, 2007; Kang &Chen, 2011).Various model fit statistics for item-fit index, for dichotomous item response theory (IRT) models had been proposed (Orlando,1997; Orlando & Thissen, 1997, 2000, 2003) to assess the appropriateness of the chosen IRT models and calibration procedure in terms of the model-data test for, 1-, 2-and 3 parameter logistic models, Wells, Wollack, and Serlin (2005) stressed that fit of model to the data must accurately portray the true relationship between ability and performance on the item. They held that model misfit has dire consequences leading to violation of invariance property. Thus, Kose (2014) emphasized that the property of invariance of item and ability parameters is the main stay of IRT that distinguishes it from CTT. The invariance property of item and ability is not dependent on the examinees distribution and characteristics of set of test items. Hence, Bolt (2002) believed that it is imperative for test developers to establish that a particular model fits the data before operationalizing a valid item.

Orlando and Thissen (2003) opined that the appropriate use of IRT models is predicated on the premise that a number of IRT assumptions are made about the nature of the data, to ensure that the model accurately represents the data. When these assumptions are not met, inferences regarding the nature of the items and tests can be erroneous, and the potential advantages of using IRT are not gained. Besides, Sinhary (2005) held that failure to ensure the appropriateness of model-data fit analysis carried the risk of drawing incorrect conclusion.

According to Hambleton and Swaminathan (1985 cited in McAlphine, 2002), the measure of model data fit should be based on three types of evidence. Firstly, the validity of the assumption of the model for the data set such as: (a) unidimensionality, (b) the test is not speeded, (c) guessing is minimal for 1 and 2PL, (d) all items are of equal discrimination for 1PL. Secondly, that the expected properties are obtained to reflect; invariance of item and ability parameter estimates. Finally, the accuracy of the model prediction should be assessed through the analysis of item residuals.

In addition, Sijtsma and Hemker (2000) and Sheng(2005) opined that the basic assumptions of IRT: unidimensionality, local independence, monotonocity and item characteristic curve should be properly assessed as standard measures to investigate model data fit analysis. Zhao (2008) recommended that making judgment about item- level and model fit to test data should be based on four major steps of evidence:

(i) Choosing software and initial classical analysis,
(ii) Checking basic assumptions of IRT,
(iii) Assessing model data fit, and
(iv) Checking model, item and ability parameter invariance.

Orlando (1997), Orlando and Thissen (2000), Stone (2000), Glas and Suarez-Falcon (2003), Stone and Zhang (2003), Dodeen (2004) and Sinharay (2005) developed a number of item-level fit statistics for use with dichotomous item response theory models. The common procedure for constructing item fit indices for the 2PL and 3PL models group respondents based on their estimated standing on the latent variable being measured by the test and obtained observed frequencies correct and incorrect each summed score for these groups.

Dodeen (2004) used several simulated data sets and real data set that employed several newer items fit statistics: The $S-\chi^2$ and $S-G^2$ statistics of Orlando and Thissen (2000), the $\chi^2$and G^2statistics of Stone (2000).Orlando &Thissen (2000) used summed score approach to form new indices: $S-X^2$, a Pearson X^2 statistics, and $S-G^2$, a likelihood ratio G^2 statistic and chi-square goodness of fit statistics. Zhao(2008) stated that the most widely used and current software packages such as BILOG, BILOG MG, BILOG, 3.0, 3.11; PARSCALE, MULTILOG, IRTPRO, GOODFIT, IRTFIT RESAMPLE, among others, provide model fit statistics analysis.

Orlando and Thissen (2003) used MULTILOG software on the utility of $S - X^2$ as an item fit index for dichotomous item response theory models. Results were based on a simulation generated and calibrated for 100 tests under each of 27 conditions (3 bad items) × (3 test lengths) × (3 sample sizes). The three non-logistic (bad) items were created and embedded in otherwise 3PL tests of length 10, 40, and 80 items for samples of size 500, 1,000, and 2,000.The item fit indices $S - X^2$ and $Q1 - X^2$ were calculated for each item. The conclusion was that the performance of $S - X^2$ improved with test length. The performance of $S - X^2$ was superior to $Q1 - X^2$ under most but not all conditions. Results from the study implied that $S - X^2$was useful tool in detecting the misfit of one item contained in an otherwise well-fitted test, lending additional support to the utility of the index for use with dichotomous item response theory models.

Also, Mokobi and Adedoyin (2014) used MULTILOG to assess item level and model fit statistics in a 3 parameter logistic model with 2010 Botswana Junior Certificate Examination Mathematics paper one. A chi-square goodness of fit statistics was employed in assessing item fit to 1PL, 2PL and 3PL models. The results revealed that 10 items fitted the 1PL, 11 items fitted the 2PL model and 24 items fitted the 3PL models. Therefore, the 3PL model was used for the analysis.

Furthermore, Dodeen (2004) used BILOG 3.11 software for fitting the 3PL model to the generated data sets and for computing the values of the χ^2 G statistic. The statistics S−χ2 and S−G^2were computed using the GOODFIT programme. The proportion significant for the S−χ^2 and χ^2 were low and close to the nominal level for all the test conditions. The statistics χ^2 and G^2 were computed using the IRTFIT RESAMPLE programme. The average item fit statistics, the proportion of item fit statistics were significant at 1percent level and the correlations between the generating item parameters and the average item fit statistics over the 100 replications under any test condition were computed under each of the nine test conditions. Furthermore, Essen (2015) examined model-data fit in 2014 in a 50 item dichotomously scored JAMB Mathematics items data with chi-square goodness of fit statistics using BILOG MG, 3.0 software programme. No item fitted the 1-parameter model,26 items fitted 2- parameter IRT model, while 3-paramater model displayed some irregularities. Therefore, the 2-paramter logistic model was best for the data.

In another study, Kang and Chen (2007) used an item-fit index, S-X^2, proposed by Orlando and Thissen (2000, 2003) for dichotomous item response theory (IRT) models, which has performed better than traditional item-fit statistics. The study extended the utility of S-X^2 to polytomous IRT models, including the generalized partial credit model, partial credit model, and rating scale model. The performance of the generalized S-X^2 in assessing item-model fit was studied in terms of empirical Type I error rates and power as compared to results obtained for G^2provided by the computer programme PARSCALE. The results showed that the generalized S-X^2 was a promising item-fit index for polytomous items in educational and psychological testing programmes.

Besides, Chon, Lee and Ansley (2007) in a study examined various model combinations and calibration procedures for mixed format tests under different item response theory (IRT) models and calibration methods. Using real data sets that consisted of both dichotomous and polytomous items, nine possible applicable IRT model mixtures and two calibration procedures 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. To assess goodness of fit, the PARSCALE's G^2 was used. In addition, two fit statistics proposed by Orlando and Thissen (2000) were extended to more general forms to enable the evaluation of fit form fixed format tests. The results indicated that the three parameter logistic models combined with the generalized partial credit model among various IRT models combinations led to the best fit to the given data sets, while the one parameter logistic model had the largest number of misfit items. In comparison of three fit statistics. Some inconsistencies were found between traditional and new indices for assessing the IRT models to data. The study revealed considerably better model fit than the traditional indices.

This study investigated item- level diagnostics and model-data fit in IRT using BILOG MG.3.0 and IRTPRO V3.0 software. The 2014 Unified Tertiary Matriculation Examination (UTME) Mathematics items was used for the analysis. Joint Admissions and Matriculation Board (JAMB) that is vested with the sole

responsibilities of conducting examination for admissions into the Nigerian Universities, Polytechnics and Colleges of Education had shifted from the CTT to IRT paradigm in test construction and development in line with the best global practices of item and person independence in educational assessment. However, the extent to which the items fit the various IRT model is the concern of the study as an essential standard condition for the use of the data.

Purpose of the study

The study investigated the extent 2014 UTME Mathematics items fitted the 1-2- and 3 parameter with the use of BILOG MG.3.0 and IRTPRO V3.0 software programmes with the use of S- X^2 and Pearson X^2 statistics. The study specifically examined:

1. The IRT model data fit statistics S-X^2 and X^2 in IRTPRO V3.0 and BILOG MG.3.0 that best diagnose 2014 UTME Mathematics items model data fit accurately
2. The IRT software programmes (BILOG MG.3.0 and IRTPRO V3.0) that best fit the 2014 UTME Mathematics items.

Research questions

The following research questions directed the study.

1. Which of the IRT fit statistics S-X^2 in IRTPRO V3.0 and X^2 in BILOG MG.3.0 best diagnose model data fit accurately?
2. Which of the IRT software programmes (BILOG MG.3.0 and IRTPRO V3.0) is appropriate for the 2014 UTME Mathematics items?

Method

The research design for this study was ex-post facto. The researcher's choice to use this method was based on the fact that the researcher had no intentions to manipulate the characteristics of the participants nor the variables involved. The population for the study consisted of 11,538 candidates who took Type L 2014 UTME Mathematics in Akwa Ibom State. Four thousand, five hundred and forty-six were females and 6,994 were males. A stratified sampling procedure was used to select 5,192 candidates' response data, comprising, 2,596 males and 2,596 females, representing 45 per cent of the candidates who took 2014 UTME Mathematics items. The 5,192 candidates' response data were subjected to BILOG-MG 3.0 and IRTPRO V 3.0 computer software calibration in a 1-, 2- and 3-parameter models. The outputs were used for analysis.

Results

The results of the data analysis are presented in Tables: 1 and 2 according to the research questions.

Research question 1

Which of the IRT fit statistics S-X^2 in IRTPRO V3.0 and X^2 in BILOG MG.3.0 best diagnose s2014 UTME Mathematics items model data fit accurately?

Table 1 shows the results obtained from two software programmes: IRTPRO V 3.0 and BILOG MG 3.0. The two programmes shows the extent 2014 Mathematics items were calibrated with S-X^2 and X^2 diagnostic indices of each item at different IRT models. Both software calibrated the data at 1-PL, 2-PL, and only IRTPRO calibrated 3-parameter logistic model. 3-parameter in BILOG MG 3.0 displayed some level of inconsistencies that did not allow for the use of the calibrated model. From the results in IRTPRO 48 items were significant at less than .05, in 1- parameter, except item no 10 with a non- significant value of .1216. In a 2-paramter, 44 items are significant at less than .05, with 5 items: 10 (.1808), 12 (.1023), 18(.0549), 19 (.1714) and 45(.0909) as non-significant. At 3-paramter model, in IRTPRO, 42 items are significant, while 7 items: 10 (.2851), 12 (.6206), 18(.1333), 19(.3215), 29 (.3878), 37(.1695) and 45(.2597) are non-significant.

Table 1: IRT item- level diagnostic fit statistics S-X^2 in IRTPRO V3.0 and X^2 in BILOGMG.3.0

Item	IRTPRO- 1PL			BILOG- 1PL			IRTPRO- 2PL			BILOG -2PL			IRTPRO -3PL		
	S-X^2	d.f	Prob.	X^2	d.f	Prob	S-X^2	d.f	Prob.	X^2	d.f	Prob	S-X^2	d.f	Prob.
2	249.52	39	.0001	187.5	8	.0000	181.29	42	.0001	22.7	9	.0068	159.74	41	.0001
3	120.31	41	.0001	66.1	8	.0000	91.10	39	.0001	21.0	9	.0125	69.84	39	.0001
5	79.61	41	.0003	52.0	8	.0000	73.66	39	.0007	15.9	9	.0690*	61.21	40	.0170
5	209.61	40	.0001	256.6	9	.0000	114.33	42	.0001	8.9	9	.4436*	110.55	42	.0001
6	907.98	39	.0001	277.5	9	.0000	607.46	41	.0001	21.9	9	.0093	572.84	40	.0001
7	91.02	40	.0001	67.4	8	.0000	76.47	39	.0003	11.2	9	.2626*	72.58	39	.0009
8	164.11	39	.0001	185.8	8	.0000	155.92	38	.0001	5.5	9	.7877*	141.53	38	.0001
9	119.44	41	.0001	63.2	8	.0000	71.39	38	.0008	19.1	9	.0243	59.59	39	.0184
10	51.71	41	.1216*	51.2	8	.0000	49.07	41	.1808*	9.7	9	.0840*	44.56	40	.2851*
11	1997.49	34	.0001	910.8	9	.0000	141.06	43	.0001	74.8	9	.0000	94.09	41	.0001
12	59.32	40	.0251	31.8	8	.0000	50.51	39	.1023*	27.1	9	.0013	36.70	40	.6206*
13	158.19	39	.0001	107.1	8	.0000	87.33	36	.0001	9.7	9	.3746*	65.47	37	.0027
14	172.89	40	.0001	110.5	8	.0000	80.32	36	.0001	15.3	9	.0825*	64.88	37	.0031
15	138.82	39	.0001	100.6	8	.0000	87.61	37	.0001	17.7	9	.0389	69.21	38	.0015
16	133.25	39	.0001	106.0	8	.0000	67.71	37	.0015	13.7	9	.1317*	56.44	38	.0273
17	2136.33	34	.0001	1067.7	9	.0000	235.45	43	.0001	107.9	9	.0000	202.12	41	.0001
18	81.87	40	.0001	69.0	8	.0000	54.06	39	.0549*	14.2	9	.1152*	50.00	40	.1333*
19	140.69	40	.0001	113.8	8	.0000	43.89	36	.1714*	11.3	9	.2562*	40.41	37	.3215*
20	352.10	40	.0001	54.0	9	.0000	501.95	39	.0001	11.0	9	.2726*	548.22	40	.0001
21	256.18	40	.0001	202.0	9	.0000	158.10	36	.0001	17.5	9	.0417	98.31	39	.0001
22	11232.38	21	.0001	1705.4	7	.0000	122.68	42	.0001	----------			1275.43	43	.0001
23	209.03	40	.0001	165.0	8	.0000	88.01	36	.0001	23.6	9	.0007	56.43	39	.0349
24	182.13	39	.0001	114.7	8	.0000	65.86	36	.0017	15.3	9	.0837*	59.36	38	.0349
25	145.70	39	.0001	97.7	8	.0000	96.91	37	.0001	17.3	9	.0440	56.37	40	.0445
26	310.61	38	.0001	152.5	9	.0000	326.20	38	.0001	28.3	9	.0009	363.98	38	.0001
27	183.33	40	.0001	165.5	8	.0000	90.73	36	.0001	27.0	9	.0014	79.92	39	.0001
28	153.96	39	.0001	157.8	8	.0000	123.83	38	.0001	4.6	9	.8662*	120.19	38	.0001
29	144.34	40	.0001	91.2	8	.0000	58.06	37	.0150	5.6	9	.7802*	41.95	40	.3878*
30	173.51	39	.0001	191.0	8	.0000	141.44	39	.0001	11.7	9	.2332*	87.44	40	.0001
31	800.83	37	.0001	191.9	9	.0000	544.12	39	.0001	18.3	9	.0320	661.00	41	.0001
32	188.17	40	.0001	39.0	8	.0000	294.59	39	.0001	5.4	9	.7990*	294.59	38	.0001
33	494.94	40	.0001	47.2	9	.0000	710.61	39	.0001	6.2	9	.7167*	737.93	41	.0001
34	247.21	40	.0001	243.3	9	.0000	122.36	36	.0001	8.7	9	.4634*	69.64	38	.0013
35	1804.09	36	.0001	998.3	8	.0000	545.03	43	.0001	230.6	9	.0000	496.95	42	.0001
36	183.56	39	.0001	286.3	8	.0000	176.07	39	.0001	5.3	9	.8106*	153.28	41	.0001
37	186.71	40	.0001	18.2	8	.0000	93.11	39	.0001	15.2	9	.0846*	47.30	39	.1695*
38	91.28	40	.0001	43.8	8	.0000	119.33	39	.0001	6.6	9	.6836*	111.89	39	.0001
39	197.54	39	.0001	199.5	8	.0000	83.10	37	.0001	25.4	9	.0026	55.56	39	.0414
40	202.44	38	.0001	272.8	8	.0000	158.11	39	.0001	24.0	9	.0043	122.98	41	.0001
41	128.29	40	.0001	51.8	9	.0000	194.86	39	.0001	19.3	9	.0224	186.50	40	.0001
42	126.44	39	.0001	82.8	8	.0000	83.19	38	.0001	32.4	9	.0002	57.47	39	.0285
43	487.99	39	.0001	48.8	8	.0000	634.84	39	.0001	3.0	9	.9660*	619.04	40	.0001
44	89.11	40	.0001	92.6	8	.0000	97.97	40	.0001	1.9	9	.9924*	88.40	40	.0001
45	81.79	40	.0001	104.8	8	.0000	51.21	39	.0909*	16.3	9	.0611*	45.30	40	.2597*
46	132.00	39	.0001	191.7	8	.0000	93.25	40	.0001	19.7	9	.0196	65.12	40	.0073
47	1331.28	36	.0001	683.0	8	.0000	481.43	44	.0001	213.6	9	.0000	429.01	43	.0001
48	163.13	39	.0001	99.4	9	.0000	180.29	39	.0001	5.0	9	.8311*	180.42	40	.0001
49	8787.90	18	.0001	1024.8	5	.0000	87.45	42	.0001	---- --		------	565.39	37	.0001
50	74.95	39	.0001	69.4	9	.0000	58.94	39	.0211	8.1	9	.5260*	58.85	40	.0275

***= non-significant items**

In BILOG MG. V 3.0, at 1-parameter model, all the 49 items are significant. At the 2-paramter model, 21 items are significant, while 26 items: 4(.0690), 5(.4436), 7 (.2626), 8(.7877), 10(.0840), 13(.3746), 14(.0825), 16(.1317) 18(.1152), 19(.2562), 20(.2726), 24(.0837), 28(.8662), 29(.7802), 30(.2332), 32(.7990), 33(.71670, 34(.4634), 36(.8106), 37(.0846), 38(.6836), 43(.9660), 44(.9924), 45(.0611), 48(.8311) and 50(.5260). However, the 3-parameter logistic model in BILOG MG V3.0 showed some inconsistencies, therefore, the calibration was not obtained. The rule of thumb in model-data fit holds that items with significant values are misfit items for the model. This implies from the results that in IRTPRO programme, 1 item fits 1-paramter; 5 items fit 2-parameter and 7 items fit 3-parameter model. In BILOG MG V3.0,

no item fits the 1-parameter mode; 26 items fit the 2-parameter model. Therefore, 2-parameter in BILOG MG V3.0 best fits the 2014 UTME Mathematics items.

Research question 2

Which of the IRT software programmes (BILOG MG.3.0 and IRTPRO V3.0) is appropriate for the 2014 UTME Mathematics items?

Table 2: IRT software programmes (BILOG MG.3.0 and IRTPRO V3.0) in 2014 UTME Mathematics items?

Software Programme	Non-sig. item per model	Remarks
IRTPRO (1- PL)	1	Not suitable
IRTPRO (2-PL)	5	Not suitable
IRTPRO (3-PL)	7	Not suitable
BILOG MG V 3.0 (1- PL)	None	Not suitable
BILOG MG V 3.0 (2-PL)	26	Most suitable

Result in Table 2 reveals that though the numbers of items shows some improvement in IRTPRO from 1 in 1-pl, 5 items in 2-pl to 7 items in 3-pl models the choice of IRTPRO software programme does not prove very suitable, as more items are not identified. Comparatively, BILOG MG V3.0 software programme show remarkable improvement in identifying 26 items that fit 2-parameter models, though no item fits 1-parameter model. The result show that the use of software is dependent on which programme indicates an improved number of items that suits a particular model. Therefore, the choice of software programme is the number of items that best show improvement in chosen software at the different models.

Results

The results from research question 1 revealed that IRTPRO V 3.0 and BILOG MG 3.0, exhibited different degrees in the use of S-X^2 and X^2 diagnostic indices of each item at different IRT models. Both software calibrated the data at 1-PL,2-PL, and only IRTPRO calibrated 3-parameter logistic model. The findings agree with a study carried out by Orlando and Thissen (2003) on the utility of $S - X^2$ as an item fit index for dichotomous item response theory models. The item fit indices $S - X^2$ and $Q1 - X^2$ were calculated for each item. The conclusion was that the performance of $S - X^2$ improved with test length. The performance of $S - X^2$ was superior to $Q1 - X^2$ under most but not all conditions. Results from the study implied that $S - X^2$ was useful tool in detecting the misfit of one item contained in an otherwise well-fitted test, lending additional support to the utility of the index for use with dichotomous item response theory models. Also, Mokobi and Adedoyin (2014) used MULTILOG to assess item level and model fit statistics in a 3 parameter logistic model with 2010 Botswana Junior Certificate Examination Mathematics paper one. The study used X^2 goodness of fit statistics in assessing item fit to 1PL, 2PL and 3PL models. The results revealed that 10 items fitted the 1PL, 11 items fitted the 2PL model and 24 items fitted the 3PL models. Therefore, the 3PL model was used for the analysis. Kose(2014) found that in a 1-, 2- and 3-paramter for assessing model data fit, 2-PL model fitted significantly better than the 3-PL model when -2Log likelihood ratio X^2 was used.

However, when Orlando and Thissen (2000) evaluated model-data fit from fixed format tests, the results indicated that the three parameter logistic models combined with the generalized partial credit model among various IRT model combinations led to the best fit to the given data sets. The one parameter logistic model had the largest number of misfit items. In comparison of three fit statistics. Some inconsistencies were found between traditional and new indices for assessing the IRT models to data. The study revealed considerably better model fit than the traditional indices. The finding implied that conducting item level diagnostics and model-data fit is imperative in using IRT models in analysis

Results from research question 2 indicated that BILOG MG V3.0 computer programme displayed greater efficiency in dictating items that fit the various IRT models than the IRTPRO programme. The results indicated that the need to use more than one software to examine model data fit. Various IRT software programmes are used to examine model-data fit, such as BILOG,

BILOG MG, MULTILOG, IRTPRO, PARSCLE, among others. These programmes provide different information concerning the model fit and comparison will show an improvement when more than one programme is compared in assessing model-data fit. Though, many studies have not considered the use of more than one software in comparing the model-data fit, this study provides the ground for more studies in this respect.

CONCLUSION

The study examined item diagnostics statistics and model-data fit in item response theory using BILOG- MG V3.0 and IRTPRO V3.0 programmes. The results indicated that χ^2 and S - χ^2 statistics showed some items that fitted the 1-, 2- and 3 parameter IRT logistic models. Also, BILOG MGV3.0 and IRTPRO V3.0 showed different degrees in locating items that fitted the various IRT models. Based on these results, the study concluded that assessing model-data fits using various statistical indices and the used of multiple IRT programmes is imperative in the use of IRT model choice analysis.

RECOMMENDATIONS

From the findings and conclusion reached, the following recommendations were made:

1. That the selection of best IRT model should depend on assessing item fit statistics as the first step to apply IRT with confidence.
2. That the use of various item fit statistics is a step to ensuring that comparison is made for informed judgment and variety of diagnostic evidences
3. That the use of more than one IRT programmes will provide the choice of the best programme that provide more useful information about the real data set.

REFERENCES

American Association of Educational Research, American Psychological Association and National Council on Measurement in Education., 2014. Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bolt, D. M., 2002. A Monte Carlo comparison of parametric and non-parametric polytomous DIF detection methods. Applied Measurement in Education, 15, 113-141.

Chon, K. H., Lee, W. C and Ansley, T. N., 2007. Assessing IRT model-data fit for mixed format tests. CASMA Research Report, 26, 1-26. Retrieved from http/www.education.uiowa.edu/casma

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

Embretson, S. E and Reise, S. P., 2000. Item Response Theory for Psychologists. Mahwah, NJ: Erlbaum.

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

Glass, C. A. W and Falcon, J. C. S., 2003. A comparison of the item-fit statistics for the three-parameter logistic model. Applied. Psychological. Measurement. 27: 87-106.

Kang, T and Chen, T. T., 2011. Performance of the generalized s-x^2 item fit index for the graded response model. Asian Pacific Education Review, 12, (10): 89-96.

Kose, I. A., 2014. Assessing model data fit of unidimensional item response theory in simulated data. Educational Research and Reviews, 9, (17): 642-649. Retrieved from: http/www.academicjournals.org/ERR.

Liu, Y and Maydeu-Olivares, A., 2014. Identifying the source of misfit in item response theory models. Multivariate Behavioral Research, 49,354-371.DOI:10.1080/00273171.2014.910744.

McAlpine, M., 2002. A summary of methods of item analysis. CAA Blue Paper, 2, 1-32.

Mokobi, T and Adedoyin, O. O., 2014. Identifying location biased items in the 2010 Botswana junior certificate examination Mathematics paper one using the item response characteristics curves.

International Review of Social Sciences and Humanities, 7(2), 63-82.

Orlando, M., 1997. Item fit in the context of item response theory. Dissertation Abstract International, 58,2175.

Orlando, M and Thissen, D., 2000. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24, 50–64.ViewArticleGoogle Scholar

Orlando, M and Thissen, D., 2003. FurtherinvestigationoftheperformanceofS-x^2:An item index for use with dichotomous item response theory models. Applied Psychological Measurement, 27,289-298.

Reise, S. P., 1990. A comparison of item-and person fit methods of assessing model data fit in IRT. Applied Psychological Measurement, 14(2),127-137.

Sheng, Y., 2005. Bayesian Analysis of Hierarchical IRT models: Comparing and Combining the Unidimensional and Multidimensional IRT models. Unpublished Doctoral Dissertation. University of Missouri-Columbia.

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

Sijtsma, K and Junker, B. W., 2006. Item response theory: Past performance, present developments and future expectations. *Behaviormetrika*, 33 (1),75-102.

Sinharay, S., 2005. Assessing fit of unidimensional item response theory models using a Bayesian approach. Journal Educational Measurement. 42, (4): 375-394.

Stone, C. A., 2000. Monte-Carlo based null distribution for an alternative fit statistic. Journal Educational Measurement, 37:58-75.

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

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

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