DIF DETECTION SENSITIVITY OF LORD'S CHI-SQUARE, RAJU'S AREA, LOGISTIC REGRESSION, MANTEL-HAENSZEL, STANDARDIZATION, AND TRANSFORMED ITEM DIFFICULTIES METHODS, IN COMPARISON, USING R.

Dr. Wokoma T. Abbott
Government Technical College, Port Harcourt, Rivers State, Nigeria.

Article DOI: https://doi.org/10.36713/epra7953
DOI No: 10.36713/epra7953

ABSTRACT

Due to psychological differences between individuals, no test item can function exactly the same manner in these individuals. Differential item functioning (DIF) will always occur as a result of these differences in the person parameter of these individuals being examined even when item parameters remain constant during testing. This postulate of item response theory (IRT) was proven in this work. This study investigated if DIF detection methods will have the same DIF detection sensitivity. Comparative research design formed the framework of the study. Transformed item difficulties (TID), Mantel-Haenszel (MH), standardization, logistic regression, Ragu's area, and Lord's chi-square methods were compared. The study used 400 vocational one students (200 male as reference group and 200 female as focal group) in Rivers state, Nigeria. The multiple choice items of 2019 computer science for the junior school certificate examination (JSCE) was adapted as the instrument for data collection, which were administered to students and scored dichotomously. Difficulty and discrimination parameters of the items were analyzed using the 2PL model of IRT with the help of ltm package. Ogives of the items were plotted with ggplot2 package. Individual DIF methods and DichoDif in DifR were used to detect DIF and compare the methods. The results revealed that all the items of the test functioned differently between the reference group and the focal group as shown in the item characteristic curves (ICCs). In comparison of the DIF detection methods, standardization method detected most of the DIF items followed by logistic regression method, and then lord's chi-square methods. Transformed item difficulties method detected more than mantel-Haenszel method. Raju's area method could not detect any. In the light of the finding, it was recommended that the best DIF detection methods (possibly combination of them) should be used to identify DIF items in tests.

KEYWORDS: ***Item response theory, differential item functioning, item characteristic curve, item parameters.***

INTRODUCTION

Generally, every item of a test will function uniquely according to the examinees trait levels on subject under assessment. Items are expected to maintain the same values in parameters among examinees having the same trait level, and shift among examinees having dissimilar trait levels on same subject. This shift shows the existence of achievement gap between groups of differing trait levels. This explains the concept of differential item functioning (DIF).

There are numerous occasions in which test items exhibit DIF. Items differently functioning for individuals of different race, gender, religious, cultural and other affiliations have been a concern in psychometrics. In education and psychology, tests are designed for variety of purposes. Items for a specific test is structured to achieve what such test is intended for. DIF items are observed to be present after test administration, it is a case of item parameter value shifting among examinee's. But if it is a case of bias or parameter drift (IPD), the consequent reduction in authenticity and acceptability will occur. If the result is an error-free observed score difference among high and low ability examinees, it is a proof of inequality in ability among them.

Classical test theory (CTT) and item response theory (IRT) are two common methodologies in psychometrics that have been adopted for over three decades in assessment practice. Optimism of some scholars about advantages of IRT has formed a point of study (Carlo, M. 2009). Despite these advantages one has over the other, the chances of errors in both theories cannot be ruled out. Reliability indices of test outcomes are reduced by errors. In educational and psychological testing, errors are either systematically or randomly introduced the process. The more as assessment is free from error the more valid the predictions from such test outcome. CTT from its definition prominently holds that the actual outcome from test/sessions are not 100% reliable. Biases and other extraneous factors sum to form the error component of the observed score (Ado, A. B., Rahimah, E., Rohaya, T., Sakinah, S., & Abdullah, B. I., 2019; Philip, E., keith, J., Barrett, F., & shannon, W., 2019).

$$Observed\ score\ (x)\ =\ True\ Score\ (T)\ +\ Error\ Score\ (E)$$

An important note concerning CCT is that the variation of observation is based on examinees group standard deviation. The standard error of estimate in IRT takes different dimension. Its measure is based on items and each examinee’s level of ability on the construct being measured (Hambleton, R. K., Swaminathan, Rogers, H. J, 1991; De Ayala, R. J., 2009, Dewars, C., 2010).

IRT is a modern psychometric methodology having related latent trait models being used for different test designs and assessment procedures. The commonly used models are one-parameter logistic model (1PLM), two-parameter logistic model (2PLM), three-parameter logistic model (3PLM), and four parameter logistic model (4PLM). These are a group of dichotomous (binary) models. Polytomous model group has a number of related models such as graded response model (GRM), nominal response model (NRM), partial credit model (PCM), and rating scale model (RSM). There are the modified and generalized models that belong to this family. Some are the modified graded response model (MGRM), and the generalized partial credit model (GPCM) (Stata 14 manual, 2015). Multiple and hybrid IRT models involving multidimensional characteristics of items and latent trait also exist (Lord, 1980; Hambleton, Swaminathan & Rogers, 1991).

From the literature, DIF also exist in polytomous and multiple models Meng, 2018; Scott, 2011; & Paula and Craig, 2013). This study was confined to dichotomous IRT models. Each of the dichotomous models have their unique individual characteristics as other family members of the IRT. All aim to measure the underlying strength of trait of examinee which produce true score. There are other fundamental concepts that are associated in determining examinees and items disposition during testing process. These are item response function (IRF), item information function (IIF) and in-variance. The dichotomous models are classified according to the number of parameter each has. 1PLM has only difficulty (or location) or (b) parameter; 2PLM has discrimination (a) and difficulty parameters; 3PLM has difficulty (b), discrimination and guessing (c) parameters; and 4PLM has difficulty (b), discrimination (a), guessing (c), and upper asymptote (d) parameters (Bartons & Lord, 1981; Xinming & Yiu-fa, 2014).

The 1PLM binary model assumes that all items in a test relate to the latent trait equally and items only vary in difficulty (Andrew, A. nd). This model explains the level on the latent trait continuum on which an examinees probability of choosing the correct option to an item is 0.5. Higer b-parameter requires higher level of trait to achieve the 0.5 probability of getting the answer correct.

Addition of discrimination (a) parameter forms the model to 2PLM. This shows the slop of the ogive. Also, higher slope values are better discriminators. If items are differently functioning due to some reasons, then differing locations and slopes will be obtained. The inclusion of the third and fourth item parameters such as guessing (c) and upper asymptote to the 2PLM will give 3PLM and 4PLM respectively (Baker, 2011; Wokoma, 2021). The mathematical functions of these models are;

For 1PLM,

$$P(X = \frac{1}{\theta,\ b})\ =\ \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}}$$

For 2PLM,

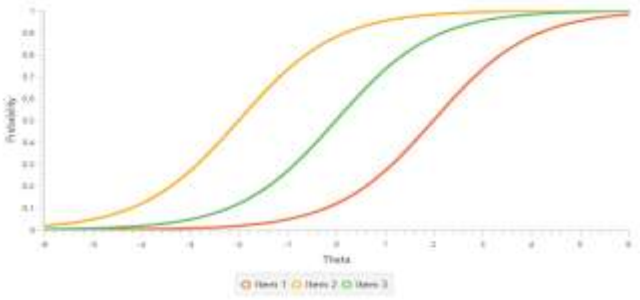
$$P(X = \frac{1}{\theta,\ b,\ a})\ =\ \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$$

For 3PLM,

$$P\left(X = \frac{1}{\theta,\ b,\ a,\ c}\right) =\ c + (1-C)\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$$

For 4PLM,

$$P\left(X=\frac{1}{\theta,\ b,\ a,\ c,\ d}\right)=\ c+(1-C)\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$$



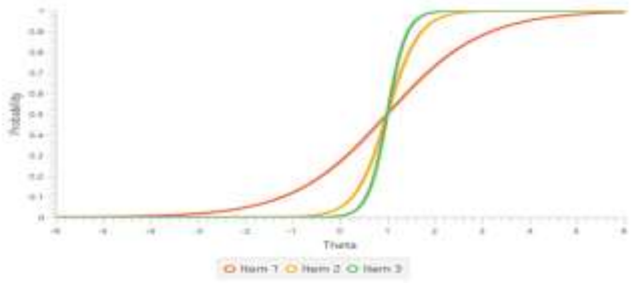


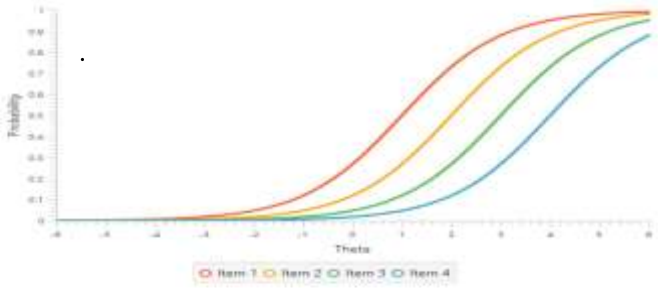
Fig. 1. ICC of items having same discrimination but different difficulties indices.

Fig. 2. ICC of items having same difficulties. but different discrimination.

DIF occurs in both IRT and non-IRT models. In the binary models, it occurs when two groups of examinees having the same trait levels do not have equal probability of choosing the correct response to a dichotomously scored test item. That is, the item appearing unequally difficult for the different groups of examinees having the same trait levels. If it occurs as a result of difference in the levels of trait between the two groups, it is normal; higher scorers are separated from lower scorers. The test is biased if items in it are extraneous to the construct being examined. Biased items can lead to erroneous measurement.

There are numerous concepts also related to DIF and its detection methods. Understanding them is necessary because in analysis of items, the related complexities are considered before choosing a method to apply. The technicalities of each method must be known.

DIF has two forms, uniform and non-uniform. There are also different methods of detecting DIF, IRT method and non-IRT method (David, Sebastian, Francis and Paul, 2010). Uniform DIF occurs if a group of examinees unfairly performs completely more than another of the same ability level on all items of a test. Also, non-uniform DIF occurs if some items of a test are fair to one group of examinees and unfair to the other group, and vice versa (Hossien, 2012). These concepts are illustrated in the item characteristic curves (ICCs) below



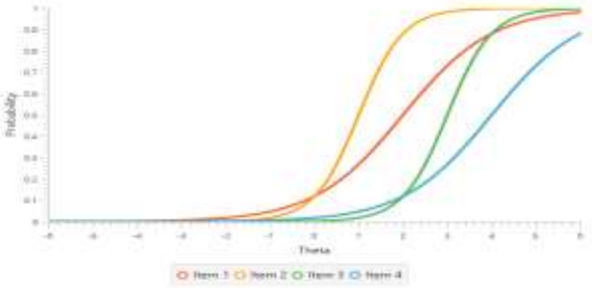


Fig. 3. ICCs for uniform DIF.

Fig. 4. ICCs for non-uniform DIF.

DIF DETECTION METHODS

Validity and reliability are central and very important to psychometricians. When we wish to carry out DIF analysis, item parameters, examinees, test model, number of groups (reference and focal groups), and other parameters are considered before adopting the right DIF method that will be suitable for these parameters. Currently there are several dichotomous DIF detection methods available, each having its own unique complexities, and DIF detection methodology. These lead to the different classes of DIF methods we have in the literature.

The methods that fall into IRT based category are designed to adopt IRT model in their DIF detection procedure. The likelihood-ratio test (LRT) (Thissen, Steinberg and Wainer, 1988); Lord's chi-square test (Lord, 1980); and Raju's area (Raju, 1990). Logistic regression (Swaminathan and Rogers, 1990), Mantel-Haenzel (MH) (Holland and Thayer, 1988), Breslow Day (Aguerri, Galibert, Attorres and Maranon, 2009), standardization (Dorans and Kullick, 1986) and transformed item difficulties (TID) (Angoff and Ford, 1973), and simultaneous item bias test (SIBTEST) (Shealy and stout, 1993) methods are classified as non -IRT based methods. The procedure adopted in these methods is more of a classical measurement theory. These classes of methods are sometimes classified as parametric and non-parametric methods by some scholars. As stated above, each of the IRT and non-IRT methods can also detect uniform and non-uniform DIF. (Yuan-Ling, 2015; Xiaoting, 2010; & Abdullah; 2017).

PURPOSE OF THE STUDY

Just as it is for physical and other measuring instrument, no two devices have the same precision. There is always a difference in their measurements. Different DIF detection methods may also differ in accuracy. The aim of this study is to compare some of the traditional DIF detection methods if they have the same DIF detection sensitivity.

METHOD

Design

Comparative research framework was used in this study. This was because it involved;

1. Comparison between the male students (reference group) and female students (focal group) performances on the subject (information and communication technology) investigated.
2. Comparison of DIF detection sensitivity between six traditional methods available in R, such as;

i. Transformed Item Difficulties (TID) method (Angoff and Ford, 1973).
ii. Mantel-Haenszel (M-H) method (Holland and Thayer, 1988).
iii. Standardization method (Dorans and kullick, 1986)
iv. Logistic regression method (Swaminathan and Rogers, 1990).
v. Lord's Chi-square method (Lord, 1980).
vi. Raju's area method (Raju, 1990).

Other methods such as Breslow-Day method (Aguerri et al, 2009, Penfield, 2003), SIBEST (Shealy and stout, 1993, Li and Stout, 1996, and Chalmers, 2018), Likelihood-ratio test method (Thissen, Steinberg and Wainer, 1988) and extensions of the traditional methods are also found in R (David, et al, 2010).

Sample

The study used 400 vocational one (Voc 1) students collected from three technical colleges out of the five in Rivers State, Nigeria. This class of students was chosen because they all study information and communication technology and use the same curriculum. The sample had 200 male and 200 female students. These students were randomly selected in proportion in accordance with their school student population. These schools are Government Technical College, Port Harcourt; Government Technical College Tombia; and Government Technical College, Ele-Ogu.

INSTRUMENT AND PROCEDURE

The 60 multiple choice items (section A) of 2019 computer science for the Junior School Certificate Examination (JSCE) was adapted for this study. Some of the items were replaced with other items having extraneous constructs different from what it should have been in regard to the curriculum for Voc 1. These items are expected to be equally very easy with low discrimination indices for both groups. The items were administered in their various schools in 60 minutes' session using paper-on-pen mode. Students responses were dichotomously scored.

STATISTICAL ANALYSES

The 2PLM of the IRT models was used to analyze the difficulty and discrimination parameters of test items responded to by the reference and the focal groups using the ltm package. ICCs of the items were also plotted using ggplot2 package which displayed the different ogives as characteristics of these items as they function in the examinees. Each of the six traditional DIF analysis was done using their dividual methods, and the final comparative analysis of all the methods were done using dichoDIF of difR.

RESULT

b is Difficulty parameter; a is Discrimination parameter.

Reference Group

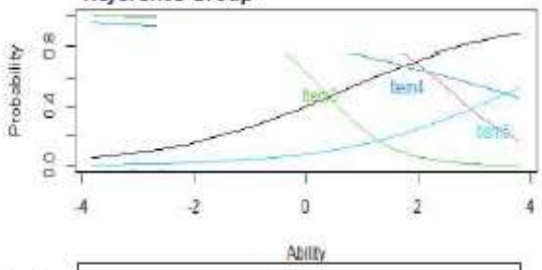


Fig. 5

	b	*a*
Item1	*0.6079*	*0.6470*
Item2	*2.5852*	*-1.3286*
Item3	*0.3303*	*-1.6456*
Item4	*3.4491*	*-0.4355*

Focal Group

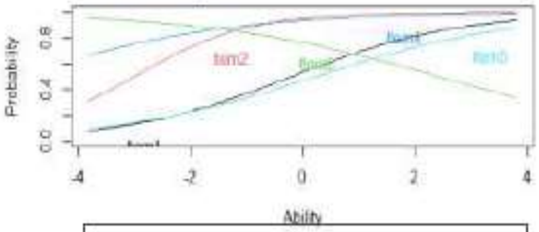


Fig. 6

	b	*a*
Item1	*-0.2058*	*0.6737*
Item2	*-2.9955*	*0.9924*
Item3	*2.4298*	*-0.4854*
Item4	*-5.1559*	*0.5194*
Item5	*0.2326*	*0.5565*

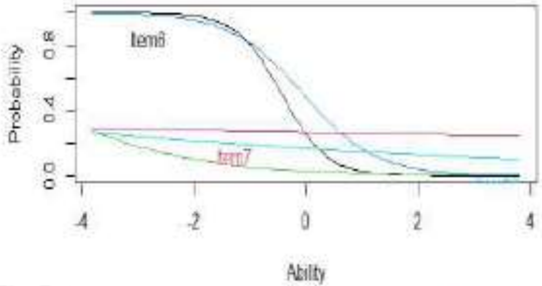


Fig. 7

	b	*a*
Item6	*-0.4281*	*-2.5185*
Item7	*-32.3626*	*-0.0311*
Item8	*-5.2020*	*-0.6791*
Item9	*-0.0440*	*-1.5595*
Item10	*-10.4013*	*-0.1536*

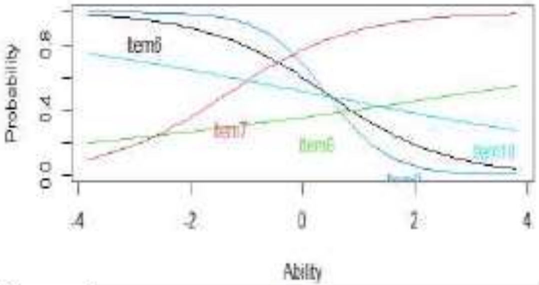


Fig.

	b	*a*
Item6	*0.4295*	*-0.9380*
Item7	*-1.3351*	*0.9049*
Item8	*2.8680*	*0.2086*
Item9	*0.4244*	*-1.8159*
Item10	*0.2181*	*-0.2729*

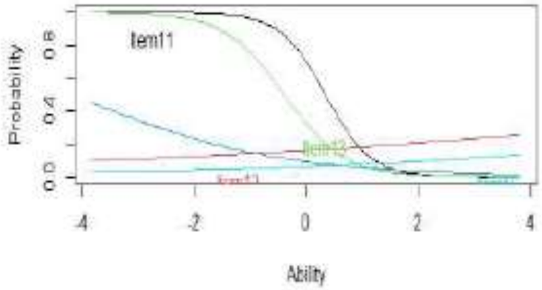


Fig. 9

	b	*a*
Item11	*0.3460*	*-2.5007*
Item12	*10.8405*	*0.1510*
Item13	*-0.3832*	*-1.8815*
Item14	*-4.1776*	*-0.5400*
Item15	*12.1134*	*0.2221*

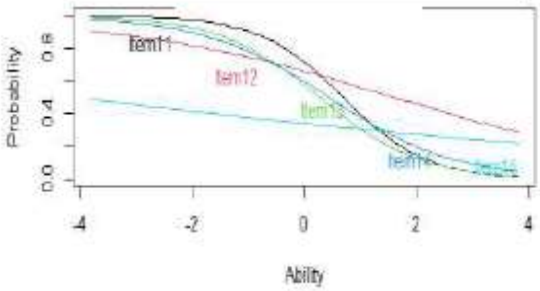


Fig. 10

	b	*a*
Item11	*0.6868*	*-1.3404*
Item12	*1.6006*	*-0.4153*
Item13	*0.2715*	*-1.1176*
Item14	*0.4328*	*-0.8913*
Item15.	*-4.1443*	*-0.1602*

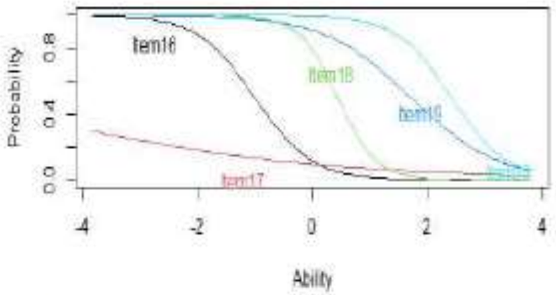


Fig. 11

	b	a
Item16.	-0.9952	-2.0000
Item17.	-6.1402	-0.3595
Item18	0.4451	-2.9645
Item19	1.7219	-1.3175
Item20	2.4230	-2.1018

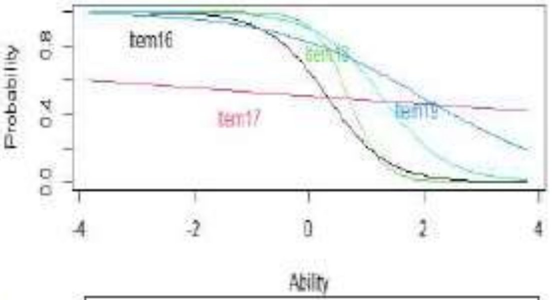


Fig. 12

	b	a
Item16	0.3184	-1.9052
Item17	0.2135	-0.0939
Item18	0.6840	-3.4068
Item19	1.9602	-0.7811
Item20	1.2711	-1.7493

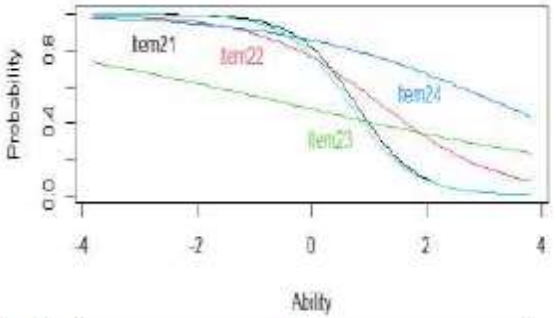


Fig. 13

	b	a
Item21	0.7758	-1.9108
Item22	1.2120	-0.9573
Item23.	-0.3366	-0.2889
Item24	3.3122	-0.5361
Item25	0.6696	-1.8264

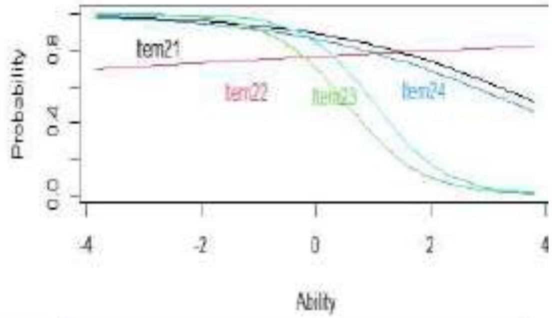


Fig. 14

	b	a
Item21	3.8950	-0.5400
Item22	-12.8540	0.0912
Item23	0.5425	-1.5372
Item24	3.4771	-0.5241
Item25	1.0284	-1.6282

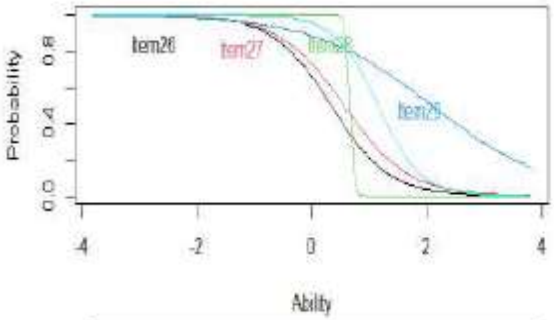


Fig. 15

	b	a
Item26	0.3356	-1.9639
Item27	0.5597	-1.7646
Item28	0.6640	-28.8061
Item29	2.0856	-0.9700
Item30	1.1525	-2.6506

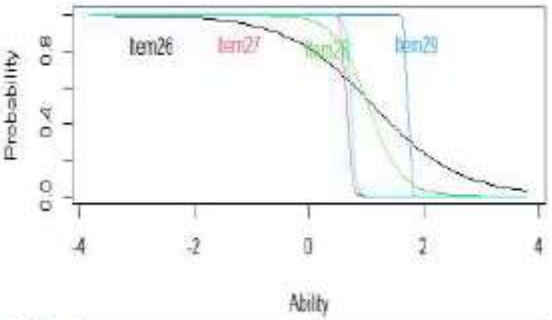


Fig. 16

	b	a
Item26	1.1177	-1.3176
Item27	0.6601	-25.7977
Item28	1.0312	-3.4650
Item29	1.7204	-51.7618
Item30	0.7311	-28.8839

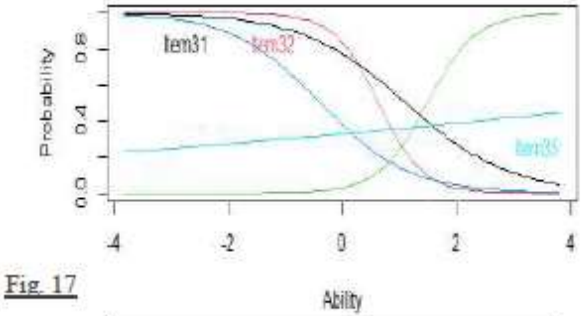


Fig. 17

	b	a
Item31	1.1169	-1.1063
Item32	0.6329	-2.5669
Item33	1.4904	2.3480
Item34	-0.3811	-1.2903
Item35	5.4505	0.1309

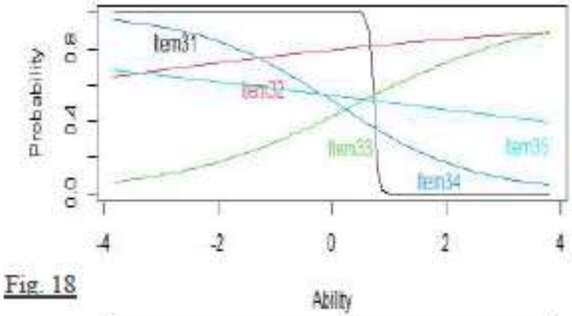


Fig. 18

	b	a
Item31	0.7340	-27.3853
Item32	-6.8203	0.1991
Item33	0.4700	0.6339
Item34	0.0602	-0.8188
Item35	1.0732	-0.1567

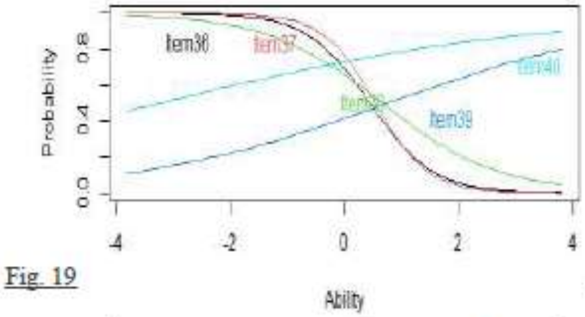


Fig. 19

	b	a
Item36	0.4451	-1.8093
Item37	0.5444	-2.1778
Item38	0.6674	-0.9827
Item39	0.7630	0.4544
Item40	-3.2110	0.3095

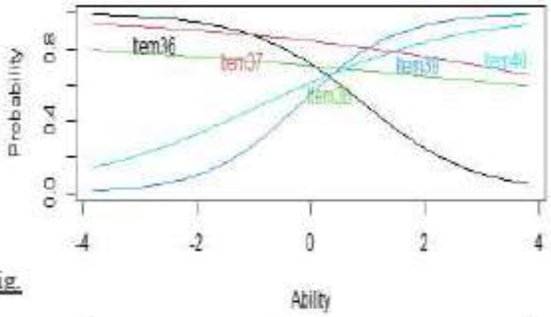


Fig.

	b	a
Item36	0.9150	-1.0188
Item37	6.1638	-0.2754
Item38	6.7568	-0.1301
Item39	-0.1636	1.2109
Item40	-0.7790	0.5910

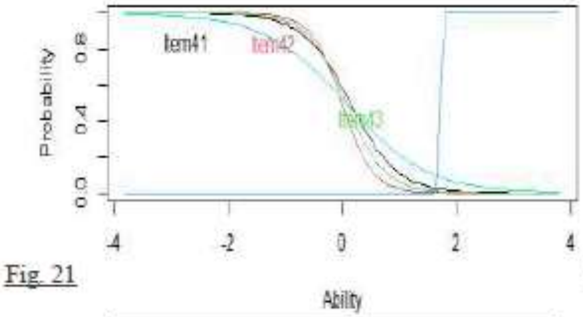


Fig. 21

	b	a
Item41	0.1434	-2.4166
Item42	-0.0017	-3.7338
Item43	0.0680	-2.9464
Item44	1.7174	76.4585
Item45	0.0668	-1.4042

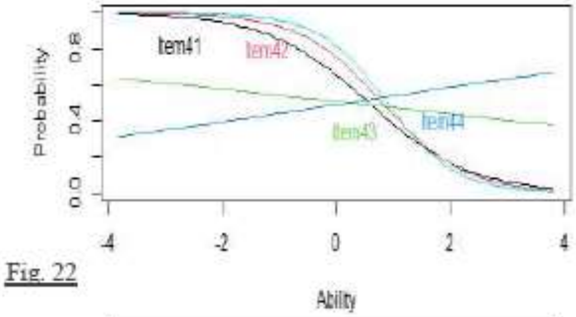


Fig. 22

	b	a
Item41	0.5720	-1.1239
Item42	0.8153	-1.3805
Item43	0.2142	-0.1394
Item44	0.2156	0.1954
Item45	0.8931	-1.6697

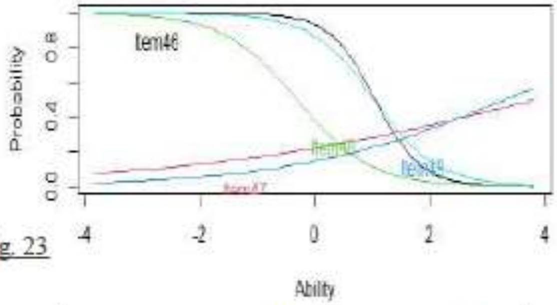


Fig. 23

	b	a
Item46	1.0501	-2.5052
Item47	3.8474	0.3283
Item48	-0.3094	-1.5640
Item49	3.3003	0.5383
Item50	1.0327	-1.8095

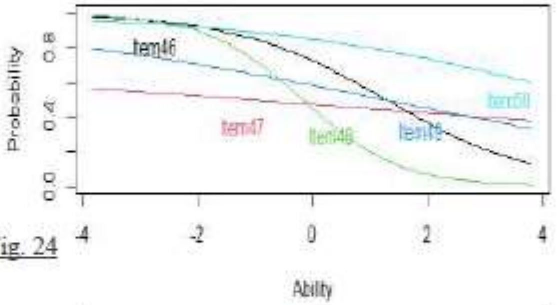


Fig. 24

	b	a
Item46	1.2786	-0.7583
Item47	-1.1255	-0.1000
Item48	-0.1722	-1.1831
Item49	1.2377	-0.2685
Item50	4.9023	-0.3534

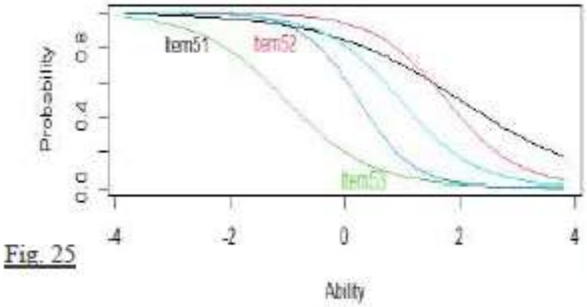


Fig. 25

	b	a
Item51	1.9905	-0.8593
Item52	1.7596	-1.5925
Item53	-1.0347	-1.3511
Item54	0.2327	-2.0985
Item55	0.9611	-1.5355

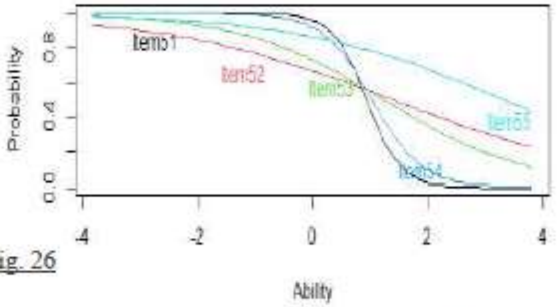


Fig. 26

	b	a
Item51	0.9386	-3.3830
Item52	1.3853	-0.5036
Item53	1.2367	-0.8015
Item54	1.0342	-2.3057
Item55	3.3211	-0.5540

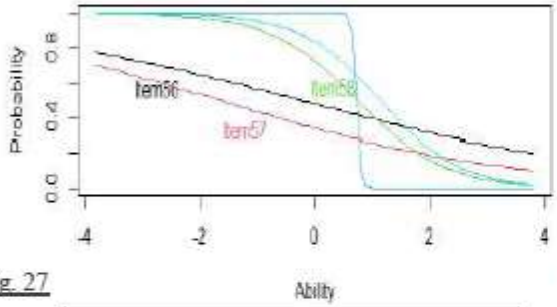


Fig. 27

	b	a
Item56	-0.2612	-0.3450
Item57	-1.6744	-0.3977
Item58	0.7382	-1.3095
Item59	0.7282	-27.4169
Item60	1.1323	-1.4370

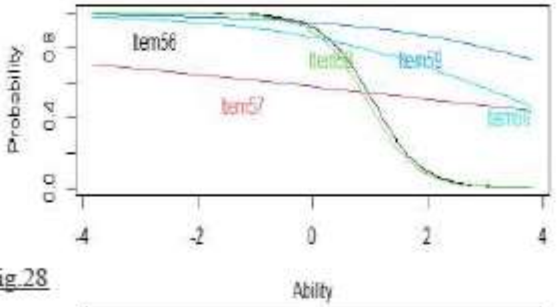


Fig.28

	b	a
Item56	1.0343	-2.3049
Item57	2.0447	-0.1475
Item58	0.9293	-2.2586
Item59	5.8535	-0.4877
Item60	3.5042	-0.5193

Comparison of DIF detection of the six methods used.

ITEMS	Angoff's Delta method(T.I.D.)	Standardization method	Raju's method	Mantel-Haenszel method	Lord's method	Logistic regression method	Number of methods that flagged DIF on the item
Item 1	0.7365	-0.2778 ***	0.1645	0.6583	0.0041 **	0.3015	2 out of 6
Item 2	-1.0945	0.0278	0.1248	0.0000 ***	0.0002 ***	0.3694	2 out of 6
Item 3	-0.4511	0.1667 **	0.4224	0.8852	0.1252	0.2688	1 out of 6
Item 4	-2.2609 ***	-0.1667 *	0.2960	1.0000	0.0086 **	0.6108	3 out of 6
Item 5	0.6310	-0.5556 ***	0.1555	0.0067 **	0.0000 ***	0.0060 **	4 out of 6
Item 6	0.4082	-0.0694 *	0.1488	0.8124	0.0579 .	0.6671	1 out of 6
Item 7	-2.0775 ***	-0.6389 ***	0.9376	0.0021 **	0.1193	0.0018 **	4 out of 6
Item 8	-0.3857	-0.1806 **	0.5009	0.0896 .	0.2241	0.0014 **	2 out of 6
Item 9	1.0387	0.2639 ***	0.3199	0.3408	0.0141 *	0.1875	2 out of 6
Item 10	0.0753	-0.1528 **	0.7607	0.2432	0.7800	0.2927	1 out of 6
Item 11	0.8749	0.3333 ***	0.1486	0.1198	0.0102 *	0.0781 .	2 out of 6
Item 12	- 1.2976	-0.4167 ***	0.6782	0.0973 .	0.1683	0.0177 *	2 out of 6
Item 13	0.8740	0.3056 ***	0.1777	0.5754	0.0241 *	0.7897	2 out of 6
Item 14	- 0.9311	0.0139	0.4410	0.2801	0.6621	0.0085 **	1 out of 6
Item 15	0.5261	-0.2500 ***	0.6029	0.2626	0.0005 ***	0.1165	2 out of 6
Item 16	0.1472	0.0556 *	0.3557	0.8897	0.4121	0.6742	1 out of 6
Item 17	-0.6310	-0.3056 ***	0.8116	0.0412 *	0.6336	0.0044 **	3 out of 6
Item 18	0.7105	0.1389 **	0.2610	0.1416	0.0028 **	0.0326 *	3 out of 6
Item 19	0.7355	0.0972*	0.4697	0.3020	0.3433	0.1273	1 out of 6
Item 20	2.2681 ***	0.3056 ***	0.5872	0.1093	0.0021 **	0.0049 **	4 out of 6
Item 21	-1.5874 ***	-0.2083 ***	0.5285	0.2561	0.1572	0.4626	2 out of 6
Item 22	- 0.0481	- 0.0278	0.8114	1.0000	0.1489	0.3219	None out of 6
Item 23	0.8855	0.0694 *	0.6180	0.7794	0.5260	0.4691	1 out of 6
Item 24	- 0.1238	-0.1250 **	0.7563	0.7548	0.8426	0.3415	1 out of 6
Item 25	0.3571	0.1389 **	0.4512	0.3367	0.0824 .	0.3294	1 out of 6
Item 26	- 0.0976	0.1111 **	0.3834	0.1530	0.1031	0.6185	1 out of 6
Item 27	1.1750	0.3056 ***	0.9945	0.0231 *	0.9340	0.0072 **	3 out of 6
Item 28	- 0.3731	0.1806 **	0.9934	0.4404	0.2633	0.1551	1 out of 6
Item 29	-2.0384 ***	0.0000	0.9696	0.6276	0.9991	0.4071	1 out of 6
Item 30	0.5130	0.1250 **	0.9987	0.1637	0.8983	0.0144 *	2 out of 6
Item 31	- 0.0481	0.1250 **	0.9952	0.4404	0.9989	0.2825	2 out of 6
Item 32	- 0.7810	- 0.0694 *	0.6151	0.6069	0.0240 *	0.0853 .	2 out of 6
Item 33	-0.0288	-0.3333 ***	0.0983	0.0272 *	0.0002 ***	0.0013 **	4 out of 6
Item 34	1.3333	0.2639 ***	0.2233	0.9049	0.5828	0.7149	1 out of 6
Item 35	0.5739	- 0.2639 ***	0.5782	0.4098	0.0460 *	0.3958	2 out of 6
Item 36	0.7105	0.3889 ***	0.2602	0.1042	0.0460 *	0.5397	2 out of 6
Item 37	- 1.3868	-0.1111 **	0.6393	0.6434	0.0963 .	0.1528	1 out of 6
Item 38	0.2026	0.1250 **	0.7629	1.0000	0.2177	0.7404	1 out of 6
Item 39	0.4299	-0.4167 ***	0.2298	0.3711	0.0043 **	0.2755	2 out of 6
Item 40	1.5153 ***	-0.0417 .	0.8762	0.6774	0.8506	0.5093	1 out of 6
Item 41	0.8855	0.3056 ***	0.1259	0.1182	0.0277 *	0.1075	2 out of 6
Item 42	- 0.0346	0.3472 ***	0.1847	0.2002	0.0608 .	0.8455	1 out of 6
Item 43	1.3389	0.1944 **	0.7008	1.0000	0.0365 *	0.0777 .	2 out of 6
Item 44	-1.6080 ***	-0.5139 ***	0.9977	0.0229 *	1.0000	0.0002 ***	4 out of 6

Item 45	-0.0574	0.2361 ***	0.6139	0.7705	0.2344	0.7929	1 out of 6
Item 46	1.3780	0.2222 ***	0.2399	0.0820 .	0.0514 .	0.0066 **	2 out of 6
Item 47	0.6484	-0.4722 ***	0.7360	0.0592 .	0.1682	0.3078	1 out of 6
Item 48	1.9421 ***	0.4028 ***	0.0886	0.3239	0.0005 ***	0.4295	3 out of 6
Item 49	- 0.6777	-0.3889 ***	0.3458	0.0237 *	0.2996	0.0432 *	3 out of 6
Item 50	- 0.7331	0.1667 **	0.5863	0.6276	0.1700	0.5107	2 out of 6
Item 51	0.8896	0.2917 ***	0.5596	0.1273	0.2262	0.0300 *	2 out of 6
Item 52	2.3782 ***	0.2917 ***	0.3087	0.1000	0.1037	0.0146 *	3 out of 6
Item 53	-1.1970	0.2361 ***	0.2763	0.2530	0.3317	0.0521 .	1 out of 6
Item 54	-0.5990	-0.0694 *	0.6232	0.8711	0.2358	0.6811	1 out of 6
Item 55	- 0.9074	-0.0694 *	0.5086	0.8875	0.2680	0.3300	1 out of 6
Item 56	- 0.7470	-0.1250 **	0.4681	0.8197	0.4123	0.4813	1 out of 6
Item 57	0.2716	-0.1806 **	0.7409	0.5040	0.7562	0.6198	1 out of 6
Item 58	0.3571	0.3056 ***	0.8145	0.0910	0.0746 .	0.0579 .	1 out of 6
Item 59	-2.4621 ***	-0.0833 *	0.9648	0.6276	0.9512	0.0552 .	2 out of 6
Item 60	-0.7331	-0.0278	0.5322	0.4292	0.2709	0.8713	None out of 6
Detection threshold	1.5	-0.1 and 0.1	-1.96 and 1.96 (significance level: 0.05)	3.8415 (significance level: 0.05)	5.9915 (significance level: 0.05)	5.9915 (significance level: 0.05)	
Signif. Codes	'***' if item is flagged as DIF	0 ' ' 0.04 '.' 0.05 '*' 0.1 '**' 0.2 '***' 1	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	
Number of DIF items detected	10	54	0	8	18	16	

RESULTS

The results are presented in graphical forms (ogives shown in fig. 5 to fig. 28) and in tabular form, (table 1). The ICCs of each item shows that each item functioned differently in the two groups of respondents. The individual DIF detection methods also showed differing statistics and p-values as well. Among the 60 items investigated using these methods only two of them (items and 60) that passed through all the methods without being detected as DIF items.

DISCUSSION

The investigation was on general DIF detection and not the type of DIF. The ICCs of fig. 5 to fig. 28 clearly showed that no two groups are exactly the same psychologically and this accounts for the different shapes of an item's curves. Although there is no datum to judge what shape showed acceptable DIF but the difference in an item's ogives from the two groups of respondents is clear. Each of the DIF detection methods has its own algorithm it uses to flag an item as DIF items. In summary, standardization method flagged 54 out of 60 items, Lord's chi-square method flagged 18 items, logistic regression method flagged 16 items, Mantel–Haenszel method flagged 8 items, and Raju's area method flagged none.

CONCLUSION

Based on the finding, the following conclusion was made;

Test items will always function differently between two groups of individuals irrespective of their similar characteristics, although the magnitude of the difference may not be up to the point of flagging the item as DIF with respect to criterion used. The standardization method detected the most items followed by Lord's chi-square, and then logistic regression method. The Angoff's TID method came fourth, followed by Mantel-Haenszel method, and then Raju's area method. On this premise, the following recommendations were made:

1. Triangulation of standardization, logistic regression and lord's chi-square methods should be used for DIF detection analyses to enhance precise identification of DIF items.
2. Items of a test should be developed with the consideration of the characteristics of all the individuals the test is designed to examine.

REFERENCE

1. *Abdullah, A. A. (2017). The Impact of Unbalanced Designs on the Performance of Parametric and Nonparametric DIF procedures: A comparison of Mantel Haenszel, Logistic Regression, SIBEST, and IRTLR Procedures. Electronic Theses, Treatises and Dissertations. The Graduate school. Florida State University Libraries.*
2. *Ado, A. B., Rahimah, E. Rohaya, T., Sakinah S., & Adbullah, B. I. (2019). Comparative Analysis of Classical Test Theory and Item Response Theory Using Chemistry Test Data. International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Volume-8, Issue-5c, May 2017 India.*
3. *Aguerri, M. E., Galibert, M. S., Attorresi, H. F; & Maranon, p. p. (2009). Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test. Quality and Quantity, 43, 35-44. doi. 10.1007//s 11135-007-91302.*
4. *Andrew, A. (n.d.). Introduction to Item Response Theory. Psy 427. Cal state Northridge. www.esun.edu/~ata20315/psy427/Topic08_IntroIRT.ppt.*
5. *Angoff, W. H. & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 2, 95-106. doi: 10.1111/j.1745.3984.1973.tb00787x.*
6. *Baker, B. F. (2001). The Basics of Item Response Theory. ERIC Clearing house on Assessment and Evaluation, ISBN 1-886047-03-0 USA.*
7. *Barton, M. A., & Lord, F. M. (1981). An upper asymptote for three-parameter logistic item response model (Research Report No. RR81-20), Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tbo1255.x.*
8. *Breslow, N. E. & Day, N. E. (1980). Statistical Methods in cancer research.: The analysis of case-control studies scientific publication No. 32. International Agency for Research on cancer, Lyon France.*
9. *Carlo, M. (2009). Demonstrating the Difference Between Classical test theory and Item Response Theory usig Derived test data. The International Journal of Educational and Psychological Assessment. April 2009. Vol. 1, Issue 1, pp. 1-11.*
10. *Chalmers, R. P. (2018). Improving the crossing SIBTEST statistic for detecting non-uniform DIF. Psychometrica B3(2) 376-386. doi. 10.1007/s11336-017-9583-8.*
11. *David, M., Sabastien, B., Francis, T., & Paul, D. B. (2010). A general framework and an R package for the detection of dichotomous differential item function. Behavioral Research Methods 2010. 42(3), 847-862 doi:10, 3758/BRM.42.3.847.*
12. *De Ayala, R. J. (2009). Theory and Practice of Item Response Theory. Oxford University Press.*
13. *Demars, C. (2010). Item Response Theory. Oxford University Press.*
14. *Dorans, N. J. & Kullick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368. doi: 10.1111/j1745-3984.1986.tb00255x.*
15. *Hambelton, R. K. Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newburg Park. CA: Sage Press.*
16. *Holland, P. W. & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Ed). Test validity, Hillsdale, NJ: Lawrence Erlbaum Associates.*
17. *Hossien, K. (2012). An Intoduction to Differential Item Functioning. The International Journal of Educational and Psychological Assessment. September 2012, Vol. 11(2).*
18. *Li, H. H., & Stout, W. (1996). A new procedure for detection of cross DIF. Psychometrika, 61, 647-677.*
19. *Lord, F. (1980). Application of Item Response Theory to practical testing problems. Hillsdale, NJ, Lawrence Erlbaum Associates.*
20. *Mens, H. (2019). Differential item functioning for polytomous response items using hierarchical generalized linear model. [PDF] Differential Item functioning for Ploytomous …-D-Scholarship@Pitt-d-scholarship.pitt.edu>MengHua.*
21. *Paula, E., & Craig S. W., (2013). Detecting DIF in polytomous items using MACS IRT and Ordinal Logistic Regression. Psicologica (2015), 34, 327-342.*
22. *Penfield, R.D. (2003). Application of the Breslow-Day test of trend in odd ratio heterogeneity to the detection of non-uniform DIF. Alberta Journal Educational Research, 49, 231-243.*
23. *Philip, E; Keith J. Barrett F., & Shannon, N. (2019). Classical test theory and item reponse theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism. Physical Review Physics Education Research 15, 100102. (2019). Department of physics, Montana state University, Bozeman, Montana 59715, USA.*
24. *Raju, N. S. (1990). Determining the Significance of estimated signed and unsigned areas between two item response functions. Applied psychology measurement. 14, 197-207. doi: 10.1177//01466216900/400208.*

25. *Scott, W. W. (2011). Differential item functioning procedure for polytomous items when examinee sample sizes are small. PhD (Doctor of Philosophy) thesis. University of lowa, 2011. https://doi.org/10.17077/etd.1th8r87.*
26. *Shealy, R. & stout, W. (1993). A model-based standardization approach that separated true bias/DIF from group ability differences and detect test bias/DIF from group abilit differences and detect test bias/DIF as well as bias/DIF. Psychometrika, 58, 159-194. doi: 10.1007/BF02294572.*
27. *Stata 14, (2015). Item response theory. Reference Manual Release 14. Stata Corp., I. P. College Station, Texas. Stata Press Publication.*
28. *Swaminathan, H. & Rogers H. J. (1990). Detecting differential item functioning using logistic regression procedure. Journal of Educational Measurement, 27, 361-370. doi: 10.1111/J.1745-3984.1990.tb00754x.*
29. *Thissen, D., Steinberg, L. & Wainer, H (1988). Use of item response theory in the study group difference in trace lines in H. Wainer and H. Braun (Eds). Test validity. Hillsdate, NJ Lawrence Erlbaum associates.*
30. *Wokoma, T. A., (2021). Detecting Differential Item Functioning in 2019 BECE Basic Science Multiple choice items Administered in Schools in Rivers State, Nigeria. EPRA International Journal of Multidisciplinary Research. ISSN: 2455-3662. Vol. 7 issue 6.*
31. *Xiaoting (2010). Differential Item Functioning: The consequence of Language, Curriculum, or Culture? Graduate School of Education. University of California, Berkeley.*
32. *Xinming, A. & Yiu-Fai, Y. (2014). Item response theory: what it is and how you can use the IRT Procedure to apply it. Paper SAS364-2014. SAS Institute Inc.*
33. *Yuan-Ling, L., (2015). When can Multiple dimensional item Response Theory (MIRT) Models be a solution for Differential Item Functioning (DIF)? A Monte Carlo Simulation Study. College of Education. University of California, Berkeley*