AFRICAN JOURNAL OF THEORY AND PRACTICE OF EDUCATIONAL ASSESSMENT (AJTPEA)

VOL 8(1), 2019: 1 – 18 Copyright © Educational Assessment and Research Network in Africa (EARNiA). ISSN 2545-5834
http://www.earnia.org; info@earnia.org; earnia13@gmail.com

COMPARATIVE ANALYSIS OF ITEM RESPONSE THEORY METHODS OF DETECTING DIFFERENTIAL ITEM FUNCTIONING IN EDUCATIONAL ASSESSMENT

Mfonobong E. Umobong & Udeme E. Tommy

Abstract

This study comparatively analysed item response theory methods of detecting Differential Item Functioning (DIF) in educational assessment. Three research questions were formulated to give direction to the study. The population comprised 32,710 senior secondary II students in the three senatorial districts of Akwa Ibom State, Nigeria from which a sample size of 3,271 students were selected through multistage sampling procedure. Data were collected using a 50-item Chemistry Achievement Test with a reliability index of 0.72 which was estimated using the Kuder-Richardson 20 formula. IRT methods of likelihood ratio test, Lord's chi-square test and Rasch b-parameter were employed in the analyses. The results revealed that likelihood ratio test had 16 items that exhibited gender DIF, Lord's chi-square test had 18 items while Rasch b-parameter had 21 items that exhibited gender DIF. The three IRT methods detected 5 common items that exhibited gender DIF. The results also showed that the IRT methods were all valid because they were able to detect a considerable number of items that exhibited gender DIF. A comparison of the IRT methods of detecting DIF showed that many items exhibited distinctive significant gender DIF based on the three methods adopted. It was recommended that DIF analysis should be incorporated in educational assessment so as to obtain valid psychometric properties of tests and valid educational assessments for making appropriate educational decisions.

Key words: Item response theory, Differential item functioning, Educational assessment

Introduction

Assessment of educational constructs in schools is made possible through the use of assessment instruments. The cognitive constructs of students, like chemistry cognitive ability can be measured with a chemistry achievement test. Items on a test ought to measure the intended ability irrespective of the parameters of subgroups of students. This is because students with equal abilities should be able to correctly answer an item at the same rate even though they are in different subgroups. If items included in the test provide more advantages for one group over another, such test items are seen as biased (Zumbo, 2015). Therefore, in developing a test, items should be examined in terms of item bias so as to guarantee the validity and reliability of such a test.

Validity and reliability are two properties that all educational instruments must have. American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (2014) claim that validity is the extent to which evidence and theory concur with interpretations derived from test results as captured in the test usage proposal. Hence, a test that is meant for different sub-populations but which gives each sub-population or a particular sub-population different approaches to responding to the test is biased but unfair to all the sub-populations of students. Such a test lacks validity because the test results would need different interpretations for each sub-population of students. Measurement experts label the issue of fairness as "bias" (Kim, 2011; Osterlind, 2013). When a test serves as a subject of interest, "*test bias"* becomes the concern that needs examination but if a single item forms the subject of interest, "*item bias"* becomes the concern that needs to be addressed. Bias in assessment has generated serious, complex and global concern which measurement experts always strive to detect and eliminate.

As empirical foundation for important judgements, tests should not be biased. They should not favour a sub-group. A biased test is made up of systematic errors that negate the validity of the test (Camilli & Shepard, 2014). Differential item functioning (DIF) and bias were interchangeable words, until Holland and Thayer (2009) gave a precise difference between the two concepts by asserting that bias is used to refer to an informed opinion that considered seriously the aim of the test as well as contextual information about groups, which can explain DIF on a given item.

DIF examination forms the beginning statistical procedure that ascertains if a test item is biased against a group or not. An additional concept that one should take into account is "impact" which refers to actual differences in attribute or ability distribution between groups (Herrera, 2008). This is important because if an item is detected by a statistical procedure as having differential item functioning, it indicates that it does not necessarily have bias characteristics. Such situations warrant identification of the reasons for which the groups score differently in the test item in line with the measurement objective. In the

case of impact, the differential item functioning is due to main ability differences (Herrera, 2008).

With the rising concerns to ensure test validity, reliability and fairness, differential item functioning has been well utilized in bias examination. According to Zumbo (2008), differential item functioning is the contemporary method of psychometric bias estimation which many studies have utilized to determine the presence of bias in assessment instruments. Differential item functioning occurs if testees of different populations exhibit diverse rates of responding correctly to a test item after prescribing the particular ability that the test item is designed to assess (Zumbo, 2015). Differential item functioning exists when there is a difference between or among test items and a likelihood of responding correctly to the test item in all strata of the trait being assessed (Embretson & Reise, 2010). DIF exists as a result of testees of a particular sample not being able to respond positively to an item compared to testees of a different sample due to certain parameters inherent in the item or measurement condition which is irrelevant to the measurement objective. In differential item functioning analyses, the abilities of various samples are evaluated based on the test items linked to demographical features like male-female in similar performance grade or students in rural-urban schools (Greer, 2004; Gomez-Benito, 2017).

DIF seems to be a *prima facie* evidence of possible presence of bias in a test. However, the presence of differential item functioning is not necessarily an indication of test bias. Even though, differential item functioning is necessary; it is not a sufficient condition for bias (McNamara & Roever, 2016). Bias occurs only when the source of differential item functioning does not form part of the trait of measurement in the test items. Differential item functioning exists if testees of similar abilities from two separate samples have various probabilities of answering a test item correctly (Clauser & Mazor, 2008). It is similar to statistical bias in which a parameter or several parameters inherent in the statistical model become either underestimated or overestimated (Camilli & Shepard, 2014). Whenever differential item functioning is present in an item, the source(s) of the variance should be investigated to ensure that it is not a case of bias. A test item that indicates differential item functioning is biased, invalid and unreliable only when the source of difference in response to the item is not related to the trait under investigation. Thus, it becomes a source of construct-irrelevant variance and the samples respond differentially to the item due to groups (Messick, 1994).

Two groups (focal and reference groups) are involved in a DIF study. The focal group is that of minorities. That is, the likely disfavoured sample. The sample that seems to be favoured by the test items is the reference group. However, classifying the samples is not usually a clear-cut issue because it is done randomly. DIF are of two kinds, uniform and non-uniform. Uniform DIF exists if a group responds better than the other group on all ability levels. In essence, nearly every member of a group outperforms nearly every

member of the other group who have similar ability levels. In non-uniform DIF, members of one group are at advantage up to a range on the ability continuum and from that point on, the relationships become reversed. In other words, an interaction between groups and ability level exists. Uniform-DIF occurs when the chance of responding to a test item positively by the focal group is greater than that of the reference group in all ability levels. When the chance of responding positively to a test item by the focal group varies from the reference group based on their ability levels, such is regarded as a non-uniform-DIF (Zumbo, 2015). According to Zumbo (2015), uniform DIF exists when the calibrated differences in the chances of getting a test item correctly by examinees in various groups but with similar ability levels is independent of the common ability level. In non-uniform DIF situations, the differences do not remain the same at all the ability levels and crossing item response curves seems to exist.

As earlier pointed out, DIF exists if two samples of similar ability levels exhibit various probabilities for answering a test item positively. A matching criterion is required for grouping the subjects for ability. Matching could be internal and external. In internal matching, the criterion refers to the observed or latent score of the test itself while for external matching, the observed or latent score of another test is taken as the criterion. External matching may pose a problem due to assumption that the supplementary test itself has no bias and assesses similar traits like the test of focus (McNamara & Roever, 2016). Therefore, DIF is an evidence of bias when the variable leading to DIF is irrelevant to the construct underlying the test. If the variable is part of the construct, it is known as impact instead of bias.

It is imperative that tests are not gender biased so as not to give test takers some advantages in answering test items. A review of literature on gender and chemistry achievement revealed inconsistent relationship between gender and chemistry attainment. Most studies found that boys consistently scored higher than girls on a number of indicators of chemistry proficiency while other studies reported that girls consistently outperformed boys (Kimball, 2009; Randhawa, 2014). Fennema (2018) did not report gender differences in the ability to solve chemistry problems. He however, found significant differences in problem-solving strategies. Girls tended to employ "concrete solution strategies like modelling and counting, while boys opted for more abstract solution strategies that reflected conceptual understanding. Leahey and Guo (2018) cautioned against the assertion that there is an evident gender difference in chemistry achievement favouring males. Could these superior performances occur as a result of DIF? Any DIF in chemistry achievement tests, either favouring male or female students, results in bias and therefore invalidates the test. DIF are done irrespective of the selected test items and abilities of examinees responding to the test items. In the words of Hambleton, Swaminathon & Rogers (2011), the likelihood of responding correctly to a test item should depend on the ability level of examinees alone. Apart from ability, other

factors like differential item functioning, lack of unidimensionality and local independence may make testees respond correctly to test items. This makes it necessary to ascertain possible characteristics of test items prior to utilizing such test items through appropriate techniques.

Techniques of checking differential item functioning are classified in line with the classical test theory (CTT) and item response theory (IRT). The major difference between CTT and IRT DIF assessment methods is that in CTT techniques, scales are subjected essentially on observable constructs like total test score and the number of subjects in the focal and reference samples who got the test item correctly or incorrectly (Clauser & Mazor, 2008). That is, the conditioning or the matching criterion is the observed score. With IRT techniques, matching is done according to the examinees' measured ability level or the latent trait which is θ. In the present research, comparing IRT techniques is useful due to IRT's sample-independent characteristic and the power to measure examinees on a similar scale. Such features lead to improved statistical evidence in differential item functioning problems (Hambleton, Swaminathan & Rogers, 2011). IRT orders test items irrespective of examinees and ability parameters as these could never be examined in CTT framework, using many unique statistical and mathematical models.

Item response theory framework is a pragmatic technique because of its mathematical and theoretical make up that make it more informative compared to CTT. Since item response theory serves as a framework that has the ability to explain the linkage amongst examinees' responses, it checks differential item functioning using similar procedure (Greer, 2004). A powerful feature of differential item functioning adopting IRT is the utilization of item response curves or item characteristic curves (Thissen, 2001). When a test item performs differently in focus groups and reference groups, that is, when test items response curves are dissimilar for the different groups, the existence of differential item functioning is certain. In the two groups, item characteristics are examined evaluated for differential item functioning based on IRT methods.

IRT uses three parameters to describe the shape of the item characteristic curve which are item difficulty, item discrimination and guessing parameter. Based on how many of these parameters are involved in the estimation of the relationship between the ability and item response patterns, there are three IRT models, namely 1-, 2- and 3- parameter logistic models. In1-parameter logistic model and the Rasch model, it is assumed that all items have the same discrimination level. The 2-parameter IRT model considers item difficulty and item discrimination. However, guessing is believed to be the same in ability levels. Finally, the 3-parameter model includes a guessing parameter together with item difficulty and item discrimination. The models provide a mathematical equation for the relation of the responses to ability levels (Baker, 2011). Thus, one possible way of detecting DIF through IRT is to compare test item characteristics in two groups. If test

item parameters are significantly different, then DIF is ensured. Likelihood ratio test, Lord's chi-square together with Rasch b-parameter are the IRT methods examined in this study.

In examining DIF with likelihood ratio test, the assumption of no presence of DIF postulated when assessing it, is that "significant difference does not exist between the item parameters estimated for focus and reference samples" Results from the compact model (CM) for assessment of the assumption and the augmented model (AM) are evaluated. Within the CM, the characteristics of every test items in focus and reference groups are required to be similar, that is, no test item is seen as exhibiting DIF. In AM, it is required that characteristics of test items from focus and reference groups can vary, and for the other test items, the characteristics should be just as it occurred in the augmented model.

A likelihood function can be derived from the compact model. Many likelihood functions for the number of test items can also be derived from the augmented model. G^2 is an index gotten through taking the logarithms of the likelihood functions for the compact model and the augmented model (Thissen, 2011). G^2 indicates the χ^2 value. Number of item characteristics is the degree of independence of the distribution. If G^2 index exceeds 3.84 (SD =1; $\alpha = 0.05$) the assumption is rejected and the existence of DIF is ensured for the item as the quantitative index of G^2 value points to the amount of DIF (Thissen, 2011). Considering Cohen's G^2 indices, Greer (2004) classifies DIF into A Level, when $3.84 < G^2 < 9.4$ indicates a negligible amount of DIF; B Level when $9.4 < G^2 < 41.9$, and a moderate amount of DIF exists, and C Level when $G^2 > 41.9$ showing a high amount of DIF.

For Lord's chi-square estimation, the item characteristics of sub-groups and covariance are calculated. Lord's Chi-square indices are derived through calibrated characteristics and covariance indices. DIF is then examined by evaluating the observed and expected indices (Camilli & Shepard, 2014). To determine uniform and non-uniform DIF, Lord (1980) suggested using the χ^2 method based on a suitable item response model (Maij de Meij, Kelderman & Van der Flier, 2010). The technique is used through comparing item characteristics among groups. The χ^2 statistic is calculated with the help of the difference between calculated item parameters and a variance-covariance matrix related to this difference (Camilli & Shepard, 2014). The obtained χ^2 statistic adheres to the chi-square distribution with 1 degree of freedom. If χ^2 index is more than the critical index, the test item contains DIF using a specified alpha level.

The Rasch model is based on the probability of responding correctly to an item by a person. In aiming to model this probability, it only considers person ability and item difficulty. Probability is a mathematical function of the variance of person ability and item difficulty. This is easily done as item difficulty and person ability are measured on the same scale in the Rasch model. The Rasch model holds that any person answering a

test item has a certain level of the trait measured by the item and all items on a test have a level of the trait. These indices work in the opposite directions. Hence, only the variance from item difficulty and person ability comes into play. The model gives item difficulty estimates that do not depend on the groups of examinees involved. Therefore, DIF exists if invariance does not arise in a specific utilization of the model which means the estimates are due to the group that responds to the test items (Engelhard, 2009; Gomez-Benito, 2017). To determine the significant DIF items, (d statistics is comparable to t-value. Meaningful indices of DIF statistics higher than or equal to 1.96 shows DIF in favour of male examinees at .05 alpha level; on the other hand, an index of d less than or equals -1.96 shows DIF in favour of female examinees at .05 level.

Hunter (2015) asserts the chi-square technique has many setbacks. The test items should be unidimensional and have high reliability so that the total test score can be a valid estimate of ability level. Also, the technique has high sensitivity when variances in the total test scores of the groups involved occur (Linn & Kessel, 2015).The likelihood ratio test is highly reliable in assessing uniform and non-uniform DIF items with many other benefits compared to other methods; thus, it assessed separately DIF items as a result of differential item difficulty or item discrimination (Linn & Kessel, 2015). At the same time, uniform and non-uniform DIF were also examined. In the IRT techniques, model-data fit is required (Hunter, 2015). Similarly, likelihood ratio test method is affected by another drawback in which only 2-parameter model, 3-parameter model and Samejima's graded model can be carried out (Linn & Kessel, 2015). To comprehend the absence of congeniality among techniques better, a close assessment of the items flagged as having DIF by just one technique may be done. Rudner (2006) reported that test items assessed as having DIF in Rasch *b*-parameter method were test items where items have the same discriminations but different item difficulties. However, Hunter (2015) and Rudner (2005) suggest that utilization of various methods of detecting DIF would help researchers to be sure of items that need to be flagged since they continuously show DIF in more than a method. In the light of this background, the study comparatively analysed IRT methods of detecting DIF in educational assessment. The specific aims of this research were to:

1. Ascertain the number of chemistry items that functioned differentially based on gender using likelihood ratio test method.
2. Determine the number of chemistry items that functioned differentially based on gender using Lord's chi-square method.
3. Assess the number of chemistry items that functioned differentially based on gender using Rasch *b*-parameter method.
4. Examine the number of common items that functioned differentially based on gender using the three IRT DIF detecting methods.

Research Questions

1. What is the number of chemistry items that functioned differentially based on gender using likelihood ratio test method?
2. What is the number of chemistry items that functioned differentially based on gender using Lord's chi-square method?
3. What is the number of chemistry items that functioned differentially based on gender using Rasch *b*-parameter method?
4. What is the number of common items that functioned differentially based on gender using the three IRT DIF detecting methods?

Methods

A descriptive survey research design that makes use of various research instruments including achievement test for collecting data for accurate and objective description of a situation was used in this study. The research was carried out in Akwa Ibom State, Nigeria. The state is among those classified as educationally advantaged states in the country as many of her citizens are exposed to all levels of education with literacy rate of 78.84% (National Bureau of Statistics, 2017). The population of the study comprised all 32,710 Senior Secondary II (SS II) Students in the 530 secondary schools in the state who wrote the 2018 SS II Chemistry Promotion Examinations in Akwa Ibom State. As of December, 2018 when the sampling was done, there were 198 schools in the urban areas while 332 schools were located in the rural areas. A sample of 3,271 students (10%)of the student population with 1,702 males and 1,569 females was selected for this research using multistage sampling procedure. The 2018 Akwa Ibom State Senior Secondary II Chemistry Promotion Examination was adopted and used for data collection. Though the examination consisted of two sections; the objective and essay sections, only the objective section was used for the study. The objective section consisted of 50 items marked dichotomously as correct (1) or incorrect (0) with each item having three distracters and a correct option. The content validity of the test was ensured through the use of table of specification. Experts in measurement and evaluation also assessed the items in order to ensure that the items were well developed. The internal consistency reliability of the examination was established using Kuder-Richardson 20 method. The high coefficient of 0.72 guaranteed that the examination was reliable and hence it was utilized in conducting the research. The marked scripts of the students were collected from Akwa Ibom State Ministry of Education by the researchers for DIF analysis using likelihood ratio test, Lords' chi-square test and the Rasch b-parameter as IRT methods of detecting DIF. Male students (group 1) were used as the reference group while female students (group 2) were used as the focal group.

Results

The collected data were analysed with the likelihood ratio test, Lords' chi-square and the Rasch b-parameter techniques. IRTLRDIF (Thissen, 2011) software was used for the likelihood ratio test analysis. The BILOG MG was used for the Lord's chi-square while the Winsteps Rasch software was employed for Rasch b-parameter analysis. The results of the analyses are presented below.

Research Question 1: What is the number of chemistry items that functioned differentially based on gender using likelihood ratio test method?

G^2 values obtained from the results of gender DIF analysis using IRTLRDIF likelihood ratio test were used to answer this research question. G^2 statistic is a chi square index obtained at 1 degree of freedom and an alpha level of 0.05. According to Thissen (2011), items with G^2 values greater than 3.84 are functioning differentially. However, Greer (2004) categorized the DIF effect size as A level when $3.84 < G^2 < 9.4$ indicates that a negligible level of DIF occurs; B level when $9.4 < G^2 < 41.9$ and a moderate amount of DIF occurs; C level when $G^2 < 41.9$ and a large amount of DIF occurs. The result of the gender DIF analysis based on IRT likelihood ratio test is shown in Table 1.

Table 1: Result of gender DIF analysis obtained with G^2 values of IRT likelihood ratio test

Item	G^2	Item	G^2	Item	G^2	Item	G^2	Item	G^2
1	22.5	11	0.0	21	0.9	31	0.0	41	0.0
2	3.7	12	0.0	22	3.2	32	0.0	42	5.0
3	1.8	13	2.5	23	0.0	33	0.3	43	4.1
4	4.2	14	1.6	24	0.1	34	21.2	44	0.0
5	0.5	15	3.3	25	0.0	35	3.0	45	0.0
6	8.5	16	4.5	26	1.8	36	4.7	46	2.7
7	5.4	17	0.9	27	12.1	37	0.0	47	0.0
8	6.2	18	3.8	28	2.2	38	2.2	48	6.9
9	0.7	19	11.6	29	2.3	39	3.5	49	0.3
10	0.0	20	13.9	30	0.0	40	7.2	50	21.7

In Table 1, the 34 items with their G^2 ranging from 0.0 to 3.8, that is, below 3.84, shows no gender DIF. The 34 items (2, 3, 5, 9, 10, 11, 12, 13, 14, 15, 17, 18, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 35, 37, 38, 39, 41, 44, 45, 46, 47 and 49) represent 68% of the 50 items while16 items (1, 4, 6, 7, 8, 16, 19, 20, 27, 34, 36, 40, 42, 43, 48, 50)representing 32% of the 50 items have their G^2 values above 3.84 to exhibit gender DIF. Ten (10) items with G^2 values ranging from 3.85to 9.4 representing 20% of the 50 items have negligible amounts of DIF. Items with moderate amount of DIF are 6with G^2 values ranging from 9.5 to 41.9; they constitute 12% of the 50 items. No item exhibited large amounts of DIF and constitute 0% of the 50 items. The significance of the G^2values indicates that the items exhibit uniform and non-uniform DIF, and favour the male students.

Research Question 2: What is the number of chemistry items that functioned differentially based on gender using Lord's chi-square method?

The chi-square values (χ^2) obtained from the analysis of gender DIF using Lord's chi-square BILOG MG was adopted in answering the second research question. To determine uniform and non-uniform DIF, the method compares the item parameters of the samples. The derived χ^2 index adheres to the chi-square distribution with 1 degree of freedom. If the χ^2 index is greater than the critical value, the test item contains DIF dependent on the prescribed alpha level. The χ^2 values for the items are presented in Table 2.

Table 2: χ^2 statistics obtained from gender DIF analysis based on Lord's χ^2 of BILOG MG with their associated probability (p) value

Item	χ^2	p-value	Item	χ^2	p-value
1	6.4	0.376	26	5.6	0.0962
2	6.3	0.455	27	5.7	0.0571
3	30.5	0.0011	28	1.6	0.4351
4	14.0	0.0009	29	0.1	0.9595
5	1.1	0.5645	30	7.4	0.0247
6	0.2	0.9111	31	5.8	0.0560
7	4.2	0.1209	32	4.9	0.0855
8	4.9	0.0850	33	3.5	0.1757
9	2.7	0.2574	34	7.6	0.0001
10	10.5	0.0053	35	10.4	0.0055
11	23.3	0.0001	36	4.3	0.1165
12	0.3	0.8543	37	9.3	0.0095
13	1.6	0.4453	38	16.1	0.0469
14	28.3	0.0001	39	2.4	0.3032
15	0.6	0.7346	40	7.8	0.0203
16	20.7	0.0001	41	10.7	0.0047
17	12.1	0.0024	42	8.0	0.0182
18	0.5	0.7878	43	0.1	0.9636
19	0.9	0.6528	44	16.1	0.0003
20	5.3	0.0706	45	1.7	0.4243
21	3.3	0.1885	46	7.4	0.0241
22	0.1	0.9589	47	1.4	0.5084
23	3.7	0.1602	48	3.0	0.2212
24	9.5	0.0086	49	2.8	0.2449
25	0.2	0.9136	50	3.0	0.2200

Results in Table 2 indicate that the χ^2 values obtained ranged from 0.1 to 47.6. The 32 items representing 64% of the 50 items with χ^2 values at probability levels above .05 are not significant. Their χ^2values ranges from 0.1 to 6.5 and includes items 1, 2, 5, 6, 7, 8, 9, 12, 13, 15, 18, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29, 31, 32, 33, 36, 39, 43, 45, 47, 48, 49 and 50. Chi-square values obtained at probability levels below 0.05 have significant DIF and ranges from 7.4 to 47.6. They are 18 in number (3, 4, 10, 11, 14, 16, 17, 24, 30, 34, 35, 37, 38, 40, 41, 42, 44, 46) representing 36% of the 50 items. These items showed uniform and non-uniform DIF, favouring the male students. Thus, 36% of the items function differentially with respect to gender while 64% of the items do not function differentially.

Research Question 3: What is the number of chemistry items that functioned differentially based on gender using Rasch *b*-parameter method?

To answer research question three, the item difficulty indices from the Rasch model obtained for the male and female groups were used. Difficulty (*b*) parameter was used because it tends to give stable indices compared to other item characteristics (Linacre, 2010). To determine the significant DIF items, DIF statistics (*d*) is comparable to *t*-value. Meaningful indices of DIF statistics higher than or equals 1.96 shows DIF in favour of male examinees at .05 alpha level; on the other hand, an index of d less than or equals -1.96 shows DIF in favour of female examinees at.05 level. The result is indicated in Table 3 below.

Table 3 : Result of gender DIF analysis using the Rasch *b*-parameter of Winsteps Rasch Software

Item	Group 2-1 DIF value	SE	*d* statistics	Item	Group 2-1 DIF value	SE	*d* statistics
1.	-0.25	0.10	-0.36	26.	-0.20	0.09	2.22
2.	-0.24	0.08	0.02	27.	0.21	0.10	2.10
3.	-0.41	0.08	0.00	28.	0.12	0.10	1.20
4.	0.22	0.09	0.07	29.	-0.02	0.09	0.22
5.	0.06	0.09	0.06	30.	0.20	0.09	2.22
6.	-0.07	0.09	-0.21	31.	0.14	0.09	1.56
7.	0.17	0.09	0.22	32.	0.14	0.09	1.56
8.	-0.29	0.09	-3.22	33.	-0.19	0.09	-2.11
9.	0.12	0.09	1.33	34.	0.32	0.09	3.56
10.	-0.24	0.08	3.00	35.	-0.27	0.09	-3.00
11.	0.19	0.09	2.11	36.	0.02	0.09	0.22
12.	0.03	0.09	0.33	37.	0.19	0.09	2.11
13.	0.06	0.09	0.67	38.	-0.23	0.09	-2.56
14.	-0.34	0.09	-3.78	39.	-0.11	0.09	-1.22
15.	0.03	0.09	0.33	40.	0.21	0.09	2.33
16.	0.21	0.09	2.33	41.	0.24	0.09	2.7
17.	-0.12	0.09	-1.33	42.	0.21	0.09	2.33
18.	-0.08	0.09	0.89	43.	0-0.05	0.09	0.56
19.	0.03	0.09	0.33	44.	0.02	0.09	0.22
20.	0.17	0.09	1.89	45.	-0.10	0.09	-1.11
21.	-0.17	0.09	-1.89	46.	0.23	0.09	2.56
22.	0.01	0.09	0.11	47.	-0.17	0.10	-1.70
23.	0.03	0.09	0.33	48.	-0.09	0.08	-1.13
24.	0.164	0.09	1.82	49.	-0.11	0.09	-1.22
25.	-0.04	0.09	0.44	50.	0.07	0.09	0.78

The result in Table 3 reveals that gender DIF values ranges from -0.41 to 0.32. Twenty eight items representing 56% of the 50 items have positive values while 22 items representing 44% have negative values. The lower the index of *b*-parameter for a sample the easier the test item for that particular sample. All items with negative DIF value imply that female examinees outperformed male examinees on the items. On the other hand, items with positive DIF value indicate that male examinees outperform female examinees. These items also exhibit uniform and non-uniform DIF. Therefore, 56% of these items were easier for the males while 44% were less difficult for the female examinees. The *b*-parameter values signifying DIF in favour of male examinees range from 0.02 to 0.32 while the *b*-parameter values signifying DIF in favour of female examinees range from -0.41 to -0.04. Twenty one items (1, 2, 3, 4, 8, 10, 11, 14, 16, 26, 27, 30, 33, 34, 35, 37, 38, 40, 41, 42, and 46) or 42% of these 50 test items reveal DIF. Differential functioning items in favour of males were 11 representing 22% of the 50test items. Differential functioning items in favour of female examinees were 10 representing

20% of the 50 items. Therefore 42% of the items functioned differentially for the students with respect to gender.

Research Question 4

What is the number of common items that functioned differentially based on gender using the three IRT DIF detecting methods?

The various IRT approaches for identifying DIF are similar in that they adopt latent ability index as a matching variable. One of the strengths of IRT methods is the ICC which provides a vivid graphical display in the uniform and non-uniform DIF. However, some drawbacks of IRT methods as Clauser and Mazor (2008) pointed out are that the data should satisfy the assumptions of the IRT model selected, and generally large samples are needed. Thus, IRT approaches are considered to be more time-consuming than other approaches. It is also difficult to examine the significance of the tests through indices like the area between ICCs (Swaminathan & Rogers, 2010). The common items that functioned differentially based on gender using the three IRT DIF detecting methods are indicated in Table 4.

Table 4 : Common items that functioned differentially based on gender using the three IRT DIF detecting methods

IRT method	No of DIF items detected	% of items detected	Common DIF items detected
Likelihood ratio test	16	32%	
Lord's chi square test	18	36%	4, 16, 34, 40, 42
Rasch b-parameter	21	42%	

The results in Table 4 reveals a comparison of gender DIF based on the three IRT DIF detecting methods namely, likelihood ratio test, Lord's chi-square test and Rasch b-parameter. Likelihood ratio test method identified 16 items (32%) as exhibiting DIF; Lord's chi-square method identified 18 items (36%) as exhibiting DIF while Rasch *b*-parameter method identified 21 items (42%) of the 50 items exhibiting DIF. However, 5 items (10%) are commonly detected to exhibit gender DIF by the three IRT methods.

Discussion of Findings

This study examines whether the chemistry achievement test items display gender DIF using IRT techniques. Derived estimates revealed the IRT methods as giving varying findings with only 10% of common items functioning differentially by gender, based on the three methods. The result of Research Question One indicated that 16 items exhibited uniform and non-uniform DIF and favoured the male students using the likelihood ratio test method. The likelihood ratio test was highly reliable in assessing a considerable

number of uniform and non-uniform items that functioned differentially because it assessed differential item functioning through differential item difficulty and item discrimination. This finding is similar to that of Linn and Kessel (2015) who found that the likelihood ratio test was good in estimating uniform and non-uniform differential functioning items through differential item difficulty as well as discrimination. The result of Research Question Two showed that using the Lord's chi-square method, 18 items exhibited uniform and non-uniform DIF and favoured the male students. A good proportion of gender differentially functioning items were also detected using the Lord's chi-square method as a result of the test items being unidimensional and having high reliability. The total test scores were valid estimates of the examinees' ability levels. Also, the technique has high sensitivity of detecting DIF when variances in the total test scores of the groups involved occur. Hunter (2015) reported in his study that the chi-square technique is very sensitive to unidimensional and reliable items. The result of Research Question Three revealed that 21 items showed uniform and non-uniform DIF and favoured the male students more using the Rasch *b*-parameter method. Similarly, it seems that the Rasch *b*-parameter method is very reliable in detecting a large number of DIF since it only focuses on only one parameter which is item difficulty. Rudner (2006) reported in his study that test items assessed as having DIF in Rasch *b*-parameter method were test items which have the same discriminations but different item difficulties. The result of Research Question Four indicated that 5 items (10%) were commonly detected to exhibit gender DIF by the three IRT methods. This number showed that the three methods were not congenial in detecting DIF among the items. The major reason for this low detection seems to arise from the fact that the indices derived from these techniques are different, even though model-data fit is met in the methods. Linn and Kessel (2015) reported theoretical reasons for lack of congeniality among pairs of techniques in examining DIF. They explained many reasons for labelling a test item as having DIF even though there is lack of DIF. They include lack of unidimensionality of test items, variance in ability levels of the groups, dissimilarity in test items' quality, guessing and absence of linearity of regression. The present finding assists to elucidate the low and moderate congeniality found in measurement write-ups on DIF techniques relating to test items identified to have DIF.

Findings from this study revealed that the three techniques identified majority of the test items to have DIF in favour of male examinees. Chemistry items having DIF in favour of males were found to involve real world references. Chemistry items indicating DIF in favour of female examinees involved items that deal with relations and functions. The content specifications of chemistry items indicating DIF in favour of male students, deal with equations. The test was anchored on a particular curricular seem to assist female students' performance. Even though male students seemed to be favoured by the chemistry achievement test, females generally seem to make better classroom grades compared to males (Linn & Kessel, 2015). In order to illustrate this inconsistency,

Kimball (2009) stated that male and female students possess distinct learning styles; females adopt more of rules learned in class, while males utilize autonomous styles which permit them to generalize knowledge to cover unknown challenges. Therefore, it is expedient upon females to perform best in test items that are related to classroom instructions. Such a high female advantage in relations and functions which is seen in DIF indices has never been noted previously. Indeed, this research gives empirical support that gender distinctions in performance on test items in chemistry exists and they vary based on contents even if contents are much linked to curricular.

Conclusion

This study comparatively analysed item response theory techniques of detecting DIF in educational assessment. These results indicated significant difference in likelihood ratio test, Lord's chi-square and Rasch *b*-parameter methods in assessing DIF in educational assessment. This finding points to the fact that researchers should not use only one method in detecting DIF. Using just a technique in assessing DIF and completely depending on derived estimates from one technique is not likely to be justified considering the numerous issues explained. Items flagged with DIF may not be the problem but rather only a symptom of differences by factors such as instruction. Moreover, with cognitive tests, although one main dimension is defined, the variation of different specific ability or skills measured by different test items could be another factor causing the DIFs. This, however, does not support the notion that different constructs are measured.

Recommendations

1. It becomes imperative that researchers, test experts and developers should utilise many methods in DIF examination. They should have a robust mechanism that does not include test items seen to show DIF through more than a technique in a test since they are likely to be justifiably disregarded.

2. Test experts and developers are encouraged to adopt item bias assessment techniques, especially IRT methods in DIF examination of educational, behavioural and psychological instruments. Test developers and users should examine the effect of differentially functioning test items on respective subjects' test scores as well as those of the groups taking the test. Test items that function differentially influence the validity of the test for the various groups and examinees for which they are meant. Therefore, it must be ensured that examinees are not favoured by the presence of differentially functioning items in tests.

3. Other sources of validity evidence to support score comparability such as dimensionality of test items should be investigated since DIF analysis alone cannot be relied upon for this purpose. Educational, behavioural and psychological tests ought not to be influenced by other properties apart from subjects' abilities; they should not be biased in favour of any of the groups involved. This makes a case for objective considerations of the basic rules of measurement in all kinds of testing.

References

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

Baker, F. B. (2011). The basics of item responses theory. Retrieved January 20, 2009, from http://eric.ed.gov/ERICDocs/data

Camilli, G. & Shepard, L. A. (2014). *Methods for Identifying Biased Test items*. California: Sage Publications.

Clauser, B. E. & Mazor, K. M. (2008). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Embretson, S. & Reise, D. (2010). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179-197.

Engelhard, G. (2009). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, 12, 29-51.

Fennema, E. (2018). New perspectives on gender differences in *chemistry*: An introduction. *Educational Researcher, 27*, 4-5.

Gómez-Benito, J. (2017). Detecting differential item functioning in behavioural indicators across parallel forms. *Psicothema, 29* (1), 91-95.

Greer, T. G. (2004). *Detection of differential item functioning (DIF) on the SATV: A comparison of four methods: Mantel-Haenszel, logistic regression, simultaneous item bias and likelihood ratio test.* Unpublished doctoral dissertation, University of Houston.

Hambleton, R. K., H. Swaminathan & Rogers, J. H. (2011) *Fundamentals of item response hierarchical generalized linear model: A comparison study with logistic regression procedure.* Unpublished doctoral dissertation, Pennsylvania State University.

Herrera, A. N. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity, 42,* 739-755.

Holland, P. W. & Thayer, D. T. (2009). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 12-145). Hillsdale, NJ: Lawrence Erlbaum and Associates, Inc.

Hunter, J. E. (2015). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education conference on test bias. Maryland.

Kim, M. (2011). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*, 89–114.

Kimball, M. (2009). A new perspective on women's achievement. *Psychological Bulletin*, 105, 198-214.

Leahey, E., & Guo, G. (2018). Gender differences in chemistry trajectories. *Social Forces, 80*, 713-732.

Linacre, J. M. (2010) Winsteps® Computer Software (Version 3.70.0). Beaverton, Oregon: Winsteps.com.

Linn, M. C. & Kessel, C. (2005). Participation in mathematics courses and careers: Climate, grades, and entrance examination scores. Paper presented the annual meeting of the American Educational Research Association, San Francisco.

Lord, M. F. (1980). *Applications of item response theory to pratical testing problems*. Lawrence Erlbaum Associates, Broadway, Hillsdile, NJ.

Maij-de Meij, A. M., Kelderman, H. & Van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research, 45*, 975-999.

McNamara, T. & Roever, C. (2016). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

National Bureau of Statistics (2017). *Data on index of Nigeria literacy*. Abuja: National Bureau Statistics.

Osterlind, S. J. (2013). *Test Item Bias*. Beverly Hills: Sage Publications.

Randhawa, B. S. (2014). Gender differences in academic achievement: A closer look at chemistry. *The Alberta Journal of Educational Research*, 17, 241-257.

Rogers, J. & Swaminathan, H. (2016). Concepts and methods in research on differential item functioning of tests items. Past, present, and future. In C.S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement* (pp. 126-142). New York: Guilford Press.

Rudner, L. M. (2006). Item and format bias and appropriateness. *Journal of Educational Measurement*, 17 (1), 143-165.

Swaminathan, H. & Rogers, H. J. (2010). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Thissen, D. (2011). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Retrieved February 10, 2008, from http://www.unc.edu/~dthissen/dl.html.

Zumbo, B. D. (2008). Statistical Methods for Investigating Item Bias in Self-Report Measures, Università degli Studi di Firenze E-prints Archive, Florence, Italy.

Zumbo, B. D. (2015). A methodology for Zumbo's third generation DIF Analysis and the Ecology of Item Responding. *Language Assessment Quarterly, 12* (1), 136-151.