International Journal of Education, Learning and Development
Vol.6, No.3, pp.80-91, March 2018
Published by European Centre for Research Training and Development UK (www.eajournals.org)

USE OF DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSIS FOR BIAS ANALYSIS IN TEST CONSTRUCTION

Iweka Fidelis (Ph.D)

Department of Educational Psychology, Guidance and Counselling, University of Port Harcourt, Rivers State, Nigeria.

ABSTRACT: *The study used differential item functioning (DIF) analysis to determine item bias in a chemistry aptitude test (CAT). A total of 270 items were administered, in paper and pencil format for item analysis purposes. Items were analysed by means of classical item analysis, IRT item analysis and DIF analysis in particular. The sample size was 554 secondary school students from SS1 and SS3. The study focused on selected secondary schools in Yoruba speaking states of the western part of Nigeria, Hausa speaking states in the northern part of Nigeria and Igbo speaking states in the eastern Nigeria. The sample was divided into subgroups which was used to investigate the DIF for the level of Education, gender, language and culture. This was used to indentify items that indicated bias in terms of gender, culture, language or level of education. Items that exceeded a predetermined amount of DIF were discarded from that final item bank, irrespective of which subgroup was being advantaged or disadvantaged. The process and results of the DIF analysis were discussed.*

KEYWORDS: Differential Item Functioning, Item Response Theory, Item Characteristic Curves, Classical Test Theory

INTRODUCTION

Item response theory (IRT) is also known as probabilistic theory since it deals with probability of possible response to a test item. Hambleton and Jones (2013) defined item response theory as a general statistical theory about examine item and test performance and how performance relates to the abilities that are measured by the items in the test. It is a psychometric theory and family of associated mathematical models that relate latent traits of interest to the probability of response to item on the assessment. A correct response depends on both the characteristics of probability of a correct response expressed as a mathematical function of the examinee ability and item characteristics - also known as the item characteristic curve (ICC). The item characteristics curve (ICC) is the primary concept in IRT, and it is a mathematical expression that connects or links a subject's probability of success on an item to the trait measured by the set of test items. It is a non-linear (logistic) regression line, with item performance regressed on examinee ability. For dichotomously scored test items (i.e, binary items scored "O" or "I") logistic functions are used to model the probability of "success"

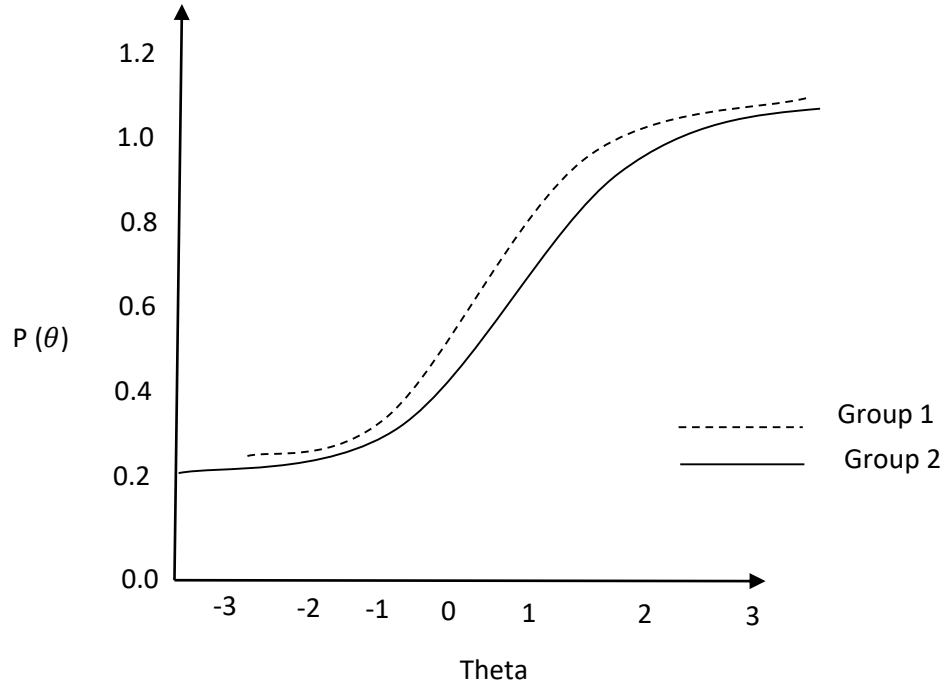
Assessing Item Bias

The way in which IRT-based ICCs are used to evaluate DIF is to compare ICCs of two groups (Osterlind, 2003). Various considerations make it extremely difficult to give one fixed magnitude at which an item should be considered bias or DIF, Visual inspection of the form of DIF, together with the magnitude of the area between the graphs of the two groups compared is usually combined to determine whether an item should be flagged as biased.

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

A distinction is made between DIF determined by different positions of location and discrimination indexes and DIF determined by varying positions of locations and same discrimination. A distriction is also made between uniform DIF and non-uniform DIF. In uniform DIF, the probability of answering an item correctly for one group is consistently lower than that of the other group. This results in the ICC for one group being below that of the other group over the entire ability range as shown in figure 2.

Figure 2: item showing uniform DIF



One of the most useful features of IRT is that the examinee's estimated ability level and item difficulty level are put on the same scale. This allows for the illustration of item difficulty and item discrimination simultaneously using ICC graphs to depict the characteristics of each item. This method provides a powerful base for assessing differential item functioning (item bias) by also using visual inspection. In IRT terms, the "overall notion is that the item characteristic curves generated for each of the two contrasting groups should be alike for an item to be considered unbiased (Mellenberg, 2004). The use of ICCS for DIF detection concerns the comparison of differences in the ICCS for different subgroups. Only two groups can be compared at a time, but a particular sample can be divided into various subgroups for such comparisons.

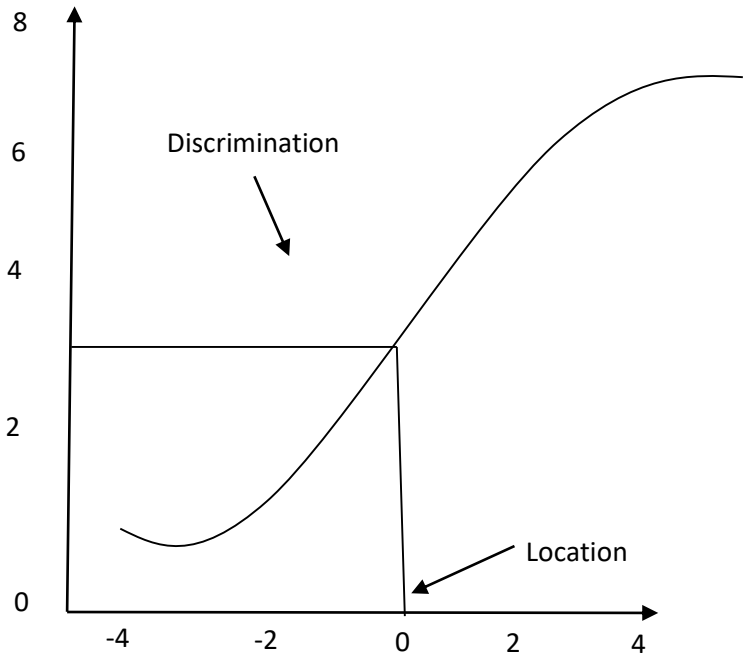
The area between the equated ICCS is an indication of the degree of bias present in a considered test item (Reeve, 2003). (ie, a "I" vs a "O"). The logistic function specifics a monotonically increasing function, such that higher ability results in a higher probability of success.

The item response function is plotted with the ability level of the examinees along the X-axis, against the probability of answering an item correctly on the Y-axis. Each examinee is considered to have an ability score which places him or her somewhere on the ability scale. An examinees ability is denoted by the Greek letter, the a (θ). At each ability level there is a certain probability that an examinee with that ability will answer the item correctly. This probability

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

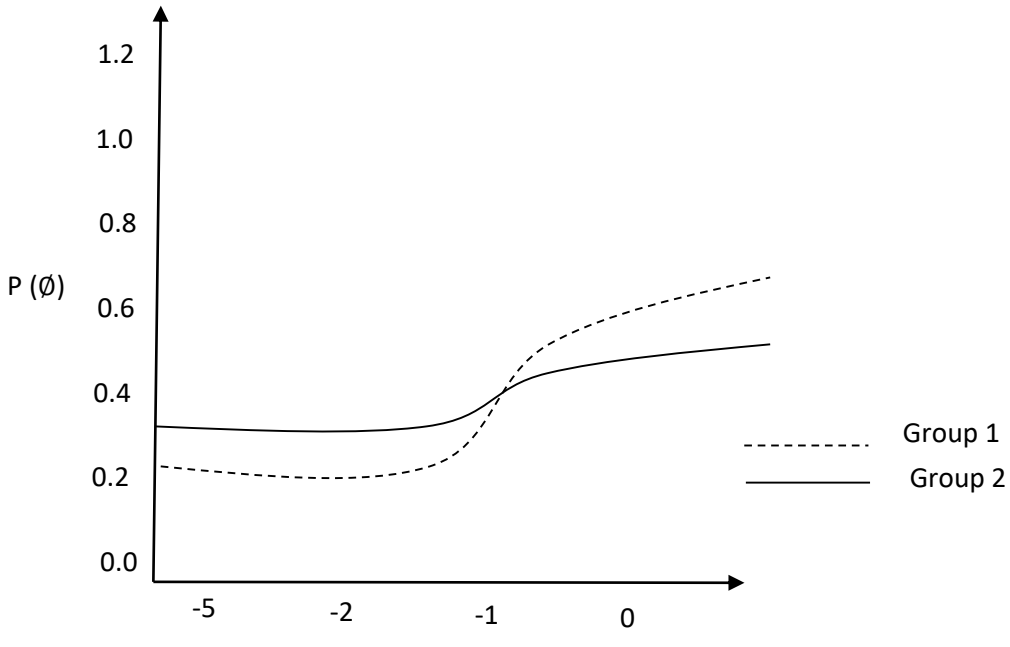
is indicated by P (θ). In typical items, this probability is smaller for those with higher ability levels. Therefore, if the probability function P(θ) is plotted against ability level, the result is the typical S-shaped form of the ICC as in figure 1.

An item characteristic curve determined by items location and discrimination.



In non-uniform DIF, the curves cross at a certain point. Whereas for one range of ability the one group has a lower probability of answering the item correctly, the reverse is true of another range of ability.

Figure 3: item showing non-uniform DIF



Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

The ideal is that there should be little difference between the ICCS of the two groups being compared as shown in figure 4

Figure 4: item showing no DIF

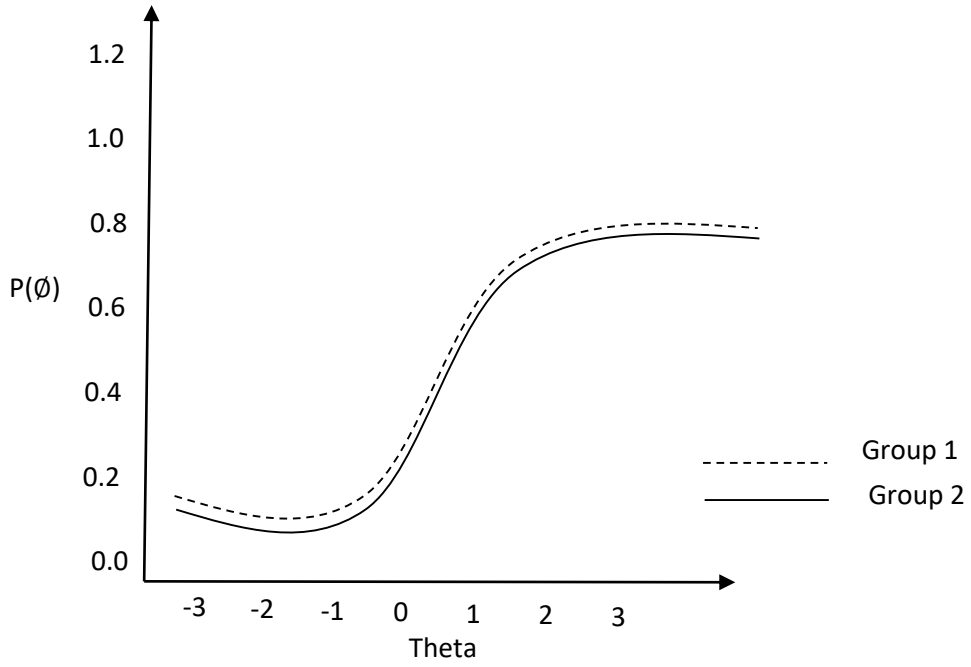
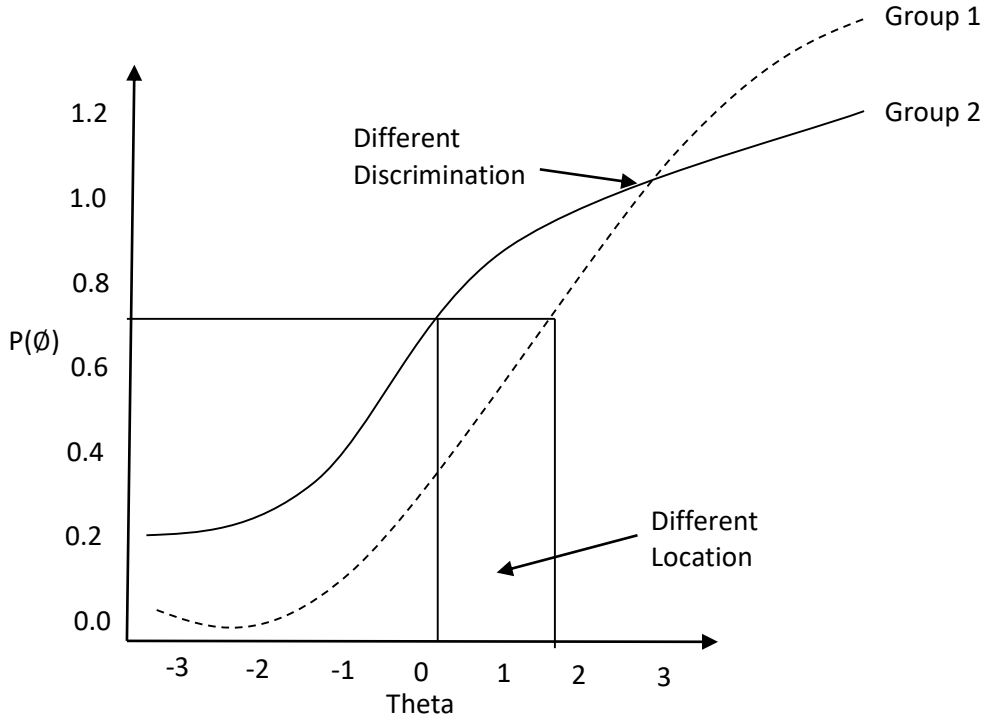
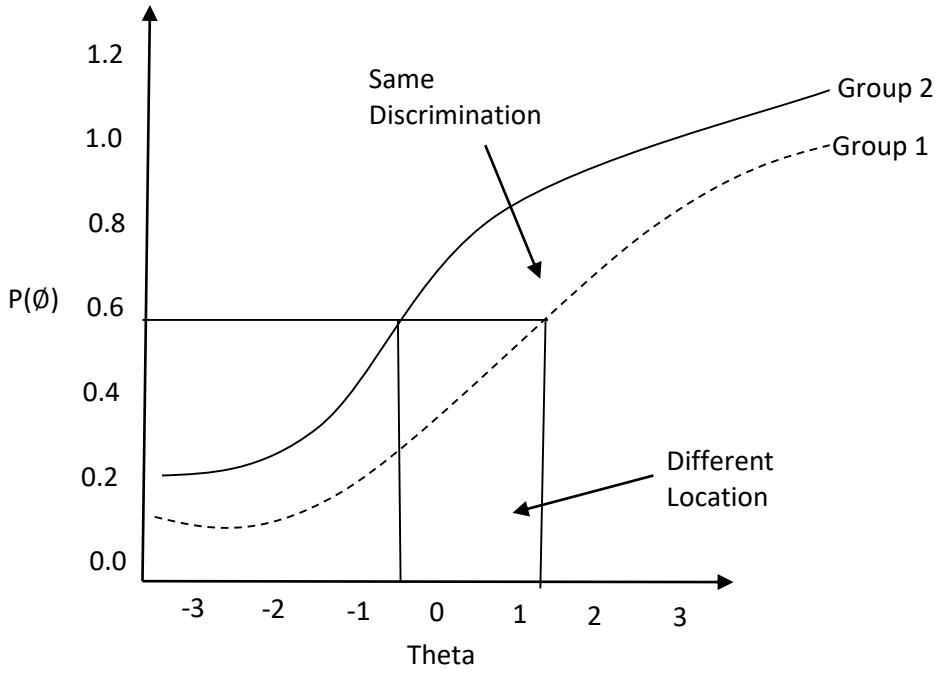


Figure 5: Shows the effect of different positions of location and discrimination indexes of test item curves for different groups.



Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

Figure 6: Shows the effect of varying positions of locations and same discrimination



METHODOLOGY

In DIF analysis, the examinee group of interest is referred to as the focal group, while the group with which its performance on the item is being compared is called the reference group. After calculating the IRT item parameters separately for the two groups, the theta scales are equated (Magis, 2007). The ICCS can then be drawn on the same graph and compared for DIF. If a test item has exactly the same item response function for each group, persons at any given level of ability will have exactly the same probability of getting the item right. This would be true even though one group may have a lower mean theta, and thus lower test scores than the other group (Lord, 1990). The basic approach to the measurement of DIF therefore lies in the difference between the probability of getting an item correct if one is a member of one (Focal) group, in contrast with what would have been the probability of a correct response if one were a member of the other (reference) group. An important precondition in DIF is that only examinees with the same ability level are compared with another (Khalid & Glas, 2013).

According to Khalid & Glas (2013), the aim of research into item bias is not simply to provide guidelines for identifying and eliminating apparently bias items, but also to identify variables or factors that may be responsible for bias with respect to specific groups. For the current research, the intention was to investigate whether some types of items were more susceptible to DIF by taking into consideration how many items of three item formats used, showed DIF and which group was advantaged in those items that were identified as indicating DIF.

Comparison groups based on level of education, gender, culture and language were used. The three – parameter IRT model was used to analyze items. Items were analyzed in terms of classical test theory criteria, IRT item parameter requirements as well as DIF.

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

Sample

DIF analysis was conducted using a sample of two 554 secondary school students from SS1 and SS3. The samples were large enough to analyze items by means of the three – parameter IRT model, which is best for analyzing multiple – choice items (McBride, 1997).

Fort – one schools were selected. These included 15 schools from the eastern Nigeria, and 14 from western Nigeria. The schools had been identified on a random basis, taking into account the Urban and rural distribution and the sizes of the school populations.

At each school, 60 students, 30 from S.S. 1 and 30 from S.S. 3 were randomly selected for testing. Furthermore, in each class group of 30 students, half the examinees were boys and half girls. In each class sample group of 30, form A and form B of the test were alternated, thereby ensuring an equal distribution of the two forms between both the gender and the class groups.

Scope: This study focused on selected secondary schools in Yeroba speaking states in the western past of Nigeria, Hayusa speaking states in the Northern Nigeria and Ibo speaking states in the Eastern Nigeria. Imo State and Abia State were randomly selected from the Eastern states, Kano and Kasina from the Northern State while Oyo and Ekiti States were sampled from the western states. The sample was divided into subgroups which was used to investigate the DIF for level of Education, gender, language and culture.

Measuring Instruments

The study concerned the investigation of item bias or DIF and the items used were constructed for the development of chemistry Aptitudes test (learning potentials of chemistry test) CAT. A total of 270 items were constricted. 90 each of the item types of figure series, figure analogies and pattern completion respectively was used. These items measure fluid ability by means of general non-verbal reasoning. The items were aimed at the average to lower – ability levels, although an attempt was made to have items of each of the three types available at all ability levels. SPSS computer software was used to conduct factor analysis of the piloted test scores in order to verify the face, content and construct validity of the test items, that is, the unidimensionality of the items.

The number of items that needed to be administered was too large to administer to examinees in a single test session. This necessitated the construction of two paper and pencil forms with sufficient anchor items – items answered by both groups – to calculate IRT item parameters on the same scale. A group of 66 anchor items (22 from each of the three item types) was used so that the IRT item parameters for the entire pool of items could be put on the same scale-despite the fact that the items were administered two forms.

These items were administered in paper and pencil format for item analysis purposes. Items were analysed by means of classical item analysis, IRT item analysis and DIF analysis in particular. Information from all three approaches was used in the selected of items for the final test item banks.

Procedure

Items were administered in two groups each containing the 66 anchor items. Anchor items are used to transform IRT item parameters for the total group of items to the same scale. For

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

classical item analysis, the items in the two forms were used separately, with the result that for the anchor items two sets of item parameters were available.

Classical test theory item analysis was done with CITAS (CITAS, 2015) (Classical item and test Analysis spreadsheet). This is a straight forward Excel work book that provides basic analysis of testing results based on classical test theory. While the three – parameter IRT item analysis was done with X caliber 4 (X-calibre 4, 2015), a Widow based soft ware of making test results' analysis based on IRT theory. This was used to calculate the item parameters for the total group thereafter, the total sample was divided into various subgroups to investigate DIF for level of education, gender, language groups and cultural groups.

Statistical Analysis

The CITAS program provides the classical test theory difficulty (F) and discrimination (ra) values for items while X caliber 4 soft ware provides the IRT difficulty (b), discrimination and guessing (c) parameters based on the three – parameter model.

Furthermore, the area between the ICC graphs for the different comparison groups was calculated to provide indices for the magnitude of DIF for both uniform and non uniform DIF items. In the case of non-uniform DIF, the two separate areas were added together despite the fact that they separated opposing patterns of which group was advantaged at different levels (Iweka, 2005).

RESULTS

The IRT parameters for the total pool of items are given in table 2

Table 2: Descriptive statistics of the parameters of the items subjected to IRT analysis

IRT parameters	N	Mean	Sd	Min	max
a-value	265	1.435	0.486	0.442	2.500
b-value	265	-0.231	0.829	-1.558	3.000
c-value	265	0.179	0.0853	0.000	0.470

Five of the 270 items were discarded during IRT item analysis in the analysis of the items, particular attention was paid to DIF analysis results for the following comparison graphs:

- Education groups S. S. 1 versus S. S. 3
- Gender groups Male versus female
- Religious groups Christians versus Muslims.
- Language groups Igbo versus Yoruba

The education group comparison was included since the chemistry Aptitude test (CAT) was intended to measure learning potential and formal level of education should not affect the general reasoning performance measured.

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

Based on the descriptive values of the resulting areas between ICC graphs for all items for the four comparison groups respectively (see table 3), as well as visual inspection, a cutoff of 0.5 was determined for flagging items as DIF. This is also in line with values used by other researchers (Kanjee & van Eeden, 1998). An item was flagged as showing DIF, if on any one or more of the four indices obtained from the DIF analysis of the four comparison groups, the magnitude of the area between the two graphs exceed the value of 0.5.

Table 3: Descriptive Statistics in Respect of the DIF areas between ICCS for different comparison groups all items included.

DIF Comparison groups	N	Mean	SD	Min	Max
Grade groups (S.S. 1 versus S.S. 3)	265	0.1789	0.1471	0.0025	1.2338
Gender groups (male versus female)	265	0.1672	0.1616	0.0089	1.4375
Religious groups (Christianity versus Islam)	265	0.3307	0.2081	0.0254	1.4050
Language groups	265	0.2336	0.1570	0.0083	0.9762

Igbo versus Yoruba).

Using the total group of items developed, on average more DIF was evident in respect of the religious and language groups than for the gender and education groups. Of the 270 items that were analyzed, 82, were discarded on the basis of various factors, five items were discarded during the IRT analysis and another 77 on the strength of the psychometric and DIF results. Items were discarded on the basis of the following criteria:

- IRT: c-values :C> 0.3
- IRT: a-values :a< 0.80
- CTT: rit < unless IRT a> 1.0
- DIF area between the ICC of any of the four DIF comparison groups > 0.5.
- The remaining 188 items which met the criteria set for inclusion, were included in the final test.
- Table 4 provides a summary of the items that were discarded and reason for this.

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

Table 4: items discarded and reasons for their being discarded

procedure	Figure series	Figure analogies	Pattern completion	Total
Item analysis (IRT & CTT)	17	15	15	47
Bias analysis	8	17	10	35
Total	25	32	25	82
Rejection categories				
IRT a<0.8	4	4	10	18
IRT C>0.3	11	6	0	17
Education DIF	0	0	1	1
Gender DIF	0	0	0	0
Religious DIF	4	12	6	22
Language DIF	2	0	1	3
DIF for 2^+ groups	2	4	2	8
IRT and DIF	2	4	2	8

The pattern of DIF indicates that more figure analogy items (N=32) were flagged as indicating DIF than for the other two item types (figure series (N=25) and pattern completion (N=25). Considering only the discarded figure analogy items, more than one third showed DIF when the religious groups were compared. Despite the content of the items having been chosen because it was considered to be the least subject to religious influence, religion nevertheless seemed to play a major role in DIF. It should be kept in mind that the direction of DIF is not considered here – and from tables 5 and 6 it can be seen that in respect of the discarded items, more favoured the Christians than the Muslims. In tables 5 and 6, summaries of the direction of DIF for included and discarded items are provided. There are some surprising results in respect of the discarded items. Firstly in the school class group category, more discarded items favoured the S.S. 1 group than the those favouring the S.S.3 group. In the gender comparison, in the discarded items, more favoured the female than the male group while in the language group comparison more of the discarded items favoured the Igbo than Yoruba language groups, these patterns were somewhat reversed in the group of items that were included (not discarded) although one should keep in mind that the amount of DIF for these items was not large enough to be flagged as DIF to merit being discarded.

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

Table 5: Direction of DIF for included items (N=188)

Item types and DIF categories	Figure series	Figure analysis	Pattern completion	Total
Garde group DIF				
Little or no DIF	43	24	33	108
Favouring S.S.1	7	13	14	34
Favouring S.S.3	12	15	6	33
Mixed	3	6	3	12
Gender group DIF				
Little or no DIF	35	38	35	108
Favouring male	11	7	11	29
Favouring female	11	7	14	32
Mixed	8	6	5	19
Religious group DIF				
Little or no DIF	14	8	15	37
Favouring Christians	17	13	23	53
Favouring muslems	25	23	16	64
Mixed	9	14	11	34
Language group DIF				
Little or no DIF	23	26	20	69
Favouring Igbo	15	9	16	40
Favouring Yoruba	15	14	21	50
Mixed	12	9	8	29

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

Table 6: Direction of DIF for Discarded items (N = 77)

Item types and DIF Categories	Figure series	Figure analogies	Pattern completion	Total
Education group DIF				
Littlr or no DIF	13	12	8	33
Favouring S.S.1	8	8	6	22
Favouring SS3	1	5	4	01
Mixed	3	5	4	12
Gender group DIF				
Little or no DIF	13	22	11	46
Favouring males	2	4	1	7
Favouring female	8	2	8	18
Mixed	2	2	2	6
Religious group DIF				
Little or no DIF	3	2	4	8
Favouring Christian	7	15	14	36
Favouring Muslims	9	12	2	23
Mixed	6	1	3	10
Language group DIF				
Little or no DIF	9	7	3	19
Favouring Igbo	7	8	11	26
Favouring Yoruba	4	8	5	17
Mixed	5	7	3	15

DISCUSSION

IRT – based DIF analysis provides useful information for assessing items. The use of this technique will probably increase over time as more researchers become familiar with it. Overtime, this will probably change as more researchers start using these techniques.

The fact that approximately one third (N=82) of the original l270 items were eventually discarded, based on their psychometric properties or because they showed more than the determined level of DIF, is in line with the general international findings (McBride, 1997). Researchers and test developers need to make use of the available techniques for DIF analysis to ensure compliance with the requirements set by the employment equity act. The IRT based DIF methods provide a useful visual representation of the bias which is easily understood. The information obtained in this manner can also be put to good use to identify patterns of bias, which can again be used as input in future test and test item development.

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300

It is important to appreciate the fact that our ultimate interest should lie with the quality of decisions based on the scores obtained from tests since it is at this level where individual lives are affected – fairly or unfairly.

REFERENCES

Hambleton, and Jones S. C. (2013). Item response theory models and testing practices: current international status and future directions. European Journal of Psychological Assessment, 13(1), 21-28.

International Congress, Bellingham, 4 Augst 1998.

Iweka, F. O. E. (2005). Using structural equation modeling to detect response shifts and true change quality of life research, 14, 857-598.

Kanjee, A. & Van Eeden, R. (1998). Item response theory and measurement equivalence in personality assessment of a South Africa sample. Paper presented at the XIVTH LACCP.

Khalid, G. C. & Glas, G. (2013). Multicultural assessment: How suitable are written test. European Journal of Psychological Assessment, 14 (1), 61.

Lord, F. M. (1990). Application of item response theory to practical testing problems. Hillsdale, N. T: Lawrence Erbaum Associates.

Magis, G. C. (2007). Cultural differences, politics and test bias in Nigeria, European Review of Applied Psychology 47(4) 297-307.

McBride, J. R. (1997). Technical perspective in W. A. Stands, B. K. Waters & J. R. McBride (Eds), Computerized adaptive testing: from inquiry to operation (pp.29-44) Washington, DC: American Psychological Association.

Mellenberg, G. J. (2004) Item bias and Item Response Theory. International journal of Educational Research, 13,127-142.

Osterlind, S. J. (2003). Test item bias. Beverly Hills: Sage.

Reeve, E. (2003) Towards an in cross-cultural assessment. European Journal of Psychological Assessment, 13(1), 29-37.

Print ISSN: ISSN 2054-6297, Online ISSN: ISSN 2054-6300