International Review of Social Sciences and Humanities
Vol. 7, No. 1 (2014), pp. 52-65
www.irssh.com
ISSN 2248-9010 (Online), ISSN 2250-0715 (Print)

Differential Item Functioning: The Implication for Educational Testing in Nigeria

A.D.E. Obinne
(Corresponding Author)
Department of Educational Foundations and General Studies
University of Agriculture, Makurdi, Benue State
ZIP 23401, Postal 2373, Nigeria
E-mail: amatheldaya@yahoo.com

A.O. Amali
Department of Educational Foundations and General Studies
University of Agriculture, Makurdi, Benue State
ZIP 23401, Postal 2373, Nigeria
Email: agbogoamali@gmail.com

(Received: 11-9-13 / Accepted: 20-11-13)

Abstract
A test is suppose to measure student's/examinees' ability/ performance or other traits irrespective of certain factors like gender, ethnicity, geographical locality, social status and others. By IRT standards, test items should not depend on the characteristics of the sample.
DIF refers to the difference in the statistical properties of an item between groups of equal ability. This study specifically determined the differential item functioning (dif) of biology examination test items administered by West African Examination Council (WAEC) and those administered by the national examination council (NECO) for the years 2000 and 2001. Senior secondary (second year) students in three education zones of Benue state were used for the study. The multi-stage stratified sampling technique was used to select one thousand eight hundred students for the study. The data collected were analyzed using the transformed difficulty technique of the *biolo-mg* computer programme and the t-test. The results of the analysis showed that some of the items in the examinations functioned differently which indicated the existence of dif effects thus measuring what they were not supposed to measure. It was recommended that the IRT system of item analysis be adopted by examination bodies in Nigeria.

Keywords: Testing, DIF effect, invariant, parameter estimate, examinees ability.

Introduction

Testing is done to determine whether or not an objective or goal has been obtained. In other words, testing concerns specific achievement of a student in terms of a given objective (Ogunniyi, 1984).

Testing consists of a set of uniform questions or tasks to which a student is to respond independently and the result of which can be treated in such a way as to provide a quantitative comparison of the performance in different students (Nworgu, 1992). As seen by Ebel (1979), test is any kind of device or procedure for measuring ability, achievement, interest and other traits.

From the description of test and testing above, a test is supposed to measure student's/examinees' ability/performance or other traits irrespective of certain factors like gender, ethnicity, geographical locality, social status and others.

By the use of Item Response Theory (IRT) framework, in the analysis of test items, psychometricians have found that some items in a test may have functions different from what the test is meant for. It means that such items have interactions with the characteristics of the sample (examinees/students) taking the test. This describes such items as having differential functions.

Differential Item Functioning (DIF), as defined by Angoff (1975), refers to the differences in the statistical properties of an item between groups of equal ability. The question is do items function in different ways for different groups of test-takers? Item functioning is intended to be invariant with respect to irrelevant aspects of the test-takers, such as gender, ethnicity and socio-economic status. Item functioning is expected to be altered by interventions targeted at those items, for instance, the use of calculators in arithmetic tests or the use of assistive device on mobility tests (Badia, Prieto and Linacre, 2002).

DIF investigates the items in a test, one at a time, for sings of interactions with sample characteristics. In the widely used Mantel-Haenszel procedure of detecting DIF (www.rasch.org/memo39.htm), reference and focal groups are identified which differ in a discernible way. These groups are stratified into matching ability levels and their relative performance on each item is quantified. The ability levels are usually determined by the total scores on the test. In this way, the DIF analysis for one item is as independent as possible of the DIF analysis of the other items. The presence of DIF may have serious consequences for the interpretation of test scores for both groups and individuals. According to Zumbo (1999), frequently, examination items are considered biased because they contain sources of difficulty that are not relevant to the construct being measured and these extraneous sources affect test-takers' performance as those items will bring about differential functions. It has been found through researches that differences in test of students achievement in some subjects like mathematics and Science subjects could be attributed to social and cultural influences which create sex role stereotypes that further reduces female achievement and interest in traditionally male-dominated subjects (Ogbebor & Onuka, 2013).

Current study evidence according to Ogbebor & Onuka (2013) implicated "ode" used in national and regional examinations as functioning differently with respect to subgroups. This only implies that whatever a student scores from such examination is based on the group he belongs not his ability. The overall impact of item DIF, accumulates across the whole test. By implication, test developers should try as much as possible to construct test items that may have minimal DIF effect.

There are several methods to detect if items have DIF effects. Some of the methods include the standard mean difference (SMD) techniques, Generalized Mantel-Haenszel (GMH) methods, chi-squares techniques, analysis of variance methods and methods of comparing plots of transformed item difficulties, factor analysis methods, correlation, logistic regression and methods based on experimental manipulations.

Item response theory (IRT) techniques are theoretically preferred procedures for detecting DIF because they least confound real mean differences in group performance with bias (Lord,

1977). They also offer a more robust solution under both uniform and non-uniform conditions (Ogbebor & Onuka, 2013).

Differential Item Functioning (DIF) in IRT has been defined as the amount of area between two item characteristic curves. It is present when the simultaneous test of the equality of parameters is rejected. it sets out to measure if examinees of equal abilities, but from different groups, have an unequal probability of answering the items correctly.

There are several advantages of using the IRT approach in testing DIF effects. IRT approaches represent an improvement over the classical approaches in latent trait parameter invariance. With the traditional approach, changes in the examinee sample yield unpredictable differences in the item statistics.

A second advantage is that item response theory is less likely to artificially label an item as biased, unlikely in the CTT approach where a large p-value difference and item by group interaction may label an item as biased when in fact no bias exists. Green & Draper (1972) were the first researchers to introduce the use of ICC for measuring bias. In their work they used test scores instead of ability estimates in obtaining the ICC. Several researchers have also used ICC in estimating item bias investigations (Lord, 1977, Rudner, 1977; Wright 1976; Ironson & Sukoviak, 1979).

Statement of the Problem

A test is supposed to measure students/examinees ability/performance or other traits of interest irrespective of certain factors such as gender, ethnicity, geographical location, social status and others. In other words, a test item by IRT standards is supposed to be invariant in nature. This is not always the case for psychometricans have often found some test items to have interactions with the characteristics of the sample (examinee/students). Why is this so? This study sought to find out the differential functioning (DIF) of the test items administered by the West African Examination Council (WAEC) and the National Examination Council (NECO).

The public opinion about the test items by these two examination bodies necessitated this study. The public (Nigerians) believe that WAEC test items are of better quality than the NECO test items because more students score high grades in NECO examinations than WAEC examination.

Objective of the Study

This study specifically determined the differential item functioning of Biology examination test items administered by West African Examination Council (WAEC) and those administered by the National Examination Council (NECO).

Research Questions

1. How do the test items of the Biology examinations conducted by NECO function with respect to sex (boys and girls)?
2. How do the test items of the Biology examinations conducted by WAEC function with respect to sex (boys and girls)?

Research Hypothesis

1. The test items of the biology examination conducted by NECO and WAEC do not differ in function among examinees with respect to sex.

Methodology

Research Design: Instrumentation research design was deemed appropriate for this study. Instrumentation research is seen as a study which aims at introducing new contents, procedures, technologies or instruments for educational practices (ICEE, 1983).

Area of the Study: The study area was Benue State, Nigeria. Benue State is within the North Central zone of Nigeria. It is made up of 23 local government areas with three educational zones (A, B and C).

Population of the Study: The population of the study comprised all year three (SS III) senior secondary school students who enrolled for the May/June/July 2006. Biology senior secondary school certificate examination of WAEC and NECO in the three education zones of Benue State. This population was chosen because it was assumed they should have covered the WAEC and NECO Biology syllabuses. There were 35,000 students that sat for WAEC 2006 Biology examination and 42,193 students for NECO Biology examination (WAEC and NECO sources).

Sample and Sampling Techniques

One thousand eight hundred (1800) students formed the sample for this study. The multi-stage stratified sampling technique was used for the study.

Instrument for Data Collection

The instruments for this study consisted of WAEC and NECO 2000 – 2001 objective Biology questions respectively. The Biology objective question for each year is made up of 60 items for both WAEC and NECO.

Method of Data Collection

The instruments were administered to the students by trained research assistants and senior Biology teachers of the selected schools under the supervision of the researcher. The instruments were administered under similar conditions as given by the examination bodies.

Method of Data Analysis

The transformed item difficulty technique (adjusted threshold) of the BILOG MG Computer Programme was used to analyze the data collected. This technique was used to answer the research question while the hypothesis was tested using the t-test analysis of the SPSS Computer programme at 0.05 level of significance. The t-test is to enable the research establish if there is a difference in the DIF of test items by the examination bodies among boys and girls.

Results

Research Question 1

How do the different items of Biology examination conducted by NECO function with respect to sex (boys and girls)?

Table 1 shows the adjusted threshold values for group differential item functioning for items in the biology examination for the year 2000 conducted by NECO. Note that for this study, any variation of between -0.05 to -3.00 and 0.05 to 3.00 is an indication of DIF.

Research Question One

Table 1: Model for Group Differential Item Functioning: Adjusted Threshold Values DIF Sex Neco 2000

ITEM	GROUP		ITEM	GROUP	
	1	2		1	2
IT01	-0.95	-1.39	IT31	1.01	1.29
IT02	1.29	1.01	IT32	0.84	0.32
IT03	0.97	0.59	IT33	1.81	0.57
IT04	0.22	0.52	IT34	1.61	0.64
IT05	-0.06	0.01	IT35	2.86	1.30
IT06	1.13	0.48	IT36	1.01	2.34
IT07	2.27	0.01	IT37	3.89	5.25
IT08	3.64	3.21	IT38	1.64	1.43
IT09	1.27	0.62	IT09	-0.47	-0.27
IT10	2.68	3.59	IT40	0.65	0.08
IT11	0.01	0.09	IT41	-0.35	-0.69
IT12	1.04	2.86	IT42	0.33	0.24
IT13	-0.10	0.38	IT43	0.73	-1.11
IT15	-0.33	0.06	IT44	0.61	0.11
IT16	1.25	1.22	IT45	2.61	1.49
IT17	1.47	0.75	IT46	1.80	2.18
IT18	1.89	0.97	IT47	1.10	1.41
IT19	-1.31	-0.31	IT48	1.36	1.62
IT20	1.66	0.65	IT49	-1.24	-1.50
IT21	1.72	2.22	IT50	0.54	0.48
IT22	0.72	0.66	IT52	1.88	3.98
IT23	1.93	1.06	IT53	1.07	1.07
IT25	1.04	0.85	IT54	2.66	4.33
IT26	3.23	3.47	IT55	1.50	2.37
IT27	1.28	1.12	IT56	1.83	1.18
IT28	3.33	3.63	IT57	0.27	0.25
IT29	0.33	-0.33	IT58	0.47	-0.39
IT30	0.32	1.17	IT59	1.66	2.74

Key: 1 = Boy, 2 = Girl

As seen from Table, 1 several items (48%) had variations in their adjusted threshold values, which indicated DIF effects. Other items either had their values equally high or equally low (31% and 29% respectively).

The adjusted threshold values for items for the year 2001 are shown in Table 2. Thirty (30) (51%) items out of the 60 items in the test had variations in their adjusted threshold values. The remaining 30 items had their values either high (32%) for both groups or low (31%) for both groups as seen for items 2 and 14 respectively.

Table 2: DIF Gender Neco 2001 Model for Group Differential Item Functioning: Adjusted Threshold Values

ITEM	**GROUP**		**ITEM**	**GROUP**	
	1	2		1	2
IT01	-2.05	-1.71	IT31	1.08	1.59
IT02	1.77	1.59	IT32	0.99	0.25
IT03	1.17	0.77	IT33	4.26	2.85
IT04	0.10	-0.35	IT34	-1.83	-1.83
IT05	3.21	5.64	IT35	0.19	0.05
IT06	0.89	1.72	IT36	0.28	0.77
IT07	2.23	1.34	IT37	0.19	0.35
IT08	0.54	1.34	IT38	2.61	1.86
IT09	1.36	0.88	IT39	-0.59	-0.25
IT10	1.17	3.06	IT49	-2.17	-0.77
IT11	1.88	0.66	IT41	2.75	0.99
IT12	1.56	2.16	IT42	-1.32	-1.16
IT13	1.27	1.59	IT43	-1.32	-1.26
IT14	0.62	0.56	IT44	1.99	1.72
IT15	2.11	2.85	IT45	2.90	2.85
IT16	-1.52	-1.05	IT46	2.48	2.15
IT17	0.80	1.22	IT47	-0.85	0.15
IT18	1.88	0.15	IT48	-0.33	-1.37
IT19	3.05	2.31	IT49	1.37	1.59
IT20	4.01	2.66	IT50	1.67	2.66
IT21	-0.50	-1.48	IT51	-0.59	-1.05
IT22	1.46	1.72	IT52	1.46	1.72
IT23	0.99	0.25	IT53	1.46	2.15
IT24	2.11	1.59	IT54	1.67	2.31
IT25	1.99	1.72	IT55	1.56	2.31
IT26	-0.07	0.05	IT56	-0.41	-1.37
IT27	4.54	4.54	IT57	-1.42	-0.85

IT28	2.48	2.66	IT58	2.23	2.85
IT29	1.88	0.77	IT59	0.28	0.46
IT30	0.89	1.47	IT60	2.11	2.00

1 = Boy
2 = Girl

Research Question 2

How do the different items of Biology examination conducted by WAEC function with respect to sex (boys and girls)?

Table 3 shows the adjusted threshold values for group differential item functioning for items in the Biology examination conducted by WAEC for the year 2000.

The result shows that about 45 items (77%) in the test had variations in their adjusted threshold values of boys and girls, as seen for items 2, 4, 5, 6, 7, 10 and others. These indicated DIF effects. A total of 15 items (26%) did not have significant differences in their adjusted threshold values. These items included items 1, 3, 6, 8, 9, 17 and others.

Table 4 revealed the result of the adjusted threshold value for items for the year 2001. It showed that 48 items (82%) in the test had variations in their adjusted threshold values for boys and girls. This could be seen for items 1, 2, 4, 5, 7, 8 and others. Seventeen (17) items (29%) had no significant variations in their adjusted threshold for boy and girls as seen for items 3, 7, 10, 12, 16, 21, 22 and others.

Research Question Two

Table 3: Model for Group Differential Item Functioning (Sex) Adjusted Threshold Values (Waec 2000)

ITEM	GROUP 1	GROUP 2	ITEM	GROUP 1	GROUP 2
IT01	-1.87	-0.83	IT31	1.78	2.48
IT02	4.72	2.31	IT32	-2.88	-3.33
IT03	-0.40	-0.43	IT33	3.75	4.76
IT04	-3.13	-4.06	IT34	-0.62	-1.01
IT05	3.36	0.94	IT35	3.97	2.31
IT06	0.65	0.76	IT36	-2.80	-4.16
IT07	-0.01	0.76	IT37	-2.96	-5.11
IT08	-0.79	-0.66	IT38	2.53	1.91
IT09	-2.50	-2.30	IT39	-0.07	0.09
IT10	-2.57	-3.25	IT40	3.27	3.14
IT11	-5.09	-3.86	IT41	-3.31	-2.17
IT12	-4.02	-3.09	IT42	-5.09	-4.26
IT13	-1.87	-2.50	IT43	-2.57	-3.25

IT14	-4.25	-3.96	IT44	-3.50	-3.25
IT15	1.59	2.94	IT45	-1.81	-1.60
IT16	-1.19	-2.37	IT46	1.78	2.39
IT17	1.29	1.07	IT47	-3.31	-4.60
IT18	0.94	2.22	IT49	5.02	2.48
IT19	3.18	3.25	IT49	-4.64	-3.42
IT20	2.46	2.48	IT50	-5.09	-4.06
IT21	1.23	1.34	IT51	-3.50	-3.68
IT22	2.69	2.84	IT52	-5.62	-3.68
IT23	-4.02	-3.86	IT53	-6.55	-7.67
IT24	-1.94	-2.04	IT54	-6.84	-4.72
IT25	-2.57	-3.96	IT55	-3.40	-4.16
IT26	-3.80	-4.16	IT56	-5.09	-3.42
IT27	-2.96	-2.50	IT57	-5.09	-3.96
IT28	-3.40	-5.57	IT58	-5.25	-4.06
IT29	-2.35	-0.89	IT59	3.09	2.56
IT30	3.80	-4.85	IT60	-6.04	-3.59

Key: 1 = Boy 2 = Girl

Table 4: DIF Sex Waec 2001 Model for Group Differential Item Functioning: Adjusted Threshold Values

ITEM	**GROUP**		**ITEM**	**GROUP**	
	1	2		1	2
IT01	5.16	7.12	IT31	9.61	10.16
IT02	-7.85	-4.57	IT32	4.51	4.98
IT03	6.81	6.04	IT33	7.15	6.76
IT04	4.19	8.23	IT34	7.49	6.40
IT05	1.07	0.56	IT35	10.73	7.49
IT06	-7.17	-5.63	IT36	3.24	6.40
IT07	10.73	10.97	IT37	11.89	13.08
IT08	10.35	8.23	IT38	7.49	5.33
IT09	7.83	9.77	IT39	0.45	3.94
IT10	4.51	4.29	IT40	11.89	11.80
IT11	11.89	9.38	IT41	4.51	4.98
IT12	-1.40	-1.13	IT42	-3.26	-4.21

IT13	3.56	5.33	IT43	5.81	5.33
IT14	6.81	4.98	IT44	0.14	1.23
IT15	9.25	10.16	IT45	548	6.04
IT16	1.38	1.90	IT46	8.18	9.77
IT17	12.29	11.38	IT47	-4.85	-5.99
IT18	-9.63	-7.10	IT48	0.14	0.89
IT19	7.15	10.16	IT49	9.25	8.61
IT20	2.62	4.63	IT50	11.50	9.38
IT21	-3.89	-3.86	IT51	-2.33	-3.17
IT22	-0.16	0.89	IT52	10.35	12.65
IT23	5.81	-0.45	IT53	-6.49	-2.48
IT24	10.35	5.69	IT54	-3.26	-7.85
IT25	-4.53	-3.17	IT55	-7.51	-10.20
IT26	-0.16	1.23	IT56	10.73	10.16
IT27	5.50	-7.10	IT57	12.29	12.65
IT28	-9.99	-9.40	IT58	9.25	9.38
IT29	-2.33	-1.13	IT59	-2.33	-1.80
IT30	4.19	1.56	IT60	1.38	-3.86

Key: Group 1 = Boy
Group 2 = Girl

Hypothesis One

The test items of the Biology examinations conducted by NECO and WAEC do not differ in function among examinees with respect to sex.

Table 5: DIF Boys (Group 1) 2000 t–Test: Two–Sample Assuming Equal Variances

	Mean	**Variance**	**Observation**	**Pooled Var**	**df**	**t-stat**	**P (T<=t) one**	**t Critical one**	**P (T<=t) two**	**t Critical two**
Variable 1 NECO	1.10	1.30	60	5.83	118	5.64	5.91	1.66		
Variable 2 WAEC	-1.40	10.37	60	5.83	118	5.64			1.18	1.98

On Table 5 are t-Test results of differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (boys) for the year 2000.
The result showed that the t-statistic (t-stat = 5.64) was higher than the t-critical (t-critic = 1.98). The null hypothesis of no significant difference is rejected. By this result, there is a significant difference in the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (boys) in 2000.

Table 6 shows the results of t-Test on the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (girls) in 2000.

Table 6: DIF Girls (Group 2) 2000 t–Test: Two –Sample Assuming Equal Variances

	Mean	**Variance**	**Observation**	**Pooled Var**	**df**	**t-stat**	**P (T<=t) one**	**t Critical one**	**P (T<=t) two**	**t Critical two**
Variable 1 NECO	1.03	1.97	60	5.30	118	6.07	7.90	1.66		
Variable 2 WAEC	-1.52	8.62	60	5.30	118	6.07			1.58	1.98

The t-Test results showed that the t-statistic (t-stat = 6.07) was higher than the t-critical (t-critic = 1.98). The null hypothesis of no significant difference is rejected. Therefore, there is a significant difference in the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (girls) in 2000.

The results of the t-Test on the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (boys) in 2001.

Table 7: DIF Boys (Group 1) 2001 t–Test: Two–Sample Assuming Equal Variances

	Mean	**Variance**	**Observation**	**Pooled Var**	**df**	**t-stat**	**P (T<=t) one**	**t Critical one**	**P (T<=t) two**	**t Critical two**
Variable 1 NECO	1.07	2.34	60	2.50	118	-1.03	0.15	1.66		
Variable 2 WAEC	9.50	5.00	60	2.50	118	-1.03			0.31	1.98

From the results on Table 9, it was observed that the t-statistic (t-stat = 1.03) was lower than the t-critical (t-critic = 1.98). The null hypothesis of no significant difference is accepted. It then means that there is no significant difference in the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (boys) in 2001.

On Table 8 are the results of the t-Test on the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (girls) in 2001.

Table 8: DIF Girls (Group 2) 2001 t–Test: Two –Sample Assuming Equal Variances

	Mean	**Variance**	**Observation**	**Pooled Var**	**df**	**t-stat**	**P (T<=t) one**	**t Critical one**	**P (T<=t) two**	**t Critical two**
Variable 1 NECO	1.07	2.41	60	2.10	118	-2.84	0.00	1.66		
Variable 2 WAEC	3.45	4.00	60	2.10	118	-2.84			0.01	1.98

The results revealed that the t-statistic (t-stat = 2.84) was higher than the t-critical (t-critic = 1.98). The null hypothesis of no significant difference is rejected. By this result, there is a significant difference in the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among examinees (girls) in 2001.

Discussion on Research Questions One and Two

The purpose of the research questions was to find out how the test items of Biology examinations conducted by NECO and WAEC functioned with respect to sex (boys and girls).

The adjusted threshold value scores of the girls and boys sub-groups were used to determine the differential item functioning of the tests (DIF). From the results of the research question one test for 2000, there was clear evidence of strong differences in the threshold values between the boys and the girls. These differences which are the DIF effects were observed with most (95%) of the items such as items 1, 5, 6, 7, 9, 12 among others. However, some items were extremely very easy for the girls. These included items 1, 5, 19, 29, 39, and 58.
DIF effect was, also, evident among test items for the year 2001. About 58% of the items had DIF effects as seen for items 1, 3, 6, 10, 11 among others. Extremely easy items for girls (female sub-group) were very few, precisely only item 4 (7%) was exceptionally easy for females (items 5, 10, 12 and 53).

The results of research question two showed that test items of WAEC Biology for 2000 had evidence of DIF effects. Forty five (45) items (77%) had variations in their adjusted threshold values for the male sub group and the female sub-group. Some items among these 45 were easier for the males (39%) than for females, while some other were easier for females (37%) than males. However, the differences in the threshold values of these items were very small.
The year 2001 had its items, also, having DIF effects with minimal variations in their threshold values. Only items 23 and 60 were extremely easy for the female sub-group. These results agree with the results of other researchers like Akindele (2003), VI – Nhuanle (1999), and Stage (1998), Gierls (1999) and Ogbebor & Onuka (2013).

Probable reasons for items having DIF effects according to Akindele (2003) include the unidimensionality assumption, and the fit of the data especially when the validity of the items were only marginally met, as reflected in the test construction procedures. Secondly, some of the items tend to be complex and ambiguous. This factor might lead to a greater chance of translation error on the part of the examinees. Thirdly, according to Badgell,G.R. (1995), identification of items with significant DIF may be related to sample size. Studies in which sample sizes were relatively small were identified as having significant DIF.

Lastly according to Ogbebor & Onuka (2013), learners not being familiar with the content and vocabularies of the test items, makes them unable to cpmprehend and understand that are presented on the test items.

Hypothesis

The purpose of this hypothesis was to compare the differential item functioning of items in the Biology examinations conducted by NECO and WAEC, with respect to sex (boys and girls).

The test for difference in the differential item functioning of the items in the Biology examinations conducted by NECO and WAEC among boys showed that there was a significant difference in the functioning of the items of the Biology examinations, in 2000 (Table 7). This difference was evident in the adjusted threshold values of these items. The

WAEC items had more negative values of adjusted threshold than NECO items. This indicated that more WAEC items were easier for boys in 2000.

Among girls in 2000 (Table 8), the test of difference showed that there was a significant difference in the differential item functioning of items in the Biology examinations of NECO and WAEC. This could be seen in the adjusted threshold values of the items where the WAEC items had more negative values than the NECO items.

However, WAEC items had higher values of adjusted threshold among girls than NECO items. Evidence of these differences was, also, reflected in the results of the difficulty parameters of these items which showed that there were differences in the difficulty parameters of NECO and WAEC items in 2000.

For items in 2001 Biology examinations for NECO and WAEC, there was no significant difference in the differential item functioning of the items among boys (Table 7).

The test of difference on the differential functioning of items among girls showed that there was a significant difference in the differential item functioning of items of the Biology examinations conducted by NECO and WAEC among girls in 2001 (Table 8). From the results on the adjusted threshold of the items differences were observed. There was, also, a significant difference in the difficulty parameters of items in 2001.

In summary, WAEC items in the Biology examination for 2000 were easier for boys and girls than NECO items for that year. In 2001, the items from both examination bodies had equal strength among boys while WAEC questions were easier for girls.

Conclusion

From the data analyzed in this study the conclusions were that the items from WAEC Biology examinations were easier for boys than for girls. It therefore implies that DIF effects existed among boys and girls. This means that some of the Biology test items functioned differently from what they were supposed to measure and this can make the validity of such items questionable.

Recommendation

The following recommendations are made based on the findings of this study.

1. Test developers should be trained on the use of IRT analysis techniques.
2. Test developers should develop test items with minimal DIF effects.
3. Examination bodies should check on the DIF of their test items before administration on examinees.
4. IRT analysis should be adopted by all examination bodies in Nigeria so that our measurement and assessment problems could be solved to a large extent.

Acknowledgement

The authors are grateful to University of Agriculture, Makurdi for making research funds available for carrying out this study.

References

[1] B.P. Akindele, The development of item bank for selection tests into Nigerian universities: An exploratory study, *Unpublished Ph. D Thesis of University of Ibadan*, (2003), Nigeria.

[2] A. Anastasi, *Psychological Testing (4th Edition)*, (1976), Macmillan Publishing Co, Inc., New York.

[3] W.H. Angoff and S.F. Ford, Item-race interaction on a test of scholastic aptitude, *Journal of Education Measurement*, 10(1973), 85-105.

[4] W.H. Angoff, The investigation of test bias in the absence and outside criterion, *Paper Presented at the NIE Conference on Test Bias*, (1975).

[5] X. Badia, Prietal and J.M. Linacre, Differential item and test functioning (DIF and DTF), *Rasch Measurement Transactions*, 16(3) (2002), 889.

[6] G.R. Budgell, Analysis of differential item functioning in translated assessment instruments, *Applied Psychological Measurement*, 19(4) (1995), 309-321.

[7] C. Cardall and W.E. Coffman, A method for comparing the performance of different groups on the items of a test, *Research Bulletin*, (1964), 64-61, Princeton, New Jersey: Educational Testing Service.

[8] T.A. Cleary and T.L. Hilton, An investigation into item bias, *Educational and Psychological Measurement*, 28(1986), 61-75.

[9] T.A. Cleary, Test bias: Prediction of grades of Negro and White students in integrated colleges, *Journal of Educational Measurement*, 5(1968), 115-124.

[10] R.L. Ebel, *Essential of Educational Measurement (Third Edition)*, (1979), Prentice-Hall Inc., New Jersey.

[11] G. Echternacht, A quick method for determining test bias, *Educational and Psychological Measurement*, 34(1974), 271-280.

[12] D.R. Green and J.F. Draper, Exploratory studies of bias in achievement test, *Paper Presented at the Meeting of the American Psychological Association*, (1972).

[13] D.R. Green, Reducing bias in achievement tests, *Paper Presented at the Annual Meeting of the American Educational Research Association*, (1976), San Francisco.

[14] G.H. Ironson and M.J. Subkoviak, A comparison methods of assessing item bias, *Journal of Educational Measurement*, 16(4) (1979), 209-225.

[15] G.H. Ironson, Using item response theory to measure bias, In R.K. Hambleton (Ed), *Applications of Item Response Theory*, (1982), B.C Vancouver, Educational Research Institute of British Columbia.

[16] R.L. Linn and C.E. Wert, Considerations for studies of test bias, *Journal of Educational Measurement*, 8(1971), 14.

[17] F.M. Lord, Statistical adjustments when comparing preexisting groups, *Psychological Bulletin*, 72(1969), 336-337.

[18] F.M. Lord, A study of item bias and using item characteristic curve theory, In Y.H. Poortinga (Ed), *Basic Problems in Cross-Cultural Psychology*, (1977), Amsterdam: Swets and Zeiblinger.

[19] W.A. Mehrens and I.J. Lehamm, *Measurement and Evaluation in Education and Psychology (3rd ed)*, (1984), Holt, Rinehart and Winston, New York.

[20] W.R. Merz, Factor analysis as a technique in analyzing item bias, *Paper Presented at the Annual Meeting of the California Educational Research Association*, (1973), Los Angeles, California.

[21] W.R. Merz, Test fairness and test bias: A review of procedures, *Paper Presented at the Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation*, May (1976), Reston, Virginia.

[22] B.G. Nworgu, *Educational Measurement and Evaluation: Theory and Practice*, (1992), Awka: Hallman Publishers.

[23] U. Ogbebor and A. Onuka, Differential item functioning of economics question papers of national examinations council in Delta State, Nigeria, *Nigerian Journal of Educational Research and Education*, 12(1) (2013), 45-60.

[24] M.B. Ogunniyi, *Educational Measurement and Evaluation*, (1984), Longman Groups Ltd, Hong-Kong.

[25] D.G. Ozenne, N.C. Van Gelder and A.J. Cohen, *Emergency School Aid Act (ESAA), National Evaluation, Achievement Test Standardization*, (1974), Santa Monica, California: Development Corporation.

[26] N.S. Peterson, Bias in the selection rule: Bias in the test, *Paper Presented at the Third International Symposium on Educational Testing*, (1977), University of Leydon, The Netherlands.

[27] L.M. Rudner, An approach to biased item identification using latent trait measurement theory, *Paper Presented at the Annual Meeting of the American Educational Research Association*, (1977), New York.

[28] J. Scheuneman, A new method of assessing bias in test items, *Paper Presented at the Meeting of the American Educational Research Association*, (1975), Washington.

[29] J. Scheuneman, Validating a procedure for assessing bias in test items in the absence of an outside criterion, *Paper Presented at the Meeting of the American Educational Research Association*, (1976), San Francisco.

[30] J. Scheuneman, Ethnic group bias in intelligence test items, *Paper Presented at the American Educational Research Association Convention*, (1978), Toronto.

[31] J. Scheuneman, A method of assessing bias in test items, *Journal of Educational Measurement*, 16(1979), 143-152.

[32] C. Stage, A comparison between item analysis based on item response theory and on classical test-theory: A study of the Swesat Word, *Educational Measurement No.29*, (1998), UMEA University, Department of Educational Measurement.

[33] J.R. Veale and D.I. Foreman, Cultural variation in criterion referenced tests: A global item analysis, *Paper Presented at the Annual Meeting of the American*, (1976).

[34] Vi-Nhuan Le, *Identify Differential Item Functioning on the NELS: 88 History Achievement Test CSE Technical Report 511*, (1999), University of California, Los Angeles.

[35] R.L. Williams, Black pride, academic relevance and individual achievement, *The Counseling Psychologist*, 2(1970), 18-22.

[36] R.L. Williams, Abuses and misuses of testing black children, *Counseling Psychologist*, 2(1971), 62-73.

[37] B.D. Wright, Solving problems with rasch model, *Journal of Educational Measurement*, 14(1977), 97-116, www.rasch.org/memo39htm, Accessed 4th Dec. (2007).

[38] B.D. Zumbo, *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Odelling as a Unitary Framework for Binary and Likert-like (Ordinal) Item Scores*, (1999), Ottowa Canada: Directorate of Human Resources Research and Evaluation.