Differential Item Functioning in English Language Test Using Item Response Theory for Ethnic Groups

Ogbogo Stella   Dr. Opara, Ijeoma Margaret
Department of Educational Psychology, Guidance And
Counselling, Faculty of Education, University of Port Harcourt Rivers State, Nigeria

Abstract
The study investigated detecting differential item functioning using item response theory in West African Senior School Certificate English language test in South-South Nigeria. Four research questions and one hypothesis were formulated to guide the study. Descriptive research survey design was employed for the study. The population of the study was 117845 Senior Secondary 3 students in Edo, Delta, Rivers and Bayelsa states. A sample of 1309 (604 males, 705 females) drawn through multi stage sampling technique was used for the study. One valid instrument titled: WASSCE/SSCE English language objective test (ELOT) was used to collect data for the study. The reliability index of the instrument was estimated using Richard Kuderson 20 with coefficient value of .84 for the English Language objective test Chi square and Lord Wald test statistics statistical technique employed by Item Response Theory for Patient Reported Outcome (IRTPRO) was the technique used in data analysis which provided answers to the research questions and Chi Square test to test the hypothesis at.05 level of significance. On analysis, the result revealed that significantly between Ijaw, and Bini, 20 items were flagged as exhibiting DIF, between Esan and Ijaw, 20 items showed DIF, between Isoko and Ijaw, 15 items showed DIF, between Urhobo and Ijaw, 12 items were flagged as showing DIF and the number of items that function differently was significantly dependent on ethnic groups. This thus shows a total of 95% based on ethnicity indicating large DIF and items that are potentially biased. Based on the findings recommendation were made and one among others was that Item Response theory should be used as DIF detection method by large scale public examination and test developers.
Keywords; Differential item functioning, Item Response Theory, Ethnicity, English language, Examination
DOI: 10.7176/JESD/10-6-02
Publication date:March 31st 2019

Introduction

Identification of biased item is of extreme necessity in the creating of equitable and content valid test or examination. Thus of great value or importance is the subject of fairness of test items to test takers. Examination and test are processes intricately tied to education and education is one powerful instrument for the development of man and society. According to Enamiroro (2007), education has as its aim manpower development and is geared at national growth and development. And as such, it has been embraced and welcomed by most government of the world and the Nigeria government has not exempted itself from using this valuable tool in developing its nation. In pursuit of national growth and development, the Nigeria government has set up objectives that will guide education. These objectives are stated in the Federal Republic of Nigeria (2004) National Policy on Education and one of these objectives is the creating of a just and egalitarian society upon which the foundation of education will operate on among other stated objectives. In pursuit of this, there emerged the development of a national curriculum that provides the platform in which interest of children from different social-cultural background, race, gender and ethnic group, are catered for to bring about equal developmental opportunity. There exists as well as terminal examinations conducted by different bodies such as West Africa Examination Council(WAEC), National Examination Council (NECO) and other bodies in which different examinees of same ability and trait from different language, culture, gender, race, ethnical group, geographical location, cultural and socio-economic background are subjected to.

These examinations are supposed to measure what they are designed for and are not be unfair to examinees from different groups. According to Roever (2005) a fair test is one that has comparable validity for all groups of individuals and that gives all test taker same opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the test's purpose. If a test favours one group of test-takers over another, the principle of test fairness may appear to have been violated and this has great implication as test or examination results play a great role in the Milieu of high stakes decision-making, which include, admission aptitude achievement and certification and so on. Items that make up a test are neutral inconsequential tools and it remains so until some sort of significance is assign or attributed to the results derived from them. Once this significance is attached to a person's score, the individual will experience some repercussions, ranging from superficial to life changing. These repercussions may be fair or unfair, helpful or disastrous, appropriate or misguided depending on the significance attached to the test score (Gregory, 2004). Therefore, it is imperative and essential that fairness principle is evident in a test to all and not biased against any group. However, in some instances, items in a test are biased based on the fact that characteristics not part of a construct being measured forms aspect of test and the

fallout is that performance is affected. (Abiam 1996: Nworgu & Odili 2005; Uwhekadom 2014)

Bias as seen by Brown (2012) happens when examinees of one groups are likely to answer an item correctly (or endorse an item) than examinees of another group due to of some characterizes of the test item that is not material to the construct being measured. Bias as defined by Hambleton and Rodgers (1995) is the presence of some features of an item that results in differential performance for person of the equal ability but from different ethnic, sex, cultural or religious group. This issue regarding test bias has become the subject of enormous research and a method called differential item functioning (DIF) has evolved to become a new standard in scientific analysis of bias (Zumbo 1999). In the psychometric literature, the term (DIF) was created to define concerns about item bias existing in the context of test bias (Lee 2015). The term (DIF) explains the scientific or empirical evidence used to refute or back up bias. Abedalaziz (2011), views that "DIF" is a collection of statistical methods utilized to determine if examination items are appropriate and fair for testing the knowledge of various subgroups of examinee. Caroll, (2015) viewing posited that when individuals from different groups (gender, majority/minority, SES, etc.) perform differently on a test item, this difference in the item score, above and beyond group differences on the construct, is referred to as DIF. Broadly defined, DIF refers to the presence of differences in individual item characteristics across groups, which can be viewed graphically as differently sloped or horizontally shifted item characteristic curves (ICCs) when the item parameters for each group are plotted (Lord, 1980; Thissen, Steinberg, & Wainer, 1988). These differences in the parameters of the items are seen to show empirical differences in an item's tendency to accurately or otherwise estimate an examinee's standing on a latent trait (often ability) among different groups. Thus, such differences may be sources of bias in the form of unfair advantages for some groups and unfair disadvantages for other groups (Carroll, 2015).

Differential Item Functioning (DIF) is a violation of the invariance assumption in Item Response Theory (IRT) models and happens when the probability of endorsing an item for test takers of equal ability level varies in different groups (Battuaz, 2015). DIF refers to difference in item functioning after groups have been matched with respect to ability (Wiberg 2007). DIF essentially deals with individuals from different subgroups or membership with the equal ability or proficiently having different probability of correctly answering an item (Anastasi & Urbina, 2007; Kamata & Vaughn 2004; Magis & Falcon 2011).. It deals with the differences in the functioning of items across groups with dissimilar cultural or experiential backgrounds (often times demographic) which are matched on the underlying trait being measured by the items (Anastasi & Urbina, 1997; Camili, 2006).

DIF describes the empirical evidence used to back bias. DIF is "Prima facie" evidence that the possibility that the test is biased exists (Karami &Nodoushan 2016). As such DIF helps in the identification of test items that are potentially biased (Perrone 2006). That an item is flagged or has evidence of displaying DIF does not necessary mean bias. Alongside the statistical evidence of DIF, the process of sensitivity of test content review can lead to the identification of Bias.

In DIF, there are at least two groups, (usually labeled focal and reference group); In general, DIF is examined by comparing item responses for these groups of examinees. In most applications, these groups represent types of examinees based on demographic characteristics such as gender or race, the focal group is the potentially disadvantaged and the referenced group is the group which is assumed to be potentially advantaged by the test, however, naming or terming the group is not always clear cut it can be arbitrary (Finch & French, 2008; Karami & Nodoushan 2011 ; & McNamara & Roever, 2006).

Over the years, various methods have been applied in flagging DIF and widely of these methods is the context of the item response theory (IRT) approach. IRT detection methods like the estimating item parameters (Lord Wald Chi test), likelihood function, Area methods and so on which are used in identifying items that are potentially biased provides alongside p-values, ICC curves or trace lines depending on the parameter model. (Finch & French, 2007; Kim & Cohen, 1995; Lee, 2015; Oshima & Morris, 2008; Woods, Cai, & Wang, 2013). IRT approaches propose a latent trait or ability that underlies the item responses and shares the use of the estimate of the latent trait as a matching variable rather than the observed score. Ertuby and Russele (1999) suggested that because of their greater sophistication, IRT, procedures provide the best results for identifying items that are biased. IRT usually offers a robust approach using item characteristics curve (ICC) for identifying DIF. The ICC reveals for each item, the differing probabilities of answering a particular way contingent on the degree of the latent construct. ICCS are obtained from a function that identifies the relationship between an individual's location on a latent trait and the probability of the individual getting an item on a test (Siebert, 2013). Wiberg (2007) viewed that if the item does not show DIF, the ICC would be identical while DIF is present when ICCS for the two group differs.

DIF is usually in magnitude and it varies in degree which can be measured by examining parameters or statistics linked with the method used in detecting DIF. Levels of DIF magnitude ranges from 0.25, 0.50, 0.75 and 1.00 as the largest magnitude of DIF representing small to large DIF magnitude (Fidalgo, Ferreres & Muniz 2004; Hidalgo & Lopez-Pina 2004; Parshall & Miller 1995; Stephen-Bonty 2007). Empirical studies have shown that the percentage of DIF items ranges from quite small 1.5% to overwhelming large (64%). Studies considers it a small amount when a test contains less than 10% DIF, a medium amount of DIF when a test contains 10 to 30%, large DIF when it exceeds 30% when the parentage of DIF exceeds 10% closer attention should be paid to it

(Hambleton & Roger, 1989; Raju 1989; Xioting 2010).

The analysis of DIF provides a convenient starting point for the study of item bias. These assertions reveal a general agreement that an item that functions differentially may not be biased but biased item must function differentially (Brown 2012; Cohen 2006, Hambletbon & Rodgers 1995 Karami & Nedoushan 2011; Schumacker 2005; Perrone 2006; Williams 1997; Zumbo 1999; 2007; Wiberg, 2007).

Items can function differentially for any subject matter including English language. English Language is a well-known language spoken by many that occupies a primal function at the work place, serves as a vehicle of instruction in schools and the language of text-books (Aina, Ogudele & Olanipekun, 2013: Adekola, Shoaga&Lawal 2015). The status assigned English in Nigeria Education is specified in the 2004 National Language Policy on Education and this states that English shall be the medium of instruction in the upper primary, secondary and tertiary/higher level of education. (Adekoya, Shoaga & Lawal 2015). Thus the status of English is enhanced as it's not only a subject of study in school but also the language of instruction. Also the Government of Nigeria (2004) stipulates in her National Policy on Education the significance of English language as one of the core subjects that will enable an individual to gain admission into any higher institution thereby making it a course that holds the key to further academic progress. To have the chance to study any discipline in the University, Polytechnic or College of Education, an individual must have obtained a credit in English language in WASSCE or its equivalent.

English language although a foreign language to Nigeria, in the sense that it is not indigenous, is widely used in Nigeria as the second language (L2), ever since it was introduced being that Nigeria is a multilingual country comprising of diverse ethnic groups with diverse languages as their first language (L1) this however has in no way diminished the significance of English Language as earlier stated.

Despite all the roles and immense significance of English Language as earlier stated, there has been a consistency poor performance in English language over time in large scale examination like WAEC with 29.99% (2009), 23.36% (2010), 30.9% (2011), 38.81% (2012), 36.57% (2013), 31.28% (2014), and 38.68% (2015) obtaining five credit and above in English Language and mathematics (Vanguardonline 6/8/16; Daily Trust 21/8/2014). This shows that a high percentage of students enrolling for English Language records or obtains below credit in the past few years. As one of the examination subject in WEAC, the English Language paper is usually divided into three sections requiring examinees to respond to. This covers the basic skills of reading, speaking, writing, and listening that Daladi established in Umar Sa'ad & Usman (2014) as the basis skills that English Language comprises of. The essay and objective test format is usually adopted by West Africa Examination Council in setting questions in English Language.

The Multiple choice test items which is a type of objective test or a form that requires testees to choose or select one response from a set of many alternatives (Orluwene, 2012; Opara, 2016). Assessment of Students via this form is one form that all schools not only large scale examinations like WEAC employs in assessing testees who takes English Language as subject of study and also in other subjects as it is the language of instruction in all educational institutions irrespective of one's group membership. Important is that English Language examinations regardless of the body conducting it should have items that are fair, unbiased and not functioning differentially for any group from different socioeconomic status, gender, race, geographical location, ethnic group and so on, on any subject matter..

Ethnicity as a concept involves some form of metaphorities of kinship, especially the notion of common ancestry and blood relationship. It involves some form of identification, individual identify themselves as belonging to a certain group and the group recognizes individual as belonging to that group. Bottomley (1997) opined that cultural practices such as the language and religion define the particularities of various ethnic groups. Thus each ethnic group has a native tongue which is considered as (L1) and English language (L2) which is the language of learning and commerce. Various ethnic groups exist in within the Nigeria environment numbering up to 350. Some of the ethnics groups that exist include Ijaw, Yoruba, Isoko, Bini, Igbo, Esan, Urhobo, Yoruba, Efik and so on. Peculiarities in words pronunciation and words usage of some ethnic may cause variation in academic performance especially English language which is a second language that most students are been exposed to after being exposed to their mother tongue which is first language. Thus Fatemi & Khaghanzhad, (2011) asserted that ethnic difference may cause great variation in test taker performance on test of English language in large scale examination This variation can be in the form of items functioning differently for some ethnic subgroups. Ethnic difference may cause great variation in test taker performance on test of English language in large scare examination where test takers take it as (L2) (Fatemi & Khaghanzhad 2011). An individual who grows up within a particular ethnic group will absorb the native tongue of that group which may influence proficiency in English language.

Large scale examination like WASSCE is taken by individuals from diverse ethnic groups and research though foreign has revealed that performance of test takers from different ethnic groups differs thus it's imperative that performance of students of same ability from different ethnic groups who take English language as a subject be examined.

It is imperative that items do not function differently for any group. This is because test results are literally seen to be strong indicators of people's ability levels and performance in subjects (Whitmore and Schumaker, 1999).

However different studies conducted have shown that some tests used in some public examinations in Nigeria and other large scale examination contain items that exhibit DIF to the advantage or disadvantage of different subgroups. For instance, in a study by Reuben & Akorede (2016) titled differential item functioning technique for detecting item bias in economics among secondary school students in Abuja metropolis, using a sample of 750 through muti -stage sampling were able to identify 3 items in the NECO (SSCE) 2013 economics objective test items that functioned differentially for groups from 2 different socio-economic status, five (5) items with regards to gender and twenty- one (21) items show DIF in regards to school location.

Ogbebor & Onuka (2013) investigating differential item functioning method as an item bias indicator in Delta state using a sample size of 447 SS3 students showed that the National Examination Council (NECO) Economic questions for 2010 had 18 items that functioned differentially for examines based on school type and school location. Uwhekadom (2014) reported that chemistry multiple choice question used by WAEC, in the 2009, 2010 and 2011 SSCE contains test items that significantly functions differentially for students of different gender, high and low socio-economic status urban and rural geographical location test takers.

Queeensoap (2014) in a study application of differential item functioning in detecting item bias in chemistry achievement test in Nigeria used a sample size of 400 administered a chemistry achievement test. Statistical and content analysis was done with logistic regression statistics; Mantel-Haenszel's adjusted DIF, Scheunneman modified chi-square, and item characteristics curve. The modified Scheunneman Chi-square showed 30 items exhibiting DIF, the Mantel-Haenszel adjusted DIF statistics flagged all items as showing significant DIF between the focal group (Ijaw) and reference group (Yoruba, Hausa & Igbo).

With the foregoing, the subject of items functioning differentially and as such, a potential root of bias in high stakes examinations is of great concern. Furthermore, item functioning differentially for individuals of various subgroups of same ability/trait level has great implication on policy, administration and classroom levels where test results constitute the ground for decision making. Test with different functioning items could bring about low achievement for a minority groups in a subject matter and this can mutilate the meaning of test result and decision that is hinged on it for some groups especially core subject like English which is compulsory criteria for further educational advancement. This thus is a concern as most often public dismay welcome the release of Senior School Certificate Examination (SSCE) as a consistent poor performance in English Language examinations has been observed over time with 29.99% (2009), 23.36% (2010), 30.9% (2011), 38.81% (2012), 36.57% (2013), 31.28% (2014), and 38.68% (2015) obtaining five credit and above including English language in WASSCE in recent years from examinees from different subgroups groups. The problem of this study as seen by the researchers is, are there items in English language multiple choice test used by WEAC in the Senior School Certificate Examination that differentially functions for candidates with equal ability from different, ethnic group such that they contribute to poor performance in the subject?

The following research questions guided the study

1.   Which are the items in the English language multiple choice that function significantly differentially between the focal group (Ijaw) and reference group (Bini)?
2.   Which are the items in the English language multiple choice that functions significantly differentially between the focal group (Ijaw) and reference group (Esan)?
3.   Which are the items in the English language multiple choice that functions significantly differentially between the focal group (Ijaw) and reference group (Isoko)?
4.   Which are the items in the English language multiple test that function significantly differentially between the focal group (Ijaw) and reference group (Urhobo) students in the English language multiple choice?

The null hypothesis was tested at 0.05 level of significance:

1.   The numbers of items that function differently in English language multiple choice test are not significantly dependent on the ethnic groups

Method

The descriptive survey research design was used for this study. The population of the study consisted of one hundred and seventeen thousand, eight hundred and forty-five (117,845) Senior Secondary three (3) students in 1190 public Secondary Schools who are studying English language as a certificate subject in the 2016/2017 academic session in Delta, Bayelsa, Rivers and Edo State. The sample size of the study was 1309 students of the population who are studying English language as a certificate subject. A multistage sampling technique was employed for the study at different stages and several sampling techniques like simple random, cluster and stratified, was employed. The instrument for the study has two sections with the first section yielding information on the demographic information of the respondents on such information like sex, state, ethnicity, name of school. While the second part of the instrument for the study was English Language Objective Test (ELOT) which was

based on the WASSCE/SSCE English language paper one used in the 2016 examination. This contains 70 multiple choice type questions constructed by subject experts and developed by WAEC into test form. This instrument was employed to detect items that differentially functions. The face and content validity of English language WAEC / SSCE used in 2016 has been established based on the fact that the questions were owned by an examining body WAEC/SSCE and has been validated by experts of the examining body through statistical techniques. The reliability of the coefficient of the instrument English Language Objective Test (ELOT) was established using the Kuder-Richardson (KR20) with internal consistency coefficient of 0.84 been the coefficient obtained. The instruments for the study were directly administered to the respondents on individual basis in their classes and were retrieved on the spot. After administration and retrieval, of the 1400 instruments administered to the 1400 initial sample size only 1309 instruments were properly filled this was what was used for analysis and the final sample size was 1309.

Every statistical model requires assumptions about data to obtain viable parameter estimates in the data analysis procedures. In checking for the assumptions, IRTPRO first conducts a unidimensional IRT analysis to check for unidimensional assumption of IRT using confirmatory factor analysis. On meeting this assumption, local independence assumption holds as data that meets unidimensionality assumption also meets local independence assumption.

For IRT assessment of Model-Data Fit and item calibration, the 2-Pl was used as seen by the. Chi-square likelihood ratio Goodness of fit statistics.. To equate scale, in the first equating the Item Response Theory for Patient Reported Outcome (IRTPRO) software equates the ijaw as the focal group while the urhobo the reference group, in the second equating, the Ijaw was the focal group and Isoko the reference group, in the third equating, the Ijaw was the focal group and the Bini the reference group, in the fourth equating, the Ijaw was the focal group and the Esan the reference group. On item calibration (estimation of item parameters and examinee ability), Lord Wald Chi test which is used by IRTPRO software developed by Cai, Thissen & Toit (2011) was employed to detect items that functions differentially between the focal and reference groups as seen by the P-value (tested at 0.05 sig level) for both the reference groups and focal groups and their χ^2 value.

Results
Research Question 1: Which are the items in the English language multiple choice that functions significantly differently between the focal group (Ijaw) and reference group (Bini) of students

To answer this research question, the 70 items were subjected to DIF analysis using IRTPRO. Items that exhibited DIF were items that differentiate significantly between group membership and this is represented in table.1

Table 1; Differential Item Functioning Statistics for Bini and Ijaw

Item numbers in:										
Group 1(bini)	Group 2(ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c\|a}$	d.f.	P
1	1	2.0	2	0.3704	0.5	1	0.4612	1.4	1	0.2299
2	2	0.0	2	1.0000	0.0	1	1.0000	0.0	1	1.0000
3	3	0.0	2	1.0000	0.0	1	1.0000	0.0	1	1.0000
4	4	4.4	2	0.1122	4.4	1	0.0369	0.0	1	0.8923
5	5	2.8	2	0.2495	2.5	1	0.1134	0.3	1	0.6069
6	6	0.2	2	0.8860	0.2	1	0.6617	0.1	1	0.8226
7	7	2.5	2	0.2903	0.1	1	0.7153	2.3	1	0.1264
8	8	0.8	2	0.6831	0.6	1	0.4461	0.2	1	0.6705
9	9	4.5	2	0.1077	4.3	1	0.0386	0.2	1	0.6733
10	10	3.8	2	0.1519	2.5	1	0.1128	1.3	1	0.2630
11	11	10.5	2	**0.0053***	0.2	1	0.6598	10.3	1	0.0013
12	12	8.8	2	**0.0120***	8.8	1	0.0030	0.0	1	0.8514
13	13	4.6	2	0.1020	1.7	1	0.1942	2.9	1	0.0912
14	14	14.7	2	**0.0007***	11.7	1	0.0006	3.0	1	0.0852
15	15	6.4	2	**0.0400***	6.4	1	0.0113	0.0	1	0.8658
16	16	9.8	2	**0.0076***	9.7	1	0.0019	0.1	1	0.7872
17	17	4.3	2	0.1176	0.4	1	0.5395	3.9	1	0.0482
18	18	3.7	2	0.1546	2.4	1	0.1244	1.4	1	0.2416

Item numbers in:		Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c|a}$	d.f.	P
Group 1(bini)	Group 2(ijaw)									
19	19	5.0	2	0.0816	4.9	1	0.0265	0.1	1	0.7702
20	20	3.2	2	0.1980	2.4	1	0.1203	0.8	1	0.3646
21	21	1.9	2	0.3929	0.2	1	0.6250	1.6	1	0.2021
22	22	7.4	2	**0.0242***	6.3	1	0.0121	1.1	1	0.2855
23	23	9.5	2	**0.0084***	3.2	1	0.0725	6.3	1	0.0119
24	24	0.0	2	1.0000	0.0	1	1.0000	0.0	1	1.0000
25	25	4.5	2	0.1034	3.2	1	0.0755	1.4	1	0.2412
26	26	2.9	2	0.2397	2.1	1	0.1457	0.7	1	0.3906
27	27	2.1	2	0.3439	1.9	1	0.1712	0.3	1	0.6109
28	28	4.4	2	0.1082	3.7	1	0.0540	0.7	1	0.3926
29	29	15.1	2	**0.0005***	15.1	1	0.0001	0.0	1	0.9376
30	30	5.4	2	0.0658	4.6	1	0.0328	0.9	1	0.3473
31	31	1.7	2	0.4348	0.2	1	0.6623	1.5	1	0.2249
32	32	8.0	2	**0.0182***	0.0	1	0.9900	8.0	1	0.0047
33	33	4.7	2	0.0938	4.1	1	0.0420	0.6	1	0.4407
34	34	9.9	2	**0.0071***	0.6	1	0.4447	9.3	1	0.0023
35	35	8.5	2	**0.0144***	6.7	1	0.0098	1.8	1	0.1794
36	36	0.6	2	0.7584	0.5	1	0.4994	0.1	1	0.7563
37	37	0.6	2	0.7300	0.6	1	0.4356	0.0	1	0.8849
38	38	3.5	2	0.1775	3.5	1	0.0624	0.0	1	0.9903
39	39	1.7	2	0.4256	1.5	1	0.2182	0.2	1	0.6629
40	40	5.4	2	0.0681	5.3	1	0.0216	0.1	1	0.7611
41	41	4.1	2	0.1270	4.1	1	0.0423	0.0	1	0.9650
42	42	4.6	2	0.0989	3.5	1	0.0616	1.1	1	0.2883
43	43	2.3	2	0.3129	0.2	1	0.6623	2.1	1	0.1446
44	44	0.1	2	0.9452	0.0	1	0.8667	0.1	1	0.7712
45	45	10.6	2	**0.0050***	7.4	1	0.0066	3.2	1	0.0727
46	46	4.5	2	0.1062	0.6	1	0.4515	3.9	1	0.0478
47	47	3.4	2	0.1884	0.0	1	0.8615	3.3	1	0.0684
48	48	1.9	2	0.3901	0.0	1	0.9192	1.9	1	0.1716
49	49	7.3	2	**0.0254***	0.3	1	0.6084	7.1	1	0.0078
50	50	3.6	2	0.1663	0.5	1	0.4877	3.1	1	0.0773
51	51	5.7	2	0.0574	1.3	1	0.2524	4.4	1	0.0360
52	52	3.3	2	0.1908	0.8	1	0.3824	2.6	1	0.1106
53	53	0.9	2	0.6475	0.6	1	0.4287	0.2	1	0.6227
54	54	0.8	2	0.6641	0.4	1	0.5420	0.4	1	0.5042
55	55	4.4	2	0.1111	1.4	1	0.2405	3.0	1	0.0827
56	56	6.4	2	**0.0403***	3.2	1	0.0737	3.2	1	0.0725
57	57	10.2	2	**0.0060***	3.2	1	0.0758	7.1	1	0.0078
58	58	4.5	2	0.1039	4.5	1	0.0334	0.0	1	0.9713
59	59	1.6	2	0.4571	1.6	1	0.2140	0.0	1	0.8902
60	60	4.8	2	0.0921	2.1	1	0.1504	2.7	1	0.1021
61	61	9.7	2	**0.0078***	8.9	1	0.0029	0.8	1	0.3592
62	62	4.9	2	0.0851	0.0	1	0.8472	4.9	1	0.0270
63	63	4.6	2	0.1023	4.6	1	0.0327	0.0	1	0.9927

| Item numbers in: | | | | | | | | | | |
Group 1(bini)	Group 2(ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c\mid a}$	d.f.	P
64	64	0.2	2	0.9197	0.2	1	0.6828	0.0	1	0.9872
65	65	6.5	2	**0.0384***	0.7	1	0.4088	5.8	1	0.0157
66	66	9.9	2	0.0070	7.3	1	0.0069	2.6	1	0.1060
67	67	8.2	2	0.0165	1.7	1	0.1992	6.5	1	0.0105
68	68	10.8	2	**0.0046***	1.4	1	0.2381	9.4	1	0.0022
69	69	2.8	2	0.2470	0.8	1	0.3774	2.0	1	0.1558
70	70	1.1	2	0.5854	0.6	1	0.4527	0.5	1	0.4768

*Asterisks * shows DIF items. Critical χ2 value = 5.99 at df (2) at 0.05 sig level.*

From table 1, it can be seen that items item 11, Wald χ2 (2) = 10.5, item 12 χ2 =8.8, item 14, χ2 = 14.7, item 15, χ2 =6.4, item 16, χ2 = 9.8, item 22, χ2 =7.4, item 29, χ2 = 6.5, item 29 χ2 =15.1, item 32, χ2 =8.0, item 34, χ2 = 9.9, item 35, χ2 = 8.5, item 45, χ2 = 10.6, item 49, χ2 =7.3, item 56, χ2 = 6.4, item 57, χ2 =10.2, item 61, χ2 =9.7, item 65, χ2 = 6.5 and item 68, χ2 = 10.8 shows DIF as seen from their Lord Wald χ2 value.

The result in table 1 further reveals the p-values for the Wald x2 statistic revealing items that significantly functions differentially at 0.05 levels of significance. From the table, it can be viewed that Item 11 has p =.0053, p< 0.05, item 12 has p = .0120, p, < 0.05, item 14 has p = .0007, p< 0.05, Item 15 has p =.0400, p< 0.05, Item 16 has p =.0076, p,< 0.05, Item 22 has p =.0242, p< 0.05, Item 23 has p =.0084, p,< 0.05, Item 29 has p =.0005, p< 0.05, Item 32 has p =.0182, p< 0.05, Item 34 has p =.0071, p< 0.05, Item 35 has p =.0144, p< 0.05, Item 45 has p =.0050, item 49 has p= 0254, p< 0.05, Item 56 has p =.0403, p< 0.05, Item 57 has p =.0060, p< 0.05, Item 61 has p =.0078, p< 0.05, Item 65 has p =.0384, p< 0.05, and Item 68 has p =.0046, p< 0.05showing that these items significantly differentially functions between the focal group Ijaw and the reference group Bini. Thus the DIF items are functioning significantly against the focal group Ijaw in the English Language achievement test between Bini and Ijaw.

Research Question 2: Which are the items in the English language multiple choice that functions significantly differentially between the focal group (Ijaw) and reference group (Esan) of students? To answer this research question, the 70 were subjected to DIF analysis using IRTPRO. Items that exhibited DIF were items that differentiate significantly between group membership and this is represented in table 2

Table 2; Differential Item Functioning Statistics for Esan and Ijaw

| Item numbers in: | | | | | | | | | | |
Group (Esan)	Group 2(Ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c\mid a}$	d.f.	P
1	1	0.0	2	1.0000	0.0	1	1.0000	0.0	1	1.0000
2	2	3.4	2	0.1866	0.7	1	0.4149	2.7	1	0.1011
3	3	0.0	2	1.0000	0.0	1	1.0000	0.0	1	1.0000
4	4	0.1	2	0.9750	0.0	1	0.9064	0.0	1	0.8479
5	5	21.3	2	**0.0001***	3.1	1	0.0805	18.2	1	0.0001
6	6	3.2	2	0.2069	3.1	1	0.0768	0.0	1	0.8578
7	7	5.0	2	0.0824	3.8	1	0.0511	1.2	1	0.2772
8	8	3.3	2	0.1948	3.3	1	0.0700	0.0	1	0.9747
9	9	2.9	2	0.2357	0.1	1	0.7274	2.8	1	0.0966
10	10	0.3	2	0.8679	0.2	1	0.6279	0.0	1	0.8264
11	11	13.8	2	**0.0010***	0.1	1	0.7681	13.7	1	0.0002
12	12	1.5	2	0.4642	1.3	1	0.2567	0.2	1	0.6195
13	13	3.5	2	0.1783	3.4	1	0.0645	0.0	1	0.8299
14	14	2.5	2	0.2946	1.9	1	0.1717	0.6	1	0.4488
15	15	20.3	2	**0.0001***	8.5	1	0.0036	11.8	1	0.0006
16	16	9.9	2	**0.0070***	9.9	1	0.0016	0.0	1	0.9066
17	17	7.6	2	**0.0227***	4.1	1	0.0436	3.5	1	0.0614
18	18	1.2	2	0.5533	1.2	1	0.2823	0.0	1	0.8705
19	19	5.4	2	0.0683	5.3	1	0.0210	0.0	1	0.8482
20	20	3.4	2	0.1862	2.8	1	0.0976	0.6	1	0.4341

Item numbers in:											
Group (Esan)	Group 2(Ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c\|a}$	d.f.	P	
21	21	8.7	2	0.0129	7.4	1	0.0064	1.3	1	0.2621	
22	22	10.4	2	**0.0054***	10.4	1	0.0012	0.0	1	0.9121	
23	23	5.9	2	0.0525	2.2	1	0.1390	3.7	1	0.0546	
24	24	10.4	2	**0.0054***	10.4	1	0.0012	0.0	1	0.9148	
25	25	0.0	2	1.0000	0.0	1	1.0000	0.0	1	1.0000	
26	26	2.4	2	0.3019	1.9	1	0.1646	0.5	1	0.4973	
27	27	2.9	2	0.2380	1.7	1	0.1945	1.2	1	0.2760	
28	28	13.4	2	**0.0013***	8.7	1	0.0032	4.7	1	0.0306	
29	29	9.2	2	**0.0098***	8.8	1	0.0030	0.4	1	0.5145	
30	30	3.8	2	0.1511	3.7	1	0.0554	0.1	1	0.7200	
31	31	8.4	2	**0.0148***	5.6	1	0.0183	2.9	1	0.0916	
32	32	3.8	2	0.1543	0.6	1	0.4539	3.2	1	0.0740	
33	33	11.5	2	**0.0031***	10.4	1	0.0013	1.1	1	0.2899	
34	34	2.2	2	0.3306	0.3	1	0.5857	1.9	1	0.1666	
35	35	0.1	2	0.9484	0.1	1	0.7738	0.0	1	0.8786	
36	36	9.1	2	**0.0106***	6.1	1	0.0132	3.0	1	0.0869	
37	37	0.7	2	0.6924	0.2	1	0.6899	0.6	1	0.4481	
38	38	0.3	2	0.8707	0.3	1	0.6096	0.0	1	0.8995	
39	39	1.1	2	0.5736	0.9	1	0.3305	0.2	1	0.6859	
40	40	1.0	2	0.6162	0.7	1	0.4163	0.3	1	0.5795	
41	41	1.7	2	0.4360	1.5	1	0.2284	0.2	1	0.6487	
42	42	4.0	2	0.1322	0.5	1	0.4602	3.5	1	0.0614	
43	43	0.6	2	0.7481	0.2	1	0.6340	0.4	1	0.5522	
44	44	1.2	2	0.5580	0.5	1	0.4846	0.7	1	0.4105	
45	45	4.4	2	0.1092	4.4	1	0.0354	0.0	1	0.9780	
46	46	0.7	2	0.6968	0.5	1	0.4593	0.2	1	0.6763	
47	47	5.5	2	0.0649	1.0	1	0.3147	4.5	1	0.0348	
48	48	1.8	2	0.3992	0.3	1	0.5601	1.5	1	0.2215	
49	49	0.4	2	0.8245	0.0	1	0.8607	0.4	1	0.5514	
50	50	9.1	2	**0.0106***	3.9	1	0.0481	5.2	1	0.0228	
51	51	7.7	2	**0.0216***	2.7	1	0.0998	4.9	1	0.0261	
52	52	3.1	2	0.2109	3.0	1	0.0812	0.1	1	0.7722	
53	53	4.1	2	0.1284	0.8	1	0.3827	3.3	1	0.0676	
54	54	0.6	2	0.7427	0.0	1	0.9775	0.6	1	0.4411	
55	55	0.8	2	0.6628	0.4	1	0.5141	0.4	1	0.5290	
56	56	0.5	2	0.7635	0.3	1	0.6031	0.3	1	0.6040	
57	57	1.2	2	0.5373	0.1	1	0.7794	1.2	1	0.2810	
58	58	2.6	2	0.2762	0.3	1	0.5658	2.3	1	0.1345	
59	59	9.0	2	**0.0109**	5.8	1	0.0160	3.2	1	0.0719	
60	60	11.5	2	**0.0032***	6.6	1	0.0103	4.9	1	0.0271	
61	61	6.2	2	0.0455	4.4	1	0.0349	1.7	1	0.1891	
62	62	1.1	2	0.5741	0.0	1	0.9137	1.1	1	0.2950	
63	63	12.7	2	**0.0018***	5.9	1	0.0151	6.8	1	0.0093	
64	64	7.5	2	0.0236	6.7	1	0.0095	0.8	1	0.3827	
65	65	1.1	2	0.5753	0.4	1	0.5215	0.7	1	0.4048	

Item numbers in:											
Group (Esan)	Group 2(Ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c\|a}$	d.f.	P
66	66	1.6	2	0.4555	1.5	1	0.2171	0.0	1	0.8278
67	67	4.3	2	0.1145	0.7	1	0.4075	3.6	1	0.0562
68	68	3.0	2	0.2228	1.4	1	0.2364	1.6	1	0.2059
69	69	1.7	2	0.4205	1.4	1	0.2437	0.4	1	0.5419
70	70	0.5	2	0.7667	0.0	1	0.8643	0.5	1	0.4788

*Asterisks * shows DIF items. Critical χ2 value = 5.99 at df (2) at 0.05 sig level.*

Table 2 shows that items 3, 14, 15, 17, 22, 28, 29, 33, 51,60, 63, 66, and 67 shows DIF as seen from their Lord Wald χ2 value. Item 5, Wald χ2 (2) = 21.3, item 11, χ2 (2) = 13.8, item 15, χ2 (2) =20.3, item 16, χ2 (2) = 9.9, item 17, χ2 (2) = 7.9, item 22, χ2 (2) = 10.4, item 24, χ2 (2) =10.4, item 28, χ2 (2) = 13.4, item 29, χ2 (2) =9.2, item 31, χ2 (2) =8.4, item 33, χ2 (2) = 11.5, item 36, χ2 (2) =9.1, item 50, χ2 (2) =9.1, item 51, χ2 (2) = 7.7 item 59, χ2 (2) = 9.0, item 60, χ2 (2) = 11.5, and item 63, χ2 (2) = 12.7 has the above as their χ2 value.

The result in table 2 further reveals the p-values for the Wald χ2 statistics that tests the difference between reference (Esan) and focal (Ijaw) group item parameters (a* & b) revealing items that significantly functions differentially at 0.05 level of significance. From the table, it can be seen that Item 5 has p =.0001, p< 0.05, item 11 has p = .0010, p, < 0.05, item 15 has p = .0001, p< 0.05, Item 16 has p =.0070, p< 0.05, Item 17 has p =.0227, p< 0.05, Item 22 has p =.0054, p< 0.05, Item 24 has p =.0054, p< 0.05, Item 28 has p =.0013, p< 0.05, Item 29 has p =.0098, p< 0.05, Item 31 has p =.0148, p< 0.05, Item 33 has p =.0031, p< 0.05, Item 36 has p =.0106, p< 0.05, Item 50 has p =.0106, p< 0.05, Item 51 has p =.0210, p< 0.05, Item 59 has p =.0109, p< 0.05, Item 60 has p =.0032, p< 0.05 and Item 63 has p =.0018, p< 0.05 revealing that these items significantly differentially functions between the focal group Ijaw and the reference group Esan. Implication therefore is that the DIF items are functioning significantly against the focal group Ijaw in the English Language achievement test

Research Question 3: Which are the items in the English language multiple choice that functions significantly differently between the focal group (Ijaw) and reference group (Isoko) of students?

To answer this research question, the 70 items were subjected to DIF analysis using IRTPRO. Items that exhibited DIF were items that differentiate significantly between group membership and this is represented in table 3

Table 3; Differential Item Functioning Statistics for Isoko and Ijaw

Item numbers in:											
Group 1(Isoko)	Group 2(Ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c\|a}$	d.f.	P
1	1	0.8	2	0.6756	0.8	1	0.3821	0.0	1	0.8893
2	2	0.3	2	0.8663	0.3	1	0.6007	0.0	1	0.9095
3	3	3.8	2	0.1531	3.5	1	0.0622	0.3	1	0.5891
4	4	13.5	2	**0.0012***	7.7	1	0.0055	5.8	1	0.0164
5	5	3.8	2	0.1543	3.5	1	0.0627	0.3	1	0.5910
6	6	4.4	2	0.1096	1.4	1	0.2397	3.0	1	0.0816
7	7	0.7	2	0.7100	0.6	1	0.4493	0.1	1	0.7377
8	8	0.3	2	0.8421	0.3	1	0.5792	0.0	1	0.8495
9	9	0.3	2	0.8565	0.3	1	0.5819	0.0	1	0.9363
10	10	11.5	2	**0.0031***	5.3	1	0.0212	6.2	1	0.0126
11	11	4.0	2	0.1330	1.0	1	0.3255	3.1	1	0.0800
12	12	4.5	2	0.1076	4.3	1	0.0386	0.2	1	0.6729
13	13	4.6	2	0.1018	2.4	1	0.1219	2.2	1	0.1424
14	14	7.7	2	**0.0211***	5.2	1	0.0222	2.5	1	0.1155
15	15	6.0	2	**0.0501***	5.8	1	0.0157	0.1	1	0.7020
16	16	3.1	2	0.2128	2.8	1	0.0961	0.3	1	0.5728
17	17	1.1	2	0.5859	0.5	1	0.4716	0.6	1	0.4582
18	18	0.9	2	0.6452	0.5	1	0.5006	0.4	1	0.5159
19	19	3.5	2	0.1753	0.6	1	0.4588	2.9	1	0.0871
20	20	1.8	2	0.4169	0.0	1	0.8931	1.7	1	0.1885
21	21	9.0	2	**0.0113***	0.9	1	0.3449	8.1	1	0.0045

| Item numbers in: | | | | | | | | | | | |
Group 1(Isoko)	Group 2(Ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c\|a}$	d.f.	P
22	22	1.1	2	0.5884	0.8	1	0.3718	0.3	1	0.6086
23	23	3.4	2	0.1864	3.3	1	0.0676	0.0	1	0.8551
24	24	3.7	2	0.1569	3.4	1	0.0642	0.3	1	0.5867
25	25	3.4	2	0.1845	0.3	1	0.6048	3.1	1	0.0771
26	26	1.2	2	0.5607	0.9	1	0.3569	0.3	1	0.5794
27	27	3.7	2	0.1606	3.6	1	0.0562	0.0	1	0.8662
28	28	5.4	2	0.0677	3.8	1	0.0525	1.6	1	0.2035
29	29	7.8	2	**0.0201***	7.8	1	0.0053	0.0	1	0.8259
30	30	4.4	2	0.1113	4.0	1	0.0444	0.3	1	0.5547
31	31	9.8	2	0.0075	6.8	1	0.0093	3.0	1	0.0819
32	32	13.2	2	**0.0013***	1.2	1	0.2802	12.1	1	0.0005
33	33	7.8	2	**0.0206***	4.5	1	0.0330	3.2	1	0.0726
34	34	1.7	2	0.4186	0.2	1	0.6966	1.6	1	0.2077
35	35	5.4	2	0.0659	4.6	1	0.0325	0.9	1	0.3528
36	36	1.2	2	0.5511	1.1	1	0.3024	0.1	1	0.7216
37	37	1.8	2	0.4040	1.2	1	0.2800	0.6	1	0.4222
38	38	2.5	2	0.2946	1.7	1	0.1913	0.7	1	0.3913
39	39	1.8	2	0.4145	0.8	1	0.3695	1.0	1	0.3284
40	40	1.1	2	0.5649	0.2	1	0.6950	1.0	1	0.3204
41	41	2.8	2	0.2416	1.0	1	0.3099	1.8	1	0.1788
42	42	1.8	2	0.4128	0.3	1	0.5766	1.5	1	0.2276
43	43	4.4	2	0.1092	4.3	1	0.0387	0.2	1	0.6933
44	44	2.5	2	0.2899	2.5	1	0.1166	0.0	1	0.9293
45	45	6.0	2	0.0491	4.5	1	0.0339	1.5	1	0.2171
46	46	7.3	2	**0.0265***	0.7	1	0.4170	6.6	1	0.0102
47	47	4.8	2	0.0907	0.8	1	0.3844	4.0	1	0.0444
48	48	0.0	2	0.9980	0.0	1	0.9574	0.0	1	0.9740
49	49	3.3	2	0.1968	1.3	1	0.2619	2.0	1	0.1582
50	50	1.4	2	0.5073	1.3	1	0.2481	0.0	1	0.8834
51	51	4.1	2	0.1294	3.6	1	0.0591	0.5	1	0.4682
52	52	3.9	2	0.1454	2.3	1	0.1310	1.6	1	0.2095
53	53	2.2	2	0.3321	2.1	1	0.1488	0.1	1	0.7332
54	54	12.5	2	**0.0019***	4.1	1	0.0429	8.4	1	0.0037
55	55	1.1	2	0.5846	1.0	1	0.3136	0.1	1	0.8112
56	56	4.9	2	0.0849	4.9	1	0.0273	0.1	1	0.8019
57	57	8.3	2	**0.0159***	5.7	1	0.0173	2.6	1	0.1070
58	58	3.9	2	0.1466	3.8	1	0.0514	0.1	1	0.8018
59	59	7.3	2	**0.0265***	7.0	1	0.0084	0.3	1	0.5811
60	60	1.0	2	0.6187	0.1	1	0.7446	0.9	1	0.3557
61	61	4.5	2	0.1040	4.5	1	0.0336	0.0	1	0.9242
62	62	1.9	2	0.3809	0.1	1	0.7900	1.9	1	0.1731
63	63	5.4	2	0.0684	2.0	1	0.1634	3.4	1	0.0647
64	64	3.4	2	0.1847	0.0	1	0.9772	3.4	1	0.0655
65	65	7.3	2	**0.0262***	6.6	1	0.0104	0.7	1	0.3977
66	66	0.8	2	0.6710	0.7	1	0.3885	0.1	1	0.8173

Item numbers in:										
Group 1(Isoko)	Group 2(Ijaw)	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c|a}$	d.f.	P
67	67	1.4	2	0.4865	0.2	1	0.6628	1.3	1	0.2637
68	68	5.3	2	0.0721	3.5	1	0.0613	1.8	1	0.1859
69	69	1.0	2	0.5999	0.3	1	0.5802	0.7	1	0.3977
70	70	1.5	2	0.4705	0.3	1	0.6102	1.3	1	0.2643

*Asterisks * shows DIF items. Critical χ2 value = 5.99 at df (2) at 0.05 sig level.*

From table 3 it can be seen that items item 4, Wald χ2 (2) = 13.5, item 10 χ2 =11.5, item 14, χ2 = .7.7, item 15, χ2 =6.0, item 21, χ2 =9.0, item 29, χ2 =7.8, item 32, χ2 = 13.2, item 33, χ2 =7.8, item 46, χ2 =7.3, item 54, χ2 = 12.5, item 57, χ2 = 8.3, item 59, χ2 = 7.3, and item 65, χ2 =7.3, shows DIF as seen from their Lord Wald χ2 value.

The result in table 3 further reveals the p-values for the Wald χ2 statistic revealing items that significantly functions differentially at 0.05 level of significance. From the table, it can be observed that Item 4 has p =.0012, p< 0.05, item 10 has p = .0031, p, < 0.05, item 14 has p = .0211, p< 0.05, Item 15 has p =.0501, p< 0.05, Item 21 has p =.0113, p< 0.05, Item 29 has p =.0201, p< 0.05, Item 32 has p =.0013, p < 0.05, Item 33 has p =.0206, p< 0.05, Item 46 has p =.0265, p< 0.05, Item 54 has p =.0019, p< 0.05, Item 57 has p =.0159, p< 0.05, Item 59 has p =.0265, and item 65 has p= 0265, p< 0.05,revealing that these items significantly differentially functions. Therefore, between the focal group Ijaw and the reference group Isoko, these items differentially functions significantly and it is functioning against the focal group Ijaw.

Research Question 4: Which are the items in the English language multiple choice that functions significantly differently between the focal group (Ijaw) and reference group (Urhobo) of students?

To answer this research question and hypothesis, the 70 items were subjected to DIF analysis using IRTPRO. Items that reveal DIF were items that differentiate significantly between group membership and this is represented in table 4

Table 4; Differential Item Functioning Statistics for Urhobo and Ijaw

Item numbers in:										
Group 1(Uhbo)	Group(Ijaw) 2	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c|a}$	d.f.	P
1	1	2.7	2	0.2662	2.1	1	0.1438	0.5	1	0.4765
2	2	4.8	2	0.0924	4.4	1	0.0364	0.4	1	0.5377
3	3	4.1	2	0.1296	0.3	1	0.5729	3.8	1	0.0523
4	4	7.9	2	**0.0191***	7.3	1	0.0071	0.7	1	0.4167
5	5	4.1	2	0.1292	0.3	1	0.5730	3.8	1	0.0521
6	6	0.5	2	0.7612	0.5	1	0.4716	0.0	1	0.8697
7	7	2.5	2	0.2881	2.5	1	0.1167	0.0	1	0.8849
8	8	0.5	2	0.7984	0.4	1	0.5494	0.1	1	0.7623
9	9	1.9	2	0.3923	1.0	1	0.3277	0.9	1	0.3396
10	10	1.1	2	0.5717	0.5	1	0.4728	0.6	1	0.4378
11	11	11.3	2	**0.0036***	0.1	1	0.7575	11.2	1	0.0008
12	12	0.4	2	0.8238	0.0	1	0.9292	0.4	1	0.5379
13	13	0.9	2	0.6332	0.7	1	0.4055	0.2	1	0.6382
14	14	2.2	2	0.3306	0.1	1	0.7161	2.1	1	0.1495
15	15	5.5	2	0.0638	0.4	1	0.5047	5.1	1	0.0245
16	16	7.0	2	**0.0294***	2.1	1	0.1454	4.9	1	0.0265
17	17	2.7	2	0.2617	1.4	1	0.2341	1.3	1	0.2609
18	18	1.3	2	0.5204	0.7	1	0.4117	0.6	1	0.4267
19	19	3.2	2	0.1998	3.1	1	0.0768	0.1	1	0.7477
20	20	0.8	2	0.6843	0.3	1	0.5721	0.4	1	0.5076
21	21	1.2	2	0.5414	0.0	1	0.8583	1.2	1	0.2746
22	22	4.4	2	0.1103	4.2	1	0.0409	0.2	1	0.6349
23	23	2.9	2	0.2357	1.5	1	0.2257	1.4	1	0.2331

| Item numbers in: | | | | | | | | | | | |
Group 1(Uhbo)	Group(Ijaw) 2	Total X^2	$d.f.$	P	X^2_a	$d.f.$	P	$X^2_{c|a}$	$d.f.$	P
24	24	4.1	2	0.1272	0.3	1	0.5743	3.8	1	0.0510
25	25	9.1	2	**0.0107***	4.8	1	0.0278	4.2	1	0.0397
26	26	1.0	2	0.5931	0.6	1	0.4288	0.4	1	0.5180
27	27	5.0	2	0.0809	4.9	1	0.0273	0.2	1	0.6937
28	28	8.9	2	**0.0118***	8.9	1	0.0029	0.0	1	0.9553
29	29	4.2	2	0.1197	2.8	1	0.0970	1.5	1	0.2256
30	30	1.3	2	0.5336	1.1	1	0.2931	0.2	1	0.6988
31	31	1.6	2	0.4533	0.4	1	0.5303	1.2	1	0.2759
32	32	1.0	2	0.5944	0.0	1	0.9391	1.0	1	0.3094
33	33	16.3	2	**0.0003***	6.9	1	0.0085	9.4	1	0.0022
34	34	5.0	2	0.0819	0.3	1	0.5684	4.7	1	0.0306
35	35	2.2	2	0.3345	1.9	1	0.1705	0.3	1	0.5788
36	36	0.5	2	0.7888	0.3	1	0.5649	0.1	1	0.7056
37	37	0.1	2	0.9373	0.0	1	0.9157	0.1	1	0.7310
38	38	0.2	2	0.9006	0.2	1	0.6476	0.0	1	0.9882
39	39	1.1	2	0.5671	1.1	1	0.2886	0.0	1	0.9354
40	40	2.7	2	0.2581	2.4	1	0.1227	0.3	1	0.5712
41	41	2.5	2	0.2847	0.2	1	0.6813	2.4	1	0.1262
42	42	4.0	2	0.1396	0.5	1	0.4908	3.5	1	0.0621
43	43	2.4	2	0.2990	0.5	1	0.4774	1.9	1	0.1673
44	44	2.2	2	0.3292	0.0	1	0.8675	2.2	1	0.1389
45	45	6.5	2	**0.0386***	5.4	1	0.0202	1.1	1	0.2922
46	46	3.8	2	0.1480	0.4	1	0.5124	3.4	1	0.0649
47	47	6.5	2	**0.0392***	3.5	1	0.0617	3.0	1	0.0851
48	48	0.2	2	0.9206	0.1	1	0.8064	0.1	1	0.7456
49	49	6.6	2	**0.0369***	4.2	1	0.0400	2.4	1	0.1238
50	50	4.4	2	0.1107	0.2	1	0.6778	4.2	1	0.0398
51	51	8.4	2	**0.0148***	6.1	1	0.0132	2.3	1	0.1326
52	52	4.3	2	0.1171	0.2	1	0.6371	4.1	1	0.0438
53	53	3.3	2	0.1925	3.1	1	0.0786	0.2	1	0.6435
54	54	1.4	2	0.4988	0.1	1	0.8150	1.3	1	0.2480
55	55	2.2	2	0.3331	0.5	1	0.4661	1.7	1	0.1969
56	56	1.2	2	0.5369	0.4	1	0.5102	0.8	1	0.3684
57	57	5.6	2	0.0602	2.4	1	0.1252	3.3	1	0.0712
58	58	7.3	2	**0.0264***	7.2	1	0.0071	0.0	1	0.8672
59	59	8.0	2	**0.0180***	8.0	1	0.0046	0.0	1	0.8966
60	60	0.0	2	0.9998	0.0	1	0.9923	0.0	1	0.9851
61	61	2.8	2	0.2433	2.8	1	0.0936	0.0	1	0.9288
62	62	0.3	2	0.8763	0.3	1	0.6108	0.0	1	0.9440
63	63	3.5	2	0.1790	3.2	1	0.0752	0.3	1	0.5911
64	64	1.4	2	0.4958	0.4	1	0.5058	1.0	1	0.3273
65	65	5.6	2	0.0596	3.1	1	0.0799	2.6	1	0.1097
66	66	5.3	2	0.0703	4.0	1	0.0456	1.3	1	0.2529
67	67	11.4	2	**0.0033***	2.0	1	0.1628	9.5	1	0.0021
68	68	5.8	2	0.0541	5.5	1	0.0193	0.4	1	0.5506

Item numbers in:											
Group 1(Uhbo)	Group(Ijaw) 2	Total X^2	d.f.	P	X^2_a	d.f.	P	$X^2_{c|a}$	d.f.	P	
69	69	2.7	2	0.2667	2.1	1	0.1459	0.5	1	0.4686	
70	70	0.9	2	0.6440	0.8	1	0.3723	0.1	1	0.7729	

*Asterisks * shows DIF items. Critical χ2 value = 5.99 at df (2) at 0.05 sig level.*

Table 4. reveals that items item 4, Wald χ2 (2) = 7.9, item 11 χ2 =11.3, item 16, χ2 = .7.0, item 16, χ2 =7.0, item 25, χ2 =9.1, item 28, χ2 =8.9, item 33, χ2 = 16.3, item 45, χ2 =6.5, item 47, χ2 =6.5, item 49, χ2 = 6.6, item 51, χ2 = 8.4, item 58, χ2 = 7.3, and item 59, χ2 =8.0, shows DIF as seen from their Lord Wald χ2 value.

The result in table 4 further reveals the p-values for the Wald χ2 statistic revealing items that significantly functions differentially at 0.05 level of significance. From the table, it can be viewed that Item 4 has p =.0191, p< 0.05, item 11 has p = .0036, p, < 0.05, item 16 has p = .0294, p< 0.05, Item 25 has p =.0107, p< 0.05, Item 28 has p =.0118, p,< 0.05, Item 33 has p =.0003, p< 0.05, Item 45 has p =.0386, p,< 0.05, Item 47 has p =.0386, p< 0.05, Item 49 has p =.0369, p< 0.05, Item 51 has p =.0148, p< 0.05, Item 58 has p =.0264, p < 0.05, and item 59 has p= 0180, p< 0.05, revealing that these items significantly differentially functions. Therefore, between the focal group Ijaw and the reference group Urhobo, these items differentially functions significantly against the focal Ijaw group favoring the Urhobo reference group.

Hypothesis 1:

1. The numbers of items that function differently in English language test is not significantly dependent on ethnic groups
2. To test the null hypothesis, data was subjected to Chi Square test and the result is represented in table 5

Table 5; Chi Square test of the items that differentially functioned based on ethnic groups

			DIF		
			DIF	NO DIF	Total
Ethnicity	Ijaw & Bini	Count	67	3	70
		Expected Count	16.8	53.3	70.0
		% within Ethnicity	95.7%	4.3%	100.0%
		% within DIF	100.0%	1.4%	25.0%
		% of Total	23.9%	1.1%	25.0%
		Residual	50.3	-50.3	
	Ijaw & Esan	Count	0	70	70
		Expected Count	16.8	53.3	70.0
		% within Ethnicity	0.0%	100.0%	100.0%
		% within DIF	0.0%	32.9%	25.0%
		% of Total	0.0%	25.0%	25.0%
		Residual	-16.8	16.8	
	Ijaw & Isoko	Count	0	70	70
		Expected Count	16.8	53.3	70.0
		% within Ethnicity	0.0%	100.0%	100.0%
		% within DIF	0.0%	32.9%	25.0%
		% of Total	0.0%	25.0%	25.0%
		Residual	-16.8	16.8	
	Ijaw & Urhobo	Count	0	70	70
		Expected Count	16.8	53.3	70.0
		% within Ethnicity	0.0%	100.0%	100.0%
		% within DIF	0.0%	32.9%	25.0%
		% of Total	0.0%	25.0%	25.0%
		Residual	-16.8	16.8	
Total		Count	67	213	280
		Expected Count	67.0	213.0	280.0
		% within Ethnicity	23.9%	76.1%	100.0%
		% within DIF	100.0%	100.0%	100.0%
		% of Total	23.9%	76.1%	100.0%

Chi-Square Tests

	Value	Df	Asymptotic Significance (2-sided)
Pearson Chi-Square	264.225[a]	3	.000
Likelihood Ratio	283.374	3	.000
Linear-by-Linear Association	157.969	1	.000
N of Valid Cases	280		

The table 5 shows that the expected counts for all the ethnics groups are greater than 5. It further shows that the residual for both the expected and observed for all the ethnic groups are same. The Chi square χ2 value of 264.22 (3) p < 0.5, i.e. p = .000 is less than 0.05 and this is statistically significant at the chosen alpha level of 0.05 for all the ethnic groups. Therefore, the numbers of items that function differentially in English language test is significantly dependent on the ethnic group as p < 0.05 i.e. p is .000 is less than .005. The null hypothesis that the numbers of items that function in English language test is not significantly dependent on the ethnic group is rejected and the alternate accepted

Discussion of Findings

Findings from this present study reveals that of the 70 items based on ethnicity that is between the Ijaw focal group and the Binis, Esan, Isoko, Urhobo reference group, 18 items functioned differentially significantly between the Ijaw focal group and Bini reference group representing 25.%, 17 items functioned differentially significantly between the Ijaw focal group and Esan reference group representing 24%, 13 items exhibited DIF significantly between the Ijaw focal group and Isoko reference group representing 18% and 12 items were identified as significantly exhibiting DIF between the Ijaw focal group and Bini reference group representing 17.1% as seen from their Wald Chi- square values which were greater than the critical value of 5.99 at df 2 as well as their p-values were all significant at .05 (p< .05).

On the whole, 40 items were flagged as showing DIF based on ethnicity this represents as well a percentage of 57% of the total percentage for the 70 items. Implication is that the English Language achievement test used in WASSCE 2016 contains items that significantly functions differently. Empirical studies have revealed that the percentage of DIF items ranges from quite small 1.5% to overwhelming large (64%).

Studies consider it a small amount when a test contains less than 10% DIF, a medium amount of DIF when a test contains 10 to 30%, large DIF when it exceeds 30% when the parentage of DIF exceeds 10% closer attention should be paid to it (Hambleton & Roger, 1989; Raju 1989; Xioting 2010). That the items contain 50% items flagged as showing DIF indicates a large amount or magnitude of DIF.

That an item that functioning differentially may not be biased but biased item must function differentially This indicates that some of the items have potentials to be biased as all biased items must differentially function and DIF is the empirical evidence used to refute or support bias (Brown 2012; Cohen 2006, Hambletbon & Rodgers 1995 Karami & Nedoushan 2011; Schumacker 2005; Perrone 2006; Williams 1997; Zumbo 1999; 2007).

Reason why there are items functioning differently for the groups could probably be disparity in the two groups of examinees exposure to the content or vocabulary contained in the item. In consonant with this is the study by Queeensoap (2014) whose research result showed 30 items exhibiting DIF, between the focal group (Ijaw) and reference group (Yoruba, Hausa & Igbo). With 1% showing DIF for Ijaw focal group and Yoruba reference group, 17% showing DIF for Ijaw focal group and Igbo reference group and 7% showing DIF for Ijaw focal group and Hausa reference group. Similar to this is the investigation by Hambleton and Rogers (1989) where the result showed that some of the items functioned differently against white test takers from the rural settlement in the mathematical and verbal component of SAT. Also in consonant with the findings of this present work is the findings of Yang & Jones (2007) whose results indicated measurement bias attributable to race which was significant for two (CES-D) items.

Result however by Engelhard, et al (2013) revealed that overall there did not appear to be any item subsets functioning in an unexpected way across the subgroups of persons (gender, race, ethnicity and best language subgroup). This divergent finding can be attributed to the fact that this current study is not foreign while that of Engelhard, et al (2013) is foreign, another reason could be that the later used Maentel Haenszel which is an observed score method technique of detecting DIF while this current study employed an IRT based method using IRTPRO that employs Wald test.

Recommendations

Based on the findings of the study, the following recommendations were made;

1.   Items identified as showing DIF in a large percentage in large scale or public examinations should be further investigated using qualitative analysis content analysis by subject matter experts. On such investigations, such DIF items can either be edited or eliminated from a test or item bank if it is ascertained

to exhibit bias Test developers, and large scale examination bodies like WAEC, should employ more IRT framework especially specialize IRT software like IRTPRO in detecting items that are differentially functioning for testees so that test items are valid, reliable and useable and thus reduce if not eliminating bias that may exist in test items for examinees of different groups WAEC and other public examination bodies should analyze their items for DIF before building it into their test bank.

2. Psychometricians, government private firms and other stake holders should implore DIF in detecting bias items. For certifications and admission into higher schools, government should make laws that encourage fair use of test scores as this will help handle the interest of examinees of matched abilities from diverse subgroups like ethnicity, socio-economic status, gender, location and the likes Test developers and examining bodies that are working within only CTT measurement framework should incorporate IRT framework into their test development framework.

References

Abedalaziz, N. (2011). Detecting dif using item characteristics curve approaches. *The international Journal of Educational and Psychology Assessment*, 8 (2), 1-15.

Adekola, B.O., Shoaga, O. &Lawal F.O. (2015). Performance of Senior Secondary School Student in English Language as correlate of academic performance in General Subjects. *Britishjournal of education, Society &Behavioural Science,* 6(4), 335-340 D10:109734/BJESBS/2015/5562.

Aina, J.K, Ogundele A.G, &Olanipekun S.S. (2013). Students' proficiency in English language relationship with academic performance in science and technical education. *American Journal of EducationalResearch* 1, (9) 355-358- D01: 10. 12691/education – 1-9-2.

Anastasi, A. & Urbina, S. (2007). *Psychological testing* (7thed). Upper saddle River, NJ: Prentice Hall.

Ariani, & M.G. &Ghafournia, N. (2015). The relationship between socio-economic status and beliefs about language learning: A study of irania post graduate EAP students. *English Language Teaching*, 8(9), 17-25.

Battuaz, M. (2017) On Wald's test on differential item functioning detection method. Retrieved on 6/4/2017 from http//www.

Breland, H.M., Bonner M.W. Dkubota M.U. (1995). Factors performance on brief impromptu essay examinations. New York; The College Board.

Bottomley, G. (1997). Identification: Ethnicity, gender, culture. *Journal ofInternational Studies*,18(1), 41-48.

Brown, A. (2012). *Measurement invariance*. Cambridge: CambridgeUniversity press

Camilli, G. (2006). Test fairness: In R.L. Brennan, (Ed), *Educationalmeasurement* (4th ed., pp. 220-256) Westport, CT: Praeger publishers.

Carroll, H-F. C. (2015). An examination of the improved walds test for differential item function detection for multiple groups. (Unpublished doctoral dissertation). University of Kansas, Chapel Hill

Cai, L., du Toit, S. H. C., &Thissen, D. (2011). IRTPRO.2.1 [Computer software]. Lincolnwood, IL: Scientific Software International.

Cohen L.A (2006). Item bias and differential item functioning. In T.M. Wood & W. Zhu (eds), *measurement theory and practice in kinesiology.* NY: Human Kinetics Publishers,

Enamiroro, E.P. (2007). Education poverty and development in nigeria: The way forward.in the 21st century. *Journal of Social Science*, 14(1), 19-24

Engelthard, G., Wind, S.A., Kobrin, J.L. &Chajewski M. (2013). Differential item and person functioning in large –scale writing assessments within the context of the SAT. College Board Research Report no. 2013-6. New York: The college Board

Engelherd, L. Gordon B., Walker, E.V., &Gabrielson S. (1994). Writing task and gender influences on writing quality of black and white students. *Journal of Education Research* 87, 197-209.

Ertuby, C. & Russell, R.J.H. (1996): Dealing with Compatibility Problems of cross Cultural Data. Paper presented at the International Congress of Psychology. Montreal. August 1996.

Fatemi, A.H. &Khaghaminezhad, M.S. (2011). The roles of ethnicity in integrative test performance of male / female Iranian English learners of different language different language proficiency levels. *Journal of Teaching Language Skills*, 3(2), 29-56.

Federal Republic of Nigeria National Policy on Education (1st Ed) 2004. Lagos: NERDC press.

Fidalgo, A.M., Ferreres, D., & Muniz, J. (2004). Utility of the mantel-haenzel procedure for detecting differential item functioning in small samples. *Educational Psychological Measurement*, 64(6), 925-936.

Finch, H., & French, B. F. (2007). Detection of crossing differential item functioning item: A comparison of four methods. *Educational and Psychological Measurement, 67*, 565-582.

Gregory, R.J. (2004): Psychological Testing, Principles and Applications. Boston. Allyn and Bacon

Haung X, (2010). Differential item functioning: The consequence of language, curriculum, or culture? Unpublished Ph.D.dissertationUniversity of California. Berkeley

Hambleton, R. & Rodgers, J. (1995): Item bias review: *PracticalAssessment, Research and Evaluation*, 4(6).

Retrived on 10/11/2016 from http://PAREonline.net/getvn.asp/v=4andn=6

Hambleton, R.K. & Rogers, J.H. (1989). Detecting potentially biased test items: Comparison of IR area and mantel-Haenszel methods. *AppliedMeasurement in Education*, 2,313-334.

Hidalgo, M.H., & Lopez-Pina, J.A (2004). Differential item functioning detecting and effect size: A comparison between logistic regression and mantel-Haenszel procedures. *Educational and PsychologicalMeasurement*, 64. 903-915.

Hossein, K, (2011): Detecting Gender Bias in Language Proficiency Test.   *International Journal of Language Studies,* Vol. 5 no. 2 pg 22-38.

Kamata, A., & Vaughn, B.K. (2004.) An introduction to differential item functioning analysis. Learning Disabilities: *A contemporary Journal*, 2, 49-69.

Karami, H. &Nodoushan, M.A. (2011). Differential item functioning dif: Current problems and future directions. *International Journal ofLanguage Studies*, 5(3), 133-142.

Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test in detection of differential item functioning. *Applied Measurement in Education*, *8*, 291–312.

Lee, S.-H. (2015). Lord Wald test for detecting DIF in multidimensional IRT model; A comparism of two approaches. Unpublished doctoral dissertation. StateSchoolUniversity of New Jersey. New Jersey.

Lord, F.M. (1980) *Application of item response theory to practical testing problems*. Hillsdale, NJ. Erlbaum.

Mattern K. Camara, W. &Kobrin, J.L. (2007). SAT writing: An overview of research and psychometrics to date (RN-32). New York: The College Board.

McNamara, T., &Roever, C. (2006). *Language testing: Thesocial dimension*. Malden, MA&Oxford: Blackwell.

Nworgo, B.G. &Odili, J.N (2005). Analysis of dIF in waec, ssce biology multiple choice test. Retrieved on 10/09/2016 from http//ww.aeafrica.orgPAPERS/effectof20%wlagmanipulation/

Ogbebor, U. &Onuka, A. (2013). Differential item functioning method as an item bias indicator. *International Research Journal*, 4(4), 367-373

Odili, J.N. (2004) Effect of language manipulation on dif in biology multiple choice test items. PhD., Dissertation, University of Nigeria, Nsuka.

Odili, J.N. (2010). Effect of language manipulation on differential item functioning of test item in biology in a multinational setting. *Journal of Education Assessment inAfrica*, 4, 264-286.

Ong, M. E. (2012). A c*omparison of differential item function Detection method for dichotomously scored items by using 2.1 IRTPRO, BILOG-MG 3 and IRTLRDIF 2.* (Unpublished doctoral dissertation). University of Georgia, Georgia.

Oshima, T. C., & Morris, S.B. (2008). An NCME instructional module on Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27*, 43-50.

Opara, I.M. (2016). *Test construction and measurement: Concepts and applications.* Owerri: Career Publishers.

Orluwene G.W. (2012) *Fundamentals of testing and non-testing tools in educational psychology*. Port Harcourt: Harey publications coy.

Parshall, C.G., & Miller, T.R. (1995). Exact versus asymptotic mantel-Haenszeldif statistics: A comparison of performance under small sample conditions. *Journal of Educational Measurement*, 32(3), 302-316.

Perrone, M. (2006). Differential item function item bias: Critical considerations in test fairness. *Teachers college, ColumbiaUniversity Working Papers in FESOL & Applied Linguistics*, 16(2), 1-3.

Satty, V. Barbuti, S., &Mattern, K. (2006). The new SAT and trends in test performance psychometrics to date (RN-32). New York: The College Board.

Queensoap, M. (2014). Application of Differential item functioning in detecting item bias in chemistry achievement test in Nigeria. Unpublished PhD thesis, University of Port Harcourt.

Raju, N.S. (1989). An Empirical assessment of the mantel-haenszel statistic for studying differential item performance. *Applied Measurement in Education*, 12(1), 1-13.

Reuben, E. &Akorede, S. (2016). Differential item functioning technique for detection of item bias in Economic among secondary school students in Abuja metropolis. *Nigerian Journal of EducationalResearch and Evaluation*.15(1), 112-125.

Roever, C. (2005). *That is not fair, bias and differential item functioning in language testing*. http:///www.hawii.edulroever/brownbag.pdf.

Schumacker, R.E. (2005). Test bias and Differential item functioning. Retrieved on 23/8/2016 from http:www.appliedmeasurementassociates.cpm/white%20Ppapers/TEST%20BIAS%appliedmeasurementassociates.com/White%20Papers/TEST%20AND%20DIFFERENTIAL%

Siebert, C. (2013). Differential item functioning identification strategy for items with dichotomous responses. Using the item information cure: A weighted Area Method (WAM). Unpublished Ph.D. Dissertation. Florida State University. Retrieved on 11/08/2016 from http://diginole.lib.fsu.edu/etd.

Steplen-Bonty, A.M. (2008) Using three different categorical data analysis techniques to detect differential item functioning. Dissertation, Georgia State University. Retrieved on 12/8/2016 from.

http://scholarworks.gsu.edu/eps _diss

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item-response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum

Umoinyang (1991). Item bias in mathematics achievement test. Unpublished M.ed. thesis, University of Calabar; Calabar.

Uwhekadom, E.E. (2014) Influence of language manipulation on differential item functioning in chemistry multiple choice test items. Unpublished Ph.D thesis, University of PortHarcourt.

Wiberg, M. (2007) Measuring and detecting differential item functioning in criterion referenced licensed test: A theoretical comparison of methods. *Educational Measurement*, 60 ,1-33

Wikipedia the free Encyclopedia (2016). Differential item functioning. Retrieved 8/10/2016 from http://en.wikipedia.org/wiki/Dif

Wikipedia the free Encyclopedia (2016). Classical test theory Retrieved 05/10/2016 from http://en.wikipedia.org/wiki/Ctt

Williams, V.L. (1997). The unbiased anchor: Bridging the gap between dif and item bias. *Applied Measurement in Education,* 10(3), 253-267.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532-547.

Yang T.M. & Jones (2008) Center for epidemiological studies-depression scale (CES.D) Item response bias found with mantel-Haenszel method successfully replication using latent variable modeling. Journal of Clinical Epidemiology,60 1195-1200, http://dx.doi.org/10.1016/j.jclinepi.2007.02.008

Zumbo, B.D. (1999) *A handbook on the theory and methods of differential item functioning (dif): Logistic Regression modeling as a Unitary framework for Binary and Likert Type (ordinal) item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly,* 4(2), 223-233.