British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

DIFFERENTIAL ITEM FUNCTIONING OF WAEC SENIOR SECONDARY CERTIFICATE EXAMINATION BIOLOGY MULTIPLE CHOICE ITEMS

Madueke Uchenna Obiageli and Prof. Casmir N Ebouh
Department of Science and computer Education, Faculty of Education, Godfrey Okoye University

Abstract: The main purpose of this study was to investigate the Differential Item Functioning of WAEC Senior Secondary Certificate Examination Biology Multiple choice items. Ex-post facto research design was used for carrying out this study because the researcher used nonmanipulative independent variables of which in this study are gender and school location. The population of the study was 85,439 which comprised all the Senior Secondary School Biology students in the 292 senior secondary schools in the 17 Local Government Areas of Enugu State. The sample size of this study was 646. This consists of 345 female and 301 male Biology students in 4 urban and 4 rural schools in the study area. The instrument adopted by the researcher for data collection for this study was the 2020 Biology Multiple Choice Questions Test (BMQT) Senior Secondary Certificate Examination (SSCE) developed by West African Examination Council (WAEC). This study shows that out of 50 items in WAEC 2020 SSCE May/June multiple choice Biology questions; with respect to gender, DIF was present in 8 items. These items revealed significant DIF between male and female students with significant level less than 0.05. These items revealed significant DIF between urban and rural students. Hence, the major educational implication of the findings of the study is that when obvious bias of test items is present in national and regional examinations as seen in the WAEC 2020 Biology Multiple Choice Items of the Senior School Certificate Examination, because of DIF, such items threaten and undermine the total test validity which could jeopardize the classification of subgroup based on their true score which negatively affect test fairness in general. It was concluded that hat the items function differentially among male and female examinees. The study recommended that Item developer should ensure that the differential item functioning effects of items are properly determined in the test instrument to avoid bias against a subgroup of the test takers

Keywords: Differential Item Functioning, WAEC Senior Secondary School Certificate Examination Biology Multiple Items, Item Response Theory, Item Fairness, Item Bias

1.1 Background to the Study

Academic and career decisions are often made based on test scores from assessment, and people's lives are affected based on these decisions. Assessments that are used to measure students' understanding of concepts in a particular discipline therefore need to demonstrate fairness and produce valid and reliable scores. According to National Council on Measurement in Education (NCME) (2014), robustness of assessments in demonstrating fairness is essential for high-stakes tests used in certification or University admissions, and it is also critical in drawing inferences about student performance on low-stakes assessments, such as those teachers made test within science classrooms to avoid test bias. Test items are considered biased when they favour the performance of one subgroup over another – irrespective of the assessment's subject. Item bias has an important impact on the fairness of psychological testing (Khalid & Glas, 2014). A complication on this quest to valid inferences is

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

termed Differential Item Functioning (DIF) (Strobl, Kopf & Zeileis, 2011). This is critical because, when studying students' performance across different subgroups or cultures, an essential aspect for appraisals is that of score comparability. In other words, if the inferences regarding performance can be regarded as valid, it is imperative that the latent variable (i.e., construct of interest) is understood and measured equivalently across all participating groups.
The psychometric property that typically must hold for scores to be equivalent when compared is acknowledged as measurement invariance, lack of bias or absence of differential item functioning. Differential Item Functioning occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure. In simple terms, DIF arises when two groups of equal ability levels are not equally able to correctly answer an item. In other words, one group does not have an equal chance of getting an item right though its members have comparable ability levels to the other group. An item does not display DIF if people from different groups have a different probability to give a certain response; it displays DIF if and only if people from different groups with the same underlying true ability have a different probability of giving a certain response. Item functioning is intended to be invariant with respect to irrelevant aspects of the test-takers, such as gender, location, ethnicity and socio-economic status (Queensoap & Orluwene, 2019). If the factor leading to DIF is not part of the construct being tested, then the test is biased.
Technically, DIF occurs when an item measures more than one underlying latent trait and when cognitive differences exist on one of these other so-called secondary latent traits. A latent trait (also known as latent knowledge, latent ability, or, more generally, latent variable) is an individual's true knowledge or understanding of the construct being measured, and it can be estimated but not directly measured.

The presence of DIF for a given item would indicate that the item may measure a secondary latent trait, either alone (completely missing the target concept) or in concert with the primary trait (which requires knowledge of the target concept and the secondary concept). For example, suppose a student needed both an understanding of the concept of homeostasis and knowledge of difficult English vocabulary (e.g., to understand the phrase "hypertension is characteristic of diabetic nephropathy"), that is not essential for understanding the concept of homeostasis. Moreover, if the focal group of students does not have the requisite knowledge of English, then the focal group would be more likely to answer incorrectly compared with those in the reference group even if the focal group has the same level of homeostasis knowledge as the reference group. Moreover, a test containing items exhibiting DIF could in turn create inaccurate observed total scores, resulting in inaccurate estimation of the focal group's primary latent trait (e.g., biological concepts).
There are two types of DIF, namely uniform and non-uniform DIF. Uniform DIF occurs when a group performs better than another group on all ability levels. That is, almost all members of a group outperform almost all members of the other group who are at the same ability levels. In the case of non-uniform DIF, members of one group are favoured up to a level on the ability scale and from that point on the relationship is reversed (Karam, 2012). That is, there is an interaction between grouping and ability level. Although the presence of DIF is a signal that an item may be biased, it does not guarantee that the item is unfair. Rather, the presence of DIF indicates the existence of a latent trait besides the one of primary interest. Fairness is established subsequently if the secondary latent trait that was detected statistically is intentionally related to the primary latent trait. It is possible that the secondary latent trait is required by the content and the test specifications, even if the reference and focal groups perform differently. As noted earlier, DIF examines

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

the probability of correctly responding to or endorsing an item conditioned on the latent trait or ability. Hence, various statistical models may be used to detect differential item functioning, such as Logistic Regression Model, the Mantel-Haenszel (MH) approach and Item Response Theory. These procedures all assume that the test takers have approximately the same abilities. Logistic regression approaches to DIF detection developed by Bock (1975) involve running a separate analysis for each item. The independent variables included in the analysis are group membership, an ability matching variable typically a total score, and an interaction term between the two. The dependent variable of interest is the probability or likelihood of getting a correct response or endorsing an item. Hence, the outcome of interest is expressed in terms of probabilities.

A common procedure for detecting DIF is the Mantel-Haenszel approach developed by Mantel and Haenszel in 1959. The MH procedure is a chi-squared contingency table based approach which examines differences between the references and focal groups on all items of the test, one by one. Item response theory (IRT) approach is another widely used method for assessing DIF. IRT allows for a critical examination of responses to particular items from a test or measure. Because IRT examines the monotonic relationship between responses and the latent trait or ability, it is a fitting approach for examining DIF (Steinberg & Thissen, 2006). Item response theory is currently one of the most widely-used methods for measuring differential item functioning in test adaptations. However, it requires a relatively large sample size.

The occurrence of differential item functioning (DIF) in external examinations used for diagnosis, selection or admission could possibly lead to severe consequences, such as inaccurate and imprecise item and attribute mastery pattern estimates (Hou, de la Torre, and Nandakumar, 2014). Given that great importance has been attached to detecting and eliminating DIF items from cognitive diagnostic tests, DIF should thus be routinely investigated in practice and application to ensure fairness and validity of the test. Although examination manuals like those of West African Examination Council's (WAEC) Senior School Certification Examination (SSCE) suggests that the factor structure of Biology items remained invariant across different national demographic samples and subgroups, there are paucity of studies reporting such to demonstrate item invariance. Thus this work has undertaken to study the response pattern of examinees, and use the methods of detecting DIF to establish whether gender of examinee and school location functioned differentially in Biology multiple choice examination administered by WAEC in Senior Secondary Certificate Examination.

Statement of Problem

Teachers and authorized examination bodies like WAEC and NECO continually assess the knowledge, skills and abilities of Biology students at senior secondary school. Nonetheless, these assessments are expected to treat all test-taker equally, but it has been observed by the researcher that more often than not, the instruments used hardly accomplish this purpose. This is because, Differential Item Functioning of multiple choice items through classical item theory is group dependent and the item statistics such as item difficulty and item discrimination are also group dependent. Based on these limitations of the instrument under classical item theory, the researcher designed this study using a relevant measurement theory to ensure objectivity in Differential Item Functioning in WAEC in analyzing Biology multiple choice items. Therefore, the question addressed is: would item response theory influence the instrument differential item functioning multiple choice items in Biology?

Purpose of the Study

The main purpose of this study was to investigate the Differential Item Functioning of WAEC senior secondary certificate examination Biology multiple choice items. Specifically, the study sought to detect:

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

1. WAEC 2020 Biology multiple choice items of the senior school certificate examination that function differentially in terms of gender.
2. WAEC 2020 Biology multiple choice items of the senior school certificate examination that function differentially in terms of school location.

Research Questions

In line with the purpose, the following research questions guided this study

1. What items in WAEC 2020 SSCE multiple choice Biology questions differentially function in terms of gender?
2. What items in WAEC 2020 SSCE multiple choice Biology questions differentially function with respect to school location?

Hypotheses

The following hypotheses were formulated and were tested at the 0.05 level of significance. **Ho_1:** Items in WAEC 2020 SSCE multiple choice Biology questions do not significantly function differentially among students based on gender.

Ho_2: Items in WAEC 2020 SSCE multiple choice Biology questions do not significantly function differentially among students with respect to school location.

LITERATURE REVIEW

1.2 Item Fairness

The fairness of an exam refers to its freedom from any kind of bias. The exam should be appropriate for all qualified examinees irrespective of race, religion, gender, or age. Fairness in assessment of student's achievement test in Biology in senior secondary school is very fundamental as Biology is the basis for studying other subjects especially in science related courses. Biology is a compulsory subject for every individual to function effectively and efficiently in today's world irrespective of one's profession and hence scores obtained by students in this subject should reflect their true ability (Githua & Mwangi, 2003) Fairness is an essential quality of a test, it's equitable treatment of all examinees during the testing process. The consequences of unfair test items can be quite serious. This is because DIF can lead to an unfair advantage or disadvantage for certain subgroups in educational and psychological testing (Strobl, Kopf & Zeileis, 2011). Although the presence of DIF is a signal that an item may be biased, it does not guarantee that the item is unfair. Rather, the presence of DIF indicates the existence of a latent trait besides the one of primary interest. Fairness is established subsequently if the secondary latent trait that was detected statistically is intentionally related to the primary latent trait. WAEC 2020 SSCE can only be considered fair when the test items favours the test takers equally across the sub-groups based on gender and location.

1.3 Item Bias

Several scholars have tried to define the term item bias, and according to Poortinga in 1989, it has been defined as differences in participant's score that do not refer to similar differences in the construct. Therefore, if there is a bias, difference in scores of individuals do not have the same meaning within and across culture. Differential Item Function (DIF) is the is the most frequently employed statistical analysis of item bias (Van de Vijiver &Tanzer 2004). The analysis of bias is mandatory before conclusions can be drawn that the groups have different scores on a target construct (van de Vijver & Matsunmoto2011). The study of bias in items and tests began at the end of the 1960s and developed exponentially over time.

Hence, when tests are labeled "biased", the accusations often have to do with the instruments chosen for a particular context, the way in which these tests are administered or the way in which the results are interpreted and/or used. According to Beer (2004), these broader issues are often far removed from the actual instrument itself and its inherent properties. Therefore, bias is not the mere presence of a score difference between two groups".

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

The term "bias" largely indicates a systematic error that stems differences in performance levels of comparison groups of the same ability level.

1.4 Validity

Validity is an important concept in testing. Messick (1989) described validity as testing acknowledged touchstone. In other words, the quality of test as a psychological measurement is judged on the basis of its validity. Validity is described in 4 components-content, criterion, concurrent and construct validity. Content validity refers to the extent to which a subject's responses to the items of a test (total score) may be considered as a representative sample of his responses to a real or hypothetical universe of situation which together constitute the area of concern to the person interpreting the test.

Criterion-related validity refers to the empirical technique of studying the relationship between the test scores and some independent external measures known as criteria. Construct validation process embraces embraces all the other types of validity. Gardner (1983) sees construct validity as embracing all the other types of validity. According to Messick (1980) construct validity is a process of validation that involves gathering evidences from several sources of such evidence made from test scores. Sources of such evidence could be content or criterion measures of a test. For a score in Biology WAEC 2020 multiple choice question examination to be valid therefore does not depend only on the degree to which test items match course content and objective. On the other hand, Biology test is valid if scores from the test are sustained by only ability in Biology, not a function of other abilities that are not related to Biology knowledge

Furthermore, when studying student performance across different subgroups or cultures, an essential aspect for comparisons is that of score comparability. In other words, if our inferences regarding performance can be regarded as valid, it is imperative that the latent variable (i.e., construct of interest) is understood and measured equivalently across all participating groups (Svetina, 2014). Thus, during the process of validation and in the quest for a valid test, developers often test for differences in performance among two or more groups of students as one way of gathering evidence of the presence or absence of test bias, such as whether men and women perform differently, whether native speakers of the testing language perform consistently better than others, or whether race/ethnicity is linked with performance. In addition, attending to the performance of different groups is critical for equity—assessments should not discriminate against any individual (Nehm, 2017). Test developers must show that performance on the assessment is not related to factors that are irrelevant to the construct being tested.

Items flagged as DIF have a strong potential to threaten the construct validity of scores if they are not further investigated, and therefore DIF analysis should be performed routinely when developing conceptual assessments. Construct validity implies that the construct to be measured is the same for all respondents of the population the test is aimed at. This is where the problem of differential item functioning (DIF) or item bias arises. According to Khalid and Glas (2014), construct validity is supported if the construct to be measured is also unidimensional and if the ordering of item difficulties imposed by the construct is reflected in the ordering of item parameters on the latent scale. Further, if it can be shown that the latent ability is unidimensional, a meaningful unidimensional variable for measuring the underlying construct can be created, and the respondent can be assigned a value on some scale. Thus, assessments scores that are used to measure students' understanding of disciplinary concepts such as Biology need to produce valid and reliable scores.

2.0 Differential Item Functioning (DIF)

Differential Item Functioning (DIF) occurs when groups (such as defined by gender, ethnicity, age, or education) have different probabilities of endorsing a given item on a

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

multi-item scale after controlling for overall scale scores. An item is labeled as having DIF when people with the same latent ability but from different groups have an unequal probability of giving a response. An item is labeled as non-DIF when people with the same latent ability have equal probability of getting an item correct, regardless of group membership. According to Zhang (2006) DIF occurs for an item when one group (the focal group) of examinees is more or less likely to give the correct response to that item when compared to another group (the reference group) after controlling for the primary ability measured in a test.

Commonly, DIF studies examine cognitive tests for the presence of item DIF or potential test bias with respect to a number of different demographic characteristics, such as gender, education, social class, ethnicity, age and so on. In particular, as the main purpose of selecting items for WAEC's SSCE this study is concentrated on DIF analysis of May/June Biology multiple choice items with respect to two main variables: gender and location. Research on sources of DIF in science by gender has been reported in many studies. Some of them focus on item format effect. Multiple choice items seem to favour male examinees and open-ended items tend to favour female examinees (Lee, 2006). Some focus on the effect of item content where they found that males seem to be advantaged over females on science items. And on the effect of item cognitive domains, some evidence was found that male examinees performed differentially better than female examinees (when matched on total test score) on items requiring spatial reasoning or visual content (Orluwene and Queensoap, 2019).

Nonetheless, when considering DIF, we must take into account three elements: (a) respondents should have the same level of the latent factor, (b) a group variable by which respondents can be divided into independent groups and (c) the latent factor which determines DIF should not be part of the measured construct. If the three conditions are cumulatively met we can consider the presence of DIF (Karami, 2012). Differential item functioning is a statistical phenomenon that can occur in any item of a test. As DIF accumulates in several items, it can produce differential functioning in clusters of items called bundles. The causes of DIF include item content, item type or format, item context, content and cognitive dimensions associated with items. The methods for detecting DIF vary depending on how students are matched. Classical methods (e.g., Mantel-Haenszel statistic and logistic regression) match students based on their total scores; methods based on item response theory (IRT) models, such as the Wald χ^2 test (also known as Lord's test) and Raju's area test, consider student ability as a latent variable estimated together with item parameters in the model (Nehm, 2017).

2.1 Biology

Biology, etymologically is from two Greek words, Bio meaning Life and Logos meaning study, therefore, Biology is the study of life. In Secondary Schools, is a subject that covers the variety of life processes and how the different organisms meet the challenges of living in their environment. It covers all life's processes such movement, respiration, nutrition, irritability, reproduction, excretion, respiration and death. According to the National Policy on Education cited in Akinjide (2018), the study of Biology would afford students the knowledge of suitable laboratory and field skills in Biology, relevant knowledge that is applicable in health, agriculture and daily life.

2.2 Gender

Gender as defined earlier is a product of the biological difference between males and females, according to their physiological and reproductive characteristics. It is the cultural, social and political attributes that accrue to being male or female. In the aspects of education, gender plays a very important role. It starts from development up until higher education (Goldberg, 2016). As mentioned earlier, there is a bit of differential treatment, for example, boys tend to be reprimanded more when they do wrong and praised more when they do right. On the part of girls, they

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

are only praised in aspects of neatness, cleanliness and artistic qualities. Also certain textbooks and materials seem to classify girls as being more helpless and dependent than boys, though this has changed over time, it contributed significantly to the disparity in the learning systems between boys and girls (Lynch, 2016; Goldberg, 2016).

2.3 Item Response Theory (IRT) (Hambleton and Swaminathan, 1985; and Lord, 1980) Item Response Theory (IRT) is a measurement theory and its focus is on the item level rather than on the total score. In IRT framework, parameters are classified into two basic component: the first is related to the examinee's ability (latent trait), second to the task (test). The assumption is that each examinee responding to a test item possesses some amount of underlying ability. This study is congruent with the Latent Test Theory of Item response Theory. There is no single model called IRT but a plethora of educational and psychological measurement concerns underlying (latent) variables of interest and involves determining how much of such a latent trait a person possesses. This theory proposes that a correct response depends on both the characteristics of the item and the person's ability. IRT models have several advantages over classical models like Mantel-Haenszel χ^2 Test, as they provide interpretation of individual performance that go beyond simple sum scores on a test. To put it simply, two individuals with the same total score but with a very different pattern of errors across individual items (or trials) may have different true levels of ability, and using IRT, researchers can build a model of the information that each item provides about the target ability and provide better ability estimates for each person IRT provides estimates of precision at both the individual and test-level and can reduce reliance on normative samples (Sunday, Lee & Gauthier, 2018). IRT methods provide ways to match individuals so that we can evaluate whether certain items on a test are biased against them in some way.

Research Method

Design of the Study

This study adopted the ex-post facto research design. This design is considered appropriate because, the researcher dealt with non-manipulative independent variables of which in this study were gender and school location

Area of the Study

The study was carried out in Enugu State. Enugu state is located within the South East geopolitical zone of Nigeria, It's capital is Enugu

Population of the Study

The population of the study consisted of all the 85,439 Senior Secondary School Biology students in the 292 senior secondary schools in the 17 Local Government Areas of Enugu State. This comprises 38,534 males and 46,905 females in the public Senior Secondary Schools in the state.

Sample and Sampling Technique

The sample size of this study was 646. This consists of 345 females and 301 males Biology students in 4 urban and 4 rural schools in the study area. Simple random sampling without replacement via balloting was first used to select two schools each from urban and rural LGAs.

This included Enugu East and South for urban and Aninri and Nkanu East for Rural. At the selected 4 schools, purposive sampling was used to select 150 females in the urban and 195 in the rural area. 150 males in urban and 151 in rural area were also random purposively sampled

Instrument for Data Collection

The instrument adopted by the researcher for data collection for this study was the 2020 Biology Multiple Choice Questions (BMQ) Senior Secondary Certificate Examination (SSCE) developed by West African Examination Council (WAEC). The instrument consists of 50 multiple choice questions, each with 4 options (A-D).

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

Validation of the Instrument

The items of the instrument were constructed and validated by experts in the department of examinations and quality control of the West Africa Examination Council (WAEC) and therefore requires no further validation.

Reliability of the Instrument

To establish the reliability of the instrument, the BMQ was administered to thirty (30) SSS3 Biology students in Nsukka Education Zone. The school used was outside the sample of the study, but has some degree of similarities with sampled schools. Their responses were scored and analyzed using Kuder-Richardson (KR-20) formula to determine the internal consistency (reliability) of the instrument. Reliability index of 0.70 was obtained to attest to the reliability of the instrument.

Method of Data Collection

The data for this study was collected through the use of Biology Multiple Choice Question Test (BMQ). The researcher visited the sampled schools to collect the data for the study. The copies of the instrument were administered to the SSS 3 Biology students through the assistance of the Biology teachers in the respective sampled schools. The test was administered to the students under a good atmosphere after been informed ahead of time about the exercise and the test lasted for 50minutes. The administration of the instrument was done once as a part of their continuous assessment and retrieved immediately for recording and analysis.

Method of Data Analysis

Logistic Regression Model approach was used in detecting and analyzing DIF of the
WAEC 2020 Biology Multiple Choice Items of the Senior School Certificate Examination.
Research questions were answered using percentages.

RESULTS

Research Question One: What items in WAEC SSCE 2020 multiple choice Biology questions differentially function in terms of gender?

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

Table 1: DIF Analysis to Detect Gender Bias on 50 Multiple Choice WAEC Biology Questions

Item	Biology	Standard Exponential	Wald(x^2)	Significance	Exponetial (B)	95.0% Confidence Intervention for Exp (B)	
1	1.11	.21	28.66	.324	.33	.22	.50
2	.50	.19	6.52	.071	.61	.42	.89
3	.02	.19	.01	.912	1.02	.70	1.49
4	-1.07	.21	27.23	.000*	.34	.23	.51
5	.14	.19	.57	.451	1.16	.79	1.68
6	.051	.19	4.40	.206	.67	.46	.97
7	.52	.19	7.11	.082	.60	.41	.87
8	.02	.19	.01	.94 3	1.02	.70	1.48
9	.22	.19	4.52	.103	.66	.45	.97
10	-.35	.19	3.38	.066	.70	.48	1.02
11	-.15	.19	.59	.441	.86	.59	1.26
12	-.57	.19	8.78	.283	.56	.39	.82
13	1.30	.21	8.02	.072	.27	.27	.41
14	1.97	.24	6.82	.532	.14	.09	.22
15	.21	.20	1.15	.283	1.24	.84	1.82
16	-.40	.19	4.21	.082	.67	46	.98
17	1.23	.22	30.20	.000*	3.43	2.21	5.32
18	-.16	.19	.70	.401	.85	.58	1.24
19	1.44	.22	41.08	.000*	.24	.15	.37
20	-.54	.20	7.19	.007*	.58	.39	.87
21	-.55	.20	7.95	.085	.58	.39	.85
22	-.37	.19	3.72	.054	.69	.47	1.01
23	.10	.21	23.36	.263	.37	.25	.55
24	-.90	.20	19.38	.000*	.41	.27	.61
25	-.06	.19	.10	.748	.94	.65	1.37
26	-.54	.20	7.02	.008*	.58	.39	.87
27	-1.23	.21	28.02	.000*	.32	.21	.49
28	-.15	.20	.57	.450	.86	.59	1.27
29	.20	.19	1.09	.296	1.22	.84	1.78

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

Item	B		S.E	Wald		Sig.	Exp (B)	95.0% C.I. for Exp (B)
30	.17	.20	15	.088	.46	.31	.67	
31	1.36	.21	41.11	.262	.26	.17	.39	
32	.00	.20	.00	.998	1.00	.68	1.47	
33	-.18	.19	.84	.359	.84	.58	1.22	
34	-.85	.21	16.93	.080	.43	.29	.64	
35	.30	.19	2.35	.125	. 1.35	.92	1.97	
36	-1.03		.20	26	.263	.36	.24	.53
37	-.41	.20	4.30	.081	.67	.45	.98	
38	-.43	.20	4.37	.067	.65	.44	.97	
39	.69	.21	11.30	.325	.50	.33	.75	
40	1.59	.25	41.91	.347	.20	.13	.33	
41	-.35	.20	3.28	.070	.70	.48	1.03	
42	.08	.19	.18	.672	1.09	.74	1.58	
43	.47	.20	. 5.29	.092	1.59	1.07	2.37	
44	.55	..19	8.02	.263	1.73	1.18	2.52	
45	-.36	.20	3.31	.069	.70	.48	1.03	
46	.28	.19	2.09	.148	1.32	.91	1.93	
47	.65	.20	11.04	.001*	1.92	1.31	2.81	
48	.11	.19	.34	.562	1.12	.77	1.63	
49	-1.62		.25	41.29	.122	.20	.12	.33
50	-1.11		.22	26.10	.334	.33	.21	.50

Data on Table 1 exposed that out of 50 items in WAEC 2020 SSCE May/June multiple choice Biology questions; with respect to gender, DIF was present in 8 items. These items are item 4, 17, 19, 20, 24, 26, 27 and 47. These items revealed significant DIF between male and female students with significant level less than 0.05. Among the 8 items that displayed significant gender DIF, 2 items (item 17 and 47) representing 4% were identified to exhibit significant gender DIF in favour of male students while 6 items (4, 19, 20, 24, 26, and item 27) representing 12% differentially functioned in favour of female students.

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

Research Question Two: What items in WAEC SSCE 2020 multiple choice Biology questions differentially function in terms of location?

Table 2: DIF Analysis to detect location Bias on 50 Multiple Choice WAEC Biology Questions

Item	B	S.E	Wald	Sig.	Exp (B)	95.0% C.I. for Exp (B)	
1	-.10	.19	.24	.621	.91	.62	1.33
2	.46	.20	5.34	.021*	1.58	1.07	2. 33
3	.58	.19	9.04	.005*	1.79	1.23	2.62
4	.16	.19	.719	.397	1.18	.81	1.71
5	.09	.19	.20	.658	1.09	.75	1.58
6	.34	.20	2.85	.022*	1.41	1.95	2.09
7	.21	.19	1.25	.263	1.24	.85	1.80
8	.19	.19	.95	.329	1.21	.83	1.76
9	.15	.20	.56	006*	1.16	1.68	1.79
10	.67	.22	4.56	.133	1.60	1.04	2.46
11	.51	.19	6.84	.030*	1.66	1.14	2.42
12	.53	.20	6.74	.011*	1.70	1.14	2.53
13	.29	.20	2.27	.032*	1.34	122	1.96
14	-.07	.19	.12	.735	.94	.64	1.37
15	-.60	.22	7.73	.005*	.55	1.36	1.84
16	.16	.20	.62	.432	1.67	.79	1.72
17	.29	.20	2.27	.432	1.34	.92	1.96
18	.02	.19	.008	.928	1.02	.70	1.49
19	.53	.20	7.10	.008*	1.70	1.15	2.52
20	-.24	.19	1.56	.001*	.79	1.14	1.25
21	.37	.19	3.72	.044*	1.45	1.99	2.12
22	.45	.20	4.91	.000*	1.57	1.05	2.33
23	.59	.20	8.57	.003*	1.80	1.22	2.68
24	.22	.20	1.31	.252	1.25	.85	1.83
25	.18	.19	.89	.345	1.20	.82	1.75
26	.16	.20	.62	.327	1.67	.79	1.72
27	.33	.20	2.57	.000*	1.38	1.53	2.06
28	.25	.19	1.67	.016*	1.31	1.28	1.87
29	.25	.20	1.57	.012*	1.29	1.87	1.91
30	-.03	.19	.02	.625	.98	.67	1.42
31	-.02	.20	.01	.925	.982	.67	1.44
32	.43	.19	4.88	.000*	1.53	1.05	2.23
33	-.22	.20	1.23	.643	.81	.55	1.18

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

34	-.09	.19	.22	.000*	.91	1.33	1.64
35	.10	.20	.24	.051	1.10	.75	1.63
36	-.38	.19	3.81	.109	.69	.47	1.00
37	.18	.21	.74	.039*	1.19	1.20	1.79
38	.41	.20	4.37	.037*	1.51	1.03	2.22
39	.83	.23	13.24	.003*	2.30	1.47	2.59

Item	**B**	**S.E**	**Wald**	**Sig.**	**Exp (B)**	95.0% C.I. for Exp (B)	
40	.24	.20	1.36	.011*	1.27	1.85	2.89
41	-.07		.19	.12	.731	.94	.64 1.36
42	.14	.20	.54	.022*	1.16	1.19	1.69
43	.20	.21	.94	.333	1.22	.81	1.85
44	.43	.23	3.62	.057*	1.54	1.66	2.39
45	.26	.19	1.87	.172	1.30	.89	1.89
46	.30	.19	2.38	.123	.123	1.34	1.96
47	.04	.20	.04	.038*	1.04	1.71	2.53
48	.30	.21	2.07	.032*	1.34	1.90	2.01
49	.13	.19	.47	.494	. 1.14	.78	1.67
50	.45	.21	4.36	.037*	1.56	1.03	2.37

Data on Table 2 exposed that out of 50 items in WAEC 2020 SSCE May/June multiple choice
Biology questions; with respect to location, DIF was present in 27 items. These items are item 2,
3, 6, 9, 11, 12, 13, 15, 19, 20, 21, 22, 23, 27, 28, 29, 32, 34, 37, 38, 39, 40, 42, 47, 48 and item 50. These items revealed significant DIF between urban and rural students with significant level less than 0.05. All the 27 items representing 52% differentially functioned in favour of urban students. **Hypotheses Testing**

Ho_1: Items in WAEC 2020 SSCE multiple choice Biology questions do not significantly function differentially among students based on gender

The differential item functioning was calculated on basis of gender. Item probability parameter procedure was used to obtain the cut off values for the different levels of significance for each item. At the Alpha level of 0.05, data on table one shows that 8 items function differentially among students based on gender. The researcher therefore rejects the null hypothesis and concludes that some items (8 items) in WAEC 2020 SSCE multiple choice Biology questions significantly function differentially among students based on gender

Ho_2: Items in WAEC 2020 SSCE multiple choice Biology questions do not significantly function differentially among students with respect to school location.

British International Journal of Education and Social Sciences
Vol.09, No.6; June-2022;
ISSN (3342 – 543X);
p –ISSN 4519 – 6511
Impact factor: 9.82

The differential Item functioning was calculated on basis of school location. Item probability parameter procedure was used to obtain the cut off values for the different levels of significance for each item. At the Alpha level of 0.05, data on table two shows that 26 items function differentially among students based on school location. The researcher therefore rejects the null hypothesis and concludes that some items (26 items) in WAEC 2020 SSCE multiple choice Biology questions significantly function differentially among students based on school location.

Discussion of Findings

WAEC 2020 Biology Multiple Choice Items of the Senior School Certificate Examination that Function Differentially in Terms of Gender

Findings of this study show that 8 items representing 16 percent of the WAEC SSCE 2020 multiple choice Biology questions differentially function in terms of gender. These items with sig value equal or less than the Alpha 0.05 value indicates significant differential functioning when compared with the cut off value set by item parameter Logistic Regression Analysis. Six (6) items were detected showing item bias against the focal group (Male) while 2 are biased against the reference group. At the 0.05 alpha level of significance, these biased items indicate that the odds of getting an item right were different for the focal/reference group and thus depend on the sub group the examinee belong. In other words, two WAEC candidates with the same primary latent trait value (ability level) but differing in other characteristics (secondary latent trait (gender)) have differing probabilities of response to the WAEC SSCE 2020 multiple choice Biology questions.

The findings of this study is analogous to those made by Ajeigbe and Afolabi (2014) study that assessed unidimensionality and occurrence of Differential Item Functioning in Mathematics and

English Language items of Osun State Qualifying (OSQ) Examination. The study of Ajeigbe and Afolabi concluded that the Examination contained considerable number of items that exhibited DIF and therefore requires adequate item quality improvement to justify its use as the inclusion or exclusion criterion of state candidate in West African Examination Council. The study of Osadebe and Agbure (2018) also revealed similar findings that there is incidence of gender, location, socioeconomic, school type and school ownership differential functioning in 2014 BECE Social Studies multiple choice test. Hence, Nehm (2017) concludede that Differential Item Functioning analysis should be a routine part of developing conceptual assessments.

This findings is however in contrast to Muray, Booth and McKenzie (2015) test which did not show evidence of differential item functioning by gender when they used a multi-group item response theory approach to assess differential item functioning across gender in a sample of 211 males and 132 females assessed in clinical and forensic settings.

Sometimes items are found to behave differently in distinct groups such as gender or (such as loading on different dimensions in a multi-dimensional factor analysis, or having largely different mean item scores but the quality of an item or test should be of utmost importance to stakeholders concerned with evaluation enterprise.

WAEC 2020 Biology Multiple Choice Items of the Senior School Certificate Examination that Function differentially in Terms of School Location

With respect to location as a source of bias, the findings of this study uncovered that out of

50 items in WAEC 2020 SSCE May/June multiple choice Biology questions, DIF was present in 27 items. These items revealed significant DIF between urban and rural students with significant level less than 0.05. All the 27 items representing 54% differentially functioned in favour of urban students. This findings is similar to those made by Ikeh, e'tal (2020) when they analysed differential item functioning in Economics multiple choice items

administered by West African Examination Council using logistic regression model. Among the 31 items, the found out that 27 items representing 54% were identified to exhibit significant location DIF in favour of urban school students while only 4 items representing 8% differentially functioned in favour of rural school students.

Since DIF occurs when a test item favours or hinders a characteristic exhibited by group members or a test taking population, this findings strongly suggest that the two groups (Urban and Rural) WAEC examinee had not had equal opportunity to learning experience related to the content of the biased items or the focal group must have been predisposed by secondary latent traits associated with underdevelopment at the rural area (lack of science teachers, poor supervision, etc) rather than factors like language or item structure. Nevertheless, logistic regression model of DIF analysis emphasized a borderline difference between item bias and differential item functioning. This posit that an item may be flagged DIF but may not be biased item while an item that is flagged biased item always show differential item functioning. But all the 27 items differentially functioned in favour of reference group (urban students). Hence, there is a scenario of Uniform Differential Item Functioning. Uniform DIF occurs when a group performs better than another group on all ability levels. That is the situation of the findings of this study because almost all members of the reference group outperform almost all members of the focal group who are at the same ability levels. Exploration of this above findings also showed that the school location DIF items are due to the fact that they contain sources of difficulty that are irrelevant or extraneous to the construct being measured (ability in Biology).Hence, these findings suggest that educational risk factors in rural areas act in concert in WAEC examinee population and that at least to some magnitude they interact to create bias that may produces unfair test scores.

Conclusions

This study investigated the Differential Item Functioning of WAEC Senior Secondary Certificate Examination Biology Multiple choice items. It was concluded that the items function differentially among male and female examinees. Also, differential Item Functioning was more apparent between urban and rural examinees in the WAEC 2020 Biology Multiple Choice Items of the Senior School Certificate Examination. Hence an important objective of drawing valid inferences about the construct that one intends to measure in Biology test is seriously susceptible to exogenous and secondary traits that can be attributed to school location.

Educational Implications

The findings of this study have obvious educational implication for teachers and test developers like WAEC. This study demonstrates that it is possible to detect items that are functioning differentially with respect to gender and school location. Hence, to ensure fairness, it becomes necessary that teachers, especially Biology teachers should set and administer items that are free from DIF. In addition, when obvious bias of test items is present in national and regional examination as seen in the WAEC 2020 Biology Multiple Choice Items of the Senior School Certificate Examination because of DIF, such items threaten and undermine the total test validity. Furthermore, when secondary traits that are irrelevant to the construct of interest (primary traits) are introduced by DIF, this could jeopardize the classification of subgroup based on their true score negatively affect test fairness in general. It therefore becomes imperative for stakeholders to properly scrutinize examinations and tests designed for heterogeneous groups with caution**.**

Recommendations

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

1. Test developers and examination bodies should take into account multiple background

risk variables of gender and school location simultaneously when collating items for administration

2. Item developer should ensure that the differential item functioning effects of items are properly determined in the test instrument to avoid bias against a subgroup of the test takers.

3. Test developers, ministry of education and examination bodies should ensure that items are free from differential item functioning (DIF).

4. Item developers should ensure that items are written in a straight forward, uncomplicated and easily read manner. Excessive wordiness can obviously prevent the examinees from responding appropriately to test items and therefore create bias in the examination.

References

Ajeigbe, T. O. & Afolabi, E. R. I. (2014). Assessing Unidimensionality and Differential Item Functioning in Qualifying Examination for Senior Secondary School Students, Osun State, Nigeria. World Journal of Education Vol 4, No 4, 30-37 .DOI: https://doi.org/10.5430/wje.v4n4p30

Ani, E. N. (2014).Application of Item Response theory in the Development and Validation of Multiple Choice Test in Economics. *An unpublished master's Thesis submitted to the department of Science Education, University of Nigeria, Nsukka*

Beer, M. (2004). Use of differential item functioning (MF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology, 2004*, 30 (4), 52-58

Bock, R. D. (1975).*Multivariate statistical methods*. New York: McGraw-Hill.

Castro, S., Cúri, M., Torman,V. & Riboldi. J. (2015). Differential item functioning in the beck depression inventory. *Rev. bras. epidemiol.* 54-67.

Crane, P. K. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIF detect and difwithpar. *Medical Care:* Vol. 44 115-123 10.1097/01.mlr.0000245183.28384.ed

DeMars, C. (2010). *item response theory*. New York: Oxford Press.

Drabinová A., Martinková, P., Detection of differential item functioning with non-linear regression: Non-IRT approach accounting for guessing. 2016 Retrieved, January, 31, 2022 from http://hdl.handle.net/11104/0259498.[Google Scholar]

Federer, M. R., Nehm, R. H., Pearl D. K., Examining gender differences in written assessment tasks in Biology: a case study of evolutionary explanations. *CBE—Life Sciences Education*.2016;15:ar2.

Fortin M. A., van F. J. R., & Poortinga, Y. H. (2013). differential item functioning and educational risk factors in guatemalan reading assessment. *Revista Interamericana de Psicología/Interamerican Journal of Psychology*, *47*(3), 422-432.

Frederickx, S. & Tuerlinckx, F. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement.* Vol. 47, No. 4, pp. 432–457

French, B. F., Finch, W. H. & Immekus, J. C. (2019). multilevel generalized mantel-haenszel for differential item functioning detection. Front. Educ., 18 June 2019 | https://doi.org/10.3389/feduc.2019.00047

Githua, B.N., and Mwangi, J.G., (2003). Students mathematics self. Concept and moltivation to learn mathematics relationship and gender differences among Kenya's secondary School Students in Nairobi and Rift valley province. *Journal of educational development*, 2(23), 487-499.

Goliath-Yarde L. & Roodt, G. (2011). Differential item functioning of the UWES-17in South Africa. *South African Journal of industrial Psychol*. vol.37 (1) 1-11

Gómez-Benito, J. Sireci, S., Padilla, J., Hidalgo, M. D., & Benítez, I. (2018).Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema, 30, 104109*doi: 10.7334/psicothema2017.183

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructedresponse science test. *Applied Measurement in Education*, 12 (3), 211-235

Hou, L., de la Torre, J. D., and Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *J. Educational. Measurement.* 51, 98–125. doi: 10.1111/jedm.12036

Ikeh, F. E., Ene, C. C., Ojobo, B., Ani, M. I., Metu. C. I., Ugwu, F. C., Owolawi, O., Omosowon, V. O., Oguguo, B. C., Ezugwu, J. I. & Agugoesi, J. O (2021). Assessment of differential item functioning to detect gender biased items in economics multiple choice questions in senior school certificate examination. *Journal of Critical Reviews. Vol, 8,* ISSUE 01. Pp, 516 – 523.

Ikeh, F. E., Ugwu, F. C., Mfon, T. E., Omosowon, V. O., Iketaku, R. I., Opa, F. A., Eze, B. A., Kalu, I. A., Ikwueze, C. C. & Ani, M. I.(2020). Analysis of differential item functioning in economics multiple choice items administered by west african examination council using logistic regression procedure. *Journal of Critical Reviews. Vol. 8,* ISSUE 01, Pp, 980 – 986.

Ikwueze, C.C. & Ani, M.I. (2020). Application of rasch model in measuring differential item functioning (dif) of students attitude to geography in south east nigeria *Journal of CUDIMAC (J-CUDIMAC).* Vol. 8, No.1.pp, 117-130.

Karam, H. (2012). An Introduction to Differential Item Functioning. *The International Journal of Educational and Psychological Assessment*. Vol. 11(2).Pp 59-65.

Khalid, M. N., &Glas, C. A. W. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement : Journal of the International Measurement Confederation (Ned.), 50,* 186-197. https://doi.org/10.1016/j.measurement.2013.12.019

Kim, J. (2010). Controlling Type 1 Error Rate in Evaluating Differential Item Functioning for Four DIF Methods: Use of Three Procedures for Adjustment of Multiple Item Testing. *Thesis*

Dissertation to *Georgia State University* for the Award of P.hD in Educational Policy Studies

Klenowski, V. (2015). Fair assessment as social practice. Assessment matter, *8,* 76-93. Doi:

10.18296/am.0005

Lee, H. Y. (2012). Evaluation of two types of Differential Item Functioning in factor mixture models with binary outcomes. *The University of Texas at Austin* educational and psychological measurement 74(5), 2014. 831-858

Liu, Y., Yin, H., Xin, T., Shao, L & Yuan, L. (2019).A Comparison of Differential Item Functioning Detection Methods in Cognitive Diagnostic Models. Front. Psychol., 17 May 2019 | https://doi.org/10.3389/fpsyg.2019.01137.

Magis, David; Béland, Sébastien; Tuerlinckx, Francis; De Boeck, Paul (2010). "A general framework and an R package for the detection of dichotomous differential item functioning". Behavior Research Methods.42 (3): 847–862. doi:10.3758/BRM.42.3.847.

Mantel, N., &Haenszel, W. (1959).Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.

Messick, s., (1989): Meaning and values in test validation: the Science and ethics of assessment. Educational Research **18**(2) 5-11.

National Council on Measurement in Education http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/

Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorDArchived 2017-07-22 at the Wayback Machine

Nenty, (2007). Assessment related obstacles to education for all (EFA) and realization of millennium development goal in African countries. 12th Biennial BOLESWANA International Symposium Educational Research, July 23-25, 2007.

Nehm, R. (2017). Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. CBE Life Sci Educ. Summer; 16(2): doi: 10.1187/cbe.16-10-0307NJ: Lawrence Erlbaum Associates Inc.

Okafor, R. N. (2015). Analysis of Gender and Ethnicity-Based Differential Item Functioning in West African Senior School Certificate Mathematics Examination. *An unpublished Thesis submitted to the department of Science Education, University of Nigeria, Nsukka in partial fulfillment of the requirements for the award of Master of Education in Measurement and Evaluation.*

Olaniyonu, S.O.A (2006). Educational Planning. Glorious Ideal and Harsh Realities. 29th Innaugral Lecture. Ojo, Lagos State University.

Omorogiuwa, K. O & Iro-Aghedo, E. P (2016). Determination of Differential Item Functioning by Gender In The National Business And Technical Examinations Board (Nabteb) 2015 Mathematics Multiple Choice Examination. *International Journal of Education, Learning and Development.Vol.4, No.10, pp.25-35.*

*Osadebe, P. U. & Agbure, B. (2018).*assessment of differential item functioning in social studies multiple choice questions in basic education certificate examination. European Journal of Education Studies. Volume 4, Issue 9.

Queensoap, M. & Orluwene, G. W. (2017).Examining Differential Item Functioning in a Chemistry Achievement Test for Students in Nigeria. International Journal of Education and Evaluation Vol. 3 No. 7. Pp, 49-57.

Steinberg, L., &Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11(4),* 402–415.

Strobl, C. Kopf, J. & Zeileis, A. (2011). A new method for detecting differential item functioning in the Raschmodel . Working Papers in Economics and Statistics, Universität Innsbruck

Sunday, M. A. , Lee, W. & Gauthier, I. (2018). Age-related differential item functioning in tests of face and car recognition ability.Journal of Vision January 2018, Vol.18, 2.doi:https://doi.org/10.1167/18.1.2

Svetina, D. (2014) Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments Large-scale Assessments in Education *volume 2, (4).*

Swaminathan, H., & Roger, H. J. (1990). Detecting Item Differential Functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thielemann, D., Richter, F., Strauss, B., Braehler, E., Altmann, U. & Berger, U. (2018). Differential Item Functioning in Brief Instruments of Disordered Eating. *European Journal of Psychological Assessment*. *Vol. 3.*https://doi.org/10.1027/1015-5759/a000472

Van de Vijver, F.J.R & Matsumoto, (2011). "Introduction to the methodological issues associated with cross-cultural research". Pp. 1-14 in Cross-cultural Research method in Psychology: Cambridge University Press.

Van de Vijver, & Tanzer, (2004). "Bias and equivalence in cross-cultural assessment. An Overview". Revn Europeenne de psychologie Appliquee **54**:119-135.

Yadegari, I., Bohm, E., Ayilara, O. F., Zhang, L., Sawatzky, R., Sajobi, T. T. & Lix, L. M. (2019). Differential item functioning of the SF-12 in a population-based regional joint replacement registry. Health and Quality of Life Outcomes *vol. 17, (114).*

Zhang, W. (2006).Detecting differential item functioning using the DINA model. The University of North Carolina at Greensboro