ASSESSMENT OF DIFFERENTIAL ITEM FUNCTIONING OF WASSCE BIOLOGY MULTIPLE-CHOICE EXAMINATION AMONG SECONDARY SCHOOL STUDENTS IN NORTH WEST, NIGERIA

[1]OLUTOLA, A.T. & [2]NURADDEEN, Y. G.

[1]Department of Educational Psychology and Counselling, Federal University Dutsin-Ma, Katsins State, Nigeria.

[2]Department of Early Childhood Care Education, Yusuf Bala Usman College of Education and Legal Studies Daura.

aolutola@fudutsinma.edu.ng , ygaladanchi2015@gmail.com

Abstract

The study analysed the school type differences in differential item functioning of WASSCE Biology multiple-choice examinations of 2021 in North West Nigeria. The study adopted descriptive survey research design. One objective, research question and corresponding hypothesis was raised and tested in the study. The population of the study consisted of 412, 323 Biology students and the sample size of 1,532 Biology senior secondary school students' (SSSIII) were drawn using multistage sampling procedure. The instrument adopted and used for the collection of data was WASSCE Biology paper III used in 2021 examinations. The reliability index of the instrument was 0.73. Binary Logistic Regression was employed in testing the hypothesis in the study at 0.05 level of significance. The results of the study indicated that WASSCE Biology multiple-choice test items used in 2021, contain test items that significantly functioned differentially for students based on school type in favour of mixed school students which placed the boys' and girls' only school at disadvantaged group in the WASSCE 2021 Biology multiple-choice test items. Based on this finding, it was recommended by the researchers that there should be more commitment by the examination bodies in using the IRT approach than the item analysis alone to ensure quality items.

Keywords: Assessment, Biology, DIF, Examination, WAEC, School Type

Introduction

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 Measurements in Education (NCME, 2014), robustness of assessments in demonstrating fairness is essentials for summative assessments used in certification or university admissions, and it is also critical in drawing inferences about student performance on formative assessment, such as teachers made test to avoid test bias. Test items are considered biased when they favour the performance of 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 the quest to avoid inferences is 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 regard as valid, it is imperative that the latent variable (that is construct of interest) is understand 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. DIF occurs when examinees from different groups show different probabilities of success on the item after matching on the underlying ability that the item is intended to measure. In simple terms, DIF arise when two groups of equal ability levels are not equally able to correctly answer an item, Lee (2012). In other words, one group does not have an equal chance of getting an item right compared with another 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 (Orluwene & QueenSoap, 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 difference exists on one of these other so-called secondary latent trait. 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.

Moreover, a test containing items exhibiting DIF could in turn create inaccurate observed total scores resulting in inaccurate estimation of the focal groups primary latent trait (example, biological concept). There are two types of DIF, namely uniformed and non-uniformed 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 (Karami, 2012). That is, there is an interaction between grouping and ability level. 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 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 (IRT). These procedures all assume that the test takers have approximately the same abilities.

The fairness of an examination refers to its freedom from any kind of bias. The examination should be appropriate for all qualified examinees irrespective of race, religion, gender, or age. Fairness in assessment of students' 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. Fairness is an essential quality of a test, its 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 established subsequently if the secondary latent trait that was detected statistically is intentionally related to the primary latent trait.

Bias indicates difference in scores of individuals do not have the same meaning within and across culture. Differential Item Function (DIF) is the most frequently employed statistical analysis of item bias (Van de Vijuer & Tanzer, 2014). The analysis of bias is mandatory before conclusions can be drawn that the groups have different scores on a target construct (Van de Vijuer & Matsunmoto, 2011). Hence, when tests are labelled “biased”, the accusations often have to do with the instruments chosen for a particular context, the way in which the results are interpreted and or used. According to Bark (2004), these broader issues are often for removed from the actual instrument itself and its inherent properties. Therefore, bias is not the mere presence of a score differences between two groups. The term “bias” largely indicates a systematic error that stems differences in performance levels of comparison groups of the same ability level.

Psychological tests can be well-developed or well-constructed, but none are perfect. The reliability of test scores can be compromised by random measurement error (unsystematic error), and the validity of test score interpretations can be compromised by response biases that systematically obscure the psychological differences among respondents (Anigbo, 2006). Psychological tests are often used to make important decisions that affect the lives of real people, which colleges (if any) will decide to accept candidate, in which class will candidate be enrolled, and will an employer decide to hire employee

Suppose you are interested in studying the possibility of school type differences existing in Biology subject ability, if you will give a reasonably reliable Biology test to a representative group of school types, and you find out that; on average mixed school have higher Biology scores than boys and girls only schools. As a researcher you would be tempted to interpret your test scores in terms of the psychological construct that they are intended to reflect that mixed school students tend to have greater Biology subject ability than the other types of school. However, it is possible that the participants' test scores interpreted is not reflecting their Biology subject ability. That is, it is possible that the test is biased in some way. For example, if the mixed school students' test scores overestimated their true Biology subject ability and the boys' girls' only school test scores underestimated their true ability, then the test is biased. In this case, the difference between the test scores might be due to test score bias, not due to a difference in their true Biology subject abilities.

According to Brown (2005) there are two general methods used to detect test biases. Roughly speaking, the two types of test bias reflect biases in the meaning of a test and biases in the use of a test. Construct bias occurs when a test has different meanings for two groups in terms of the precise construct that the test is intended to measure. Construct bias has to do with the relationship of observed scores to true scores on a psychological test. If this relationship can be shown to be systematically different for different groups, then we might conclude that the test is bias. Construct bias can lead to situations in which two groups have the same average true score on a psychological construct but different average test scores (Ary, Jacobs, & Sorensen, 2010). The second type of bias is predictive bias, which occurs when a test used has different implications for two groups of examinees. Predictive bias has to do with the relationship between scores on two different tests. One of these tests (the predictor test) is thought to provide values that can be used to predict scores on the other test (the outcome test or measure) (Brian, Daniel, & William, 2007).

Therefore, 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 assessment.

The Item Response Theory (IRT) is a theory that focuses on an individual's responses to discrete questions. Each question lends insight onto a person's position on one or more spectrums of personality traits. The main focus of IRT tests is the performance on an examination made up of many individual items (Steinberg & Thissen, 1995). One score is not given for the entire test, instead, the respondent is evaluated based on spectrums. This shows researchers their subject's strengths and weaknesses rather than giving one score for the entire test. Many major tests use the IRT approach because it allows for the creation of large test banks.

The West African Senior School Certificate Examination (WASSCE) is a standardized test taken by students in West Africa, including Ghana, Nigeria, Sierra Leone, the Gambia, and Liberia. The WASSCE Biology examination consists of multiple choice test items that are used to assess students' understanding of biological concepts and principles. However, there is a growing concern about the presence of Differential Item Functioning (DIF) in these test items, which may unfairly advantage or disadvantage certain groups of students (Ihechu, 2019). Differential Item Functioning (DIF) analysis is a statistical technique employed to assess whether test items perform differently across various groups of examinees. Looking critically at examinees response in West African Examination Council (WAEC) Senior School Certificate Examination (SSCE) which has been subsumed in the measurement theory of CTT and IRT, student's achievement in WAEC may or may not function differentially in biology between gender of students and between locations of school. Measurement theory reveals why some student respond better than others considering the inherent or latent trait possess by the individual. This trait made it possible for high ability students to consistently function higher and low ability students to also function consistent by lower in an achievement test. When examinees preparing for WASSCE examination walk into the testing hall, they bring along with them, their theta (Θ), which according to IRT, is the ability level, an amount of subject matter knowledge . The WASSCE examination on the theta scale interpret the examinees theta and produce a measurement of ability in the form of raw scores. It is these raw scores that the researcher uses to analyse the psychometric properties of the examination and in turn used these properties in the different methods of calculation DIF to determine the differential status of the examinees. Differential item functioning (DIF) represents a significant challenge in educational assessments, as it has the potential to produce skewed test outcomes and inequitable evaluations of students' competencies and knowledge. In relation to the biology multiple choice test items utilized in WASSCE for year 2021, it becomes imperative to explore possible disparities in DIF based on school type, in order to uphold the validity and reliability of the assessment. Many research findings in Nigeria have shown that there are always differences in the performance between examinee from different school type (Olutola, 2011, Olutola, 2016b, Olutola, Ihechu & Nuraddeen, 2022). Study by Amuche and Fan (2014) have indicated that out of sixty items in NECO Biology questions, ten (10) items were biased in relation to school type and eight (8) items were biased in relation to school location. In addition, a study conducted by Madueke and Casmir (2022) on Differential Item Functioning of WAEC senior secondary certificate examination biology multiple choice items showed that out of 50 items in WAEC 2020 May/June multiple choice biology questions, with respect to gender and school location, DIF were discovered in eight (8) items. These items revealed significant DIF between male and female students and with significant DIF between urban and rural students. Some of the past studies review in this study on DIF were carried out in locale different from the locale of this present study, hence this study on assessment of Differential Item Functioning of WASSCE Biology Multiple-Choice Examination Among Secondary School Students In North West, Nigeria. Specifically, the researchers aim to examine whether there are any systematic differences in item functioning between the school types (Boys' only, Girls' only and mixed school students).

Research Question

In this study, the following one research question was asked to guide the study:

1. What is the percentage of items in the 2021 Biology WASSCE multiple choice items that functioned differentially by school type?

Hypothesis

One hypothesis was formulated and tested at 0.05 alpha level of significant.

Hypothesis One: There is no significant difference in the percentage of items which functioned differentially in the 2021 WASSCE Biology multiple choice items in North West on the basis of school type.

Methodology

In carrying out this study, a descriptive survey research design was employed. All the senior secondary three (SSSIII) students who offered Biology in North West Nigeria were used as study population. This consists of 412,323 students. Multi-stages sampling procedure was used to determine the sample of the study. The sampling stages involved in this study were cluster sampling technique to select the states (Kano, Kaduna and Sokoto) involved and stratified proportionate and simple random sampling techniques to select the schools and subjects of the study. This resulted to have a sample size of 1,532 Biology students from 18 selected schools. The sample size is determined at 95% of confidence level in 2.5% merging error. The WASSCE 2021 Biology multiple choice test items was administered to the students who were not part of the sample in order to determine its reliability index. Thus, the reliability index obtained is 0.73. Binary Logistics Regression is employed in testing the null hypothesis at 0.05 level of significance.

Result

Hypotheses: There is no significant difference in the percentage of items which functioned differentially in the 2021 WASSCE Biology multiple choice items in North West based on school type.

Table 1: Summary of Binary Logistic Regression in Detecting DIF by School Type for 2021 Biology WASSCE multiple choice items

Item	**B**	**S.E**	**Wald**	**Sig**	**Exp (B)**	**95% C.I for Exp (B)**		**Decision**
						Lower	**Upper**	
1.	.019	.064	.086	.769	1.019	.898	1.156	NO.DIF
2.	-.222	.060	13.863	**.000***	.801	.713	.900	DIF
3.	-.186	.061	9.156	**.002***	.830	.736	.937	DIF
4.	.005	.065	.005	.941	1.005	.884	1.142	NO.DIF
5.	-.211	.065	10.559	**.001***	.810	.713	.920	DIF
6.	-.264	.062	18.198	**.000***	.768	.681	.867	DIF
7.	.021	.057	.132	.716	1.021	.921	1.143	NO.DIF
8.	.018	.061	.082	.775	1.018	.903	1.147	NO.DIF
9.	.054	.066	.665	.415	1.055	.927	1.201	NO.DIF
10.	.067	.074	.816	.366	1.069	.925	1.235	NO.DIF
11.	.067	.074	.816	.366	1.069	.925	.1235	NO DIF
12.	-.055	.063	.778	.378	.946	.836	1.070	NO.DIF
13.	-.102	.064	2.491	.115	.903	.796	1.025	NO.DIF
14.	-.134	.066	4.187	**.041***	.875	.769	.994	DIF
15.	-.070	.065	1.181	.277	.932	.821	1.058	NO.DIF
16.	.111	.064	3.003	.083	1.117	.985	1.266	NO.DIF
17.	.096	.069	1.923	.166	1.101	.961	1.261	NO.DIF
18.	-.145	.061	5.627	**.018***	.865	.768	.975	DIF
19.	.050	.062	.653	.419	1.052	.931	1.188	NO.DIF
20.	-.041	.060	.470	.493	.960	.853	1.080	NO.DIF
21.	-.037	.067	.308	.579	.963	.845	1.099	NO.DIF
22.	-.027	.062	.186	.666	.974	.863	1.099	NO.DIF
23.	-.072	.062	1.338	.247	.930	.823	1.051	NO.DIF
24.	-.141	.068	4.315	**.038***	.869	.761	.992	DIF
25.	-.011	.062	.029	.865	.990	.877	1.117	NO.DIF
26.	-.055	.063	.778	.378	.946	.836	1.070	NO.DIF
27.	-.111	.064	3.024	.082	.895	.789	1.014	NO.DIF
28.	.086	.069	1.554	.213	1.090	.952	1.248	NO.DIF
29.	-.060	.054	.000	.343	.942	.831	1.066	NO.DIF
30.	.020	.066	.091	.764	1.020	.897	1.161	NO.DIF
31.	-.143	.063	5/195	**.023***	.867	.767	.980	DIF
32.	-.097	.062	2.452	.117	.907	.803	1.025	NO.DIF
33.	-.123	.062	3.888	**.049***	.884	.783	.999	DIF
34.	.056	.064	.749	.387	1.057	.932	1.199	NO.DIF
35.	-.185	.066	7.728	**.005***	.831	.730	.947	DIF
36.	.123	.063	3.779	.052	1.131	.999	1.280	NO.DIF
37.	-.158	.059	7.212	**.007***	.853	.760	.958	DIF
38.	-.143	.063	5.145	**.023***	.867	.766	.981	DIF
39.	-.153	.061	6.279	**.012***	.858	.762	.967	DIF
40.	-.046	.066	.488	.485	.955	.840	1.086	NO.DIF
41.	.072	.063	1.332	.248	1.075	.951	1.216	NO.DIF
42.	-.015	.069	.045	.833	.985	.860	1.129	NO.DIF
43.	-.240	.062	14.941	**.000***	.786	.696	.888	DIF
44.	-.161	.072	5.026	**.025***	.851	.739	.980	DIF
45.	-.057	.065	.755	.385	.945	.831	1.074	NO.DIF
46.	.019	.062	.093	.760	1.019	.903	1.150	NO.DIF
47.	-.043	.064	.441	.507	.958	.845	1.087	NO.DIF
48.	-.098	.068	2.063	.151	.907	.793	1.036	NO.DIF
49.	.132	.072	3.396	.065	1.141	.992	1.313	NO.DIF
50.	-.029	.068	.188	.665	.971	.850	1.109	NO.DIF

Variables on School Type: *DIF EXIST; Item 2, 3, 5, 6, 14, 18, 24, 31, 33, 35, 37, 38, 39, 43 and 44 only.

Table 2: School Type Cross Tabulation in Group Performance of 2021 Biology WASSCE multiple choice Items Item 2, 3, 5, 6, 14, 18, 24, 31, 33, 35, 37, 38, 39, 43 and 44.

Item	Dichotomous Score	Boy's Only	Girls' Only	Mixed
2.	Incorrect	304	131	578
	Correct	191	92	**236** [a]
3.	Incorrect	316	170	587
	Correct	180	53	**226** [a]
5.	Incorrect	316	170	587
	Correct	180	53	**226** [a]
6.	Incorrect	327	137	620
	Correct	169	86	**193** [a]
14.	Incorrect	361	167	632
	Correct	135	56	**181** [a]
18.	Incorrect	316	170	574
	Correct	180	53	**239** [a]
24.	Incorrect	381	156	658
	Correct	115	67	**155** [a]
31.	Incorrect	345	147	609
	Correct	151	76	**204** [a]
33.	Incorrect	345	147	603
	Correct	151	76	**210** [a]
35.	Incorrect	375	142	660
	Correct	121	81	**153** [a]
37.	Incorrect	298	143	548
	Correct	198	80	**265** [a]
38.	Incorrect	341	165	607
	Correct	155	58	**206** [a]
39.	Incorrect	327	145	587
	Correct	169	78	**226** [a]
43.	Incorrect	322	161	610
	Correct	174	62	**203** [a]
44.	Incorrect	393	172	681
	Correct	103	51	**132** [a]

a = School Type that DIF favoured

Table 1 and 2 shows fifteen (15) items that identified significant DIF in School types of students, using binary logistic regression analysis with the aid of SPSS version 2, items 2, 3, 5, 6, 14, 18, 24, 31, 33, 35, 37, 38, 39, 43 and 44 reveals significant difference between boys, girls and mixed students with significant level less than 0.05. This represents 30% of the total 2021 Biology WASSCE multiple choice items while 70% of the items do not differentiate significantly based on school type. The results further reveals that items 2, 3, 5, 6, 14, 18, 24, 31, 33, 35, 37, 38, 39, 43 and 44 favoured mixed school students which placed the boy's and girl's school students at disadvantaged group. Therefore, there is significant difference in the percentage of items which functioned differentially in the 2021 WASSCE Biology multiple choice items in North West based on school type. Thus, the stated hypothesis is rejected.

Discussion of Finding

This study showed that, the test items used in WASSCE 2021 Biology multiple choice items contain test items that significantly functioned differentially among the students based on school type. The results also revealed that items 2,3,5,6,14,18,24,31,33,35,37,38,39,43 and 44 favoured mixed school students which placed the boy's and girl's school students only at disadvantaged group. This also represents 30% of the total 2021 Biology WASSCE multiple choice items while 70% of the items do not differentiate significantly on the basis of school type. This finding agrees with the finding of Amuche and Fan (2014), who reported that ten (10) items of NECO Biology questions for 2012 were flagged biased with respect to school type. This study is also in line with finding of Ihechu (2019), who submitted that Agricultural science multiple choice items used in NECO and NABTEB 2015-2017, contain test items that significantly differential functioned for students in relation to school type.

Conclusion

The presence of school type differences in DIF of WASSCE Biology test items raises important questions about the equity and reliability of the examination. Understanding and addressing these differences is vital for ensuring that the WASSCE accurately assesses students' knowledge and skills, regardless of their school type.

Recommendation

There should be more commitment by the examination bodies in using the IRT approach than the item analysis alone to ensure quality items. This will eliminate or reduces school type-biased items in public school examinations.

References

Amuche, C. I. & Fan, A. F. (2014). An Assessment of Item Bias Using Differential Item

Anigbo, L. C. (2006). Development and standardization of mathematics achievement test batteries for Primary Four Pupils in Nigeria. "*Unpublished doctoral dissertation*", University of Nigeria, Nsukka.

Ary, D., Jacobs, L. C., & Sorensen, C. (2010). *Introduction to research in education*. Belmont, USA: Wadswoth cengage learning.

Bark, N. (2004). Item Bias detection method for small samples. *Educational Measurement:* Issues practices, 17(1), 31-44.

Brian, F.F., Daniel, H.B., & William, E.F. (2007). The psychometric properties of the agricultural hazardous occupation order certification training program on written examinations, *Journal of Agricultural Education, 48 (4), 11-19*.

Brown, J. D. (2005). *Testing in language programs, a comprehensive guide to English* CA: Duxbury

Ihechu, K.J.P. (2019). Differential Item Functioning of National Examinations Council and National Business Technical Examinations Board Agricultural Science. Unpublished PhD. Dissertation. Michael Okpara University of Agriculture, Umudike.

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

Khalid, M. N., &Glas, C. A. W. (2014). A scale purification procedure for evaluation of differential item functioning. *Journal of Measurement, 50, 186-197.*

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), 831-858

Madueke, U. O., & Casmir, N. E. (2022). Differential Item Functioning of WAEC senior secondary certifcate examination Biology multiple choices. *BIJESS,* 9(1), 31-48.

NCME (2014). National Council on Measurement in Education.

Ogbebor, U. & Onuka, A. (2012). Differential item functioning of economics question papers of National Examinations Council in Delta State, Nigeria, *Nigerian Journal of Educational Research and Education, 12* (1), 45-60.

Olutola, A.T. (2011). *Analysis of Item Parameters of Senior School Certificate Multiple Choice Tests in Biology in Ekiti State, Nigeria*. Doctor of Philosophy (Ph.D) Thesis, University of Ilorin, Ilorin, Nigeria.

Olutola, A.T. (2016b). Assessing Students' Performance in Senior School Certificate Multiple Choice Tests in Biology. I*ssues and Ideas in Education,* 4(1), Pp. 11-20.

Olutola, A. T.; Ihechu, K. J. P. and Nuraddeen, Y. G. (2022). Assessment of Differential Item Functioning of Basic Education Certificate Examination Mathematics Multiple Choice Items. *Ilorin Journal of Education (IJE),* 42 (1), 88 - 94.

Orluwene, G. W. & Queensoap, M. (2019). Use of Mantel Haenszel Differential Item Functioning in Detecting Item Bias in a Chemistry Achievement Test in four ethnic groups in Nigeria. *Internationl Journal of Current Research, 11(3), 2665-2670.*

Steinberg, L., & Thissen, D. (1995). *Item response theory in personality research*. In P.E. Shrout & S. T. Fiske (Eds.), Personality research, methods, and theory: a festschrift honoring Donald W. Fiske (pp. 161-181). Hilldale, NJ: Erlbaum.

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

Van de Vijver, & Tanzer, (2014). "Bias and equivalence in cross-cultural assessment. An Overview" . Revn Europeenne de psychologie Appliquee 54:119- 135.

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