Assessment of Gender-related Differential Item Functioning of Teacher-Made Chemistry test.

[1]Elizabeth O. Ojo [2]Joshua O. Adeleke [3]Simeon O. Ariyo

Abstract

*A good item that will measure the intended domain is expected to be free of biases. But several studies have confirmed that some items in a test reveal biases due to a group of testees.. A generally acceptable analytical technique that can be used to discover biases in test items is the Differential Item Functioning (DIF) which Item Response Theory (IRT) offers to check differences in psychometric properties due to the groups that testees belong. Thus, this study used the DIF technique to detect gender biased items in a teacher made Chemistry test. BILOG-MG was employed using 350 (183 males and 167 females) students from 10 Senior secondary school Two (SSII), randomly drawn from Obio/Akpor Local Government Area of Port Harcourt, River State, Nigeria. The study showed that out of one hundred items, fifty-three items were biased. However, 26****(49.1%)*** *out of 53 were in favour of the female while 27****(50.9%) were in favour of the male which confirmed biases. DIF is effective in detecting group biases of test items. The study concluded that*** *Differential Item Functioning should always be used by scale developers before collating the final items for a test.*

Key words: Item Response Theory, BILOG MG, Differential Item Functioning

INTRODUCTION

Educational measurement and evaluation is a vital approach to ascertaining the desired outcome in learners. It is a way of providing tangible proof that the set objectives are achieved. It is used to ascertain the presence of an attribute or trait in a learner or a group of testees and the extent to which these traits are manifested and being able to pass a value judgment on the learner or group of testees. Banta (2000) said that outcome assessment helps teacher to know the knowledge level of students as a result of underlined instructions of the course module. The data obtained from the assessment can be used as a guide to know how successful a student has mastered the expected set module. This will also assist faculty to improve various aspects of the learning process such as instruction, course content and curricular structure. The information will help the institution to have a strong position about the credibility of a course module in producing competent and marketable graduates. Studies have been carried out on subjects areas such as physics, mathematics, biology, which are the core science subjects. Whereas, not many of such has been done on chemistry. It then becomes imperative to measure and evaluate students' capabilities and outcome performances in Chemistry as a subject.

Nworgu (2011) defined test as an instrument that consists of a set of uniform questions or tasks that a testee or a group of testees will attend to independently. The result of which can be used to provide a reliable on-comparison of different testees' performance. The most practical tool through which the extent of knowledge and skill acquisition is determined at each stage of education is in the form of examination. It is well planned and strategized in order to evaluate, assess and test knowledge and skills. An examination can also be defined as a way of ensuring what a candidate has mastered in a subject matter in a certain field of study (Maduka, 1993). Homby (1995) defined an examination as a formal test of somebody's knowledge or ability in a particular subject, especially by means of answering questions or practical exercises. The process through which students are assessed to understand the quality of what they know in a certain period of time is an examination (Balogun 1999). An examination could be oral or written, essay or objectives, theory or practical.

Whatever the means or instruments of education evaluation, the enhancement of test or examination fairness across a group of testees is very crucial, as the results of such tests or examinations will be used to pass value judgment and make important decisions about the testees.

A test that is not biased provides equal chance to all testees to demonstrate the skills and knowledge which they have acquired and which are vital to the purpose of the test. Testees or test takers of the same latent trait should respond to test items correctly irrespective of their school location, school type or gender (Ogbebor, 2012). Roever (2005) defines a fair test as one that gives all examinees equal chance to exhibit their skills and knowledge which they have acquired and which are relevant to the test's purpose.

Some group of testees could be favoured in some test outcomes while others are disadvantaged not on the basis of the trait being measured but by some factors that have no relationship to the test/examination being undertaken. . Such tests are considered to be biased against the group of testee. The presence of bias is a fundamental problem that needs to be taken care of since bias can lead to systematic errors that affect the inferences made in the classification and selection of students (Zumbo, 1999).

Item bias is the presence of some extraneous elements present in the items that cause differential performance for individuals of the same ability but from different specified subgroups. Zumbo (1999) observed that item bias occurs when examinees of one group are less likely to answer an item correctly than examinees of another group only because of some characteristics of the test item or testing situation that is not relevant to the test purpose. Several questions arise concerning whether higher average test scores by certain groups of testees (gender, age or parental background) are due to actual achievement differences, the bias in a test or some combination of both (Le, 1999).

A biased item measures attributes irrelevant to the tested construct (Williams, 1997). Frequently, examination items are considered biased because they contain sources ofdifficulty that are not relevant to the construct being measured and these extraneous sources impact test-takers' performance (Zumbo, 1999). An item might also be considered biased if it contains language or content that is differentially difficult for different subgroups of test-takers. In addition, an item might demonstrate item structure and format bias if there are ambiguities or inadequacies in the item stem, test instructions, or distractors (Hambleton & Rodgers, 1995).

Nworgu, (2011), revealed that current research evidence has implicated test used in the national and regional examinations as functioning differently with respect to different subgroups. This means that students' scores in such examinations are determined largely by the group to which an examinee belongs and not by ability. Adedoyin (2010) in his study on investigating gender-biased items in public examinations found that, out of 16 test items that fitted the 3PL item response theory statistical analysis, 5 items were gender biased

The statistical technique used to assess the presence of item bias is Differential Item Functioning. The North Central Regional Educational Laboratory (1996) referred to Differential Item Functioning analysis as a procedure used to determine if test questions are fair and appropriate in assessing the knowledge of various groups present among the testees. It is said to be based on the assumption that test takers who have similar knowledge (based on the total score) should perform in similar ways in the individual test questions regardless of their sex, race or ethnicity. However, Differential Item Functioning analysis alone is not sufficient to declare an item biased, other follow-up analysis, such as content analyses, empirical evaluation, are being employed.

Worthy of investigation, is item bias in chemistry as one of the core science subjects, as performance differences are often foundb etween gender(male and female), location(rural and urban), school type(public and private), ethnicity and the like. This study, therefore, found out the gender biased items in teacher made assessment in chemistry using Differential Item functioning approach.

Research Question

1 How do the test items of teacher-made Chemistry test function with respect to gender?

Hypothesis

H_0: There is no significant gender difference in students' response in teacher made Chemistry test.

H_1: There is a significant gender difference in students' response in teacher made Chemistry test.

METHODOLOGY

This study adopted a survey research design. Multistage sampling technique was used to select three hundred and fifty (350) SS II Chemistry students that participated in the study from Obio/Akpor Local Government Area of Rivers State in Nigeria. The instrument (teacher-made chemistry test) was administered by 4 trained research assistants. The instruments were adequately scored and coded by the researcher. Each correct response was scored '1' while each incorrect response was scored '0'. The likely Maximum Likelihood Estimation technique of BILOG MG computer programme was used to analyze the data collected. This technique was used to answer the research question while the hypothesis was tested using the t-test analysis of the SPSS Computer Programme at 0.05 level of significance.

Results

Research Question 1: How do the test items of teacher-made Chemistry test function with respect to gender?.

Table 1: IRT Analysis with respect to Gender in the selected teacher made Chemistry items.

Table 1 shows the IRT DIF statistics on examined item performance with respect to respondent gender. Column 2 of the table gives the adjusted difficulty parameter estimates on the items for male, while column 3 gives the difficulty parameter of the items for the female respondent. Column 4 gives the group difficulty differences. A difference greater than ±0.5 indicates the presence of DIF. Phase 2 of the BILOG MG model gives the statistics in column 2, 3, and 4 for group differential item functioning. 53 items out of 100 have group difficulty differences of +0.5.

Hypothesis: There is no significant gender difference in student's response in teacher made Chemistry test.

Table 2: Independent t-test showing gender difference of students' responses to the teacher made Chemistry test items

GENDER	N	Mean	Std. Deviation	df	T	sig	D
Male	183	41.09	15.73	348	-1.726	0.085	0.18
Female	167	44.34	19.42				

Table 2 shows the t-value and significant (p) value of the group. When the items were subjected to a t-test, the result showed that there was no significant difference in the testees' responses to the items for the teacher made Chemistry test. The result showed that t (348) = -1.726, p>0.05. This showed that the mean difference of male (x=41.09, S.D = 15.73), (x=44.31, S.D = 19.42) was not statistically significant and the effect size (d=0.18) indicated no effect. Therefore the null hypothesis was not rejected and the alternative hypothesis was rejected.

ITEM	GROUP			
	MALE	FEMALE	DIFFERENCE	Remark
CH001	-5.392	-4.567	0.825	Favoured male
CH002	-1.273	-1.66	-0.387	No DIF
CH003	-0.411	2.096	2.507	Favoured male
CH004	-1.007	-1.006	0	No DIF
CH005	-2.961	-3.42	-0.459	No DIF
CH006	-2.846	-1.482	1.364	Favoured male
CH007	-3.278	-1.043	2.235	Favoured male
CH008	-0.839	-1	-0.161	No DIF
CH009	-0.071	-0.439	-0.368	No DIF
CH010	-1.171	-1.742	-0.571	Favoured male
CH011	-0.056	1.54	1.596	Favoured male
CH012	6.238	9.468	3.23	Favoured male
CH013	1.802	1.337	-0.465	No DIF
CH014	-1.381	-1.847	-0.465	No DIF
CH015	-7.345	-11.013	-3.668	Favoured female
CH017	-0.506	-0.863	-0.357	No DIF
CH018	-0.694	-0.803	-0.108	No DIF
CH019	-0.123	-0.636	-0.513	Favoured female
CH020	-0.308	-0.651	-0.343	No DIF
CH021	-0.913	-0.974	-0.06	No DIF
CH022	1.177	1.127	-0.05	No DIF
CH023	12.693	10.884	-1.809	Favoured female
CH024	0.734	2.044	1.31	Favoured male
CH025	-1.388	-1.253	0.136	No DIF
CH027	2.178	2.555	0.377	No DIF
CH028	0.374	-0.123	-0.498	No DIF
CH029	0.038	10.131	10.093	Favoured male
CH030	-1.062	3.329	4.391	Favoured male
CH031	-0.659	-1.535	-0.876	Favoured female
CH032	-0.204	-0.284	-0.08	No DIF

CH033	0.427	1.645	1.218	Favoured male
CH034	0.849	2.039	1.19	Favoured male
CH035	5.864	6.329	0.465	No DIF
CH036	0.854	0.19	-0.665	Favoured female
CH037	-6.448	-5.941	0.507	Favoured male
CH038	1.46	1.139	-0.321	No DIF
CH039	0.178	-0.376	-0.554	Favoured female
CH040	1.091	1.554	0.464	No DIF
CH041	4.808	-0.913	-5.722	Favoured female
CH042	-1.179	-1.17	0.009	No DIF
CH043	1.121	1.76	0.639	Favoured male
CH044	-1.085	-1.067	0.018	No DIF
CH045	3.493	0.917	-2.576	Favoured female
CH046	-0.83	-0.877	-0.047	No DIF
CH047	-0.323	-0.145	0.178	No DIF
CH048	0.393	1.913	1.52	Favoured male
CH049	-0.57	-1.277	-0.707	Favoured female
CH050	-0.823	-1.036	-0.213	No DIF
CH051	-1.269	-1.961	-0.692	Favoured female
CH052	1.616	8.68	7.064	Favoured male
CH053	-0.245	-1.212	-0.967	Favoured female
CH054	-0.592	-1.006	-0.414	No DIF
CH055	3.824	-3.7	-7.525	Favoured female
CH056	0.836	2.071	1.235	Favoured male
CH057	0.7	1.13	0.429	No DIF
CH058	14.693	17.096	2.404	Favoured male
CH059	2.637	1.611	-1.027	Favoured female
CH060	1.38	0.346	-1.035	Favoured female
CH061	12.413	4.974	-7.439	Favoured female
CH062	1.842	0.555	-1.287	Favoured female
CH063	0.413	-0.226	-0.639	Favoured female
CH064	-0.76	-0.943	-0.183	No DIF
CH065	0.2	-0.328	-0.528	Favoured female
CH066	-0.818	0.028	0.846	Favoured male
CH067	-0.612	-0.807	-0.195	No DIF
CH068	0.218	0.614	0.396	No DIF
CH069	0.31	1.257	0.947	Favoured male
CH070	0.754	0.773	0.019	No DIF
CH071	1.229	1.681	0.452	No DIF
CH072	1.15	2.03	0.88	Favoured male
CH073	0.881	0.094	-0.788	Favoured female

CH074	-1.487	-1.042	0.445	No DIF
CH075	-0.956	-1.074	-0.117	No DIF
CH076	5.685	-5.127	-10.812	Favoured female
CH077	0.179	-1.007	-1.186	Favoured female
CH078	-0.357	-0.318	0.04	No DIF
CH079	0.816	0.32	-0.496	No DIF
CH080	-0.165	-0.496	-0.331	No DIF
CH081	1.633	1.767	0.133	No DIF
CH082	-0.324	-0.622	-0.299	No DIF
CH083	0.427	-0.264	-0.691	Favoured female
CH084	2.105	6.935	4.83	Favoured male
CH085	1.65	1.482	-0.168	No DIF
CH086	17.835	18.979	1.144	Favoured male
CH087	-0.06	0.01	0.069	No DIF
CH088	11.645	11.68	0.035	No DIF
CH089	1.754	1.903	0.149	No DIF
CH090	6.361	2.971	-3.39	Favoured female
CH091	1.664	-0.534	-2.198	Favoured female
CH092	0.98	0.126	-0.854	Favoured female
CH093	6.546	2.443	-4.103	Favoured female
CH094	-0.44	-0.182	0.258	No DIF
CH095	-0.97	-0.243	0.727	Favoured male
CH096	6.987	16.53	9.543	Favoured male
CH097	-1.824	-0.034	1.79	Favoured male
CH098	0.429	0.288	-0.141	No DIF
CH100	-0.114	1.326	1.439	Favoured male

Table 3 shows the t-value and significant (p) value of the group. When the items were subjected to a t-test, the result showed that there was no significant difference in the testees' responses to the items for the teacher made Chemistry test. The result showed that t (348) = -1.726, p>0.05. This showed that the mean difference of male (x = 41.09, S.D = 15.73) to female (x = 44.31, S.D = 19.42) was not statistically significant and the effect size (d=0.18) indicated no effect. Therefore the null hypothesis was not rejected and the alternative hypothesis was rejected.

Discussion

This study investigated gender-related Differential Item Functioning (DIF) in a teacher-made chemistry test. It was revealed that out of one hundred teacher made chemistry test administered to SS2 students, fifty-three items showed DIF. Twenty-six (26) items were more difficult for male testees whereas twenty-seven (27) items were more difficult for female testees. This finding shows that 53 of the items measured different things other than Chemistryability purported to be measured.

The finding of this study agrees with the work of Ogbebor (2012) who found that ten (10) out of sixty(60) items in NECO Economics questionsshowed DIF between the rural public schools and private schools, while the private schools were more disadvantaged. The finding further agrees with the findings of Adedoyin (2010) who investigate gender-biased items in public examinations. It was discovered that out of 16 test items that fitted the 3PLitem response theory statistical analysis, 5 items were gender biased.

Furthermore, when the teacher made Chemistry testwas subjected to t-test, it showed that there was no significant gender difference in the students' response to the items. This finding was in agreement with Adeleke and Olabode (2017) who found no significant difference in gender performance in Mathematics items of WAEC 2013 objectives examination. This finding affirms the position of Nworgu (2011) who argued that the presence of DIF in an item does not necessarily mean that the item is biased.

Based on the findings of this study, it was concluded that there was the presence of gender bias in the teacher made Chemistry test, though the difference in gender performance was not statistically significant. Notwithstanding, the fact that there was no significant difference in gender performance does not mean that there were no problems. The difference considered was a group aggregate, whereas , some individual students will be badly affected. On the basis of the findings, it is therefore recommended that test items should be subjected to DIF analysis to identify both uniform and non-uniform biased items. This will guide against items with DIF from reducing the power of the test.

References

Adedoyin O.O (2010). Using IRT Approach ToDetect Gender Biased Items In Public Examinations: Educational ResearchAnd Reviews Academic Journals. 5(7):385-399.

Balogun, J. O. (1999). Examination Malpractice and the Nigerian Society. The Jos Journal of Education4(1) 110-116

Banta T.W. & George D Kuh (2000) Faculty-student Affairs collaboration on Assessment- Lessons from the field. SAGE journals. Vol 4 pp 4-11

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

Hornby A. S. (1995) Oxford Advanced Learner's Dictionary. Oxford University Press.

Le, V.N. (1999). Identifying Differential Item Functioning on the NELS:88 history achievement test. Center for the Study of Evaluation, Los Angeles, CA: CRESST/UCLA.

Maduka, C. (1993). Examination Malpractice: Causes, Implications and Remedies.Benin- City: Ambik Press.

Nworgu B.G (2011). Differential item functioning: Acritical issue in regional quality assurance. Paper presented in NAERA conference.

Ogbebor U.C (2012). Differential Item Functioning Economics Question Paper of National Examinations Council in Delta State Nigeria. Unpublished M.ed thesis. The University of Ibadan.

Olabode J.O & Adeleke J.O (2017) An Investigation Into Differential Item Functioning of 2013 Mathematics Objectives Items conducted by WAEC. Psychological Testing and Innovative Testing Strategies (eds) . pp 64-77.

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

Zumbo BD (19[+99). A Handbook On The Theory And Methods Of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-like (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation.