E-ISSN: 1115 - 8425 P-ISSN: 3027-012X
MPRI-JAEE | INTERNATIONAL JOURNAL OF ASSESSMENT & EVALUATION IN EDU.
VOL. 13 NO. 8 – JULY. 2026

DIFFERENTIAL ITEM FUNCTIONING IN THE 2016 GENERAL STUDY PROGRAMME (GSP) USE OF ENGLISH EXAMINATION: EXAMINING THE EFFECTS OF GENDER AND DISCIPLINE AMONG STUDENTS OF BAYERO UNIVERSITY KANO, NIGERIA

***DR. IBRAHIM SHUAIBU; & **DR. BINTA ABBA**

*Department of Education, Sule Lamido University, Kafin-Hausa. ibrahim.shuaibu@slu.edu.ng **Faculty of Education, Department of Education, Bayero University, Kano. babba.edu@buk.edu.ng

DOI: https://doi.org/10.70382/mejaee.v12i8.0138

Abstract

This paper presents the Analysis of Differential Item Functioning (DIF) of 2016 Use of English (GSP 2201) Examinations of Bayero University Kano, Nigeria with reference to gender and discipline. The analysis was done on the responses of students on Use of English Examination consisting of 100 items. The population of the study consisted of 7,875 L200 undergraduate students and a sample of 1,000 students was selected using multi-stage sampling technique. An expost- facto designed was employed for the study. Four (4) research questions were answered using logistic regression and IRT-based item bias analysis respectively. The findings of the study revealed that, with reference to gender as an indicator, six items (8, 10, 16, 52, and 77) were flagged as having magnitude level DIF effect sizes. In the same vein, eleven (11) Items (i.e. 13, 18, 28, 31, 47, 52, 59, 61, 83, 84, and 93) were found as having magnitude level DIF effect sizes with regards to discipline as an indicator. Consequently, from the items, (8, 52, and 55) are in favour of male group; (10 and 16) are in favour of the female group while item 62 is not in favour of any group respectively. More so, items (13, 18, 28, 31, 47, 52, and 83) are in favour of science-based group; items (59, 84 and 93) are in favour of art-

Key Words:
Differential Item Functioning; Item/Test Bias; Logistic Regression Analysis; Item Response Theory.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

based group, while item 61 is not in favour of any group. From the findings therefore, it is concluded that majority of the items fall under negligible DIF effect sizes with regards to both indicators. However, it is recommended that analysis of students' response to test items for DIF should be encouraged as pilot testing before the final version of such examination.

Introduction

Nigerian Universities provide a number of opportunities for undergraduates to acquire new knowledge, or skills that will best enhance their ability, not only in communication, but also in solving problems and making decisions for the betterment of the society.

The introduction of the new General Study Courses (GSC) was as a result of the gaps identified from needs assessment survey conducted by National Universities Commission (NUC) in 2004. These GSC include: Peace Studies and Conflict Resolution, Entrepreneurial Studies and Languages (French and Arabic) Studies (NUC, 2007). The goal of General Studies Unit as stated in the Bench Mark for Academic Standard (BMAS) is to produce well-rounded, morally and intellectually capable graduates, with vision and entrepreneurial skills in a conducive, peaceful, and socially cohesive environment (NUC, 2007).

Bayero University, Kano (BUK) currently offers the following General Studies Courses (GSC): Use of English (GSP 1201/2201); Use of Library, Study Skills and ICT (GSP 1202/2202); Science, Technology and Society (GSP 2203); Foundations of Nigerian Culture, Government and Economy (GSP 2204); Philosophy and Logic (GSP 2205); Peace Studies and Conflict Resolution (GSP 2206); Entrepreneurship and Innovation (EEP 3201); as well as Venture Creation and Growth (EEP 4201). The courses are offered through the General Studies Unit (GSU) of the institution.

Students' learning is the main concern for schooling, which is measured by various assessment tools such as achievement tests. As such, at the end of each semester, the unit conducts an examination for each of the courses being taught to students. Typically, the items on the examinations are usually multiple-choice (MC) type with four options. Such items appear to be the most flexible and probably most effective of the objective item types. This is due to their versatility in measuring all levels of cognitive abilities. They also permit sampling of wide contents and objectives. As a result of the objectivity in scoring the MC items, scores tend to be reliable. In the same vein, the items can be scored quickly and accurately using machine. Despite these numerous advantages, MC items can provide unprepared students the opportunity to guess, and with the guesses that are right, they get credit for things they do not know. This can consequently expose students

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

to misinformation that can influence subsequent thinking about the content and thus likely to make the test not to measure what it was designed to measure (Suen & McClellan 2003).

Thus, it is therefore necessary that after constructing and reviewing test items based on some common item writing rules, to subject them to series of judgmental and statistical evaluation techniques for validation purposes. This is done in order to identify poor quality items that might be in need of revision or outright rejection. The accepted or revised items are then trial tested. The responses from the trial test would be subjected to series of statistical analyses (Item analyses). The results of the analyses would be used to further identify and/or finalize the most appropriate test items to be incorporated in the final version of the test (Suen & McClellan 2003; Sidhu, 2005; Anastasi & Urbina, 2009).

Gender differences in cognition and learning have long been examined by numerous studies in the past on the basis of language proficiency performance, especially in terms of language skills, test content/topics familiarity, and test format/response types. Regarding language skills and ability, girls tend to have greater verbal ability for e.g. in music. Song, (2014); Yousif, (2011) and Rezaee, (2009) cited studies that supported the above contention in terms of test content and topic familiarity. Other researches generally found that males appear to be at an advantage over females on physical, earth, and space science items (e.g., Brantmeier, 2003, Madu 2012).

In psychometric terms; students' endorsement of an option or answer on a given test (usually multiple examination) could be influenced by students' discipline rather than what the test measures, which means that students' discipline could be a source that the items might functions differently for the test takers (e.g. Education vs. Engineering). This by implication means some items on the test might be worded to be understood in a different way based on students' discipline.

Thus, the probability of endorsing an option or answer in examination item should be determined by the latent trait (e.g. use of English) measured by the examination that the said item comes from. Meaning, the probability of selecting an option or answer on the item for two test takers who have the same trait level is not the same and they are from different groups (different disciplines or field of study) then the item could be biased or its function is not the same across these groups, and this item is a Differential Item Functioning (DIF) item. Raju and Ellis (2002) indicate that detecting DIF means examining the degree to which two test takers with identical standing on the latent trait but from different groups (e.g. male and female) have the same probability of choosing the same option on the item.

Differential Item Functioning (DIF) occurs 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. If the factor leading to DIF is not part of the construct being tested, then the test is biased (Karami, 2012).

The need to analyze tests (examinations) for DIF at the university level arises out of three motives: In the first instance, students in the university system have different experiences in terms of knowledge, skills, norms and values. Secondly, these students are expected to be exposed to common knowledge and skills upon which they will be awarded certificates. Thirdly, tests (examinations) regarding certification should only measure that communal knowledge and skills taught to the students irrespective of gender, socio-economic background, academic discipline etc.

Differential Item Functioning (DIF)

Differential Item Functioning (DIF) occurs whenever people of the same ability level but from different groups have differential probabilities of endorsing an item (Kamata & Vaughn, 2004).

DIF analysis as opined by Ferne and Rupp (2007); Zumbo (2007), is a statistical method that explore whether test items function differentially across different groups of test takers who are matched on ability. it can be said that DIF exists when different groups of learners have differing response probabilities of either (a) successfully answering an item (i.e., in multiple choice) or (b) receiving the same item score in an assessment (Kamata & Vaughn, 2004).

There are two groups in any DIF study, i.e. focal and reference groups. The former, is a group of minorities, which is the potentially disadvantaged group and the later is a group which is considered to be potentially advantaged by the test. (McNamara & Roever, 2006).

Moreover, 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 favored up to a level on the ability scale and, from that point on, the relationship are reversed. That is, there is an interaction between grouping and ability level (Bachman, 2004).

DIF Detection with Logistic Regression (LR)

The use of binary Logistic regression (LR) was reported by Salehi and Tayebi (2012) as the most popular procedure for detecting DIF in dichotomous test items. In addition, LR is a useful technique for detecting both kinds of DIF, uniform and non uniform DIF in dichotomously scored items. LR, in a predictive context, uses regression of the external criterion on test score. Noortgate and Boeck (2005) pointed out that there are three

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

major strengths of the logistic mixed models approach for modeling DIF: "First, logistic mixed models are easy to understand and very flexible. Secondly, the logistic mixed model framework allows considering items and/or groups to be random, resulting in more economical models. Thirdly, using random DIF effects, a hypothesized explanation for DIF can be included in the model through the effects of covariates, without requiring that the explanation is perfect. The covariates can relate to items, persons, or (higher-level) groups of items or persons. The level of effect can be determined as well. To do this, the standardized regression parameters can be used. Jodoin and Gierl (2001) classified the effect levels of DIF that are determined with LR in the following ways:

A Level: If R < .035, a negligible level of DIF is present.

B Level: If .036 < R < .070, a medium level of DIF is present.

C Level: If R > .071, a magnitude level of DIF is present.

DIF Detection Using IRT

Item response theory (IRT) is another widely used method for assessing DIF. It allows for a critical examination of responses to particular items from a test or measure. as noted earlier, DIF examines the probality of correctly responding to or endorsing an item conditioned on the latent trait or ability, it is a fitting approach for examing DIF (Steinberg & Thissen 2006). In the IRT-based approach ICCs are used to evaluate DIF for two groups as cited by De Beer (2004). This means that an item is said to be unbiased if the item characteristic curves for the item measured on two groups are identical. If the item characteristics curves are not identical, then the item is biased and the area between the group ICCs serve as a measure of item aberrance (Adedoyin 2014).

Item Bias/Test Bias

Test items are biased if there are features in them which become sources of difficulty that are irrelevant or extraneous to the construct being measured as pointed out by Li and Zumbo (2009), items that give rise to greater differences in responses between two groups, and if consequent upon the combined effect of these items across the test as a whole would produce greater differences in the total test scores; these artificial differences would be due to the presence of biased items. Item bias occurs when examinees of one group are less likely to answer an item correctly than examinees of another group due to some characteristics of the test item or testing situation that is not relevant to the testing purpose. According to Kamata and Vaughn (2004), item bias is the presence of some characteristics of an item that result in different performance for individuals of the same ability but from different ethnic, sex, cultured or religious group.
Bias in test exists if the testing procedure is unfair to a group of individuals who can be defined in some way. A biased test according to Ling and Lau (2005) may be due to

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

presence of relevant, non-target constructs which are related to gender, ethnicity, race, linguistic background, socioeconomic status or handicapping conditions differences in upbringing environment, culture and daily life experiences.

To Rust and Golombok (2009), the major issues of test bias include race, linguistic minorities, sex, cultural differences and disability, although other categorizations are important in particular circumstances e.g. social class, age and physical attractiveness. It is suspected that there is test bias when a given identifiable group scores low or high on a test relative to some groups, other things being equal. Bias is a major factor for tests considered unfair, inconsistent, and contaminated by extraneous factors. Kamata and Vaughn (2004), hypothesized that, biased items may: 1. result in differential performance for individuals of the same ability from different ethic, sex, or cultural groups; 2. lower the average score of a particular group; 3. contain content or language that is differentially familiar to matched groups of examinees; 4. contain sources of difficulty that are irrelevant or extraneous to the construct being tested, thus adversely affecting test performance; 5. ask for information that disadvantaged children/students who have not had equal opportunity to learn.

In light of this, Kanjee (2007), added that, biased items may a) contain content which may be radically different from a particular subgroup of student's life experiences but the assessment results may be interpreted without taking such differences into proper consideration; b) be representing single-gender negative stereotypes, rather than a balance of gender accomplishments; c) contain clues that would facilitate the performance of one group over another; d) contain inadequacies or ambiguities in the test instructions, item stem, keyed response, or distractors.

Research Objectives

1) To investigate the extent to which 2016 Use of English Examination items displayed DIF effect size by gender
2) To investigate the extent to which 2016 Use of English Examination items displayed DIF effect size by Discipline.
3) Determine if the items on 2016 Use of English Examination are gender biased.
4) Determine if the items on 2016 Use of English Examination are Discipline biased.

Research Questions

1) To what extent do 2016 Use of English Examination items displayed DIF effect size by gender.
2) To what extent do 2016 Use of English Examination items displayed DIF effect size by Discipline.
3) Which items on 2016 Use of English Examination are gender biased?

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

4) Which items on 2016 Use of English Examination are discipline biased?

Methodology

The data used for the study consisted of students' responses on 2016 Use of English Examination of Bayero University Kano, Kano State, Nigeria. An ex- post facto design was adopted to serve as the framework in the study which seeks to find out the factors that are associated with certain occurrences of already existing condition or state of affairs and searching back in time for plausible causal factors retrospectively (Cohen, Lawrence & Morrison, 2007; Cooper & Schindler, 2001). The population of the study consisted of 7,875 L200 Undergraduate students of BUK out of which 5,291 are male students and 2,584 are female students respectively. The students were drawn from fifteen (15) faculties in Bayero University Kano; namely: Agriculture, Allied Health, Art and Islamic Studies, Biomedical Sciences, Clinical Sciences, Communication, Dentistry, Education, EES, Engineering, FCSIT, Law, Pharmaceutical Sciences, Science and SMS respectively. The sample for the study comprised of one thousand (1,000) students who sat for 2016 BUK-GSP 2201 in which 674 are male while 326 are female students respectively. A large sample size was considered to obtain a stable estimate as a requirement in IRT based analysis (Edelen & Reeve, 2007, Bond & Fox, 2007; Morizot, Ainsworth & Reise, 2007). The 2016 GSP 2201, consists of 100 multiple choice items with four (4) options (lettered A to D) prepared, administered and scored by the (SGSE), BUK. The item responses were scored in a binary format of "1" correct and "0" incorrect.

In order to analyze the data collected, the following analyses were performed:

i) Binary Logistic Regression analysis was performed in determining the class effect sizes for each item. Logistic regression analysis was chosen due to the fact that, it is used to analyze dichotomous data such as 1 for right and 0 for wrong response.

ii) IRT- based item bias analysis was done using IRTPRO (Item Response Theory for Patient- Reported Outcomes) in detecting the direction of the biased items cross gender and Discipline using ICCs. IRTPRO is a newly developed application by the Scientific Software International for item calibration and test scoring. The application is only limited to a maximum of 25 items per analysis with a sample of 1,000 and below with three dimensions (1PL, 2PL and 3PL respectively).

Results of the Analyses

The results of the data analyses are presented in tables according to the research questions.

Research Question 1: To what extent do 2016 Use of English Examination items displayed DIF effect size by gender.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Table 1: Summary of Logistic Regression Analysis with Regards to Gender

Effect Level	Item Numbers	Percentages
R < .035, a negligible level of DIF	1,2,3,4,5,6,7,9,11,12,14,15,17,18,19,20,21,22,23,24,25, 26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,4 2,43,44,45,46,47,48,49,50,51,53,54,55,56,57,58,5 9,60,61,63,64,65,66,67,68,69,70,71,72,73,74,75,7 6,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,9 3,94,95,96,97,98, 99,100	93.00
.036 < R < .070, a medium level of DIF	13	1.00
R > .071, a magnitude level of DIF	8,10,16,52,62,77	6.00
	Total	**100.00**

In the above table it is clear that 93.00% of the items have negligible level of DIF, only 1.00% was found to have moderate level DIF and 6 items (6.00%) were found to have a large amount (a magnitude level) of DIF. It is concluded that only 6 Items (i.e. 8, 10, 16, 52, 62, and 77) were found as having large amount of DIF.

Research Question 2: To what extent do 2016 Use of English Examination items displayed DIF effect size by Discipline.

Table 2: Summary of Logistic Regression Analysis with Regards to Faculty of Study

Effect Level	Item Numbers	Percentages
R < .035, a negligible level of DIF	1,2,3,4,5,6,7,8,9,10,11,12,14,16,17,19,20,21,22,23,24,25, 26,27,29,32,33,34,35,36,37,38,39,40,41,42,43,44,45,4 6,48,49,50,51,53,54,55,56,57,58,60,62,63,66,67,68,7 0,71,72,73,74,75,76,77,78,79,80,81,82,85,86,87,89,9 0,91,92,94,95,96,97,98,99	82.00
.036 < R < .070, a medium level of DIF	15,30,64,65,69,88,100	7.00
R > .071, a magnitude level of DIF	13,18,28,31,47,52,59,61,83,84,93	11.00
	Total	**100.00**

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

The table 2 above presents the summary of logistic regression analysis with reference to Discipline. It is indicated that 82.00% of the items have negligible level of DIF, 7.00% were found have moderate DIF and 11 items (11.00%) were found to have a large amount (a magnitude level) of DIF. Thus, it is concluded that 11 Items (i.e. 13, 18, 28, 31, 47, 52, 59, 61, 83, 84, and 93) were found as having large amount of DIF.

Research Question 3: Which items on 2016 Use of English Examination are gender biased?

Table 3: Summary of Item Bias Analysis with regards to Gender

Item No.	Difficulty Parameter (b)		Direction of Bias
	Male	Female	
8	-0.60	-0.33	Favours Male
10	1.36	0.62	Favours Female
16	-0.67	-3.00	Favours Female
52	-1.50	-0.83	Favours Male
62	-0.62	-0.62	Neutral
77	-0.56	-0.54	Favours Male

Table 3 above presents the summary of item bias analysis performed with IRTPRO software with regards to gender. The results from the analysis indicated that item 8, 52, and item77 are in favour of male group, item 10 and 16 are in favour of female group, while item 62 is not in favour of any group. The ICCs for item (8, 52 and 77) for each of the two curves (males and females) were not identical because the ICCs for the female students shifted more to the right. This is an indication that, the items could not discriminate well among these groups and it is more difficult for female students than their male counterparts. It can therefore, be concluded that item (8, 52 and 77) were biased towards female students. In the vein, item (10 and 16) were not identical when compared; the ICCs for male shifted more to the right indicating the items are biased towards male students. However, the ICCs for item 62 are identical, meaning that the item is not biased against any group.

Research Question 4: Which items on 2016 Use of English Examination are gender biased?

Table 5: Summary of Item Bias Analysis with regards to Faculty of Study

Item No.	Difficulty Parameter (b)		Direction of Bias
	Art-Based	Science-Based	
13	-0.21	-1.05	Favours Science-Based
18	-1.01	-2.60	Favours Science-Based
28	3.00	2.40	Favours Science-Based
31	1.08	0.69	Favours Science-Based

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Item No.	Difficulty Parameter (b)		Direction of Bias
	Art-Based	Science-Based	
47	3.29	0.54	Favours Science-Based
52	-0.73	-1.49	Favours Science-Based
59	1.97	2.32	Favours Art-Based
61	-0.98	-0.98	Neutral
83	0.00	-2.07	Favours Science-Based
84	1.25	2.40	Favours Art-Based
93	-0.69	1.15	Favours Art-Based

The table above depicts the summary of item bias analysis performed with IRTPRO software with regards to faculty of study. The outcomes of the result indicated that item 13, 18, 28, 31, 47, 52, and item 83 are in favour of science-based faculty group; item 59, 84 and 93 are in favour of art-based faculty group, while item 61 is not in favour of any group. The ICCs for item (13, 18, 28, 31, 47, 52, and 83) were not identical for both groups. This is an indication that the bias is tilling towards the art-based students due to the fact that, the ICCs for the art-based group shifted more to the right. In the vein, the ICCs for item (59, 84 and 93) were also not identical for both groups which is an indication that the items are tapping some secondary sources rather than the ability being measured. These items are particularly biased towards the science-based students as the ICCs shifted more to the right. However, the ICCs for item 61 are identical showing no bias towards any group.

Conclusion

One important and immediate conclusion based on the findings of the study is that the underlying structure of the items on the 2016 Use of English examination was some items are bias with regards to gender and Discipline in Bayero University, Kano. Further, the existence of the biasness is shifted towards the female students. The same applies to the Art-based students and Science based students. Most of the biased items were biased towards Art-based students.

Recommendations

The detection of DIF items is an important factor to be considered looking at the factors that influence students' performance in Use of English Examination. This is because if an assessment tool is biased, then obviously performance of the students will be influenced, be it in a positive or negative manner. Items developers of such examinations should bear in mind that whenever one or more items in a test are biased, then the validity of the test or examination is at stake. Thus the study recommends that analysis of students'

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

response to test items for DIF should be done as pilot testing before the final version of such examination by the GSP Unit in Bayero University, Kano. By so doing will help test developers to improve on item selection for the test by eliminating DIF and unreliable items, substituting for poor items, or by recasting poorly stated questions for better effect.

References

Adedoyin, O.O. (2014). Identifying Location Biased Items in the 2010 Botswana Junior Certificate Examination Mathematics Paper One Using the Item Response Characteristics Curves. *Journal of International Review of Social Sciences and Humanities 7*, (2), 63-82

Anastasi, A. & Urbina, S. (2009). *Psychological Testing.* 7th Edition. New Delhi: PHI Learning Private Ltd

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bond, T. G., & C. M. Fox. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a Foreign Language, 15,* 1-27.

Cohen, L. Manion, L & Morrison, K. (2007). Research Methods in Education (6th ed). New York: Taylor & Francis Group

Cooper, D.C. and Schinlder, P.S.(2001). *Business Research Methods* (7thed). New York: McGraw-Hill.

De Beer, M. (2004). Use of Differential Item Functioning (DIF) Analysis for Bias Analysis in Test Construction. *SA Journal of Industrial Psychology*, 30 (4), 52-58

Edelen, M. O. & Reeve, B. R. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and renement. *Qual Life Res*, 16, 5.

Ferne, T. & Rupp A. (2007). A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations. *An International Journal Language Assessment Quarterly, 4*, 113-148.

Jodoin, G.M.; & Gierl M.J. (2001) Evaluating type I Error and Power Rates using an Effect Size Measure with Logistic Regression Procedure for DIF Detection. *Applied measurement in Education, 14,* 329-349

Kamata, A. & Vaughn, B. (2004). An Introduction to Differential Item Functioning Analysis. Learning Disabilities: *A Contemporary Journal, 2,* (2), 49-69

Kanjee, A. (2007). Using logistic regression to defect bias when multiple groups are tested. *South African Journal of Psychology, 37,* 47-61

Karami, H. (2012). An Introduction to Differential Item Functioning. *International Journal of Educational and Psychological Assessment, 11(2)*, 59-76

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30*, 343-370.

Ling, S. & Lau, S.H. (2005). Detecting DIF in Standardized Multiple-Choice Test: An Application to IRT Using Three Parameter logistic Model. Retrieved on June 28, 2012, from, www.iiste.org/journals/index.php/JEP/article/download/442/321

Madu, B.C. (2012). Using Transformed Item Difficulty Procedure to Assess Gender-Related Differential Item Functioning of Multiple-Choice Mathematics Items Administered in Nigeria. *Journal of Research on Humanity and Social Sciences* 2,(6), 41-55

McNamara, T., & Roever, C. (2006). *Language testing*: The social dimension. New York: Blackwell publishing.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward Modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R.C.Fraley, & R. F. Krueger (Eds.), Handbook of Research Methods in Personality Psychology (pp. 407-423). New York: Guilford

Noortgate, W. V. D., & Boeck, P. D. (2005). Assessing and examining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics. 30* (40), 443-464.

NUC (2007). Benchmark Minimum Academic Standard (BMAS). Retrieved on June 26, 2012, from https://www.nuc.edu.ng

Raju, N., & Ellis, E. (2002). Differential Item and Test Functioning. In Drasgow, F and Schmitt, N (Eds). *Measuring and Analyzing Behavior in Organizations; Advance in Measurement and Data Analysis.* (pp. 123-155). San Francisco: Jossey-Bass A Wiley Company

Rezaee, A. A. (2009). Gender Differential Item Functioning Analysis of the University of Tehran English Proficiency Test. Retrieved on June 22, 2017, from *http://www. journal.ut.ac.ir/page/download-zk-xe7KD8.artal*

Rust, J. & Golombok, S. (2009). *Modern Psychometrics: The Science of Psychological Assessment,3rd ed*. Washington DC: Routledge/Taylor & Francis Group.

Salehi, M. & Tayebi, A. (2012). Differential Item Functioning: Implications for Test Validation. *Journal of Language Teaching and Research. 3*,(1), 84-92

Sidhu, S.K. (2005). *New Approaches to Measurement and Evaluation*. New Delhi: SPV Limited

Song, X. (2014). DIF investigations with Pearson Test of English Academic. Retrieved on December 20, 2017, from, *Pearsonpte.com/wp…/song_X_2014.pdf*

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),* 482-415

Suen, H.K., & McClellan, S. (2003). Item construction principles and techniques. Retrieved on December 22, 2017, from, *http://suen.ed.psu.edu*

Yousif, A.N. (2011). Comprehension Performance of Gender-Neutral Texts. Retrieved on June 28, 2012 from, *faculty.ksu.edu.sa/yousif/publication/microsoftword-Gender Difference in Reading Comprehension.pdf*

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

EDITOR-IN-CHIEF

Prof. A.I. Mteiye [MALAWI]