Global Journal of Education and Allied Research (GJEAR)
Volume.14, Number 10; October-2023;
ISSN: 2837-3707 | Impact Factor: 7.80
https://zapjournals.com/Journals/index.php/gjear
Published By: Zendo Academic Publishing

GENDER AND SOCIO-ECONOMIC STATUS-BASED ITEM BIAS IN COMPUTER-BASED EXAMINATIONS IN IGNATIUS AJURU UNIVERSITY OF EDUCATION, PORT HARCOURT, RIVERS STATE

[1]Anwuri, Owhorchukwu and [2]Omeke, Kasitem Sheila

Article Info

Keywords: Gender, Socio-Economic status-based item, Computer based examination, Ignatius Ajuru University of Education.

DOI
10.5281/zenodo.10062990

Abstract

This study is aimed at detecting gender and socio-economic-based item bias in computer-based examinations at Ignatius Ajuru University of Education in Port Harcourt, Rivers State. Grounded in the Classical Test Theory and using the instrumentation research design, a sample of 483 second-year students from various departments was analyzed. The primary tool for data collection was the General Studies Assessment Exam (GNS-AE). Results revealed specific items in the examination that favoured certain groups based on gender and socioeconomic status (SES). Specifically, items were identified that favoured male students, students from low socioeconomic backgrounds, and those residing on campus. This study underscores the need for regular item analysis to ensure fairness and accuracy in test results, advocating for the elimination or modification of biased items. The importance of test fairness in educational decision-making cannot be overemphasized, and as such, ongoing training for test developers and the incorporation of advanced analytical systems is recommended.

INTRODUCTION

Background to the Study

Test is a tool or a methodical process for obtaining a generally agreed-upon outcome by presenting a time-based set of questions or tasks to be answered, based on stipulated guidelines. According to this definition, we can perceive a test as both a tool and a process. For example, if I want to determine the best pianist in an audition, I would provide a tool (a piano) and allow each volunteer to showcase their skills based on certain criteria. In doing so, a test has been administered. In ancient times, as narrated in Judges 6:1-7:25 (KJV), Gideon also used a test as a methodical process to select his army to fight the Midianites. Initially, he asked anyone who was afraid to go home. Finally, he narrowed it down to the fine detail of how they drank water from the stream. A test is a process of determining human advancement by assigning a task or requiring a response. In ancient times, some cultures stipulated that a man could only marry a woman of his choice by winning fights. These fights represented

[1]Department of Educational Psychology, Guidance and Counselling, Faculty of Education Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt ***Email:*** ***oc4realzeal@gmail.com***
[2]Department of Educational Foundation, Faculty of Education, Rivers State University ***Email:*** ***renechisom@gmail.com***

the test he had to pass to advance, in this case, to get married. Generally speaking, a test is aimed at achieving an outcome that is generally accepted and agreed upon as right, correct, or the standard for making decisions.

In Agommuoh et al.'s (2016) study, problems associated with Computer-Based Testing (CBT) include errors due to input devices (mouse and keyboard), slow computer response, incorrectly selected options, and test anxiety. Other issues include power outages, disconnection of networked computers from the internet server, and malfunctioning computer systems. Akinola (2019) suggests that to avoid the many pitfalls plaguing tests, Close Circuit Television (CCTV) cameras should be installed to monitor the test-takers and proctors. The system should guide test-takers to the test items after biometric authentication of their fingerprints and verification of their facial appearance. This would reduce incidents of impersonation and ensure that test-takers are taking the test themselves. Another factor that could improve CBT is the randomization of the order of test items for each test-taker, and changing individual login passwords from time to time to prevent unauthorized access.

Huseyin (2018) suggests that before administering a test, the tester must first decide on the test's aim and the type of test that would be appropriate and suitable. A test can be considered appropriate and suitable if it is reliable and valid. A reliable test allows the test-taker to obtain the same or a similar test score irrespective of the number of times the test is taken. Hughes (2017) stresses that the reliability of the test should be judged taking into account situational factors (condition of the test hall), test factors (time frame of the test, font type and size of the test sheet), and test-taker factors (fatigue, stress, and health-related conditions). A valid test ensures that the construct being measured is relevant and appropriate for all test-takers.

Test results have no value until they are interpreted and used for their intended purpose. Therefore, a test aims to determine the value of the information provided by the test-taker and interpret those results to make decisions and judgments. Test bias has been a concern for educational stakeholders, especially after Jasens's (1969, 1980) publications proposed that intelligence is an inherited trait, suggesting that differences in performance can be attributed to genetic factors. This assertion led to a discussion about test performance being attributed to both nature and nurture. These assertions, however, did not fully explain the differing performance of different groups of students. The "bias or different psychometric properties" school of thought attributed the differences in performance within the same group to bias or the test itself measuring different psychometric properties unfamiliar to all test-takers. Bias essentially results in educational injustice stemming from known or unknown factors that tend to disadvantage test-takers of similar ability.

Universities in Nigeria and around the world continue to seek better and more flexible means of assessing students due to the fast-paced nature of today's world. University-based activities need to be conducted quickly, effectively, and efficiently. Computer-Based Testing (CBT) has become the "new normal" for assessing students due to its versatility and flexibility. However, just like traditional Paper-Based Testing (PBT), it is crucial to carefully review the test items, one by one, for any potential bias to ensure fair assessment and evaluation of students while eliminating items that could threaten the reliability and validity of the test

Differential Item Functioning (DIF) is a measurement phenomenon that manifests in the performance of different groups, such as male/female, proficient/non-proficient, etc. This occurs when there are varying probabilities of group members correctly responding to test items, even though they should possess the same level of ability or knowledge.

Item Response Theory-based DIF methods are currently used to detect any form of bias or DIF in tests, especially in Computer-Based Testing (CBT).

Statement of the Problem

The goal of testing in the school setting is to gather reliable and valid data on students to make critical decisions about the teaching and learning process. This goal can be achieved by using well-defined measuring parameters and devices that show a level of equivalence between the sought-after ability and the actual test scores obtained. A measuring device can be considered equivalent if there is a relationship between the measured trait and the test score across various subgroups. In reality, this is not always the case, as there are inconsistencies with measuring instruments, including tests.

Technological trends in today's world mean that school-based activities can be carried out efficiently and effectively with the use of computer systems. This has led to a shift from laborious Paper-Based Testing (PBT)

to Computer-Based Testing (CBT). However, this shift in the mode or medium of testing does not automatically address the persistent questions about test fairness. While other researchers have focused on test bias and test fairness in various subject areas like Mathematics and English Language, little research has been conducted on CBT, particularly at Ignatius Ajuru University of Education. This study aims to fill the gap in understanding test bias in computer-based examinations.

Purpose of the Study

The purpose of this study is to detect item bias in Computer-Based Examinations at Ignatius Ajuru University of Education, Port Harcourt, Rivers State. Specifically, the study aims to:

1. Determine whether the items in Computer-Based Examinations at Ignatius Ajuru University of Education function differently based on students' gender.
2. Examine whether the items in Computer-Based Examinations at Ignatius Ajuru University of Education function differently based on students' socioeconomic status (SES).

Research Questions

The following research questions guide the study:

1. To what extent do items in Computer-Based Examinations at Ignatius Ajuru University of Education function differently based on students' gender?
2. To what extent do items in Computer-Based Examinations at Ignatius Ajuru University of Education function differently based on students' SES?

Hypotheses

The following null hypotheses were tested in the study:

1. **H_0:** The items in Computer-Based Examinations at Ignatius Ajuru University of Education, Port Harcourt do not significantly function differently for male and female students.
2. **H_0:** The items in Computer-Based Examinations at Ignatius Ajuru University of Education, Port Harcourt do not significantly function differently for students with low SES and students with high SES.

Conceptual Review

Test

Onukwo, as cited in Chikwe (2017), defined a test as an instrument or device used to detect behaviors, qualities, traits, characteristics, attributes, etc., possessed by an individual, object, or thing. This means that a test is to a teacher what a stethoscope is to a doctor. If the teaching and learning process is to continue, there must be the introduction of a testing instrument. Therefore, a test forms the basis for teaching and learning. Ukwuiji (2009) defines a test as a series of questions presented to the test-taker or examinees to respond to in order to measure performance or knowledge. Based on this definition, a test is a question and response-based activity that involves the questioner (teacher) or their pre-established questions and the respondent (test-taker). Inko-Tariah & Ogidi (2017) defined a test as a task or series of questions presented to an individual or a group of individuals to assess the presence or quality of traits possessed by them. This definition focuses on the latent (unobservable) traits possessed by the test-taker, indicating that a test is a means of determining what a student has learned to do or can do. Orluwene (2019), citing Kaplan & Saccuzzo (2005), defined a test as a measurement, understanding, and prediction of behavior. A test, therefore, is a means of measuring performance, understanding, and predicting what the test-taker is capable of doing in the future. Onunkwo (2002) defined a test as an instrument that can be used to detect qualities, traits, characteristics, attributes, etc. Oku & Iweka (2018) defined a test as an instrument used to measure as accurately as possible the trait, character, personality, or behavior for which it is designed. A test is an instrument or systematic procedure for measuring a sample of behavior by presenting a set of questions in a uniform manner.

Computer Based Test (CBT)

The British Psychological Society (BPS, 2012) refers to CBT as any psychological test or assessment that involves the use of digital technology to collect, process, and report the results of the assessment. Sorana-Daniela and Lorentz (2017) explain CBT as tests that are administered by a computer, either in a stand-alone or dedicated network, or through other technology devices linked to the internet or World Wide Web, most of which use multiple-choice questions (MCQ).

There are two main types of computer-based testing. The most familiar type involves candidates filling in their responses on a paper form, which is then processed by a computer optical mark reader (OMR). This system reads the form, scores the paper, and may even provide information about test reliability. The second type of computer-based testing involves a computer interface for students to input their answers and receive feedback.

Computer Adaptive Test (CAT):

This type of test tailors or adapts the examination based on the individual test-taker's responses, item by item. In a CAT, the difficulty level of the items increases as the test-taker provides correct answers and vice versa. Brown (2014) and Hughes (2017) add that while students are responding to test items, the system calculates and updates their scores. It decides the difficulty level of the next question based on their previous responses.

Computerized Classification Test (CCT): A CCT is similar to a CAT in that test items are administered one at a time to a test-taker. After responding to an item, the computer scores it and determines whether the test-taker can be classified at that point. If they can be classified, the test ends. If not, another item is administered. This cycle continues until the test-taker is classified or another termination point is reached, such as administering all items in the item bank or reaching a maximum test length.

The CCT is a model of CBT aimed at classifying test-takers into either a dichotomous class (e.g., pass/fail) or a multiple class (e.g., on probation/fail/pass), etc.

Theoretical Review

Classical Test Theory (CTT)

Classical Test Theory (CTT) was the dominant framework for analyzing and developing tests until the 1970s when it gave way to Item Response Theory (IRT). CTT is appreciated for its simplicity and weak assumptions. Over the course of about 80 years, CTT has been instrumental in developing high-quality psychometric scales. It encompasses several theoretical aspects, including the Theory of Item Analysis, Theory of Objectivity, Theory of Validity, and Theory of Reliability.

CTT operates on the fundamental assumption that a testee's observed score (X) is the sum of their true score (T) representing their trait or capability at the time of testing and an error score (E) reflecting extraneous factors that could have influenced the testee during testing. This relationship is mathematically expressed as $X = T + E$. As pointed out by Mehrens & Lehmann (1978), CTT is primarily focused on measuring instruments that differentiate among testees at different points on the test scale. It assumes that test items should discriminate between individuals based on their level of aptitude, with those having greater aptitude expected to obtain higher scores than those with lower aptitude.

CTT, however, has limitations. It cannot predict how well a testee will perform on a test unless the test items have been previously administered. This limitation makes CTT more suitable for Norm-Referenced Testing (NRT) where comparisons are made among students to determine their relative performance. CTT does not explicitly address factors that can affect test item performance. In reality, situational and testee-specific factors, as identified by Hughes (2017), can influence test results. These factors include the condition of the test hall, test duration, font type and size of the test sheet, test-takers fatigue, stress, and health-related conditions. Nevertheless, despite these shortcomings, CTT can still be applied in constructing NRT instruments for comparing students' performance to establish their relative positions.

Empirical Review

Ling & Lau (2014) conducted a study to investigate gender Differential Item Functioning (DIF) in multiple choice and open response science items for elementary, middle, and high school students in Guangxi, China. The study included 23,096 students and used multiple choice and open response science item scales for data collection. Analysis involved the use of xcalibre 4.2.0.1 IRT item parameter extension software, Microsoft Excel, and area index statistics. The findings revealed gender-based DIF attributed to differences in content category, visual-spatial components, and dimensions of item types.

Adedoyin (2016) investigated gender-biased items in public Mathematics examinations. The study included 2,300 junior secondary school students and used the Junior Secondary School Certificate Examination (JSSCE) Mathematics Mock Examination Questions for data collection. The study employed 3PL item response theory statistical analysis and identified 16 items that fitted the analysis, with five of them exhibiting gender bias.

Darmain (2017) studied gender bias in computer-based tests at Kwara State University, Ilorin, involving 1,135 college students. The Kwara State University (KWASU) Post-Unified Tertiary Examination (UTME) served as the test instrument. The analysis, based on the two-parameter logistic model, revealed no significant difference in item functioning between genders.

Birjandi & Mohadeseh (2017) investigated gender-based item bias in computer science tests in Varanasi, India, with a population of 5,209 secondary school students. The findings, obtained using the Rasch method, indicated that seven out of 13 items in general reading comprehension favored female students, while six favored male students.

Omorogiuwa & Iro-Aghedo (2018) examined gender DIF in the National Business and Technical Examinations Board (NABTEB) 2015 Mathematics multiple choice examination. The study involved 63,584 examinees, with 17,815 examinees included in the sample. The analysis, using the Raju method of Item Response Theory, revealed that six items favored males, while 11 items favored female students.

Bichi (2016) evaluated Socioeconomic Status (SES)-based DIF in Northwest University, Kano Post-Unified Tertiary Matriculation Examination (UTME) Economics Test Items. The study included 600 students and utilized the Item Response Theory design. The findings indicated that certain items differentially favored students from low SES (LSES) or high SES (HSES).

Yarmelenko (2018) investigated SES-based item bias in Undergraduate Achievement Test (UAT) at Kharkov University in Ukraine. The study included 3,209 year-one undergraduate students, with 835 students in the sample. Mantel-Haenszel method of Item Response Theory design was employed, revealing that only a few items favored students residing in the school.

Obinne (2018) examined SES-based DIF in Biology examinations by the National Examination Council Examination (NECO) using Item Response Theory. The study involved 1,660 senior secondary year-two students from 26 secondary schools in Benue State, Nigeria. The findings indicated that more test items favored students from HSES than LSES, with only a few items favoring LSES students.

METHODOLOGY

The research design used for this study was the instrumentation research design. The population included all 5,138 year-two students at Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt. The sample for the study was selected using matrix sampling, and the test items were refined to a total of 50 items.

Validation of the draft GNS-AE instrument was carried out by four experts in Measurement and Evaluation from the faculty of Education at Ignatius Ajuru University of Education. Their input led to modifications and removal of irrelevant items. A sample of 50 students from the Department of Human Kinetics was used to establish the internal consistency coefficient of 0.83, which demonstrated the instrument's reliability.

Data collection involved administering the PBT version of the instrument to 462 students, representing 96% of the total administered copies. Data cleaning was conducted, and the coding and analysis were carried out using Statistical Product and Service Solutions (SPSS) and RStudio software packages, respectively.

DATA PRESENTATION AND ANALYSIS

Research Question One

To what extent do items of Computer-Based Examinations in Ignatius Ajuru University of Education function differentially, based on student’s gender?

Table 4.3: Raju Area Statistics and Item Probability (p) values of the 50-item GNS-AE Dataset based on Student's Gender

	Stat.	P-value			Stat.	P-value
ITM_1	1.5328	0.1253		ITM_26	-0.0004	0.9997
ITM_2	0.3164	0.8317		ITM_27	1.5328	0.1253
ITM_3	0.7111	0.4770		ITM_28	0.3164	0.8317
ITM_4	**1.7239**	**0.0847 .**		ITM_29	0.7111	0.4770
ITM_5	0.8712	0.3837		**ITM_30**	**1.7239**	**0.0847 .**
ITM_6	-1.1639	0.2445		ITM_31	0.8712	0.3837
ITM_7	1.5325	0.1254		ITM_32	-1.1639	0.2445
ITM_8	-0.4403	0.6597		ITM_33	1.5325	0.1254
ITM_9	0.0706	0.9437		ITM_34	-0.4403	0.6597
ITM_10	0.5542	0.5795		ITM_35	0.0706	0.9437
ITM_11	0.4483	0.6540		ITM_36	0.5542	0.5795
ITM_12	0.3166	0.8316		ITM_37	0.4483	0.6540
ITM_13	-0.7756	0.4380		ITM_38	-0.7756	0.4380
ITM_14	0.7109	0.4771		ITM_39	0.7109	0.4771
ITM_15	1.2266	0.2200		ITM_40	1.2266	0.2200
ITM_16	1.2977	0.1944		ITM_41	1.2977	0.1944
ITM_17	0.7106	0.4773		ITM_42	0.7106	0.4773
ITM_18	0.7108	0.4772		ITM_43	0.7108	0.4772
ITM_19	0.3165	0.8316		ITM_44	0.3165	0.8316
ITM_20	**2.4943**	**0.0126 ***		**ITM_45**	**2.4943**	**0.0126 ***
ITM_21	0.3972	0.6912		ITM_46	0.3972	0.6912
ITM_22	0.7110	0.4771		ITM_47	0.7110	0.4771
ITM_23	1.1700	0.2420		ITM_48	1.1700	0.2420
ITM_24	-0.1000	0.9204		ITM_49	-0.1000	0.9204
ITM_25	1.2977	0.1944		ITM_50	1.2977	0.1944

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Detection thresholds: -1.96 and 1.96 (significance level: 0.05)

Table 4.3 reveals the Raju statistics as well as the probability (p) value or significance value gotten by the Raju method. Also, item purification (using the latent trait model (ltm) package of R studio) was carried out with 25 iterations. After purification, the following results were obtained and reported.

Table 4.3 also indicates that the items 4, 20, 30 and 45 were flagged as DIF items based on gender. Item 4 have a probability (p) value or significance value of 0.0847, while item 20 have a probability (p) value or significance value of 0.0126. Also, item 30 have a probability (p) value or significance value of 0.0126. Item 4 and 20 are less than the significant code (0.1), while item 20 and 45 are less than the significant code (0.05) hence the items are deemed to function differentially based on gender.

Furthermore, Table 4.3 shows that apart from the four (4) items (4, 20, 30 and 45) which functioned differentially based on gender (i.e: favoured the reference group – male students), the other forty six (46) items did not function differentially based on gender.

Hypothesis One

The items of Computer-Based Examinations in Ignatius Ajuru University of Education, Port Harcourt do not significantly function differentially for male and female students.

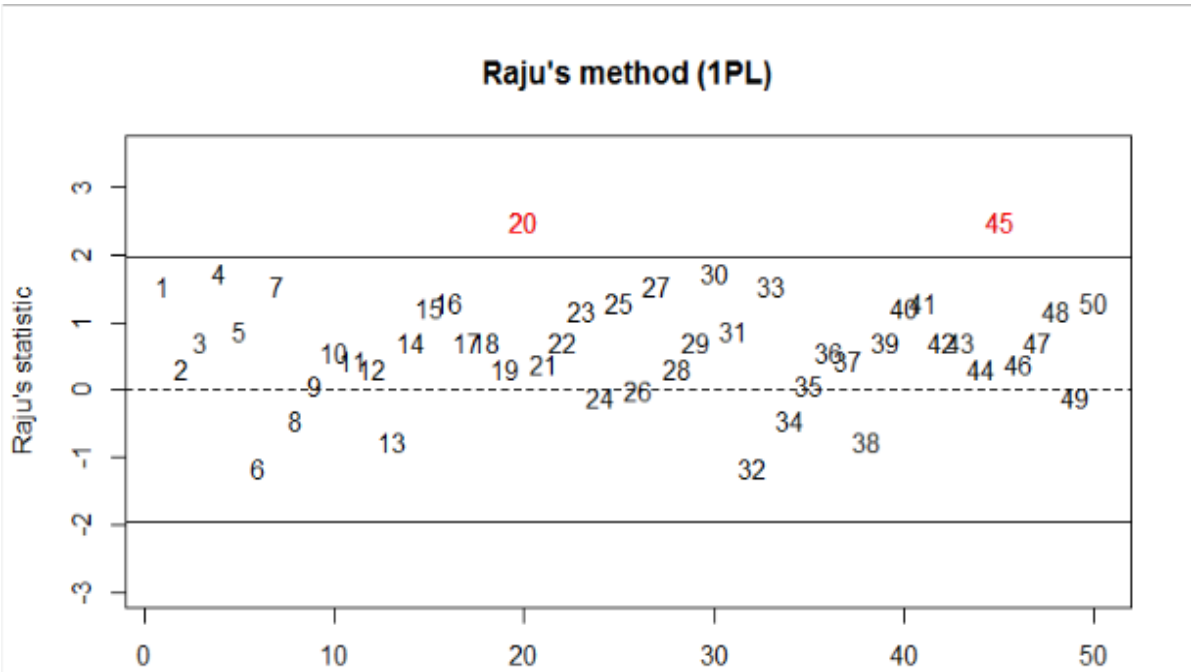

Figure 10: Raju Area Statistics indicating Detection Threshold Plot for Gender-based DIF items.

Figure 10 reveals the Raju statistics as well as the probability (p) value or significance value gotten by the Raju method. Also, item purification (using the latent trait model (ltm) package of R studio) was carried out with 25 iterations. After purification, the following graphic plot was obtained.

Figure 10 indicates that items 20 and 45 (as already indicated on Table 4.3 to have p-values less than 0.05 (level of significance) do not fall within the 1.96 and -1.96 detection threshold at 0.05 level of significance, hence are deemed to function differentially based on gender.

Figure 10 also shows that items 20 and 45 favours the reference group (male students) and disfavours the focal group (females) because it falls above the detection threshold.

Conjointly, as indicated in Appendix C (page), items 20 and 45 are categorized as having large effect sizes (-5.165). This further proves that indeed, both items significantly function differentially for the reference group (male students)

Research Question Two

To what extent do items of Computer-Based Examinations in Ignatius Ajuru University of Education function differentially, based on student's SES

Table 4.4: Raju Area Statistics and Item Probability (p) values of the 50-item GNS-AE Dataset based on Student's SES

	Stat.	P-value			Stat.	P-value	
ITM_1	0.0001	0.9999		**ITM_26**	**-7.7897**	**0.0000**	*******
ITM_2	0.0001	0.9999		ITM_27	0.0001	0.9999	
ITM_3	**-7.9914**	**0.0000**	*******	ITM_28	0.0001	0.9999	
ITM_4	0.0001	0.9999		**ITM_29**	**-7.9914**	**0.0000**	*******
ITM_5	0.0001	0.9999		ITM_30	0.0001	0.9999	
ITM_6	0.0001	0.9999		ITM_31	0.0001	0.9999	
ITM_7	**-8.0824**	**0.0000**	*******	ITM_32	0.0001	0.9999	
ITM_8	**-7.8670**	**0.0000**	*******	**ITM_33**	**-8.0824**	**0.0000**	*******
ITM_9	-0.0012	0.9991		**ITM_34**	**-7.8670**	**0.0000**	*******
ITM_10	**-7.9071**	**0.0000**	*******	ITM_35	-0.0012	0.9991	
ITM_11	0.0001	0.9999		**ITM_36**	**-7.9071**	**0.0000**	*******
ITM_12	0.0001	0.9999		ITM_37	0.0001	0.9999	
ITM_13	**-8.2388**	**0.0000**	*******	**ITM_38**	**-8.2388**	**0.0000**	*******
ITM_14	**-7.9914**	**0.0000**	*******	**ITM_39**	**-7.9914**	**0.0000**	*******
ITM_15	0.0001	0.9999		ITM_40	0.0001	0.9999	
ITM_16	0.0001	0.9999		ITM_41	0.0001	0.9999	
ITM_17	**-7.9914**	**0.0000**	*******	**ITM_42**	**-7.9914**	**0.0000**	*******
ITM_18	0.0001	0.9999		ITM_43	0.0001	0.9999	
ITM_19	**-7.9486**	**0.0000**	*******	**ITM_44**	**-7.9486**	**0.0000**	*******
ITM_20	0.0001	0.9999		ITM_45	0.0001	0.9999	
ITM_21	**-7.8280**	**0.0000**	*******	**ITM_46**	**-7.8280**	**0.0000**	*******
ITM_22	**-7.9913**	**0.0000**	*******	**ITM_47**	**-7.9913**	**0.0000**	*******
ITM_23	0.0001	0.9999		ITM_48	0.0001	0.9999	
ITM_24	0.0001	0.9999		ITM_49	0.0001	0.9999	
ITM_25	**-8.1316**	**0.0000**	*******	**ITM_50**	**-8.1316**	**0.0000**	*******

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Detection thresholds: -1.96 and 1.96 (significance level: 0.05)

Table 4.4 indicates that the items 3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36 38, 39, 42, 44, 46, 47 and 50 were flagged as DIF items based on SES. Items 3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36 38, 39, 42, 44, 46, 47 and 50 all have a probability (p) value or significance value of 0.000. This probability (p) value or significance value is less than the significant code (0.001), hence the items are deemed to function differentially based on SES.

Furthermore, Table 4.4 shows that apart from the twenty three (23) items (3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36 38, 39, 42, 44, 46, 47 and 50) which functioned differentially based on SES (i.e: favoured the focal group – LSES students), the other twenty seven (46) items did not function differentially based on SES.

Hypotheses Two

The items of Computer-Based Examinations in Ignatius Ajuru University of Education, Port Harcourt do not significantly function differentially for LSES students and HSES students.

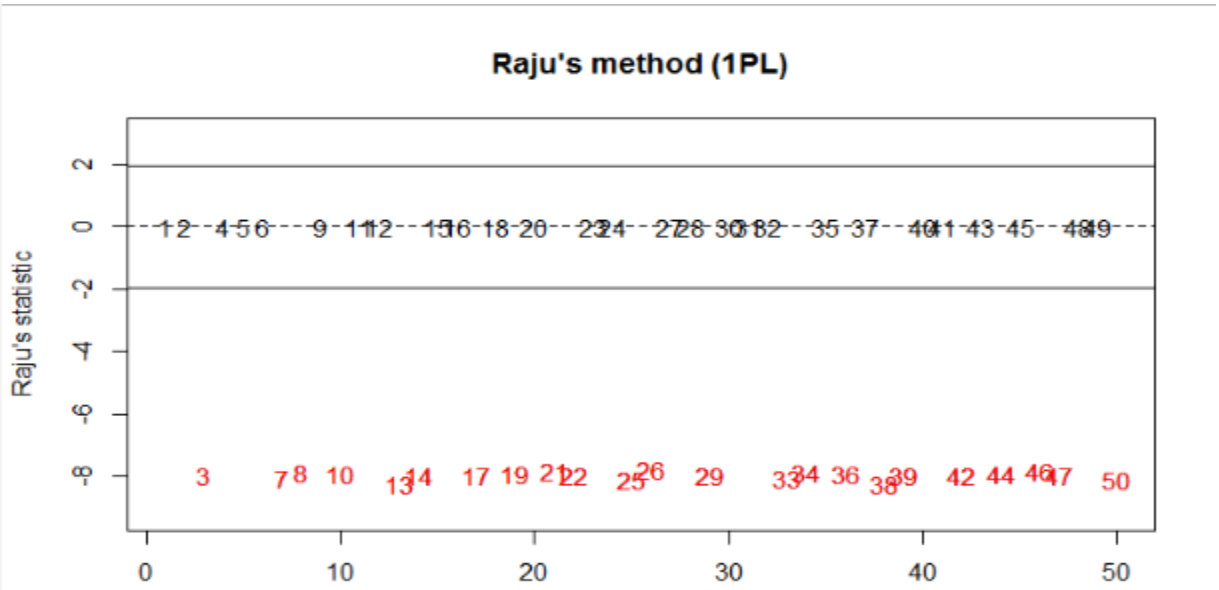

Figure 11: Raju Area Statistics indicating Detection Threshold Plot for SES-based DIF items.

Figure 11 indicates that twenty three (23) items (3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36 38, 39, 42, 44, 46, 47 and 50 (as already indicated on Table 4.4 to have p-values less than 0.05 (level of significance)) do not fall within the 1.96 and -1.96 detection threshold at 0.05 level of significance, hence are deemed to function differentially based on SES.

Figure 11 also shows that items 3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36 38, 39, 42, 44, 46, 47 and 50 favours the focal group (students from LSES) and disfavours the reference group (students from HSES) because it falls below the detection threshold.

Conjointly, as indicated in Appendix C (page) items 3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36 38, 39, 42, 44, 46, 47 and 50 are categorized as having large effect sizes (25.5229, 26.1203, 24.7335, 24.985, 27.2018, 25.5229, 25.5229, 25.2479, 24.4905, 25.5222, 26.4525, 25.5229, 25.5229, 26.1203, 24.7335, 24.9852, 27.2081, 25.5229, 25.2479, 24.4905, 25.5222 and 26.4525 respectively). This further proves that indeed, these items significantly function differentially for the focal group (students from LSES)

Discussion of Findings

The study explored the detection of item bias in computer-based examinations of Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt, Rivers State. Before the GNS-AE Dataset was fitted into any of the IRT models, it was first tested to determine if it fulfills the unidimensionality assumption. The dataset proved so, hence the analyzed.

The findings from Research Question One reveal that two (2) items (items 20 and 40) favoured the male students in the expense of the female students in Ignatius Ajuru University of Education Computer-based Examinations. This finding is in line with that of Ling & Lau (2014), Adedoyin (2016), Birjandi & Mohadesh (2017) which also showed gender-based bias. On the other hand, the finding is in contrast with the findings of Darmain (2017) which did not find any gender bias in the research carried out in that area.

The findings from Research Question Two reveal that twenty two (22) Items 3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36, 38, 39, 42, 44, 47 and 50 favoured students from LSES in the expense of the students from HSES Ignatius Ajuru University of Education's Computer-based Examinations. This is partially in agreement with that of Omolara (2016), Bichi (2016), Yarmelenko (2018), Obinne (2018), and Kwado (2019) which also showed that some items favoured students from LSES, as well as other items favouring students from HSES in that area.

Conclusion

The Findings of the study showed that Items 1, 3, 7, 8 and 10 favoured male students, items (3, 7, 8, 10, 13, 14, 17, 19, 21, 22, 25, 26, 29, 33, 34, 36 38, 39, 42, 44, 46, 47 and 50 favoured students from LSES, items 2, 21, 28

and 46 favours the reference favoured students who reside on campus, while item 22 and 47 favours students who reside off campus.

The essence for test administration is to be able to make decisions which would affect the student, curriculum, school, teacher, parents and other stakeholders in education. The presence of test bias would defeat this essence, thus bias items should be detected, modified or eliminated, so that all students (irrespective of their cultural attributes) would get results that they merit. This will in turn accurately influence the decisions taken about the student, curriculum, school, teachers, etc.

Recommendations of the Study

The following Recommendations are made by the researcher,

1. All tests which are to be administered must go through item analysis to ensure that no bias item(s) exists which could affect test results.
2. Test experts and Psychometricians should act as watch dogs for schools to ensure that test fairness is ensured at all times.

References

Adedoyin, O. O. (2010). Using IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics. Educational Research and Review, 5(7), 385-399.

Agommuoh, N. (2006). Validity of Nigerians unified tertiary matriculation examination, physics computer-based tests: Threats and Opportunities. Journal of Research & Methods in Education, 3(5), 33-38.

Akinola, W. T. (2019). Test security and educational development: A critical issue in quality assurance. Paper presented in NAERA conference.

Bridgeman, B., & Cline, F. (2000). Variations in mean response times for questions on the computer adaptive graduate record examination general test: implications for fair assessment. Wiley.

Brown, C. G. (2004). Computer-assisted assessment in higher education.

Chikwe, C. K. (2017). Fundamentals of test, measurement, and evaluation in education. Emmanest Ventures.

Hughes U. G. (2017). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. Applied Psychological Measurement, 15, 142-187.

Huseyin, R. T. (2018). Examination malpractice: Causes, implications, and remedies. Educational Research and Review Academic Journals, 6(7), 125-133.

Inko-Tariah, D. C., & Ogidi, R. C. (2017). Fundamentals for psychological testing for psychologists, counselors, and educationists. Rodi Printing and Publishing Company.

Iweka, F. (2014). Comprehensive guide to test construction and administration. Chifas Publications.

Ling, S. E., & Lan, S. H. (2004). Detecting DIF in standardized multiple-choice tests: An application of IRT using three-parameter logistic model. Journal of Applied Psychology, 94(7), 452-459.

Ojerinde, D. (2012). Introduction to item response theory, parameter models, estimation, and application. Lagos State University Press.

Oku, K., & Iweka, F. (2018). Development, standardization, and application of chemistry achievement test using the one-parameter logistic model (1-PLM) of IRT. American Journal of Educational Research, 6(3), 238-257.

Omorojuwa, H. P., & Iro-Aghedo, B. H. (2016). The evaluation of the differences in test performance of two or more groups. Educational and Psychological Measurement, 3(4), 807-816.

Onunkwo, G. I. N. (2002). Fundamentals of educational measurement and evaluation. Cape Publishers International.

Orluwene, G. W. (2019). Detecting item bias with Scheuneman chi-square in chemistry achievement test in Nigeria. International Journal of Innovative Social Science Educational Research, 7(1), 88-101.

Parshall, C. G. (2002). Principal considerations in computer-based testing. Oxford University Press.

Sorana-Daniella, B., & Lorentz, J. (2007). Computer-based testing in physical chemistry topics. International Journal of Education and Development using Information and Communication Technology, 3(1), 94-95.

Zumbo, B. D. (2007). Three generations of DIF analysis: considering where it has been, where it is now and where it is going. Language Assessment Quarterly, 4(3), 223-233.