IAIIEA
International Association For Innovations
In Educational Assessment

JOURNAL OF INNOVATIONS IN EDUCATIONAL ASSESSMENT

Vol. 7, No. 2, June 2025.

´Revue des Innovations en Evaluation Pédagogique

ISSN: 2705-3830 (Print)

ISSN: 2705-3857 (Online)

Exploring Item Bias Analysis Methods for Enhanced Digital Assessment in the Tertiary Education Sector

by

Inko-Tariah, D. C.[1]

and

Anwuri, Owhorchukwu[2]

[1, 2] Department of Educational Psychology, Guidance and Counselling, Faculty of Education,

Ignatius Ajuru University of Education, Rumuolumeni, Rivers State.

dorothy.inkotariah@iauoe.edu.ng +2348033397091

and

oc4realzeal@gmail.com +2348051115727

Abstract

This paper explored Item Bias Analysis Methods for Enhanced Digital Assessment in the Tertiary Education Sector. With the growing adoption of digital platforms for university examinations, there are increasing concerns about the psychometric quality of Computer-Based Tests (CBTs). The Lord Raju method, grounded in Item Response Theory (IRT), offers a sophisticated approach to evaluating item discrimination and difficulty, enhancing precision and validity. In contrast, the Mantel Haenszel method, which relies on contingency tables, provides a more straightforward and accessible alternative. This paper compared these methods to assess their applicability, reliability, and suitability for tertiary assessments. Conducted with 3,851 third-year undergraduates at Ignatius Ajuru University of Education, the study utilized matrix sampling techniques to select 800 students and 50 test items. Data were gathered using the validated General Studies English Language Performance Test (GNSELPT), which has a reliability coefficient of 0.84. The findings indicate that both methods effectively identified Differential Item Functioning (DIF) items, though their consistency varied with gender and mode of entry. The study recommends regular monitoring and evaluation of test items using multiple methods as part of a quality assurance process.

Keywords: Item analysis methods, digital assessment, tertiary education

Exploring Item Bias Analysis Methods for Enhanced Digital Assessment in the Tertiary Education Sector

In the pursuit of enhancing efficiency, reliability, and security in assessment processes, educational institutions worldwide are increasingly turning to digital platforms for examinations. This global trend reflects the desire to leverage technological advancements to streamline assessment administration and improve the integrity of examination processes (Brown and Kennedy, 2019). Digital assessments offer numerous advantages over traditional pen-and-paper methods, which include reduced examination malpractice, quick feedback, easy examination administration, which collectively enhance the overall educational experience (Anwuri, 2022).

Rivers State, Nigeria, characterized by its diverse population and vibrant tertiary education sector, is no exception to this trend. As a key player in Nigeria's educational landscape, Rivers State is committed to adopting innovative solutions to maintain and elevate educational standards. The state's tertiary institutions are increasingly adopting digital assessments to address issues such as examination malpractices, logistical inefficiencies, and the need for timely feedback (Hartoyo et al., 2020). Despite these efforts, the challenge of ensuring the quality and fairness of digital assessments, particularly regarding the psychometric properties of test items remains.

The Lord Raju Differential Item Functioning (DIF) method is based on Item Response Theory (IRT). This sophisticated method evaluates the performance of test items by examining item discrimination and difficulty across different groups (Embretson & Reise, 2013). In the context of using R software, the latent trait model (ltm) package is often employed for IRT-based analyses, including the Lord Raju method. The Lord Raju method involves several steps in R which include: Model Fitting, i.e.: fitting an IRT model (typically a 2PL or 3PL model) to the data using the latent trait model (ltm) or multidimensional item response trait (mirt) package. Estimation of Parameters, i.e.: estimating the item parameters (discrimination and difficulty) for each item. DIF Analysis, i.e.: conducting DIF analysis by comparing item parameters across different groups using the differential item functioning R(difR) package, which includes functions specifically designed for IRT-based DIF detection. This method allows for a sophisticated understanding of item performance and can identify items that function differently for various groups based on underlying ability levels (Anwuri 2022).

The Mantel Haenszel (MH) method is a classical approach to detecting DIF that uses contingency tables to compare item performance across different groups (Anwuri 2022). This method is relatively straightforward and is often implemented using the differential item functioning R(difR) package in R, which provides functions for MH DIF analysis. The Mantel Haenszel method involves the following steps in R: Data Preparation, i.e.: preparing the data by grouping responses based on different subgroups. Contingency Tables, i.e.: constructing contingency tables for each item. DIF Analysis, i.e.: using the dif MH function from the differential item functioning R(difR) package to perform the MH DIF analysis. The MH method provides an odds ratio for each item, indicating the likelihood of different groups responding differently to the same item. This method is particularly valued for its simplicity and ease of interpretation (Anwuri 2022).

The MH DIF method seems a more practical method, not just in terms of basic manual computation but also in terms of teaching test constructors. It's easier to understand each step. However, its simplicity may not allow for more sophisticated approaches that could control for DIF sensitivity. This is where the Lord Raju Method outweighs the MH DIF method, with respect to it's focus on computing the area between the ICCs (curves showing the probability of a correct response across different ability levels) of the focal and reference groups across all ability levels. If the area is significant, the item exhibits DIF. This kind of estimation involves more in-depth mathematical knowledgeability.

However, despite the seeming mathematical constraint, the R Software can perform these tasks with the use of several dedicated codes. This is the approach considered for this study. In R Software, both methods can be implemented using the differential item functioning R(difR) package, with additional support for IRT modeling provided by packages such as latent trait model (ltm) and multidimensional item response trait (mirt) for the Lord Raju method. The Mantel Haenszel method's implementation is more straightforward, requiring only the differential item functioning R(difR) package.

Gender-based Differential Item Functioning (DIF) refers to the occurrence when test items exhibit different levels of difficulty or favor one gender over another, even when individuals from both genders have equivalent abilities. Several studies have been carried out on this area of Differential Item Functioning (DIF), for instance, Yao and Chen (2020) conducted a study titled "Gender-related Differential Item Functioning Analysis on an ESL Test," focusing on whether test items in the General English Proficiency Test for Kids (GEPT-Kids) are biased by gender. The population comprised 492 participants from five Chinese-speaking cities. The sample was analyzed using the Mantel-Haenszel method to detect gender DIF, and items showing DIF were further examined through content analysis by three experienced reviewers. The findings revealed that three items exhibited moderate gender DIF according to statistical methods, and three items were potentially biased based on expert judgment.

Furthermore, Annan-Brew (2020) focused on examining Differential Item Functioning (DIF) in the West African Senior School Certificate Examination (WASSCE) core subjects (English Language, Mathematics, Integrated Science, and Social Studies) in Southern Ghana from 2012-2016. The study aimed to determine whether the exam items exhibited gender DIF using a cross-sectional design. A sample of 36,035 candidates was selected from those who took the exams. The findings from the study indicated that 43 items favoured males, with 38 favouring females in the first year investigated, while 31 items favoured males, with 37 favouring females in the second year investigated. Also, 32 items favoured males, with 27 favouring females in the third year, while 35 items favoured males, with 22 favouring females in the fourth year. Finally, 28 items favoured males, with 27 favouring females in the fifth year. Similarly, mode of entry Differential Item Functioning (DIF) identifies items that favour either the JAMB entry undergraduates or direct entry undergraduates. This area has barely been explored in the context of universities. Anwuri (2022) carried out a study titled: Detection of item bias in Computer-Based Examinations in Ignatius Ajuru University of Education, Port Harcourt, Rivers State. This study's findings, though focused on Peace and Conflict Resolution indicated that no item functioned differentially for either students admitted through direct entry or students admitted through JAMB. It was also found that two items were flagged as (significantly) differentially functioning items (based on student's gender) which favoured the reference group (male students).

Comparing different DIF methods can help evaluate the consistency and robustness of the results. When multiple methods consistently flag the same item(s) as exhibiting DIF, it raises confidence in the findings. On the other hand, inconsistencies between methods may indicate a need for further investigation to understand the underlying reasons for the differences. In the same vein, Abbott (2021) explored the sensitivity of various methods for detecting Differential Item Functioning (DIF), including Lord's chi-square, Raju's area, logistic regression, Mantel-Haenszel (MH), standardization, and transformed item difficulties (TID). The study, conducted with 400 vocational students in Rivers State, Nigeria, used multiple-choice items from the 2019 Junior School Certificate Examination in computer science. These items were analyzed using the 2PL model from Item Response Theory (IRT) and assessed for DIF using different methods implemented in R. Key findings indicated that the standardization method identified the most DIF items, followed by logistic regression and Lord's chi-square, which also detected a significant number of DIF items. The TID method detected more DIF items than the Mantel-Haenszel method, while the Raju's area method did not detect any DIF.

Based on Complexity and Sophistication, the Lord Raju method, being IRT-based, provides a more detailed and sophisticated analysis by modeling item characteristics and examining how these characteristics interact with examinee abilities. The Mantel Haenszel method, on the other hand, is simpler and focuses on comparing proportions of correct responses across groups using contingency tables. Based on data requirements, the Lord Raju method requires fitting an IRT model, which involves more complex data requirements and assumptions about the underlying ability distribution. The Mantel Haenszel method can be applied with simpler data structures and fewer assumptions, making it more accessible for basic DIF detection. Based on interpretation of results, the results from the Lord Raju method provide detailed information on item discrimination and difficulty parameters, which are useful for understanding how items function across different levels of ability. The Mantel Haenszel method provides an odds ratio, which is straightforward to interpret but offers less sophisticated insights compared to IRT-based (Anwuri 2022).

By systematically comparing these two methodologies, this study aims to shed light on their applicability, reliability, and suitability within the context of tertiary institution-based assessments in Rivers State. This sophisticated examination contributes to the ongoing discourse on improving the quality and effectiveness of assessments, guiding educators in discerning the most appropriate approach to suit their assessment objectives and contextual constraints.

Statement of the Problem

The increasing apprehension regarding assessments in tertiary education, particularly Computer-Based Tests (CBTs), revolves around the possibility that they might be solely devised by instructors without the rigorous evaluation required to verify the psychometric attributes of each item. This concern underscores a significant issue: merely transitioning to a digital testing format does not inherently enhance the caliber of the test items. Without thorough item analysis, digital assessments may fail to provide accurate, reliable, and fair measures of student performance.

A singular approach to item analysis may prove inadequate in identifying and rectifying biased test items, which can lead to inequities and compromise the validity of assessment outcomes. The Lord Raju method, through its IRT-based framework, offers a detailed analysis of item characteristics, but its complexity can be a barrier to widespread adoption. On the other hand, the Mantel Haenszel method, while simpler, may not provide the same depth of analysis but is easier to implement. It is on this premise that the study investigated Item Bias Analysis Methods for Enhanced Digital Assessment in the Tertiary Education Sector.

Research Questions

The following research questions guided the study:

1. Which item(s) of the Computer-Based Test (CBT) of Ignatius Ajuru University of Education was (were) flagged as DIF using Lord Raju and Mantel-Haenszel's Methods, based on Gender?

2. Which item(s) of the Computer-Based Test (CBT) of Ignatius Ajuru University of Education was (were) flagged as DIF using Lord Raju and Mantel-Haenszel's Methods, based on Mode of Entry?

Hypotheses

The following null hypotheses were tested at 0.05 level of significance.

1. There is no significant difference in the items flagged as Differential Item Functioning (DIF) in the Computer-Based Test (CBT) of Ignatius Ajuru University of Education when analyzed using Lord Raju and Mantel-Haenszel methods based on Gender.
2. There is no significant difference in the items flagged as Differential Item Functioning (DIF) in the Computer-Based Test (CBT) of Ignatius Ajuru University of Education when analyzed using Lord Raju and Mantel-Haenszel methods based on Mode of Entry.

Method

The study adopted the instrumentation and comparative research designs. The Instrumentation research design involves the selection and use of specific tools or instruments (in this case GNS CBT of Ignatius Ajuru University of Education) for data collection, assessment and measurement of potential biases in the test items. The comparative methods are the Lord Raju and Mantel Haenszel Methods. The population of this study consist of 3,851 years three undergraduates in Ignatius Ajuru University of Education (2023 Student Enrolment List from the Various Faculties in IAUE). Matrix sampling technique was used in selecting the sample for the study as well as selection of testees for research purpose. The sampled test items are fifty (50) altogether from an item pool of two hundred and eighty-three (283). The sample of the study consisted of both male and female three hundred (300) level students from the six faculties (Education, Humanities, Social Sciences, Management Sciences, Natural and Applied Sciences and Vocational and Technical Education) in the University. The minimum sample size was determined using Taro Yamene formula, yielding 362. However, the researcher purposively increased the sample size to 450.

The instrument for data collection is the General Studies English Language Performance Test (GNSELPT). It was designed using the GNS 2-Communication Skills in English II. The instrument was segmented into two sections, A, B. Section A contains the demographic information such as: gender, mode of entry. Section A contained the fifty (50) test items with four options (A-D). The scoring of the responses was dichotomous (correct = 1, wrong = 0)

The General Studies English Language Performance Test (GNSELPT) items was given to the researcher's supervisor, along with two language experts in the Department of English Language and Communication and two Measurement and Evaluation experts in the Department of Educational Psychology, Guidance and Counselling, all in Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt, Rivers State

The instrument was administered to 20 undergraduates who were not sampled for the study, and Cronbach Alpha reliability coefficient was calculated from their responses to the items. The reliability coefficient was 0.84. The research instrument was administered directly. Later, 800 copies of the instrument were retrieved out of the 780 representing 98 percent of the total administered copies. Lord Raju and Mantel-Haenszel DIF detection criteria was used in answering and testing the null hypotheses.

The reliability of the General Studies English Language Performance Test (GNSELPT) was determined using Kuder-Richardson20 method (KR20), by means of Microsoft Excel package of the Microsoft Office Suite, after being administered to 20 three hundred level students in the Department of Human Kinetics who did not make up the sample for the final study.

Lord Raju and Mantel-Haenszel DIF methods were comparatively used to determine the indices of DIF of all the test items, with the differential item functioning R(difR) package of the R software (via R) was used in determining the indices of DIF (with respect to the Mantel Haenszel and Lord Raju Methods).

Results

To answer the research questions, the Mantel-Haenszel method with continuity correction using the Multiple comparisons made with Benjamini-Hochberg adjustment of p-values and without item purification were used to detect DIF in the dataset. A follow up hypothesis testing was carried out to verify if the number of items flagged for gender-based DIF by the two methods are significant.

Research Question One

Which item(s) of the Computer-Based Test (CBT) of Ignatius Ajuru University of Education was (were) flagged as DIF using Lord Raju and Mantel-Haenszel's Methods, based on gender?

Hypothesis One

There is no significant difference in the items flagged as Differential Item Functioning (DIF) in the Computer-Based Test (CBT) of Ignatius Ajuru University of Education when analyzed using Lord's chi-square, Raju's area measure, and the Mantel-Haenszel method based on gender.

Table 1:

Comparison of DIF detection of the two methods used based on students' gender.

ITEMS	Mantel-Haenszel's Method (Adjusted p-value)	Decision	Lord Raju's method (p-value)	Decision	Methods that flagged item as DIF
ITEM_1	0.8342	No DIF	0.9480	No DIF	None
ITEM_2	0.9516	No DIF	0.9516	No DIF	None
ITEM_3	0.1847	No DIF	0.4015	No DIF	None
ITEM_4	0.4100	No DIF	0.6176	No DIF	None
ITEM_5	0.3550	No DIF	0.5546	No DIF	None
ITEM_6	0.0373*	No DIF	0.1274	No DIF	None
ITEM_7	0.1964	No DIF	0.4092	No DIF	None
ITEM_8	0.6194	No DIF	0.7758	No DIF	None
ITEM_9	0.1784	No DIF	0.4015	No DIF	None
ITEM_10	0.0252*	No DIF	0.0968	No DIF	None
ITEM_11	0.0000***	DIF	0.0000***	DIF	LR/MH
ITEM_12	0.7524	No DIF	0.8958	No DIF	None
ITEM_13	0.7969	No DIF	0.9267	No DIF	None
ITEM_14	0.1008	No DIF	0.2653	No DIF	None
ITEM_15	0.018*	No DIF	0.0750	No DIF	None
ITEM_16	0.2957	No DIF	0.4987	No DIF	None
ITEM_17	0.2476	No DIF	0.4584	No DIF	None
ITEM_18	0.0000***	DIF	0.0000***	DIF	LR/MH
ITEM_19	0.0048**	No DIF	0.0269 *	No DIF	None
ITEM_20	0.9425	No DIF	0.9516	No DIF	None
ITEM_21	0.3092	No DIF	0.4602	No DIF	None

ITEMS	Mantel-Haenszel's Method (Adjusted p-value)	Decision	Lord Raju's method (p-value)	Decision	Methods that flagged item as DIF
ITEM_22	0.1640	No DIF	0.0099**	DIF	MH
ITEM_23	0.0000***	DIF	0.9516	No DIF	LR
ITEM_24	0.6505	No DIF	0.7758	No DIF	None
ITEM_25	0.8955	No DIF	0.6242	No DIF	None
ITEM_26	0.0382*	No DIF	0.7758	No DIF	None
ITEM_27	0.4199	No DIF	0.6242	No DIF	None
ITEM_28	0.2577	No DIF	0.7758	No DIF	None
ITEM_29	0.0014**	No DIF	0.0099	No DIF	None
ITEM_30	0.9472	No DIF	0.9516	No DIF	None
ITEM_31	0.6361	No DIF	0.7758	No DIF	None
ITEM_32	0.4370	No DIF	0.6242	No DIF	None
ITEM_33	0.6300	No DIF	0.7758	No DIF	None
ITEM_34	0.0009***	DIF	0.0075**	No DIF	LR
ITEM_35	0.6240	No DIF	0.7758	No DIF	None
ITEM_36	0.4930	No DIF	0.6662	No DIF	None
ITEM_37	0.1766	No DIF	0.4015	No DIF	None
ITEM_38	0.3067	No DIF	0.4987	No DIF	None
ITEM_39	0.2469	No DIF	0.4584	No DIF	None
ITEM_40	0.015*	No DIF	0.0745	No DIF	None
ITEM_41	0.0704	No DIF	0.1956	No DIF	None
ITEM_42	0.0000***	DIF	0.0001***	DIF	LR/MH
ITEM_43	0.0045**	No DIF	0.0269*	No DIF	None
ITEM_44	0.9482	No DIF	0.9516	No DIF	None
ITEM_45	0.2344	No DIF	0.4584	No DIF	None
ITEM_46	0.0164*	No DIF	0.0745	No DIF	None
ITEM_47	0.4835	No DIF	0.6662	No DIF	None
ITEM_48	0.0472*	No DIF	0.1409	No DIF	None
ITEM_49	0.9164	No DIF	0.9516	No DIF	None
ITEM_50	0.0479*	No DIF	0.1409	No DIF	None

N/B: Focal Group (Male) = 1, Reference Group (Female = 2), Significance codes, (abs. values): 0 ' ' 0.04 '.' 0.05 '*' 0.1 '**' 0.2 '***'

Table 1 shows the Mantel-Haenszel method with continuity correction using the Multiple comparisons made with Benjamini-Hochberg adjustment of p-values and without item purification (using the difMH function from the differential item functioning R(difR) package of R). From the table it can be observed that five items (11, 18, 23, 34, 42) were flagged as DIF.

Table 1 also revealed the Raju probability (p) value or significance value gotten by the Raju method after item purification (using the latent trait model (latent trait model (ltm)) package of R) was carried out with 25 iterations the items that were flagged as DIF based on gender. From the table it can be observed that based on Lord Raju DIF method using the obtained p-value four items (11, 18, 22, 42) were flagged as DIF.

To test the null hypothesis, the chi-square test of goodness-of-fit was conducted to find out if there lies a significant difference in the number of items flagged to function differential for male and female students using both DIF detection methods.

Table 2:

Chi-Square Goodness-of-Fit Results for MH and LR DIF Detection Methods, based on Gender

Test Statistics		
		DIF Detected
Chi-Square		.111[a]
Df		1
Asymp. Sig.		.739
DIF Methods	MH (Observed)	5
	(Expected)	4.5
	LR (Observed)	5
	(Expected)	4.5

With chi-square value of 1.111, the assumption significance value of 0.739 is greater than 0.05, hence, the already stated null hypothesis is valid. This implies that there is no significant difference in the items flagged as Differential Item Functioning (DIF) in the Computer-Based Test (CBT) of Ignatius Ajuru University of Education when analyzed using Lord Raju and Mantel-Haenszel methods based on gender.

Research Question Two

Which item(s) of the Computer-Based Test (CBT) of Ignatius Ajuru University of Education was (were) flagged as DIF using Lord Raju and Mantel-Haenszel's Methods, based on Mode of Entry?

Hypothesis Two

There is no significant difference in the items flagged as Differential Item Functioning (DIF) in the Computer-Based Test (CBT) of Ignatius Ajuru University of Education when analyzed using Lord Raju and Mantel-Haenszel methods based on Mode of Entry.

Table 2:

Comparison of DIF detection of the two methods used based on students' Mode of entry

ITEMS	Mantel-Haenszel's Method (Adjusted p-value)	Decision	Lord Raju's method (p-value)	Decision	Methods that flagged item as DIF
ITEM_1	0.3221	No DIF	0.2433	No DIF	None
ITEM_2	0.0546	No DIF	0.0450	No DIF	None
ITEM_3	0.9703	No DIF	0.9703	No DIF	None
ITEM_4	0.0544	No DIF	0.0544	No DIF	None
ITEM_5	0.0078**	DIF	0.0454*	DIF	LR/MH
ITEM_6	0.3400	No DIF	0.2720	No DIF	None
ITEM_7	0.0213	No DIF	0..0113	No DIF	None
ITEM_8	0.0455*	No DIF	0.0524*	No DIF	None
ITEM_9	0.3400	No DIF	0.2702	No DIF	None
ITEM_10	0.0127	No DIF	0.0061	No DIF	None
ITEM_11	0.6595	No DIF	0.5540	No DIF	Both
ITEM_12	0.0272	No DIF	0.0152*	No DIF	None
ITEM_13	0.1049	No DIF	0.0671	No DIF	None
ITEM_14	0.0046	No DIF	0.0619	No DIF	None
ITEM_15	0.8525	No DIF	0.7843	No DIF	None
ITEM_16	0.2592	No DIF	0.1814	No DIF	None
ITEM_17	0.0000***	DIF	0.0301***	DIF	LR/MH
ITEM_18	0.0602	No DIF	0.0456	No DIF	Both
ITEM_19	0.1581	No DIF	0.1075	No DIF	None
ITEM_20	0.0213*	DIF	0.0115*	No DIF	LR/MH
ITEM_21	0.4704	No DIF	0.0325**	DIF	MH
ITEM_22	0.0643	No DIF	0.6102	No DIF	One

ITEMS	Mantel-Haenszel's Method (Adjusted p-value)	Decision	Lord Raju's method (p-value)	Decision	Methods that flagged item as DIF
ITEM_23	0.0043	No DIF	0.0316	No DIF	One
ITEM_24	0.0102	No DIF	0.2160	No DIF	None
ITEM_25	0.3075	DIF	0.5214	No DIF	LR
ITEM_26	0.8299	No DIF	0.7469	No DIF	None
ITEM_27	0.0030***	DIF	0.0000***	DIF	LR/MH
ITEM_28	0.0514*	No DIF	0.0445**	No DIF	None
ITEM_29	0.0000***	DIF	0.0000***	DIF	LR/MH
ITEM_30	0.0544	No DIF	0.0703	No DIF	None
ITEM_31	0.0920	No DIF	0.0714	No DIF	None
ITEM_32	0.0471*	No DIF	0.0292*	No DIF	None
ITEM_33	0.0532	No DIF	0.0721	No DIF	None
ITEM_34	0.0000***	DIF	0.0000***	DIF	LR/MH
ITEM_35	0.3221	No DIF	0.2448	No DIF	None
ITEM_36	0.0029**	No DIF	0.0010***	No DIF	None
ITEM_37	0.0000***	No DIF	0.0000***	No DIF	None
ITEM_38	0.0000***	No DIF	0.0000***	No DIF	None
ITEM_39	0.6671	No DIF	0.5737	No DIF	None
ITEM_40	0.0000***	No DIF	0.0000***	No DIF	None
ITEM_41	0.3782	No DIF	0.3101	No DIF	None
ITEM_42	0.9172	No DIF	0.8622	No DIF	Both
ITEM_43	0.0142*	No DIF	0.0071**	No DIF	None
ITEM_44	0.0360*	No DIF	0.0209*	No DIF	None
ITEM_45	0.1348	No DIF	0.0892	No DIF	None
ITEM_46	0.9204	No DIF	0.9010	No DIF	None
ITEM_47	0.0001***	DIF	0.0000***	DIF	LR/MH
ITEM_48	0.0316**	DIF	0.0702*	No DIF	LR
ITEM_49	0.0005***	DIF	0.0001***	DIF	LR/MH
ITEM_50	0.0506	No DIF	0.0612	No DIF	None

N/B: Focal Group (JAMB) = 1, Reference Group (Direct = 2), Significance codes, (abs. values):

0 ' ' 0.04 '.' 0.05 '*' 0.1 '**' 0.2 '***'

Table 2 shows the Mantel-Haenszel method with continuity correction using the Multiple comparisons made with Benjamini-Hochberg adjustment of p-values and without item purification (using the difMH function from the differential item functioning R(difR) package of R). From the table it can be observed that ten items (5, 17, 20, 25, 27, 29, 34, 47, 48, 49) were flagged as DIF.

Table 2 also reveals the Raju probability (p) value or significance value gotten by the Raju method after item purification (using the latent trait model (latent trait model (ltm)) package of R) was carried out with 25 iterations the items that were flagged as DIF based on mode of entry. From the table it can be observed that based on Lord Raju DIF method using the obtained p-value eight items (5, 17, 21, 27, 29, 34, 47, 49) were flagged as DIF.

To test the null hypothesis, the chi-square test of goodness-of-fit was conducted to find out if there lies a significant difference in the number of items flagged to function differential for JAMB and direct entry students using both DIF detection methods. The result below was generated.

Table 3

Chi-Square Goodness-of-Fit Results for MH and LR DIF Detection Methods, Based on Mode of Entry

Test Statistics

		DIF Detected
Chi-Square		.222[a]
Df		1
Asymp. Sig.		.637
DIF Methods	MH (Observed)	10
	(Expected)	9
	LR (Observed)	8
	(Expected)	9

With chi-square value of 0.222, the assumption significance value of 0.637 is greater than 0.05, hence, the already stated null hypothesis is valid. This implies that there is no significant difference in the items flagged as Differential Item Functioning (DIF) in the Computer-Based Test (CBT) of Ignatius Ajuru University of Education when analyzed using Lord Raju and Mantel-Haenszel methods based on mode of entry.

Discussion of Findings

Gender-based DIF in Computer-Based Test (CBT) of Ignatius Ajuru University of Education

The result indicated that both DIF methods identified 3 items (11, 18, 42) as DIF and varied in detection of three items (22, 23, 34). The collective and consistent detection of items 11, 18, 42 as DIF by the two methods increases confidence in this result. However, the isolated detection of items 22, 23 and 34 as DIF indicate there may be a need to further investigate the underlying reasons for the differences. This may also mean utilizing additional DIF detection method(s). The findings of this study align with those of Yao and Chen (2020) and Annan-Brew (2020), who examined gender-related Differential Item Functioning (DIF) in an ESL test using the and West African Senior School Certificate Examination (WASSCE) core subjects (English Language, Mathematics, Integrated Science, and Social Studies) in Southern Ghana from 2012-2016, respectively. Their study found that only three items exhibited moderate DIF, indicating that most test items did not significantly favor one gender over another. Similarly, the present study on the Computer-Based Test (CBT) of Ignatius Ajuru University of Education revealed no significant difference in the items flagged as DIF across Lord Raju, and Mantel-Haenszel methods. This suggests that the assessment items used in the CBT were well-constructed and free from systematic gender bias. Despite this agreement, Yao and Chen (2020) still identified a few test items with moderate DIF, while the present study found no significant DIF in any item. This difference may be due to variations in test content and subject matter, as language assessments often involve cultural or contextual elements that could introduce subtle biases. Additionally, differences in sample characteristics and statistical thresholds used for defining DIF may have contributed to this variation.

Mode of entry-based DIF in Computer-Based Test (CBT) of Ignatius Ajuru University of Education

The result indicated that both DIF methods identified seven items (5, 17, 27, 29, 34, 49 and 49) as DIF and varied in detection of three items (21, 25, 48). The collective and consistent detection of items 11, 18, 42 as DIF by the two methods increases confidence in this result. However, the isolated detection of items 5, 17, 27, 29, 34, 49 and 49 as DIF indicate there may be a need to further investigate the underlying reasons for the differences. This may also mean utilizing additional DIF detection method(s).

The findings of this study align with Anwuri (2022), who found no evidence of DIF in Computer-Based Examinations at Ignatius Ajuru University of Education based on mode of entry. Similarly, this study shows no significant difference in the items flagged as DIF using Lord Raju, and Mantel-Haenszel methods based on gender. This agreement suggests that the test items may have been well-constructed and fairly developed, ensuring that they do not favour any subgroup. Standardized test development procedures and rigorous item validation processes could explain the consistency in findings.

Despite this agreement, it disagrees with Anwuri (2022) findings, which although flagged 10 items and 8 items for MH and LR methods respectively. Although it showed they were not significant, it still indicates variation in findings. This can be attributed to different type of test. The present study considered English Language Test, while the previous study considered Peace and Conflict Resolution.

Conclusion

The study Explored Item Analysis Methods for Enhanced Digital Assessment in the Tertiary Education Sector. The comparison of DIF detection methods reveals a notable consistency in the identification of Differential Item Functioning (DIF) across different approaches, while also highlighting some discrepancies that warrant further investigation. For Research Question One, the Mantel-Haenszel method with continuity correction and the Benjamini-Hochberg adjustment of p-values identified five items (11, 18, 23, 34, 42) as exhibiting

DIF, whereas the Lord Raju method flagged four items (11, 18, 22, 42) as DIF. Both methods consistently identified three items (11, 18, 42) as exhibiting DIF, suggesting a robust finding. However, they differed in the detection of items 22, 23, and 34, indicating a need for further analysis to understand these discrepancies. For Research Question Two, the Mantel-Haenszel method with the Benjamini-Hochberg adjustment flagged ten items (5, 17, 20, 25, 27, 29, 34, 47, 48, 49) as exhibiting DIF, while the Lord Raju method identified eight items (5, 17, 21, 27, 29, 34, 47, 49) as DIF. Both methods consistently identified seven items (5, 17, 27, 29, 34, 47, 49) as exhibiting DIF, reinforcing the reliability of these findings. However, they varied in the detection of items 21, 25, and 48, suggesting areas where further investigation is needed to understand the underlying reasons for these differences.

Overall, the consistency in the identification of several items across both methods strengthens the confidence in these findings. At the same time, the discrepancies observed indicate that a comprehensive approach, potentially incorporating multiple DIF detection methods, may provide a more complete understanding of item functioning and ensure robust conclusions.

Recommendations

Based on the findings, the following recommendations are made:

1. The first result proves that CBTs are likely to have DIF items, if there is no proper check of its psychometric properties as well as DIF detection. The university and other institutions of higher learning should regularly subject their crafted test items to a psychometrician to test for difficulty, discrimination level as well as DIF before handing them over to the CBT administrator. Additionally, periodic student feedback sessions could help identify any hidden concerns that the CBT miss.
2. The second result proves that CBTs can contain bias items based on mode of entry. Based on the result, it is recommended that higher institutions should provide preparatory modules for students admitted through direct entry and JAMB to ensure that all students, regardless of entry mode, have an equal shot at succeeding in the CBT test.

References

Annan-Brew, R, (2020) *Differential item functioning of West African Senior School Certificate Examination in core subjects in Southern Ghana,* Published Doctoral Thesis, University of Cape Coast, Sam Jonah Library.

Abbott, W.T, (2023), DIF Detection Sensitivity of Lord's, Chi-Square, Raju's Area, Logistic Regression, Mantel-Haenszel, Standardization, And Transformed Item Difficulties Methods, In Comparison, Using R. *Journal of Language Testing & Assessment. 3*(1)*,* 5-19. *EPRA International Journal of Multidisciplinary Research (IJMR), 7(*7), 1-12.

Bassman, M, (2023), A Comparison of the efficacies of differential item functioning detection methods. *Journal of Language Testing & Assessment. 3*(1)*,* 5-19. *International Journal of Assessment Tools in Education, 10(*1), 145-159.

Anwuri, O, (2022) *Detection of item bias in computer-based Test of Ignatius Ajuru University of Education,* Unpublished Master's Thesis, Ignatius Ajuru University of Education.

Brown, G. T., & Kennedy, K. J. (2019). *Computer-based testing: Practices, advantages, challenges, and practical guidelines. Cham*: Springer.

Don Y, & Kayla, C. (2020). Gender-related differential item functioning analysis on an ESL Test. *Journal of Language Testing & Assessment. 3*(1)*,* 5-19.

Hartoyo, V., Putra, I. E., & Syafryadin, S. (2020). Utilization of online quiz as a learning media to improve student learning outcomes. *Journal of Physics: Conference Series, 1467*(1), 012046.