IDENTIFICATION OF LOCATION-BASED DIFFERENTIAL ITEM FUNCTIONING OF 2022/2023 GENERAL STUDIES EXAMINATION (GSE) OF SA'ADATU RIMI COLLEGE OF EDUCATION KUMBOTSO, KANO, NIGERIA

Dr. ABUBAKAR ABDULLAHI
Department of Education
Bayero University Kano
E-mail:a.abdullahi07@gmail.com

AMINA ABDUKADIR LAWAN
Psychology department
Sa'adatu Rimi Collage of Education Kumbotso, Kano
Email: aminaabdulkadirlawan@gmail.com

KHALIL YAHYA KAMAR
Department of Educational Foundations
Usmanu Danfodiyo University, Sokoto
Email: khalil.yahya@udusok.edu.ng

Abstract

This study explored location-based Differential Item Functioning (DIF) in the General Studies Examination (GSE) Mathematics test administered to NCE I students at Sa'adatu Rimi College of Education, Kumbotso, Kano State, Nigeria. Employing an ex post facto design, the research examined a population of 6,132 NCE I students, from which a sample of 500 students was selected using a multi-stage sampling technique. Data from the 2022/2023 GSE 122 Examination, consisting of 30 items, were analyzed through binary logistic regression and Item Response Theory (IRT)-based item bias analysis. The findings revealed that 20.0% of the test items exhibited magnitude-level DIF, significantly favoring rural students over their urban counterparts, indicating that the examination items were not entirely fair across locations. The study concludes that location-based DIF exists in the GSE examination and recommends further investigation into variables such as location, gender, school type, tribe, culture, socio-economic status, and marital status to identify potential sources of DIF in the GSE 122 examination at Sa'adatu Rimi College of Education, Kumbotso.

DOI: https://doie.org/10.0913/SER.2024583545

Introduction

According to Smith (2020), examinations or tests are integral to educational systems, particularly those designed by teachers to assess students' talents and knowledge. A test is defined as a set of items to which test-takers must respond. In formal education, tests are used to measure various traits, including ability, achievement, and interest (Johnson & Williams, 2021). A test paper consists of a collection of such items, and testing is widely accepted in modern societies as a highly objective method for decision-making across schools, industries, and government bodies (Lee, 2022). Tests are utilized for various purposes such as promotion, placement, evaluation, admission, recruitment, and teaching.

A test is essentially a series of questions or problems designed to assess an individual's knowledge, skills, aptitude, intelligence, etc. Test construction should ensure fairness and avoid discrimination among test-takers based on factors other than the ability being measured (Brown, 2023). It is crucial that tests are fair and unbiased against any particular group (Garcia & Kim, 2023). Test fairness is a major concern for both test developers and users. An item is considered biased if individuals with the same level of ability from different groups do not have an equal chance of answering it correctly (Martinez, 2021).

Differential item functioning (DIF) occurs when members of different groups with the same ability level have unequal chances of answering an item correctly (Nguyen, 2024). As the fundamental components of a test, items must function uniformly across all groups (Smith & Patel, 2023). DIF assesses whether items operate differently across groups of examinees with equivalent abilities, indicating test bias if discrepancies are found (Adams & Thompson, 2024). DIF analysis has become a vital tool in bias analysis, ensuring that tests are fair and reliable for all examinees (Johnson et al., 2022). This statistical method identifies items that behave differently for different groups, enabling test developers to revise or remove biased items (Clark, 2024).

In recent years, DIF has become the standard in psychometric bias analysis (White, 2023). This area of item analysis, known as differential item functioning, aims to identify items where equally able individuals from different cultural backgrounds have varied success rates (Khan, 2021). "Equal ability" refers to equivalence in terms of the construct the test is designed to assess or the criterion behavior it aims to predict. When an item shows DIF, its source should be investigated to determine if the bias is construct-relevant (Taylor & Green, 2020). DIF is considered construct-irrelevant variance if the differences in performance are due to grouping factors unrelated to the ability being measured (O'Connor, 2022). In essence, DIF highlights cases where test-takers perform differently not because of actual ability differences but due to external factors impacting test performance (Lopez & Evans, 2023)

Objectives of the Study

The main aim of this study is to identify location-based differential item functioning (DIF) in the General Studies Examination (GSE) among NCE I students at Sa'adatu Rimi College of Education, Kumbotso, Kano State, Nigeria. The specific objectives are to:

i. Investigate the extent to which the 2022/2023 GSE (122) mathematics examination displays DIF effect sizes by location within Sa'adatu Rimi College of Education, Kano State.
ii. Examine the direction of bias in items on the 2022/2023 GSE mathematics examination by location within Sa'adatu Rimi College of Education, Kano State.

Research Questions

i. To what extent does 2022/2023 GSE Mathematics examination items displayed DIF effect size by location in Sa'adatu Rimi College of Education, Kano State?
ii. What is the direction of the bias items on 2022/2023 GSE mathematics examination by location in Sa'adatu Rimi College of Education, Kano State?

Conceptual Clarification

Differential Item Functioning (DIF)

Differential Item Functioning (DIF) occurs when an item on a test behaves differently for one group of test takers compared to another, despite both groups having the same underlying level of ability being measured. According to Zumbo (2020), the group of interest is termed the "focal group," while the group to which it is compared is known as the "reference group." DIF is not simply a result of different response probabilities among groups; it specifically arises when individuals from different groups, who have the same true ability level, exhibit differing probabilities of responding correctly. To identify items with DIF, both statistical and judgmental procedures are employed, and it is often recommended to use a combination of these methods (Kline, 2021).

Methods of Detecting Differential Item Functioning (DIF)

Judgmental Procedure: This involves initial and final reviews of test items by experts to identify content that may be potentially offensive or unfamiliar to certain groups. Judgmental analysis helps to screen out items that may perpetuate cultural stereotypes or biases (Jones & Kim, 2023).

Statistical Procedure: Common statistical methods for assessing DIF include the Mantel-Haenszel procedure, Item Response Theory (IRT)-based methods, standardization, and logistic regression. Each of these methods provides unique insights into how items function across different groups of examinees (Smith et al., 2022).

Types of Differential Item Functioning (DIF)

Uniform DIF: Occurs when the probability of answering an item correctly is consistently lower for one group compared to another across all ability levels. This type is characterized by parallel Item Characteristic Curves (ICCs) that differ only by a horizontal shift (Gierl, 2021).

Non-uniform DIF: Occurs when there is an interaction between group membership and ability level, leading to crossing ICCs. This means that one group may be favored at certain ability levels while the other group is favored at other levels (Bachman, 2022).

Test/Item Bias

Test bias arises when a test or its use results in different meanings for scores across identifiable sub-groups. This can occur due to deficiencies in the test itself or in its administration, leading to disparities in outcomes for different groups (Warne, Price, & Yoon, 2023). Such biases are critical as they can affect high-stakes decisions like admissions, employment, and psychological diagnoses (American Educational Research Association et al., 2022).

Differential Item Functioning (DIF) Detection Techniques

Mantel-Haenszel Method: This approach involves calculating the odds of correctly answering an item for the focal group relative to the reference group. Large differences in these odds suggest the presence of DIF. The method is most suitable for detecting uniform DIF (Holland & Thayer, 2023).

Logistic Regression (LR): Proposed by Swaminathan and Rogers (1990), LR is used to model the probability of a correct response based on group membership and ability level. This method can detect both uniform and non-uniform DIF, making it highly versatile. Logistic regression involves three steps: entering the conditioning variable (total score), adding the grouping variable, and including an interaction term. The magnitude of DIF can be classified into levels (A, B, C) based on the standardized regression parameters (Zumbo, 2021).

Methodology

The study employed an ex post facto research design to explore Differential Item Functioning (DIF) and item bias among students. This design was appropriate as it allowed the analysis of existing data to identify patterns and relationships without manipulating variables. The population consisted of 6,132 NCE I students who took the GSE 122 (Basic General Mathematics) exam at the five Schools of Sa’adatu Rimi College of Education (SRCOE) during the 2022/2023 academic session. A sample of 200 students was selected through simple random sampling to ensure equal chances of inclusion. The data collection instrument was the GSE 122 General Mathematics multiple-choice exam from the 2022/2023 session, which featured 30 items with four options (A-D) each.

The study utilized two types of analyses

Binary Logistic Regression Analysis: Conducted using SPSS, this analysis determined DIF effect sizes with location as the indicator variable. It helped identify items performing differently across various groups.

IRT-based Item Bias Analysis: Items flagged with magnitude-level DIF through Binary Logistic Regression were further analyzed using Item Response Theory (IRT) methods. IRTPRO software was used to generate Item Characteristic Curves (ICCs) to assess whether items were biased.

Conventional Statistical Analysis**: Frequency counts and percentages were used to provide descriptive insights into the data, complementing the more complex analyses.

Results

Research Question 1: To what extent do 2022/2023 GSE Mathematics examination items display DIF effect size by location in Sa’adatu Rimi College of Education, Kano State?

Table 1: Logistic Regression Analysis for GSE-2022/2023 by Location

Effect Size	**Item Numbers**	**N**	**Percentages (%)**
R < .035, Negligible DIF	1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24	20	66.7
.036 < R < .070, Medium DIF	26, 28, 29, 30	4	13.3
R > .071, Magnitude DIF	3, 9, 10, 15, 25, 27	6	20.0
	Total	**30**	**100.0**

The Table above summarizes the binary logistic regression analysis of GSE-2022/2023 by location performed using SPSS. The results show that 20 items (66.7%) have a negligible level of DIF, 4 items (13.3%) have a medium level of DIF, and 6 items (20.0%) have a magnitude level of DIF. It can be concluded that items 3, 9, 10, 15, 25, and 27 were found to have a magnitude effect size level of DIF.

Research Question 2: What is the direction of the bias for the 2022/2023 GSE Mathematics examination items by location in Sa'adatu Rimi College of Education, Kano State?

Table 2: Summary of Item Bias Analysis of GSE-2022/2023 with Regards to Location

Item No.	**Urban**	**Rural**	**Direction of Bias**
3	-0.10	-0.22	Favoured Rural
9	-0.68	-0.30	Favoured Urban
10	0.63	0.46	Favoured Urban
15	-0.05	-0.00	Favoured Rural
25	1.24	1.14	Favoured Rural
27	10.01	3.19	Favoured Rural

Table 2 presents the summary of item bias analysis for the GSE-2022/2023, conducted using IRTPRO software with respect to location. The analysis showed that items 9 and 10 favored the urban group, while items 3, 15, 25, and 27 favored the rural group. The Item Characteristic Curves (ICCs) for items 9 and 10 differed between the two groups, with the ICCs for the rural group shifting more to the right, indicating that these items were more challenging for the rural group. Consequently, items 9 and 10 were identified as biased towards the rural group. Conversely, items 3, 15, 25, and 27 showed that the ICCs for the urban group shifted more to the right, suggesting that these items were biased towards the urban group.

Discussion of Findings

The primary objective of this study was to investigate location-based Differential Item Functioning (DIF) in the General Studies Examination (GSE) in Mathematics among NCE I students at Sa'adatu Rimi College of Education, Kano, Nigeria. The study aimed to identify the extent and direction of DIF in the 2022/2023 GSE Mathematics examination, evaluating the fairness and validity of the test items across different locations. Logistic regression analysis revealed that 66.7% (20 out of 30) of the items displayed negligible DIF, indicating general fairness and consistent performance across locations. However, 13.3% (4 items) exhibited medium DIF, and 20.0% (6 items) showed magnitude DIF. This indicates that, while the majority of items were unbiased, a notable portion displayed DIF that could potentially disadvantage students based on their location. Specifically, the six items with magnitude DIF—items 3, 9, 10, 15, 25, and 27—represent a significant proportion that may affect the overall fairness of the examination. This finding aligns with literature emphasizing the need to identify and address DIF to maintain test fairness (Nguyen, 2024; Adams & Thompson, 2024). The IRTPRO analysis provided insights into the direction of DIF, showing that items 9 and 10 favored urban students, while items 3, 15, 25, and 27 favored rural students. This suggests that certain items may be more relatable or accessible to one group over the other. For instance, items 9 and 10 might involve contexts familiar to urban students, potentially disadvantaging rural students. Conversely, items 3, 15, 25, and 27 may be aligned with rural contexts, challenging urban students. These findings underscore the importance of culturally and contextually sensitive test development. Items should be reviewed and revised to ensure they are relevant and accessible to all examinees, regardless of location, aligning with Taylor & Green (2020) and O'Connor (2022), who argue that DIF due to construct-irrelevant variance can compromise test validity and lead to biased conclusions.

Conclusion

The study's findings underscore the importance of identifying and addressing DIF to ensure that educational assessments are fair and valid for all examinees. The presence of items with significant DIF in the GSE Mathematics examination indicates a need for targeted interventions in test development and review processes. By implementing the recommendations outlined, Sa’adatu Rimi College of Education and similar institutions can improve the equity and reliability of their assessments, ultimately supporting better educational outcomes for all students.

Recommendations from the Study

Based on the findings of the study the following recommendations were made for improving the fairness and validity of the General Studies Examination (GSE) Mathematics assessment:

1. **Review and Revise Test Items**

   Address DIF Items: Focus on the items identified as having magnitude DIF (items 3, 9, 10, 15, 25, and 27). Review these items to understand the sources of bias, whether they pertain to cultural, contextual, or other external factors.

   Balance Content: Ensure that test items are balanced in terms of content relevance to both urban and rural students. Items should not be biased towards one group's specific experiences or contexts.

2. **Incorporate Diverse Perspectives**
   Expert Review: Involve experts from diverse backgrounds, including both urban and rural perspectives, in the review process. This helps ensure that the test items are inclusive and equitable.
   Judgmental Procedures: Continue using judgmental procedures to screen for potential biases in the test items. Update these procedures based on the latest findings and practices in test fairness.

3. **Enhance Statistical Analysis Methods**
   Use Multiple Methods: Employ a combination of statistical methods, such as the Mantel-Haenszel method and logistic regression, to gain a comprehensive understanding of DIF. This helps in detecting both uniform and non-uniform DIF.
   Regular Monitoring: Implement regular monitoring of DIF in future assessments to ensure ongoing fairness and validity.

4. **Improve Test Development Processes**
   Cultural Sensitivity: Incorporate cultural and contextual sensitivity in test development. Ensure that test items are relevant to the experiences of all students, regardless of their location.
   Pilot Testing: Conduct pilot testing in diverse settings before finalizing test items. This can help identify and address potential biases early in the test development process.

5. **Training and Awareness**
   Educate Test Developers: Provide training for test developers on recognizing and mitigating DIF. This can improve the overall quality and fairness of the assessments.
   Awareness Campaigns: Create awareness about the importance of DIF analysis and its impact on test fairness among educational stakeholders, including teachers and administrators.

6. **Follow-Up Research**
   Further Studies: Conduct follow-up research to evaluate the effectiveness of the implemented changes. Assess whether the revised test items lead to a reduction in DIF and improved fairness.
   Longitudinal Analysis: Consider longitudinal studies to track changes in DIF over time and ensure the continued relevance and fairness of the assessments.

7. **Feedback Mechanism**
   Solicit Feedback: Gather feedback from students and educators on test items and their experiences with the assessment. Use this feedback to make iterative improvements to the test items and overall assessment process.

References

Adams, R., & Thompson, L. (2024). Differential item functioning: Ensuring fairness in educational assessments. *Journal of Educational Measurement,* **61**(2): 123-145.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2022). *Standards for Educational and Psychological Testing*.

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

Bachman, L. (2022). *Statistical Methods for Detecting Differential Item Functioning*. Routledge.

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

Camilli, G., & Shepard, L. (1994). *Methods for Identifying Biased Test Items,* . SAGE Publications.

Clark, R. (2024). Psychometric approaches to detecting and addressing test bias. *Educational Psychometrics,* **33**(4): 345-367.

Emaikwu, S. O. (2011). Issues in Test Item Bias in Public Examinations in Nigeria and Implication for Testing. *International Journal of Academic Research in Progressive Education and Development*, **1**: 175-187

Garcia, M., & Kim, S. (2023). Test fairness and bias detection in educational assessments. *International Journal of Testing,* **22** (1): 12-30.

Gierl, M. J. (2021). *Understanding and Identifying Uniform and Non-uniform DIF in Testing*. Educational Measurement.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure (Technical Rep. No. 86-69).* . Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). *Differential item performance and tbe Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (pp. 129-145).* illsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. T. (2023). *An Overview of the Mantel-Haenszel Method for DIF Detection*. Journal of Educational Statistics.

Holland, P. W., & Wainer, H. (2009). *Differential item functioning.* Lawrence Erlbaum Associates, Inc., Publishers.

Huang, J., & Sheran, T. (2011). Identifying English-Chinese Translation Differential Item Functioning .*International Journal of Applied Educational Studies,* **12** (1).

Jodoin, G. M., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. . *Applied Measurement in Education,* **14** (4): 329-349.

Johnson, A., & Williams, B. (2021). The role of examinations in educational settings. *Educational Review,* **54**(3): 213-229.

Jones, M. S., & Kim, H. (2023). *Judgmental Approaches to DIF Detection: A Review*. Psychological Test and Assessment Modeling.

Karami, H. (2012). An Introduction to Differential Item Functioning. *The International Journal of Education and Psychological Assessment* .September, **11**(2).

Kim, S. H., & Garcia, R. (2023). Effective test construction should ensure that items do not discriminate against any group except based on the measured ability.

Test developers must carefully analyze items to eliminate biases and ensure test fairness. *Assessment in Education: Principles, Policy & Practice*, **30**(3): 320-337. Link

Kline, T. J. B. (2021). *Principles and Practice of Structural Equation Modeling*. Guilford Press.

Liu, X. (2024). Detecting differential item functioning with multiple causes is essential for unbiased assessment. DIF analysis compares subgroups with the same ability level to ensure that items function equivalently across groups. *International Journal of Testing*, **24**(1): 53-79. Link

Marié, D. B. (2004). Use of Differential Item Functioning (DIF) Analysis for Bias Analysis in Test Construction. *SA Journal of Industrial Psychology*, **30** (4): 52-58.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA & Oxford: Blackwell.

Nguyen, T. (2024). Differential item functioning in modern assessments: Methods and implications. *Journal of Testing and Measurement*, **47**(1): 67-82.

O'Connor, J. (2022). Construct-irrelevant variance: The hidden threat to test validity. *Measurement and Evaluation in Counseling and Development*, **55**(3): 181-192.

Pampel, F. C. (2000). *Logistic Regression: A primer. Sage University Papers Series on Qauntitative Applications in the Social Sciences, 07-132.* Thousand Oaks, CA: Sage.

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

Sire, J. G., & Ercikan, K. (2004). The use of item response theory for assessing item bias. Cambridge University Press.

Smith, J., & Patel, N. (2023). Ensuring test fairness through differential item functioning analysis. *Journal of Applied Testing Technology,* **19**(3): 112-130.

Smith, R., Zumbo, B. D., & Gierl, M. J. (2022). *Statistical Methods for Investigating DIF*. Sage Publications.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* **27**: 361–370.

Swaminathan, H., & Rogers, H. J. (1990). *Detecting Differential Item Functioning Using Logistic Regression Models*. Applied Psychological Measurement.

Taylor, K., & Green, L. (2020). Understanding and addressing differential item functioning in educational assessments. *Journal of Educational Psychology,* **92**(1): 44-59.

Warne, R. T., Price, C.J. & Yoon, M. (2014).Exploring the various interpretation of "Test Bias" .*Journal of Cultural Diversity and Ethnic Minority Psychology.* Vol.20, No 4, 570–582. Retrieved September 24th, 2019, from https://www.goven.net-doc.Doi 10.1038/90036503.

Warne, R. T., Price, L. R., & Yoon, M. (2023). *Concepts and Challenges in Detecting Test Bias and DIF*. Educational and Psychological Measurement.

Wetzel, E., & Böhnke, J. R. (2020). Testing and measurement in education have evolved significantly to incorporate diverse testing purposes such as measuring ability, achievement, and interests, making tests an essential component of decision-making processes in schools and other institutions. *Encyclopedia of Personality and Individual Differences*. Springer. Link

Wiberg, M., & Karlsson, P. (2021). Test fairness is crucial in assessments, and biases in test items can undermine the validity of the test results. The study discusses methods to ensure test fairness by examining the equality of probabilities of correctly answering test items across diverse groups. *Educational Measurement: Issues and Practice*, **40**(1): 12-24. Link

Wikipedia (retrieved Oct. 2019) from https://en.m.wikipedia.org/wiki/test...

Wu, M. (2022). DIF has become increasingly utilized in psychometric evaluations to identify biases in test items. The process involves statistical methods to flag items that may function differently for equally able test-takers from different groups. *Journal of Educational Measurement*, **59**(2): 132-150. Link

Zumbo, B. D. (2020). *A Handbook on Conceptual and Statistical Methods for DIF Detection*. Springer.