



**AN ANALYSIS OF GENDER AND LOCATION BASED DIFFERENTIAL ITEM
FUNCTIONING (DIF) IN AI-GENERATED MATHEMATICS QUESTIONS IN
CALABAR MUNICIPALITY LOCAL GOVERNMENT
AREA OF CROSS RIVER STATE, NIGERIA”**

**¹Prof. I. E. Idaka, ¹Anele, Ezebunwo, ¹Ekereke, Aidam Benjamin, ¹Beshel, Ignatius
Akwagiobe, ¹Emmanuel King Ekong & ²Ogba, Uwaoma Flora Ph.D**

idaka.idakaegbe@gmail.com, aneleezebunwo2022@gmail.com, aidamekereke@gmail.com,
emmanuelkingekong@gmail.com, beshel-ignatius@gmail.com & ogbosouwaoma@yahoo.com

¹Department of Health Information Management, College of Health Technology, Calabar,

²Department of Educational Psychology, Faculty of Educational Foundation Studies,
University of Calabar, Calabar.

Abstract

The study examined gender and location-based differential item functioning (DIF) in AI-generated mathematics multiple-choice examination questions in the Calabar Municipality Local Government Area, Cross River State, Nigeria. Two objectives were developed that direct the study in accordance with the research questions. The study used a survey research design using a stratified random sampling approach to collect a sample of 400 secondary school students from the Calabar Municipality Local Government Area. Data for analysis was collected using AI-generated mathematics multiple-choice questions. Kuder-Richardson 20 was used to establish the reliability of the study and yielded a coefficient of .81. The research questions were subsequently tested using Bilog-MG 3. The findings revealed that there were ability disparities between male and female counterparts in AI-generated mathematics multiple-choice tests, and there were ability disparities between urban and rural counterparts in AI-generated mathematics multiple-choice questions. It was determined that there was an incidence of gender and school location “Differential Item Functioning (DIF)” in an AI-generated mathematics multiple choice test questions in Calabar Municipality Local Government. Based on the research findings, it was suggested, among other things, that schools and governments embrace AI and validate AI-generated test questions for classroom usage.

Keywords: Gender, School Location, Differential Item Functioning (DIF), Artificial Intelligence (AI).

Introduction

The introduction of artificial intelligence (AI) has altered several areas of endeavours, including the educational sector. AI-generated evaluations provide personalised and adaptable learning experiences. However, the efficacy of these assessments should be scrutinised, especially aimed at the possibility of “differential item functioning (DIF)”. DIF arises when individuals from

various groups (e.g., sexes) with comparable skill levels have differing chances of properly answering a test item (Bao, Dayton, & Hendrickson, 2019). According to Baker and Hawn (2022), this phenomenon can compromise the fairness and validity of educational evaluations, resulting in biased judgments regarding students' abilities.

At the setting of secondary education in Cross River State in Nigeria, the consequences of DIF are particularly crucial. Gender inequalities in learning outcomes have been widely reported worldwide, with various cultural, social, and economic factors resulting in these variations (UNESCO, 2020). In mathematics, traditionally associated with male dominance, the performance differences between male and female students have raised questions among educators and policymakers.

The worth of analysing DIF in AI-generated objects cannot be overestimated. As AI systems grow increasingly prominent in the educational sector, assessing whether they reinforce current prejudices or contribute to new kinds of unfairness is critical. Previous research has shown that conventional examinations might be gender biased, preferring male students (Miller & Stoeckel, 2019). AI, although providing creative solutions, may unintentionally repeat these biases if not carefully constructed and assessed. "The differential Item Functioning (DIF) refers to theoretical or empirical evidence used to disprove or support the presence of prejudice". "Differential Item Functioning (DIF) is a violation of the invariance assumption in Item Response Theory (IRT) models that occurs when the chance of endorsing an item for test takers with equal skill levels vary between courses" (Battuz, 2017). After groups have been balanced for ability, DIF is defined as the disparities in question functionality. Subsequently controlling for the latent capacity that the item is designed to compute, an unexplained discrepancy occurs between the two groups (Bauer, 2023).

Additionally, Akorede (2021) "revealed that Junior Secondary School Certificate Examination exhibited differential item functioning (DIF) in mathematics based on sex, school location, and school ownership". Their study revealed 10 items that functioned differently between male and female students, 12 that functioned differently between urban and rural schools, and six that functioned differently between private and public-owned schools. Differential Item Functioning (DIF) in educational assessments can significantly negatively affect students' psychological well-being. Increased stress, damaged self-esteem, an inequitable learning environment, reduced motivation, and long-term mental health consequences are among the detrimental outcomes associated with DIF.

Many specialists have discovered differential item functioning during high-stakes assessments, such as the West African Examination Council's English examination. Eteng-Uket's (2021) "study found differential item functioning in the West African Senior School Certificate English language assessment in South-South Nigeria". Eteng-Uket (2021) discovered that 13 questions functioned differently for male and female students, while 23 questions did for high and low socioeconomic groups. The purpose of bias investigations was to determine whether the reasons for group differences are real (that is, an accurate reflection of true performance differences) or not (caused by the measuring process itself). Favoritism may occur at the test level (test bias) and the test level (test bias). Test bias occurs when a test favours one set of candidates over another, whereas bias in tests happens when one group of testees outperforms the other. Favoritism must exist in the examination environment, either against the target population (minority) or in favour of the reference group (majority) (Cesario, 2022).

Test bias occurs when one group finds an item challenging than another due to unrelated information on the exam. Item bias arises when one set of candidates is less inclined to accurately

answer an item than another due to test features or circumstances that are unrelated to the test goal (Effiom, 2021). A test can be considered biased when it favours a particular group over another (Faleye & Rasheed, 2020). This indicates that the test includes features that result in differential performance for testees with the same ability but belong to different groups (Faleye & Rasheed, 2020). To be biased, an item must undergo a substantial evaluation with evidence that it performs differently, has a different meaning, or assesses an undesired nuisance element for one group compared to another (Faleye & Rasheed, 2020).

Different approaches for identifying DIF have been utilised over time, all based on the IRT approach (Lang & Tay, 2021). IRT techniques produce the greatest results for detecting biased items due to their larger complexity since IRT provides a rigorous way for recognising DIF using an item characteristics curve (ICC) (Acciarini, Brunetta, & Boccardelli, 2021). According to Battuaz (2017), if the item does not exhibit DIF, the ICC will be the same; nevertheless, if DIF is present, the ICC(s) for the two classes will differ. IRT detection approaches include calculating or comparing object parameters (Lord Wald test), probability functions, area methods for finding possibly biased items, p-values, ICC curves, and tracing lines based on the parameter model.

According to Svetina Valdivia et al. (2024), the proportion of DIF products varies from a small 1.5 percent to a large percentage of the total (64 percent). The DIF is usually measured in its amount as well as the change in degree, which may be calculated through evaluating parameters or statistics depending on the process employed to identify DIF. Studies classify it as a small quantity of DIF a significant level of DIF is present when a test contains less than 10% DIF between 10 and 30 percent DIF, and a substantial quantity of DIF when it hits 30% (Wang et al., 2020). DIF magnitude levels start at 0.25

Additionally, the educational environment in Cross River State brings distinct difficulties and possibilities. Despite progress in incorporating technology within education, the Nigerian Ministry of Education reports persistent gaps in resource availability (2021). AI-generated test items should be optimised to serve all students evenly, minimising unequal item functionality. Studies on DIF in assessments for mathematics have primarily focused on traditional test questions, leaving a vacuum in knowing how AI-generated questions function among diverse demographics.

This study intends to provide insights into creating more equitable AI-driven evaluations by investigating gender and location-based DIF in various scenarios. This present study seeks to address the question: Could AI-generated mathematics test questions function differently based on gender and school location among students in the Calabar Municipality Local Government Area of Cross River State?

Research questions

The research questions guiding the study were as follows:

1. Which questions in the AI-generated mathematics multiple-choice test performed differently among male and female students?
2. Which AI-generated mathematics multiple-choice test questions performed differently in urban and rural schools?

Methodology

The study used a survey research design. The study's population consists of 1787 public senior secondary school two (SSSII) students from Calabar Municipality Local Government Area in Cross River State. A sample of 400, representing 22.34% of 1787 senior secondary school two

students, was determined using Taro Yamani's (1964) sample size determination approach and selected using basic random selection procedures. The data was collected using an AI-generated mathematics multiple-choice achievement test. The items, which were 30 in number, covered the different topics in Mathematics. Each item comprised a stem and four options from which a student was expected to select the correct option. Each student wrote the answers on the question paper within one hour and 30 minutes (1hr, 30m). The instrument was subjected to both content and face validation by psychometrician experts. A Kuder-Richardson estimate was used to establish the reliability of the research instruments, which yielded coefficients of .81 for AI-generated mathematics. Bilog-MG 3 was utilised to answer the study questions with a significance level 0.05. To determine the item that indicates the presence of Differential Item Functioning (DIF), any items in the Table with a DIF-value of .05 and above indicate the presence of DIF. In contrast, items with a DIF-value below .05 indicate no DIF. Also, to identify the group that an item favours, any item with a lower group value and DIF-value above .05 indicated DIF in favour of the participants.

Results

Research question one: Which AI-generated mathematics multiple-choice test questions performed differently among male and female students?

Table 1 Model intended for group DIF for AI-generated mathematics regarding students' gender

Items	Groups		DIF	Item 14	0.708*	0.489*	
	Male	female			-6.563	-6.439	
Item 01	-2.829	-2.605	0.225	Item 15	0.635*	0.607*	0.022
	0.357*	0.342*			-5.517	-5.540	
Item 02	-3.966	-4.674	0.707	Item 16	0.517*	0.517*	
	0.409*	0.444*			-5.104	-6.439	-1.335
Item 03	-4.331	-4.277	-0.054	Item 17	0.493*	0.606*	
	0.425*	0.419*			-5.995	-4.430	1.565
Item 04	-4.254	-5.030	-0.775	Item 18	0.566*	0.429*	
	0.449*	0.476*			-2.858	-2.990	0.132
Item 05	-5.408	-5.224	0.185	Item 19	0.334*	0.340*	
	0.527*	0.496*			-5.009	-5.030	-0.021
Item 06	-4.652	-5.224	-0.572	Item 20	0.462*	0.471*	
	0.477*	0.495*			-5.517	-5.892	-0.375
Item 07	-5.995	-5.224	0.771	Item 21	0.515*	0.538*	
	0.576*	0.498*			-6.593	-5.630	-0.963
Item 08	-6.756	-5.408	-1.347	Item 22	0.518*	0.628*	
	0.527*	0.660*			-5.518	-6.019	-0.502
Item 09	-6.127	-6.020	-0.107	Item 23	0.512*	0.558*	
	0.599*	0.570*			-6.725	-5.770	0.955
Item 10	-5.747	-6.593	-0.846	Item 24	0.647*	0.537*	
	0.554*	0.628*			-5.995	-5.770	0.025
Item 11	-5.303	-4.847	0.457	Item 25	0.559*	0.529*	
	0.520	0.469*			-6.725	-5.652	1.072
Item 12	-6.563	-6.439	0.125	Item 26	0.645*	0.519*	
	0.632*	0.612*			-5.869	-5.325	0.031
Item 13	-7.076	-5.224	1.853		0.537*	0.489*	

Item 27	-6.266 0.589*	-5.125 0.485*	1.141	Item 29	-5.202 0.491*	-5.430 0.498*	-0.228
Item 28	-5.325 0.496*	-5.104 0.473*	-0.221	Item 30	-5.223 0.479*	-4.826 0.452*	-0.398

*standard error

The result presented in Table 1 revealed ability disparities among male and female counterparts in an AI-generated mathematics multiple-choice test. The male students had superior abilities on 14 items, which constituted 46.6 percent. These items are item-2 with the ability difference of (0.707), item-4 (-0.775), item-5 (0.185), item-6 (-0.572), item-10 (-0.846), item-12 (0.125), item-13 (1.853), item-16 (-1.335), item-18 (0.132), item-20 (-0.375), item-22 (-0.502), item-23 (0.955), item-25 (1.072), and item-29 (-0.228). Similarly, the female students had superior abilities on 12 items, constituting 40 percent. These items are; item-1 (0.225), item-3 (-0.054), item-7 (0.771), item-8 (-1.347), item-9 (-0.107), item-11 (0.457), item-14 (0.125), item-17 (1.565), item-21 (-0.963), item-27 (1.141), item-28 (-0.221) and item-30 (-0.398). Lastly, no ability disparity was witnessed on only four items, constituting 13.33 percent. These items are 15, 19, 24 and 26, respectively. In conclusion, the result revealed that 86.67 percent (46.67% was in favour of male while 40% was in favour of female students) of the items showed disparity, with only 13.33% of the items showing no disparity among the groups of students in AI-generated mathematics multiple-choice questions.

Research question two: Which questions in AI-generated mathematics multiple choice test performed differently among urban and rural schools?

Table 2 Model intended for group DIF for AI-generated mathematics regarding students' location

Items	Groups			DIF	Item	Groups		
	Urban	Rural				Urban	Rural	DIF
Item 01	-5.76	-6.115	0.409	Item 10	0.496	0.477	1.428	
	0.573	0.499			-5.347	-6.775		
Item 02	-5.556	-5.348	0.027	Item 11	0.638	0.467	-0.018	
	0.530	0.477			-5.161	-5.263		
Item 03	-5.523	-6.237	0.714	Item 12	0.582	0.471	1.490	
	0.590	0.487			-5.454	-3.964		
Item 04	-5.454	-5.434	0.019	Item 13	0.518	0.396	-0.365	
	0.521	0.481			-5.343	-5.069		
Item 05	-5.180	-5.256	0.077	Item 14	0.488	0.480	1.185	
	0.511	0.462			-2.626	-3.811		
Item 06	-6.115	-6.102	0.013	Item 15	0.411	0.349	-0.701	
	0.576	0.531			-5.018	-4.312		
Item 07	-5.523	-5.345	-0.169	Item 16	0.446	0.450	-0.832	
	0.510	0.485			-4.028	-4.890		
Item 08	-5.801	-6.496	0.695	Item 17	0.475	0.404	-0.832	
	0.622	0.496			-4.712	-3.880		
Item 09	-5.434	-5.162	-0.273					

	0.424	0.427			0.374	0.383	
Item 18	-4.395	-3.718	0.037	Item 25	-3.483	-2.539	-0.914
	0.443	0.385			0.373	0.375	
Item 19	-3.416	-3.426	-0.010	Item 26	-3.778	-2.431	-1.347
	0.405	0.374			0.367	0.386	
Item 20	-3.480	-3.599	-0.009	Item 27	-2.527	-2.168	-0.358
	0.395	0.383			0.358	0.353	
Item 21	-2.830	-2.704	0.126	Item 28	-3.779	-2.485	-1.294
	0.376	0.361			0.773	0.391	
Item 22	-3.370	-2.874	-0.496	Item 29	-1.589	-1.865	0.276
	0.376	0.383			0.353	0.337	
Item 23	-3.370	-2.932	-0.438	Item 30	-2.934	-3.049	0.114
	0.384	0.381			0.396	0.362	
Item 24	-3.718	-2.649	-1.069				

*standard error

The result presented in Table 2 revealed ability disparities among urban and rural counterparts in an AI-generated mathematics multiple-choice test. The urban students had superior abilities on 9 items, constituting 30 percent. These items are item 1 with the ability difference of (0.409), item-3 (0.714), item-5 (0.077), item-8 (0.695), item-10 (1.428), item-14 (1.185), item-16 (0.862), item-29 (0.276), and item-30 (0.114). Similarly, the rural students had superior abilities on 14 items, constituting 46.67 percent. These items are; item-7 (-0.169), item-9 (-0.273), item-12 (1.490), item-13 (-0.365), item-15 (-0.701), item-17 (-0.832), item-21 (0.126), item-22 (-0.496), item-23 (-0.438), item-24 (-1.069), item-25 (-0.914), item-26 (-1.347), item-27 (-0.358) and item-28 (-1.294). Lastly, no ability disparity was witnessed on only seven items, constituting 23.33 percent. These items are 2, 4, 6, 11, 18, 19, and 20, respectively. In conclusion, the result revealed that 76.67 percent (30% was in favour of urban while 46.67% was in favour of rural students) of the items showed disparity, with only 23.33% of the items showing no disparity among the groups of students in AI-generated mathematics multiple-choice questions.

Discussion of findings

This study analysed sex and location-based DIF of AI-generated mathematics multiple-choice questions. Out of thirty items in the test administered to senior secondary school students, 26 showed DIF based on gender. Furthermore, 23 items showed DIF based on school location. The results of this study are consistent with the findings of Eteng-Uket (2021), who discovered differential item functioning in the West African Senior School Certificate English language test in South-South Nigeria. The study discovered that 13 questions were achieved differently for male/female students, while 23 items were achieved differently for upper and lower socioeconomic-level subgroups. In addition, “Ndifon, Umoinyang and Idiku (2010) found that junior secondary school certificate mathematics examination in Southern Educational Zone of Cross River State exhibited differential item functioning (DIF) based on gender, school geographical location/school ownership”. Their study revealed 10 items that functioned differently between male/female students, 12 between urban and rural schools, and six between private and public-owned schools.

Conclusion and Recommendations

In line with the research's results, it was concluded that AI-generated test questions are encountered by DIF based on both sex and location in Mathematics. Specifically, AI-generated mathematics multiple-choice tests showed DIF in terms of students' gender (male/female) and participants' location (urban/rural) schools. In line with the findings and implications of this study:

1. The school authorities should organise training for teachers on AI use.
2. Teachers and school administrators should review and revise AI-generated test items before they are used in the school.
3. Schools and the government should incorporate AI and validate AI-generated test items before they are used in schools.

References

- Acciarini, C., Brunetta, F., and Boccardelli, P. (2021). A comprehensive literature review explores cognitive biases and decision-making methods during transition periods. *Management Decision*, 59(3): 638–652.
- Akorede, M. (2021). *Pre-School Teachers' Knowledge and Use of Authentic Assessment in the Kwara Central Senatorial District* (Master's thesis, Kwara State University, Nigeria).
- Baker, R. S., and Hawn, A. (2022). Algorithmic prejudice in education. *International Journal of Artificial Intelligence in Education*, 1–41.
- Bao, H. C. M. Dayton, and A. B. Hendrickson. (2019). A reading test uses differential item functional amplification and cancellation. *Practical Assessment, Research, and Evaluation*, 14(1), 19.
- Battuz, M.(2017). Using Wald's test to discover differential item functioning. *American International Journal of Research in the Humanities, Arts, and Social Sciences*, 14(84), 95–100.
- Bauer, Daniel J. (2023). Improving measurement validity in varied populations: Contemporary ways to assessing differential item functioning. *British Journal of Mathematical and Statistical Psychology*, 76(3), 435–461.
- Cesario J. (2022). What can experimental studies of prejudice inform us about real-world group differences? *Behavioural and Brain Sciences*, 45(e66).
- Effiom, A.P. (2021). Using item response theory, we assessed the fairness and differential item functioning of a mathematics performance exam for senior secondary students in Cross River state, Nigeria. *Global Journal of Educational Research*, 20(1), 55–62.
- Eteng-Uket, S. (2021) Assessment of Socio-Economic Status and Sex Related Differential Item Functioning Using Item Response Theory Approach. *Asian Journal of Education and Social Studies* 16(2).
- Faleye, B. A.; Rasheed, A. A. (2020). Differential Item Functioning of 2015/2016 Biology Multiple-Choice Questions on the Osun State Joint Advancement Examination: *Al-Hikmah Journal Of Education*, Vol. 7, No. 1.Lang, J. W., & Tay, L. (2021). The science and practice of item response theory in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, 8(1), 311-338.
- Lee, P., Joo, S. H., and Stark, S. (2021). Detecting DIF in multidimensional forced choice assessments with the Thurstonian item response theory model. *Organisational Research Methods*, 24(4), 739–771.
- Miller, D. I.; Stoeckel, M. (2019). Gender and Mathematics: A Written Review. *Journal of Educational Psychology*, 111(3), 455–472.

- Nigerian Ministry of Education (2021). Education in Cross River State: Challenges and Opportunities.
- Nussbaum, M. (2021). The gender gap in mathematics: A worldwide view. *International Journal of Educational Research*, 107, 101–115.
- Svetina Valdivia, Huang, and Botter, P. (2024). Identifying differential item functioning in the context of multilevel data: *Do approaches that account for the multilevel data structure make a difference?* In *Frontiers of Education* (Vol. 9, p. 1389165). Frontiers Media S.A.
- UNESCO (2020). The International Educational Monitoring Report 2020: Inclusion and Education.
- Wang, F., Hou, H., Luo, Y., Tang, G., Wu, S., Huang, M., and Sun, Z. (2020). Laboratory testing and host immunity of COVID-19 patients with varying degrees of disease. *JCI Insight*, 5(10).