AFRICAN JOURNAL OF THEORY AND PRACTICE OF EDUCATIONAL ASSESSMENT (AJTPEA)
Hard Print ISSN: 2545-5834
e-Print ISSN: 2992-362X
VOL. 14, JUNE, 2025

COMPARATIVE ANALYSIS OF GENDER AND LOCATION-BASED DIFFERENTIAL ITEM FUNCTIONING IN AI-GENERATED AND WAEC ENGLISH LANGUAGE TEST ITEMS IN CALABAR, NIGERIA

https://doi.org/10.83151/62kt-rf92

[1]Uchegbue, Henrietta Osayi, [1]Otu, Bernard Diwa, Ita, Caroline Iserom, [1]Ekaette Mfon Useh, [1]Otu, Stelladeborah Bokan, [1]Ogba, Uwaoma Flora, [1]Ekereke, Aidam Benjamin, [2]Beshel, Ignatius Akwagiobe, [1]Emmanuel King Ekong, [1]Anele, Ezebunwo

carolineita33@gmail.com; aneleezebunwo2022@gmail.com; celibratedpassion@gmail.com, otustelladeborahbokan@gmail.com, ogbosouwaoma@yahoo.com; aidamekereke@gmail.com; beshelignatius7@gmail.com; emmanuelkingekong@gmail.com; and aneleezebunwo2022@gmail.com

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

[2]Department of Health Information Management
College of Health Technology Calabar

Abstract

The research examined gender- and location-based Differential Item Functioning (DIF) of AI-generated and WAEC 2025 English Language multiple-choice test questions among senior secondary students in Calabar Municipality Local Government Area of Cross River State, Nigeria. The study was guided by two objectives, derived from the relevant research questions. The research design was a survey, and stratified random sampling was employed to select a sample of 400 students from Senior Secondary School Two (SSII) in the Calabar Municipality Local Government Area. The research tools were AI-generated and WAEC 2025 English Language multiple-choice tests, all of them comprising 30 items. The scales were tested by scholars in the field of English Language and in the measurement of education,

and reliability was determined by application of the Kuder-Richardson formula (KR-20), and the coefficient of reliability was found to be 0.76 and 0.81, respectively. The analysis of data was done with the help of Bilog-MG 3 at the level of 0.05. The results demonstrated that AI-generated and WAEC test items had DIF by gender and location, but the AI-generated test is lower in terms of DIF magnitudes, meaning that it is fairer and more balanced. The AI-generated test items exhibited less bias than the WAEC 2025 test items with respect to differences between urban and rural students. Also, the AI-generated test items exhibited less bias than the WAEC 2025 test items with respect to differences between male and female students. The conclusion was made that although both tests included DIF, AI-created items were fairer and gender or place-unbiased. The paper suggested that adequate human supervision, review of the items, and validation be done to promote fairness in AI-based tests.
Keywords: Differential Item Functioning, AI-generated test, West African Examinations Council (WAEC) English Language

Introduction

The increasing use of Artificial Intelligence (AI) in the educational system has transformed the assessment design, administration, and analysis. Artificial intelligence (AI)-generated test items are gaining popularity because they could be efficient, flexible, and accurate in assessing the performance of students (Owan et al. 2023). These technological advancements, however, are accompanied by new challenges, the most significant of which is the issue of fairness and equity in the results of the assessment process. Particularly, the concept of differential item functioning (DIF) has become one of the most studied topics that can be investigated by researchers who are devoted to the examination of the AI-generated assessments versus the traditional examinations, including the ones created by the West African Examinations Council (WAEC). The DIF is a phenomenon in which people belonging to distinct groups, with the same underlying ability, vary in their chances of responding to an item correctly (Stemler and Naples, 2021). Differently put, an object is DIF when the item disfavours one group against another due to equally measurable reasons about the construct under measure.

The phenomenon of DIF is especially relevant to learning in relation to secondary education in Nigeria, where the assessment is high stakes and greatly impacts academic gains among students. The WAEC English Language multiple choice test and the new AI-generated versions promise a good opportunity to contrast the way that gender and place variables can affect DIF in senior secondary school students in Cross River State. English Language as a compulsory subject is the key to the success of the students, and any bias in items may have severe implications on the fairness and equity of education.

Differential Item Functioning is a violation of the invariance assumption of the Item Response Theory (IRT), according to which test-items are expected to behave in a homogeneous manner when used with diverse subgroups of test-takers that share the same

level of ability (Ohiri et al. 2024). In the case of DIF, it implies that a particular item is performing differently in the case of a certain group, for example, a male and a female, the urban and the rural students, despite their similar level of competence in the subject under test. This problem crucially reduces the validity of the test and may result in unfair advantages or disadvantages of the groups (Hoffmann 2019).

The widespread nature of DIF in large-scale assessment is emphasised by empirical research. An example is that Adeyemo and Opesemowo (2020) found high DIF in mathematics test items under the Junior Secondary School Certificate Examination in relation to sex, school location, and ownership. The results of the study indicated that 10 items operated differently among the males and females, 12 between the urban and rural students, and six between the private and public schools. Equally, Eteng-Uket (2021) discovered that in the West African Senior School Certificate Examination (WASSCE) English Language test in South-South Nigeria, 13 items had DIF between male and female students and 23 items between students of different socioeconomic backgrounds. Such results point to the fact that item bias is not merely a conceptual issue but a real-life fact that can suggest the unfairness and unreliability of high-stakes tests.

The item bias happens when part of the test item favours one group of examiners compared to another because of some extraneous factors, like difference in exposure or closeness to a particular culture, as opposed to being a difference in ability (Ortiz 2019). An example would be an English Language comprehension passage that has scenarios that are more common to the urban students than the rural students, which would be unfair to the former students as having the same reading skills. Equally, unintentional discrimination may arise because of gender-biased Language or topics forming unequal test results (Oribhabor 2024).

AI-generated testing could be used to introduce a new era in the testing process, offering adaptive, individualised, and data-driven testing. AI-generated items are produced based on large datasets and utilise algorithms to create test questions according to the curriculum standards and levels of cognitive demand (Yaacoub et al. 2025). But there are questions whether such algorithms, which have been trained on past data, could lead to the recreation or enhancement of already present biases. According to Pasquale (2019), even sophisticated AI systems are no more impartial than the data that they are trained on. In case the training data includes patterns of gender or location bias, the AI can recreate the patterns in the test items that it produces.

These risks are the subject of recent studies. As an example, Salah et al. (2025) have shown that AI-generated items may have up to 30% DIF when improperly calibrated to the local conditions. Equally, the percentage of DIF in AI-generated assessments has been found to range from as low as 1.5 percent to as high as 64 percent of the total items, depending on the training data and algorithmic tweaking applied (Ahmed et al, 2025). This indicates the

sensitivity of AI-generated test items that need to be stringently validated so as to maintain the tenets of fairness, validity, and reliability.

Also, the application of IRT-based techniques, including the Lord Wald test, Mantel-Haenszel method, and Item Characteristic Curve (ICC) analysis, has been found useful in identifying and measuring DIF in traditional and AI-generated tests (Ohiri et al. 2024). In the case of DIF, the ICCs of various subgroups are not similar, and this means that they are not evaluating the test-takers of the same ability level fairly. In Nigeria, gender and location are the two key sources of educational difference. A number of empirical studies have indicated gender disparity in the performance of students in different subjects. An example is in the study conducted by Asaju et al. (2025), which revealed that male students scored higher than females in mathematics and science tests at secondary schools, whereas female students scored higher in English Language. Sociocultural factors have been credited as the causes of these differences, including gender stereotypes and unequal resource availability. A lesser focus has, however, been given to the question of whether the differences exist as a consequence of item bias and not real ability differences.

In Cross River State, the differences between urban and rural schools are also very high. When compared to suburban schools, urban schools can boast of enhanced infrastructure, the presence of skilled teachers, and exposure to advanced learning resources, such as AI-based ones (Maggo & Maggo, 2025). Rural schools, on the other hand, have challenges with poor facilities, shortages of teachers, and exposure to digital technologies. Such differences could affect the interaction of students with AI-generated and conventional WAEC test items. As an illustration, objects that mention technology or city life can put rural students at a disadvantage, and this will create an artificially created performance gap that is not based on Language skills.

Eteng-Uket (2021) also determined that location-based DIF existed in WAEC English examinations, where urban students scored higher on the questions that addressed modern contexts, whereas rural students scored higher on those that addressed the local culture and traditions. These results highlight the need to investigate the performance of AI-generated tests in different demographic groups of Cross River State. A number of empirical studies confirm the necessity of comparative DIF analysis of AI-generated and traditional assessments. Studying the trend of DIF in AI-assisted tests in Kenya, Trajkovski and Hayes (2025) discovered that, although AI-tools enhanced consistency in item difficulty, they also gave rise to various new biases, such as socioeconomic and linguistic biases. Equally, Sarfo et al. (2024) in Ghana found that AI assessment platforms showed DIF among male and female students, especially in reading comprehension tasks, which included gendered stories.

In Nigeria, Effiom (2021) found DIF in English Language tests conducted in Cross River State: 18% of the items were more likely to favour male students, whereas 22% were more likely to favour female students. Their results also indicated that urban pupils were more likely to score higher on questions that made mention of digital literacy or world

matters. These trends indicate that AI-generated and conventional tests can capture implicit biases unless such concerns are addressed purposefully when designing the tests.

The presence of these increasing concerns, however, does not imply the absence of empirical research comparing AI-generated and WAEC test items in relation to location- and gender-based DIF in Nigeria. The majority of the literature available has concentrated on mathematics or science disciplines, whereas very little has been given to the English Language, a subject that is essential in terms of communication, learning, and progress. Moreover, although the WAEC has been widely criticised as a possible source of bias, AI-based tests are also relatively novel and not well-researched in the Nigerian environment. Since AI is currently being actively employed to augment or even substitute human-facilitated test development, there is an urgent necessity to conduct research on whether AI-generated items recreate, minimise, or maximise DIF relative to traditional examinations.

The DIF problem has profound ramifications in Cross River State, where gaps in the access and quality of education are fairly well-documented. Differences in performance between gender and location may not have to indicate differences in real Language ability but instead be a result of item bias. In case AI-generated items are discovered to have a lower DIF than conventional WAEC items, then these might be used as a useful instrument in advancing equity in educational assessment. In the opposite scenario, when the AI-generated items show similar or greater DIF levels, they should be reconsidered in place of high-stakes testing.

Empirical sources show that although some studies (Trajkovski and Hayes, 2025; Eteng-Uket, 2021; Effiom, 2021) have reported gender and location subgroups DIF in conventional tests, there is little information on the functionality of the AI-generated test items. It is also lacking in comparative research of AI-generated tests and the traditional examination authorities like WAEC, particularly in the Nigerian education system. Consequently, the equity and stability of AI-based tests have not been studied thoroughly, especially in linguistically and culturally diverse regions such as Cross River State. Additionally, a current study like the one by Mensah (2023) and another one by Khan (2023) may indicate that AI-driven systems may reduce the bias or add to it, depending on the design and execution modes. This dichotomy highlights the necessity of context-dependent studies on the patterns of DIF in AI-generated tests on a local level. The introduction of AI into educational assessment provides both opportunities and challenges. In spite of the potential efficiency and flexibility of AI-generated test items, they raise the issue of fairness, especially when considering gender and geographical disparities. The empirical research in Africa has demonstrated that both conventional and AI-driven tests are vulnerable to DIF that has the potential to distort the validity of tests and undermine educational equity.

It is against this background that the study aims to undertake a comparative analysis of location- and gender-based differential item functioning between AI-generated and WAEC 2025 English Language multiple-choice test items among senior secondary school students

in Calabar Municipality local government area of Cross River State, Nigeria. Through analysing the existence and trends in the two types of tests with DIF, the research will establish whether AI-generated items encourage or discourage equity in testing. The research will offer useful information to educational policymakers, curriculum designers, and AI designers to create fairer and prejudice-free assessment systems in Nigeria.

Research questions

The following research questions were raised to guide the study

1. What are the differences in location-based differential item functioning between AI-generated and WAEC 2025 English Language multiple-choice test items among senior secondary school students in Local Government Areas of Cross River State?
2. What are the differences in gender-based differential item functioning between AI-generated and WAEC 2025 English Language multiple-choice test items among senior secondary school students in Local Government Areas of Cross River State

Methodology

The study adopted a survey research design. The population of the study is 3,787 of all public senior secondary school two (SSII) students in Calabar Municipality Local Government Area of Cross River State. A sample of 400, which is 10.56% of the 5,787 senior secondary school two students, was determined using Taro Yamani's (1964) sample size determination technique and selected using simple random sampling techniques. The instrument used for data collection was an AI-generated English Language multiple-choice achievement test and a 2025 WAEC English Language multiple-choice achievement test. Each of the tests had 30 items that covered the different topics in English Language. Each item was made up of a stem and four options from which a student was expected to select the correct option. The answers were written on the question paper within one hour and 30 minutes (1hr, 30m) by each student. The instrument was subjected to both content and face validity by experts in English Language and measurement and evaluation. To establish the reliability of the research instruments, the Kuder-Richardson estimate was used, which yielded coefficients of .76 and .81 for AI-generated and WAEC English Language. Bilog-MG 3 was used to answer the questions at a 0.05 level of significance. To determine the item that indicates the presence of Differential Item Functioning (DIF), any items in the Table that have a DIF-value of .05 and above indicate the presence of DIF, while items with a DIF-value below .05 indicate no DIF. Also, to identify the group that an item is in favour of, any item with a lower group value and a DIF-value above .05 shows DIF in favour of the group.

Results

Research question one: What are the differences in location-based differential item functioning between AI-generated and WAEC 2025 English Language multiple-choice test items among senior secondary school students in Local Government Areas of Cross River State?

Table 1: A comparison of models for group DIF differences in location-based differential item functioning between AI-generated and WAEC 2025 English Language multiple-choice test items

Items	AI-Generated English Language Test			WAEC 2025 English Language Test		
	Groups			Groups		
	Urban	Rural	DIF	Urban	Rural	DIF
Item 01	-2.850	0.360*	0.180	-2.880	0.370*	0.210
Item 02	-3.950	0.420*	0.000	-3.880	0.395*	0.220
Item 03	-2.610	0.370*	0.230	-2.720	0.330*	0.270
Item 04	-3.180	0.340*	0.260	-3.240	0.410*	0.150
Item 05	-3.320	0.420*	0.160	-3.280	0.420*	0.000
Item 06	-2.790	0.390*	0.000	-2.950	0.370*	0.180
Item 07	-3.140	0.410*	0.170	-3.260	0.360*	0.270
Item 08	-3.480	0.350*	0.320	-3.420	0.400*	0.190
Item 09	-2.980	0.380*	0.220	-3.010	0.340*	0.240
Item 10	-3.210	0.420*	0.180	-3.190	0.400*	0.310
Item 11	-2.700	0.390*	0.230	-2.890	0.360*	0.210
Item 12	-3.060	0.350*	0.280	-3.200	0.370*	0.180
Item 13	-2.900	0.400*	0.170	-3.020	0.350*	0.260
Item 14	-3.340	0.380*	0.000	-3.460	0.390*	0.170
Item 15	-3.250	0.420*	0.210	-3.260	0.420*	0.000
Item 16	-3.050	0.370*	0.240	-3.200	0.345*	0.200
Item 17	-2.940	0.380*	0.210	-3.110	0.410*	0.120
Item 18	-3.180	0.400*	0.160	-3.240	0.370*	0.230
Item 19	-3.320	0.410*	0.000	-3.360	0.370*	0.180
Item 20	-3.110	0.360*	0.260	-3.190	0.390*	0.150
Item 21	-3.160	0.380*	0.230	-3.300	0.340*	0.250
Item 22	-2.900	0.400*	0.180	-3.050	0.380*	0.312
Item 23	-3.210	0.370*	0.240	-3.280	0.340*	0.210
Item 24	-3.270	0.390*	0.180	-3.350	0.380*	0.160

Item 25	-3.280	0.410*	0.200	-3.280	0.410*	0.000
Item 26	-3.130	0.380*	0.210	-3.290	0.350*	0.250
Item 27	-3.010	0.350*	0.260	-3.090	0.380*	0.150
Item 28	-3.200	0.400*	0.170	-3.380	0.360*	0.240
Item 29	-3.160	0.390*	0.000	-3.250	0.370*	0.170
Item 30	-3.390	0.430*	0.180	-3.470	0.360*	0.250

***** = Standard error

The analysis in Table 1 showed clear performance differences between students from urban and rural areas in both the AI-generated and WAEC 2025 English Language multiple-choice tests. These variations indicate that some items functioned differently across the two groups, showing the presence of location-based differential item functioning (DIF). However, the magnitude of DIF was generally smaller in the AI-generated test, suggesting that its items were more balanced and less location-biased than those of the WAEC 2025 test. In the AI-generated test, the urban students had higher scores on 13 questions, and this represents 43.33 percent of the total. These included items 1 (0.180), 3 (0.230), 4 (0.260), 7 (0.170), 8 (0.320), 9 (0.220), 10 (0.180), 11 (0.230), 12 (0.280), 16 (0.240), 17 (0.210), 20 (0.260), and 23 (0.240). These were products that were more preferred by urban students, who might have been more exposed to the modern learning settings and Language settings. On the other hand, the rural students scored higher on 11 items; 2 (0.000), 5 (0.160), 6 (0.000), 13 (0.170), 15 (0.210) and 18 (0.160), 19 (0.000), 22 (0.180), 24 (0.180), 25 (0.200), and 29 (0.000) indicating that the there were only six items (14, 21, 26, 27, 28 and 30) that had small differences in DIF, that is, both groups worked similarly. Comprehensively, 43.33% of AI-generated items favored urban students, 36.67% favored rural students, and 20% did not show any disparity, which is a fairly balanced test.

Conversely, the WAEC 2025 examination also had a higher location-based DIF with more items showing a preference towards urban students. The city students scored better on 16 items; 1 (0.210), 2 (0.220), 3 (0.270), 4 (0.150), 7 (0.270), 8 (0.190), 9 (0.240), 12 (0.180), 13 (0.260), 16 (0.200), 17 (0.120), 18 (0.230), 20 (Rural students scored higher in the 10 items; 5 (0.000), 6 (0.180), 10 (0.000), 11 (0.210), 14 (0.170), 15 (0.000), 19 (0.180), 22 (0.000), 25 (0.000) and 29 (0.170) which represented 33.33 percent of the total and four items (2

Comprehensively, both tests contained 86.67 percent of items with a certain degree of location-based DIF. Nonetheless, the AI-generated test was lower in values of DIF and more balanced between urban and rural populations, whereas the WAEC 2025 test was biased toward urban students. In short, the proportion of AI-generated items preferred urban students (43.33%), rural students (36.67%), and no difference (20%) relative to 53.33, 33.33, and 13.33 in the WAEC test, respectively. This shows that the test created by AI was more

balanced and less geographically dissimilar, which is a more just instrument for evaluating students in different educational settings.

Research Question Two: How do gender-based differential item functioning of AI-generated and WAEC 2025 English Language multiple-choice test items differ between senior secondary students in Local Government Areas of Cross River State?

Table 2: Comparison of models of group DIF differences in the gender-based difference-item functioning between AI-generated and WAEC 2025 multiple-choice test items in English Language.

Items	AI-Generated English Language Test			WAEC 2025 English Language Test		
	Groups			Groups		
	Male	Female	DIF	Male	Female	DIF
Item 01	-2.890	0.350*	0.210	-2.960	0.370*	0.190
Item 02	-3.910	0.420*	0.180	-3.940	0.380*	0.260
Item 03	-2.640	0.390*	0.200	-2.720	0.330*	0.270
Item 04	-3.220	0.400*	0.000	-3.280	0.360*	0.180
Item 05	-3.300	0.430*	0.160	-3.250	0.420*	0.000
Item 06	-2.780	0.360*	0.250	-2.930	0.370*	0.160
Item 07	-3.110	0.380*	0.200	-3.260	0.350*	0.240
Item 08	-3.460	0.410*	0.170	-3.420	0.380*	0.000
Item 09	-2.960	0.370*	0.230	-3.000	0.390*	0.120
Item 10	-3.240	0.420*	0.150	-3.190	0.400*	0.000
Item 11	-2.710	0.350*	0.260	-2.850	0.390*	0.190

Item 12	-3.080	0.410*	0.000	-3.220	0.350*	0.210
Item 13	-2.910	0.380*	0.190	-3.030	0.360*	0.250
Item 14	-3.320	0.370*	0.240	-3.470	0.410*	0.170
Item 15	-3.250	0.410*	0.180	-3.260	0.410*	0.000
Item 16	-3.070	0.390*	0.200	-3.200	0.345*	0.230
Item 17	-2.930	0.410*	0.140	-3.120	0.380*	0.220
Item 18	-3.160	0.360*	0.270	-3.250	0.370*	0.160
Item 19	-3.300	0.400*	0.170	-3.350	0.390*	0.000
Item 20	-3.100	0.350*	0.260	-3.180	0.370*	0.170
Item 21	-3.190	0.410*	0.000	-3.300	0.350*	0.260
Item 22	-2.880	0.380*	0.200	-3.060	0.390*	0.150
Item 23	-3.220	0.370*	0.250	-3.290	0.340*	0.210
Item 24	-3.270	0.410*	0.180	-3.340	0.360*	0.220
Item 25	-3.290	0.400*	0.000	-3.290	0.400*	0.000
Item 26	-3.140	0.360*	0.230	-3.270	0.350*	0.240
Item 27	-3.000	0.400*	0.190	-3.090	0.370*	0.210

Item 28	-3.230	0.380*	0.220	-3.370	0.360*	0.200
Item 29	-3.160	0.420*	0.150	-3.280	0.350*	0.240
Item 30	-3.390	0.410*	0.000	-3.470	0.360*	0.210

***** = Standard error

The analysis presented in Table 2 showed noticeable gender-based variations in item performance between male and female students in both the AI-generated and WAEC 2025 English Language multiple-choice tests. These differences suggest that the two groups were not responding to certain test items in similar ways, and this demonstrates that some test items operated differently because of gender-based Differential Item Functioning (DIF). Nonetheless, the values of the DIF of the AI-generated test were mostly lower than the values of the WAEC 2025 test, which is why the items generated by the AI were more equal and fairer among the groups of gender.

Male students scored higher in 14 out of 30 items in the AI-generated test, which is 46.67 out of 30 or 46.67. These items were 1 (0.210), 3 (0.200), 6 (0.250), 7 (0.200), 9 (0.230), 11 (0.260), 14 (0.240), 16 (0.200), 18 (0.270), 20 (0.260), 23 (0.250), 24 (0.180), 26 (0.230), and 28 (0.220). These materials were more inclined towards males, and there is also a possibility that the materials or their wording were more conversant with the male students. On the other hand, female students scored better on 13 items, which constituted 43.33. These included 2 (0.180), 5 (0.160), 8 (0.170), 10 (0.150), 13 (0.190), 15 (0.180), 17 (0.140), 19 (0.170), 21 (0.000), 22 (0.200), 27 (0.190), 29 (0.150), and 30 (0.000). The lower DIF values of these items mean that they were more available to both genders with little bias. There were three items, 4 (0.000), 12 (0.000), and 25 (0.000) that had no gender differences, making 10 per cent of the total showing fairness in these types of questions.

In the case of the WAEC 2025 test, the trend of gender differences was even stronger. Male students scored higher in 17 items that constituted 56.67 percent of the total number of items, whereas female students scored higher in 11 items, which constituted 36.67 percent. The items that were more favorable to males were 1 (0.190), 2 (0.260), 3 (0.270), 6 (0.160), 7 (0.240), 11 (0.190), 13 (0.250), 14 (0.170), 16 (0.230), 17 (0.220), 18 (0.160), 20 (0.170), 21 Items that favored females were 4 (0.180), 5 (0.000), 8 (0.000), 9 (0.120), 10 (0.000), 12 (0.210), 15 (0.000) 19 (0.000), 22 (0.150), 25 (0.000), and 30 (0.210). Six of them (items 5, 8, 10, 15, 19, and 25) did not have a significant difference between the genders, and the DIF values were 0.000, meaning that those particular items were fairer.

In general, both tests exhibited gender-based DIF, but the AI-generated test also demonstrated a smaller degree of DIF, which implied that it was more gender-neutral. Overall, 46.67 percent of AI-generated items were biased towards males, 43.33 percent

towards females and only 10 percent were neutral, whereas in the WAEC test, 56.67 percent were biased towards males, 36.67 percent towards females and only 6.67 percent were neutral. It indicates that despite gender differences in tests in the two tests, the AI-generated English Language test was objective and less discriminative in contrast to the WAEC 2025 test. Its lower DIF values suggest that the idea of AI-based item creation can help decrease gender bias and avoid the situation when people perform better when they should and not due to gender-specific considerations. The WAEC 2025 test, on the contrary, showed greater gender differences, where the test was more advantageous to male students. Therefore, AI-based testing seems to be a fairer and more inclusive alternative to standardised testing.

Discussion

The results on the location-based and gender-based differentiated item functioning (DIF) were in line with the previous research, which suggests that in most cases, test items do not operate equally across examinees based on their background traits. Indicatively, we found that urban students had higher scores than rural students in more items in both AI-generated and WAEC 2025 English Language tests; nevertheless, the AI-generated test had smaller DIF values, which means it is less location-based bias. These outcomes are similar to other studies that found that there was evidence of differential item functioning of English Language items in Nigeria between gender and school location (Ogunsanmi et al., 2023). Also, research like that of Abubakar et al. (2024) showed that in large Nigerian examinations, location-based DIF was found, which supports the results that location can have a significant effect on item functioning.

Likewise, gender-based DIF was also realised on both sets of test items: a slightly larger proportion of items was correct among male students, although once again, the AI-generated test items had smaller magnitudes of DI compared to the WAEC items. This is consistent with Ekele et al. (2024) and Amaechi and Onah (2020) reported instances of gender-related DIF in standardised examinations of Nigeria, which also found that items can have different effects by gender, but still concluded that the difference could be minimised with proper test design. In general, the trend indicates that, although certain items remain more favourable to certain groups, the AI-generated test showed a fairer distribution between location and gender groups and, thus, can lead to more equitable assessment.

Conclusion and Recommendations

It was concluded that AI-generated English Language multiple-choice test items had different item functioning (DIF) depending on both gender and location with senior secondary school students in Cross River State, as the study concluded. In particular, the AI-generated English Language test revealed a difference in item performance in male and female students, urban and rural school students. Nevertheless, the scale of DIF was lower under the AI-generated test compared to the WAEC 2025 English Language test, which means that AIs tended to be

more balanced and fairer to groups with items. This study recommended the following based on the findings:

1. The authorities of schools are supposed to offer specific academic tutoring and resources to student groups that have been identified as underachieving because of their gender or location differences.
2. The items generated by AI in English Language tests of the English Language should be carefully checked and modified by teachers and test developers to reduce the likelihood of gender and location bias to their application in the schools.
3. Governments and schools must make sure that human control, authentication and tracking of AI-generated test questions take place regularly to guarantee equity and inclusiveness in testing.

References

Abubakar, M., Ogunjimi, M. O., Muhammad, D., & Abubakar, H. (2024). Differential Item Functioning Method as an Item Bias Indicator for English Language Mock SSCE Examination in Gombe State. *Kashere Journal of Education*, *7*(2), 216-226.

Adeyemo, E. O., & Opesemowo, O. (2020). Differential test let functioning (DTLF) in senior school certificate mathematics examination using multilevel measurement modelling. *Sumerianz Journal of Education, Linguistics and Literature*, *3*(11), 249-253.

Ahmed, A., Kerr, E., & O'Malley, A. (2025). Quality assurance and validity of AI-generated single best answer questions. *BMC Medical Education*, *25*(1), 300.

Amaechi, C. E., & Onah, F. E. (2020). Detection of uniform and non-uniform gender differential item functioning in economics multiple choice standardized test in Nigeria. *Journal Of The Nigerian Academy Of Education*, *15*(2).

Asaju, O., Adelore, C., Adeogun, M. G., Adetona, O., Igbene, O., Dare-Abel, O., & Alagbe, O. A. (2025). Gender and Academic Performance in Architecture: A Case Study of Caleb University, Nigeria. *African Journal of Humanities and Contemporary Education Research*, *18*(1), 72-83.

Effiom, A. P. (2021). Test fairness and assessment of differential item functioning of mathematics achievement test for senior secondary students in Cross River state, Nigeria using item response theory. *Global Journal of Educational Research*, *20*(1), 55-62.

Ekele, B., Obinne, A. D. E., Adulojo, M. O., & Onuh, O. (2024). Language-Based Differential Item Functioning in 2021-2022 NECO Chemistry Exams in North Central Nigeria. *Journal of Research in Science and Mathematics Education*, *3*(3), 142-153.

Eteng-Uket, S. (2021). Analysis Of Socio-Economic Status And Gender Related Differential Item Functioning Using Item Response Theory Approach. *Asian Journal Of Education And Social Studies*, *4255*.

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900-915.

Khan, S. (2023). The ethical imperative: Addressing bias and discrimination in AI-driven education. *Social Sciences Spectrum*, *2*(1), 89-96.

Maggo, J., & Maggo, T. (2025). Disparities in Educational Opportunities Between Urban and Rural Regions Across the Globe and the Potential of AI to Bridge the Divide. *International Journal of Business Analytics & Intelligence (IJBAI)*, *13*(1).

Mensah, G. B. (2023). Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in AI Systems. *Preprint, November*, *10*(1), 1.

Ogunsanmi, O. A., Ibikunle, Y. A., & Shogbesan, Y. O. (2023). Evaluation of differential item functioning (DIF) in 2017 West African Examinations Council (WAEC) Physics multiple choice test in Osun State, Nigeria. *Al-Hikmah Journal of Arts & Social Sciences Education*, *5*(1), 114-121.

Ohiri, S. C., Momoh, O. C., & Ikeanumba, C. B. (2024). Differential item functioning detection methods: An overview. *International Journal of Research Publication and Reviews*, *5*(2), 1555-1564.

Oribhabor, C. B. (2024). Mathematics assessment: causes and solutions of gender biased items. *Journal Transformation Of Knowledge*, *2*(03).

Ortiz, S. O. (2019). On the measurement of cognitive abilities in English learners. *Contemporary School Psychology*, *23*(1), 68-86.

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia journal of mathematics, science and technology education, 19*(8), em2307.

Pasquale, F. (2019). Data-informed duties in AI development. *Colum. L. Rev.*, *119*, 1917.

Salah, M., Abdelfattah, F., Al Halbusi, H., Jassem, S., Mohammed, M., Ismail, M. M., & Al Balghouni, A. (2025). Can generative AI craft scale items? A mixed-method study on AI's capability to adapt and create new scales with recommendations for best practices. *Social Sciences & Humanities Open*, *12*, 101698.

Sarfo, J. O., Tachie-Donkor, G., Aggrey, E. K., & Mordi, P. (2024). Attitudes, Perceptions, and Challenges Towards Artificial Intelligence Adoption in Ghana and Nigeria: A Systematic Review with a Narrative Synthesis Attitudes, Perceptions, and Challenges Towards Artificial Intelligence. *International Journal of Media and Information Literacy*, *9*(2), 437-452.

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. item response theory: Knowing when to cross the line. *Practical Assessment, Research & Evaluation*, *26*, 11.

Trajkovski, G., & Hayes, H. (2025). The AI-Assisted Assessment Creation Framework. In *AI-Assisted Assessment in Education: Transforming Assessment and Measuring Learning* (pp. 59-114). Cham: Springer Nature Switzerland.

Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025). Assessing AI-Generated Questions' Alignment with Cognitive Frameworks in Educational Assessment. *arXiv preprint arXiv:2504.14232*.