**Jurnal Bidang Pendidikan Dasar**
Vol 8 No 2, June 2024, pp 136 - 150
Available at:
http://ejournal.unikama.ac.id/index.php/JBPD

# Analysis of differential item functioning in agricultural science examination across southwestern nigeria's senior schools

**Oluwaseyi Aina Gbolade Opesemowo[1]*, Temitope Babatimehin[2], Temitope Sarah Ogungbaigbe[2]**

[1]*University of Johannesburg, Auckland Park, 2006, South Africa*
[2]*Obafemi Awolowo University, Ile-Ife, P.M.B 13, Nigeria*
*oopesemowo@uj.ac.za**

***Abstract:*** *Differential Item Functioning poses a threat to test fairness and validity in educational assessments. To this end, this study probes the incidence of DIF in the Senior School Certificate Agricultural Science Examination (SSCASE) across school locations in Southwestern Nigeria. It determined the magnitude of DIF using item parameters in SSCASE. The study applies one, two and three-parameter logistic models to analyze the dichotomous data responses of 818 randomly selected students from urban and rural locations. The ex-post facto design adopted the 2015 National Examination Council SSCE in Agricultural Science as its instrument (α =0.887) in this study. The study population comprised SSCE candidates from Osun, Ondo, and Oyo. The results showed the magnitude of DIF in SSCASE was large and moderate (1PLM=0.52 significant DIF, 2PLM=0.45 moderate DIF, 3PLM=0.47 moderate DIF) across school locations. The study concluded that 1, 2 and 3 PLMs can produce fair items in SSCASE.*

***Keywords****: Differential Item Functioning; Public Examination; One-Parameter Logistic Model; Two-Parameter Logistic Model; Three-Parameter Logistic Model*

## Introduction

Public examinations are external examinations that are open to the public, and these examinations are usually conducted by examination bodies using standardized tests. In Nigeria, examples of public examinations are the West African Senior School Certificate Examination (WASSCE), conducted by the West African Examination Council (WAEC); National Business Certificate (NBC)/National Technical Certificate Examinations (NTCE), conducted by the National Business and Technical Examination Board (NABTEB), The University Matriculation Examination (UTME) and Polytechnics/Colleges of Education (PCE) conducted by Joint Admission and Matriculation Board (JAMB) and Senior School Certificate Examination (SSCE) conducted by National Examination Council (NECO). The WAEC and NECO conduct the Senior School Leaving Certificate Examination in Nigeria. The purpose of the public examination was for placement and selection of the test takers into groups or places where they ought to be.

The importance of public examinations cannot be overemphasized (Wiggins et al., 2023) as it helped teachers focus on syllabus content and train their students to pass tests, thereby using teaching methods that are not useful for every student's learning (Mitana, Muwagga, & Ssempala, 2019). However, it was established that an increase in test scores might be due to teachers' and students' greater familiarity with the tests rather than an increase in learning (Jang, Pashler, & Huber, 2014; Yeng, Ali, & Adzifome, 2023). many teachers did not necessarily resent the amount and kind of testing. Indeed, most saw tests and examinations as advantageous rather than disadvantageous; teachers and students relied on tests and

https://doi.org/10.21067/jbpd.v8i2.10204 

examinations to ensure learning. The need to pass examinations drove students' learning and teachers' teaching; tests and examinations were strong partners to didactic, textbook-driven methods, drill, rote learning and memorization, superficial learning, student passivity and spoon-feeding.

Faremi and Jimoh (2022) claimed that some public examinations unfairly favour examinees of groups, e.g., cultural or linguistic groups, to the extent that it is now believed that a specific section of the country performs most woefully in these national examinations. A critical look at people's perception of such national examinations in Nigeria indicates the severe nature of item bias (Effiom, 2021). The issue of mass failure in public examinations has become a common phenomenon in Nigeria (Jerome, 2023; Oluwatimilehin & Opesemowo, 2019) (Jerome, 2023; Oluwatimilehin & Opesemowo, 2019); this is a result of the unfairness of test items among examinee sub-groups. If test items are fair enough to be used in public examinations, the issue of bias will not be present. The issue of Differential Item Functioning (DIF) in testing is currently appearing in public, including courts of law, and decisions are being made that have an impact on critical issues such as who shall be educated and who shall be employed (Bundsgaard, 2019; Yavuz Temel, 2023).

DIF occurs when test takers of the same ability are given the same items to respond to, and the items turn out to be more difficult for one group than the other group after the overall differences in knowledge of the subject tested are considered (Chalmers, Counsell, & Flora, 2016; Song, Gadermann, Zumbo, & Richardson, 2022). In educational and psychological testing, DIF means that the probability of a correct response among equal ability test takers differs for various racial, ethnic, gender or subgroups. It occurs when people from different groups with the same ability systematically respond differently to specific test items. In this case, examinees of equal ability in different groups, such as ethnicity, gender, and race, have different probabilities of correct responses to the items in a test (Warne, Yoon, & Price, 2014). DIF can also be seen as a collection of statistical methods utilized to determine if examination items are appropriate and fair for testing the knowledge or ability of different groups of examinees. It can also be explained as the statistical methodology for determining whether item bias creates an unfair advantage due to race, ethnicity, religion, gender and culture (Bauer, 2023; Effiom, 2021; Wallin, Chen, & Moustaki, 2024). However, an item displays DIF when the difficulty parameter (b), the discrimination parameter (a) or the lower asymptote parameter (c) estimated by the item differ across groups. Thus, when one or more item parameters differ across groups, then the item is said to display DIF. It indicates diverse behaviour by an item, showing evidence that people from different groups (such as gender and ethnicity) with the same latent trait have a different probability of answering the item correctly.

DIF in items is a significant threat to the validity of the instruments that measure the traits of members from different populations or groups because instruments containing such items may reduce validity for between-group comparisons (Jumadi, Sukarelawan, & Kuswanto, 2023; Opesemowo, Ayanwale, Opesemowo, & Afolabi, 2023). Their scores may indicate attributes other than those the scale intends to measure (Thissen, Steinberg, & Wainer, 2013). To determine whether the differences in performance of an item between

two demographic groups are due to differences in ability or some form of unfairness in the item is a more complex task for a polytomous item because of its many score categories than for a dichotomous item (Kuzu & Gelbal, 2023). DIF exists within all score categories or specific subsets of score categories within the item; hence, testing for DIF at each score level is required (Kristjansson, Aylesworth, Mcdowell, & Zumbo, 2005; Wallin et al., 2024).

Examination bodies often carry out empirical verification to detect DIF in their respective examinations, redeem and exclude items found to be biased so that all the examinees can be assured of equity in the examination and ensure that examinees' abilities are reliably assessed. Examination bodies are expected to construct test items in such a manner that test items are free from writing errors such as wordiness, irrelevancy, offensiveness, and excessive stimulations so that when an inadequacy exists between groups' examination items scores, the disparity will be attributed to true differences in whatever the test purports to measure in the examinees (Awopeju, Afolabi, & Opesemowo, 2017; Faleye & Dibu-Ojerinde, 2006; Ojerinde, Popoola, Onyeneho, & Egberongbe, 2016). Similarly, reasons were given for the occurrence of DIF in public examinations such as the NECO SSCE. It was based on the belief that DIF comes into testing due to the language used, which is not too clear; the questions raised, which are not direct to the point, are too wordy and ambiguous. In addition, test items displaying DIF have an item structure for a differentially complex format for subgroups of examinees. In this case, test takers lack the knowledge and skills of test items and are not familiar with the content of the test items (Adedoyin, 2010; Jimoh, Opesemowo, & Ferami, 2022).

In support of the forgoing discussion, the following reasons were also adduced for test items having DIF effects, and they include the unidimensionality assumption (Opesemowo et al., 2023) and the fit of the data, especially when the validity of the items was only marginally met, as reflected in the test construction procedures (Akindele, 2003). Secondly, some of the items tend to be complex and ambiguous. These factors might have led to a greater chance of translation error on the part of the examinees. The third reason is identifying items with significant DIF that may be related to sample size. Studies with relatively small sample sizes were identified as having significant DIF (Kaya, Leite, & Miller, 2016; Lu et al., 2022; S. Zhou & Shen, 2022). Some studies (Adeyemo & Opesemowo, 2020; Chankseliani, Gorgodze, Janashia, & Kurakbayev, 2020; Española, 2022) have highlighted differentials in students' performance in rural and urban areas. The differentials in their performance are sometimes predicated on the level of exposure of the two groups of test-takers in terms of their experience in practical agriculture.

In the rural areas, because of the agrarian economy that is prevalent therein, most students are involved in agricultural practices as they accompany their parents to the farms during the holidays. These students tend to have practical insight into Agricultural Science at school when taught, which past studies have not explored. When questions are asked, they quickly understand and can respond well to the questions raised. Unlike the students who reside in urban areas and are not exposed to agricultural practicals except the ones taught in school, they will not be able to apply the knowledge gained from practical agriculture to theoretical Agricultural Science. This, in turn, makes the questions not favor test-takers in

urban areas but rather favor test-takers from rural areas. This study aims to analyze DIF in the Agricultural Science examination of Southwestern Nigeria's senior schools. The specific objectives are to analyze DIF in the Senior School Certificate Agricultural Science Examination regarding school location using one, two and three Parameter Logistic Models (PLM) and determine the magnitude of bias/DIF across school locations. However, the study was guided by the following research questions. Which items showed DIF across school locations using one, two, and three PLMs? To what extent did the items show bias/DIF across school locations?

## Method

This study examines the occurrence of DIF across school locations in senior school certificate agricultural science examinations in Southwestern Nigeria. This section unveils the research design, research instrument, and data analysis. The research design used in the study was an ex-post-facto design. However, the student's responses to multiple-choice items in the Agricultural Science in the 2015 NECO Senior School Certificate Examination constituted the data for the study. The study sample covered three states in southwestern Nigeria. The study population comprised all students enrolled in Southwestern Nigeria's 2015 NECO Senior School Certificate Examination (SSCE). This group (students) comprised male and female senior secondary school students in class three (SSS 3), with their school locations either rural or urban. The sample size for the study comprised 818 senior SSS 3 students in Southwestern Nigeria who were prepared to write the NECO SSCE examination. These samples were selected using purposive sampling, focusing on the examinees' school location.

The instrument for the study is the NECO Senior School Certificate Agricultural Science Examination (SSCASE) 2015. The examination consists of 60 multiple-choice items, each with five options, lettered A-E, from which the test takers were to indicate the correct option for each item. The reliability and validity of the instrument were determined using the Cronbach Alpha coefficient of 0.887.

The students' responses to each item were dichotomously scored right or wrong, and the total of each student's scores was coded for correct responses on the dependent variable as 1. In contrast, incorrect responses were coded as 0. For the independent variables, rural is 1, and urban is 2. The Mantel-Haenzel method of analysis DIF in Xcaliber 4.2.2 statistical software was utilized to establish the 1PLM, 2PLM and 3PLM (Ojerinde & Ajeigbe, 2021) as well as answering the research questions.

## Results and Discussion

The analysis results addressed the research questions by providing valuable insight into the topic under investigation.

*Research Question 1: Which Items showed DIF by school locations using one, two, and three PLMs?*

To answer this question, the scores of 818 examinees in the Agricultural Science examination conducted by NECO from three states (Ondo, Osun and Oyo states) were initially

calibrated and estimated using X-caliber 4.2.2.0. The final version of the calibrated analysis was used to establish the 1PLM, 2PLM and 3PLM for detecting DIF across school locations.

**Table 1.** Summary Statistics for all Calibrated Items IPLM

| Parameter | Items | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| *b* | 60 | 0 | 1 | -1.654 | 2.8217 |

Table 1 shows the summary for all calibrated items across school locations. It has a mean of 0.00 and standard deviation of 1.00. The minimum value is -1.654, and the maximum is 2.8217. From the table above, the model did not consider other parameters except the difficulty of the test items.

**Table 2.** Summary Statistics for Total Scores

| Test | Items | Alpha | Mean | SD | Skew | Min | Q1 | Median | Q3 | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Full Text | 60 | 0.8873 | 23.911 | 10.07 | -0.029 | 1 | 14 | 26 | 32 | 43 | 18 |

Table 2 shows the summary statistics for the 1PLM scores. The statistics gave an alpha value of 0.8873, which shows that the reliability and the internal consistency of the 60 items were high and good enough for the analysis. The mean of the 60 items was 23.911. It has a standard deviation of 10.07. The table showed that the scores were moderately skewed (-0.0267). This implies that they are slightly skewed, and the values are asymmetrical.

**Table 3.** Summary Statistics for Theta Estimate for 1PLM

| Test | Examinees | Mean | SD | Skew | Min | Q1 | Median | Q3 | Max | 1QR |
|---|---|---|---|---|---|---|---|---|---|---|
| Full Text | 818 | -0.5837 | 0.8992 | -0.184 | -3.299 | -1.43 | -0.363 | 0.1195 | 1.088 | 1.545 |

Table 3 shows the summary statistics for the theta estimates for 1PLM. It has 818 examinees who responded to the 60 test items, with a mean of -0.5837 and a standard deviation of 0.899. It is slightly skewed, and theta is monotonically increasing.

**Table 4**. 1PLM in the detection of DIF across School Location

| **Number in Urban Group** | | **445** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Number in Rural Group** | | **373** | | | | | | | |
| **S/N** | **Item ID** | **M-H** | **M-HD** | **M-HSE** | **z-Test** | **P** | **Bias Against** | **Theta 1 Odds-Ratio** | **Theta 2 Odds-Ratio** |
| 1 | ITEM 01 | 0.9825 | 0.0414 | 0.316 | 0.0558 | 0.9555 | | 2.6006 | 0.357 |
| 2+ | ITEM 02 | 0.24 | 3.3533 | 0.292 | 4.895 | 0* | U | 0.4197 | 0.124* |
| 3 | ITEM 03 | 0.9988 | 0.0043 | 0.254 | 0.0071 | 0.9943 | | 0.4287 | 1.496 |
| 4- | ITEM 04 | 2.38 | -2.0376 | 0.293 | -2.964 | 0.003* | R | 0.5746 | 4.204** |
| 5 | ITEM 05 | 0.9222 | 0.1903 | 0.27 | 0.3003 | 0.764 | | 0.4678 | 1.209 |
| + | + | + | + | + | + | + | + | + | + |
| + | + | + | + | + | + | + | + | + | + |
| 56+ | ITEM 56 | 0.3148 | 2.7161 | 0.266 | 4.344 | 0* | U | 0.2703 | 0.349* |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 57 | ITEM 57 | 1.2707 | -0.5629 | 0.268 | -0.893 | 0.3719 | 1.4396 | 1.237 |
| 58 | ITEM 58 | 0.757 | 0.6541 | 0.265 | 1.0524 | 0.2926 | 0.3157 | 1.047 |
| 59 | ITEM 59 | 0.7015 | 0.8333 | 0.412 | 0.8617 | 0.3888 | 0.6968 | 0.707 |
| 60 | ITEM 60 | 0.7174 | 0.7805 | 0.413 | 0.8049 | 0.4209 | 0.2493 | 0.956 |

U + = Urban, R- = Rural, *P<0.05 and it odd ratio <1.00, ** odd ratio > 1.00

Table 4 shows the test items in the Mantel Haenszel coefficient (M-H) column; in the M-HD column, items whose values are positive and have their p-value be less than 0.05 with their odd ratio to be less than 1.00 are endorsed for bias and are grouped as the urban bias group. Those items whose values are under the M-HD with negative values and have their p-values less than 0.05 but have an odd ratio greater than 1.00 are endorsed for bias and are grouped as the rural bias group.

In Table 4, 1PLM detected 30 items (50%) out of 60 as biased since their p-values were less than 0.05. Such items included items 2, 6, 8, 14, 17, 20, 21, 25, 28, 29, 35, 38, 40, 42, 43, 45, 47, 48, 49, 53 and 56, which were biased against urban schools since the odd ratio for these items was less than 1.00. The rural schools also endorsed nine items as bias items. The items are 4, 9, 22, 27, 36, 44, 46, 51 and 55. These items were said to be biased because they have their p-values less than 0.05 and their odd ratio values greater than 1.00. It can be concluded that the nine items indicate that the rural group was more likely to be correctly endorsed than the urban group, and so the items are biased against rural schools. However, item 23 was not included as a bias item against the rural school location because it has a value of 1.96 and is at the undecided boundary.

**Table 5**. Summary Statistics for all Calibrated Items for 2PLM

| Parameter | Items | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| *a* | 60 | 1.0948 | 0.625 | 0.1697 | 2.4631 |
| *b* | 60 | 0.9501 | -1.4167 | -1.037 | 3.6448 |

Table 5 shows the summary statistics for all calibrated items across school locations for 2PLM. The *a*-parameter (discrimination) has a mean of 1.0948 with a standard deviation of 0.1697. In contrast, the *b*-parameter, also known as the difficulty, has a mean of 0.9501 with a standard deviation of -1.4167. The parameter has a minimum of 0.1697 and a maximum of 2.4631. At the same time, the *b* parameter has a minimum calibrated item of -1.0366 and a maximum calibrated item of 3.644. The table shows that the model considered the test items' difficulty and discrimination parameters.

**Table 6**. Summary Statistics for Total Scores for 2PLM

| Test | Items | Alpha | Mean | SD | Skew | Min | Q1 | Median | Q3 | Max | 1QR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Full Text | 60 | 0.8873 | 23.911 | 10.075 | -0.029 | 1 | 14 | 26 | 32 | 43 | 18 |

Table 6 demonstrates the summary statistics for the 2PLM scores. The statistics gave the alpha value of 0.887, which shows that the reliability and internal consistency of the 60 items were high. The mean of the 60 items was 23.911, with a standard deviation of 10.075.

The table showed that the scores were moderately skewed (-0.029). This implies that they are slightly skewed, and the values are asymmetrical.

**Table 7**. Summary Statistics for Theta Estimates for 2PLM

| Test | Examinees | Mean | SD | Skew | Min | Q1 | Median | Q3 | Max | 1QR |
|---|---|---|---|---|---|---|---|---|---|---|
| Full Text | 818 | -0.583 | 0.899 | -0.184 | -3.299 | -1.425 | -0.363 | 0.1195 | 1.088 | 1.545 |

Table 7 shows the summary statistics for the theta estimates for 2PLM. It has 818 examinees who responded to the 60 test items, with a mean of -0.583 and standard deviation of 0.899, and it is slightly skewed with a value of -0.184.

**Table 8**. 2PLM in the detection of DIF across School Location

| **Number in Urban Group** | | **445** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Number in Rural Group** | | **373** | | | | | | | |
| **S/N** | **Item ID** | **M-H** | **M-HD** | **M-HSE** | **z-Test** | **P** | **Bias Against** | **Theta 1 Odds-Ratio** | **Theta 2 Odds-Ratio** |
| 1 | ITEM 01 | 0.9301 | 0.173 | 0.32 | 0.2264 | 0.82 | | 2.201 | 0.409 |
| 2+ | ITEM 02 | 0.2714 | 3.0649 | 0.3009 | 4.334 | 0* | U | 0.448 | 0.158* |
| 3 | ITEM 03 | 0.9817 | 0.0433 | 0.255 | 0.0724 | 0.94 | | 0.396 | 1.457 |
| 4- | ITEM 04 | 2.3959 | -2.053 | 0.2989 | -2.923 | 0* | R | 0.671 | 3.805** |
| 5 | ITEM 05 | 0.9819 | 0.0429 | 0.2773 | 0.0658 | 0.95 | | 0.596 | 1.172 |
| + | + | + | + | + | + | + | + | + | + |
| + | + | + | + | + | + | + | + | + | + |
| 56+ | ITEM 56 | 0.3314 | 2.5957 | 0.2706 | 4.0814 | 0* | U | 0.329 | 0.333* |
| 57 | ITEM 57 | 1.3201 | -0.653 | 0.2729 | -1.018 | 0.31 | | 1.603 | 1.272 |
| 58 | ITEM 58 | 0.7588 | 0.6486 | 0.2693 | 1.025 | 0.31 | | 0.346 | 0.993 |
| 59 | ITEM 59 | 0.6323 | 1.0773 | 0.4174 | 1.0981 | 0.27 | | 0.625 | 0.641 |
| 60 | ITEM 60 | 0.7664 | 06252 | 0.4123 | 0.6453 | 0.52 | | 0.201 | 1.068 |

U +=Urban, R- = Rural, *P<0.05 and * odd ratio <1.00, ** odd ratio > 1.00

Table 8 shows the 2PLM for detecting bias items. 27 (45%) items were flagged as biased since their p-values were less than 0.05. eighteen items were flagged as biased items against urban schools. The flagged items against urban schools are items 2, 6, 8, 14, 17, 21, 25, 28, 29, 35, 38, 40, 42, 43, 45, 48, 49 and 56. This is because their odd ratio was less than 1.00, and their p-values less than 0.05. Similarly, nine items were flagged as biased against the rural schools because their odd ratio was greater than 1.00 and had p-values less than 0.05. The nine items flagged as biased against rural schools are as follows: items 4, 9, 22, 27, 36, 44, 46, 51 and 55 respectively.

**Table 9**. Summary Statistics for all Calibrated Items for 3PLM

| Parameter | Items | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| *a* | 60 | 2.692 | 1.1148 | 0.5501 | 6 |
| *b* | 60 | 1.042 | 0.9114 | -0.119 | 2.5053 |
| *c* | 60 | 0.215 | 0.0575 | 0.1201 | 0.3809 |

Table 9 shows the mean scores for the 3PLM for school location were 2.69, 1.042 and 0.215, respectively. The 3PLM standard deviations are 1.1148, 0.9114 and 0.0575. This parameter considered the difficulty, discrimination, and guessing parameters.

**Table 10**. Summary Statistics for Total Scores for 3PLM

| Test | Items | Alpha | Mean | SD | Skew | Min | Q1 | Median | Q3 | Max | 1QR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Full Text | 60 | 0.887 | 23.911 | 10.075 | -0.0287 | 1 | 14 | 26 | 32 | 43 | 18 |

Table 10 shows the summary statistics for the 3PLM scores. The statistics gave an alpha value of 0.887, which shows that the reliability and internal consistency of the 60 items were high. The mean of the 60 items was 23.911, and the standard deviation of 10.074. This implies that the scores obtained by the test takers were close. The table showed that the scores were negatively skewed (-0.0287). This means that they are moderately skewed, and the values are asymmetrical.

**Table 11**. Summary Statistics for Theta Estimates for 3PLM

| Test | Examinees | Mean | SD | Skew | Min | Q1 | Median | Q3 | Max | 1QR |
|---|---|---|---|---|---|---|---|---|---|---|
| Full Text | 818 | -0.01 | 1.0023 | 0.834 | -7 | -1.16 | 0.314 | 0.643 | 1.638 | 1.798 |

Table 11 displays the summary statistics for the theta estimates for the 3PLM. It has 818 examinees who responded to the 60 test items, with a mean of -0.0159 and a standard deviation of 1.0023. The standard deviation is 1.00 and is negatively skewed.

**Table 12**. 3PLM in the detection of DIF across School Location

| **Number in Urban Group** | | **445** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Number in Rural Group** | | **373** | | | | | | | |
| **S/N** | **Item ID** | **M-H** | **M-HD** | **M-HSE** | **z-Test** | **P** | **Bias Against** | **Theta 1 Odds-Ratio** | **Theta 2 Odds-Ratio** |
| 1 | ITEM 01 | 0.957 | 0.1022 | 0.3163 | 0.1375 | 0.8906 | | 1.97 | 0.43 |
| 2+ | ITEM 02 | 0.241 | 3.3412 | 0.2926 | 4.8594 | 0* | U | 0.351 | 0.156* |
| 3 | ITEM 03 | 0.923 | 0.1893 | 0.246 | 0.3275 | 0.7433 | | 0.287 | 1.594 |
| 4- | ITEM 04 | 2.116 | -1.762 | 0.2831 | -2.6478 | 0.008* | R | 0.475 | 4.124** |
| 5 | ITEM 05 | 0.864 | 0.3444 | 0.2638 | 0.5555 | 0.5785 | | 0.456 | 1.119 |
| + | + | + | + | + | + | + | + | + | + |
| + | + | + | + | + | + | + | + | + | + |
| 56 | ITEM 56 | 0.306 | 2.7842 | 0.2646 | 4.4782 | 0 | U | 0.29 | 0.317* |
| 57 | ITEM 57 | 1.284 | -0.588 | 0.2708 | -0.9239 | 0.3555 | | 2.377 | 1.146 |
| 58 | ITEM 58 | 0.71 | 0.8036 | 0.2606 | 1.3124 | 0.1894 | | 0.254 | 1.078 |
| 59 | ITEM 59 | 0.582 | 1.2706 | 0.4142 | 1.3055 | 0.1919 | | 0.671 | 0.465 |
| 60 | ITEM 60 | 0.681 | 0.9038 | 0.4143 | 0.9283 | 0.3532 | | 0.28 | 0.881 |

U+=Urban, R- = Rural, *P<0.05 and * odd ratio <1.00, ** odd ratio > 1.00

Table 12 showed that the 3PLM detected 28 (46.7%) items flagged as biased against rural and urban schools with p-values of less than 0.05. Eighteen items were flagged as biased

against the urban school because they have p-values of less than 0.05 and an odd ratio of less than 1.00. The eighteen items that flagged bias in the urban schools are as follows: items 2, 6, 8, 14, 17, 21, 25, 28, 29, 35, 38, 40, 42, 43, 45, 48, 49 and 56. On the other hand, ten items, such as items 4, 9, 22, 23, 27, 36, 44, 46, 51 and 55, flagged bias against rural schools since their p-values are less than 0.05 but have their odd ratio to be greater than 1.00. The ten items indicate that the focal group was more likely to endorse the items correctly.

*Research Question 2: To what extent did the items show bias/DIF across school locations?*

To answer this question, the extent of bias/DIF items was determined using the methods' three Item Response Theory (IRT) parameter models.

**Table 13**. The extent of biased/DIF items detected by the various IRT-parameter models in SSCASE

| Methods | The extent of bias/DIF items detected by the various IRT-parameter models |
|---|---|
| 1PLM | 0.52 Large DIF/BIAS |
| 2PLM | 0.45 Moderate DIF/BIAS |
| 3PLM | 0.47 Moderate DIF/BIAS |

Table 13 reveals the three IRT parameter models that detect bias/DIF. When 1PLM was used to determine the extent of error in school location, 1PLM showed a large bias/DIF at a value of 0.52; this indicated that the extent of bias/DIF in the test items for school location was enormous. In the same vein, the 2PLM was also used to determine the extent of bias/DIF in school locations, and it showed that the extent of bias/DIF was moderate at a value of 0.45. The 3PLM was used to determine the extent of bias/DIF in the test items, which has a value of 0.47. It implies that the presence of bias/DIF by school location was moderate.

In IRT, item parameters are mathematical models that describe the performance or behaviour of each test item in a test, influencing how individuals interact with those items based on their latent abilities. It specifies the probability of a discrete outcome, such as a correct response to an item, in terms of persons and item parameters. The three kinds of IRT models based on the parameters used are difficulty, discrimination, and the guessing parameter. Item parameters include 1PLM, 2PLM and 3PLM. The 1PLM is also the b parameter that considers the difficulty of the test items only, 2PLM considers both the difficulty of the items and the discrimination of the test items, and 3PLM puts both the difficulty, discrimination of the test items and guessing into consideration. In this study, Xcaliber 4.2.2.0 was used for the estimation and calibration of 60 test items, which establish the 1PLM, 2PLM and 3PLM for the final analysis in the detection of bias/DIF in the Agricultural Science examination conducted by NECO across school location.

Investigating the occurrence of DIF across school locations in senior school certificate Agricultural Science examination in Southwestern Nigeria conducted by NECO using 1PLM, 2PLM and 3PLM were used to detect bias/DIF and the magnitude items exhibit DIF. The findings showed that there was bias/DIF. Specifically, 1PLM showed 30 items flagging bias/DIF. Using 1PLM, the results showed that the test items were tricky for the test takers who could not answer correctly. Some test takers with an average trait level or response ability have a 50/50 chance of answering the questions correctly. The result aligns with (Jumadi et al., 2023), a 50/50 chance of answering the questions correctly. The result aligns

with Jumadi et al. (2023) ), who reported item bias in the four-tier heat and temperature diagnostic test (4T-HTDT), and it was stated that 35% of the items (i.e., 7 out of 20) across five concept groups showed bias when evaluated using the Rasch model. Conversely, for an item difficulty of zero, an individual with a high trait level has a higher chance of answering the item correctly. Similarly, an individual with a low trait level is less likely to answer the item correctly (Y. Zhou & Jia, 2023).

This study also uncovered that the 2PLM flagged 27 items as displaying bias/DIF. This model took into consideration both the difficulty and discrimination parameters. From the findings, items with M-HD having positive values indicated that the items were tricky, and low-ability test takers could not answer the test items correctly. These items were endorsed for bias against urban school locations. This finding supports the result of Perez and Loken (2023), whose report shows that utilizing standard fitting algorithms for 2PL IRT could significantly improve the accuracy of discrimination parameter estimates. Researchers can obtain more precise and reliable item discrimination estimates by incorporating these more advanced techniques into the fitting process. This leads to better overall mode fit and more accurate inferences about individual abilities.

In addition, this advanced algorithm can help identify and address potential sources of misfit at both the person and item levels, ultimately improving the overall quality of the IRT model. Similarly, items with M-HD having negative values showed that the items were too easy for the test takers to answer. This means even test takers with low ability could answer the items correctly. Conversely, 2PLM is very vital in detecting both difficulty and discrimination of the test items since the discrimination parameter is essential for adaptive testing (Aybek, 2023; Kishida, Fuchimoto, Miyazawa, & Ueno, 2023), a popular method of administering test items because some test items are tailored to the individual's ability. The item discrimination parameter can depict an item's effectiveness in discriminating between poor and good test takers. The more discrimination of an item, the less the range of item difficulty, and there will also be a steeper curve in items of the item characteristic curve (Metsämuuronen, 2023; Sweeney, Sinharay, Johnson, & Steinhauer, 2022).

The 3PLM showed 28 items flagging bias/DIF as it added the guessing parameter to the 2PLM. Despite some examinees performing poorly, 3PLM acknowledges that some will respond correctly (Omale, Dike, & Chibundum, 2023). This model can take on values different from random guessing on a multiple-choice test, indicating that the individuals demonstrate some understanding or knowledge of the subject matter. The causes of DIF were assumed to be too much wordiness of test items or ambiguity of the test items on the part of the constructors. Similarly, some items were too complex and vague for the test takers from the various school locations who responded to the items. All this could have led to a greater chance of translation error on the part of the test takers. Another reason for the display of DIF in the examination was a result of the unidimensionality assumption and the fits of data, especially when the validity of the test items was only marginally met, reflected in the construction procedure (Martin, Tarantino, & Levy, 2023). Another reason for the display of bias/DIF in the NECO SSCE 2015 could be attributed to poor teaching of agricultural science at different school locations where the examinees are located.

Based on the magnitude of DIF in the Agricultural Science examination conducted by NECO 2015 for location, the result revealed that DIF was found at various magnitudes or degrees in the test items. When the 1PLM was used to determine the magnitude of DIF, findings revealed that the magnitude of DIF was significant, which shows that the extent of bias in the test items was very high. However, this finding aligns with the reports of Joo, Ali, Robin, and Shin (2022); Khoeruroh and Retnawati (2020), revealing that when using 1PLM to assess the extent of DIF, research findings indicated a substantial magnitude of DIF, signifying a high level of bias present in the test items. Similarly, the outcome was moderate when 2PLM and 3PLM were used to verify the presence of bias/DIF in the test items. The study of Jodoin and Gierl (2001) supported this finding and classified the magnitude of DIF into three levels when they used logistic regression to detect DIF. The classification levels are as follows: A level, also known as the negligible DIF; B level, the moderate DIF; and C level, the large or high DIF. Also, Roussos and Stout (1996) classified the magnitude of DIF into three levels: negligible, moderate, and high. Negligible DIF occurs when the impact of DIF on test scores is minimal and does not significantly affect the validity of the test. Moderate DIF indicates a moderate level of bias in the test items, which may need further investigation and potential adjustment. High DIF suggests a significant impact on test scores, including a need for thorough analysis and revision of the test items to ensure fairness and accuracy in assessment outcomes.

## Conclusion

This study employed IRT models (i.e., 1PLM, 2PLM, and 3PLM) to analyze DIF in the Agricultural Science examination of Southwestern Nigeria's senior schools. Specifically, 1PLM identified 30 items with bias/DIF, indicating difficult items for some examinees. While the 2PLM flagged 27 items, considering both item difficulty and discrimination, highlighting items biased against urban schools. Steps should be taken to address any potential bias that may be present in the test items to create a more equitable testing experience for all examinees, irrespective of their location. The 3PLM included a guessing parameter that identified 28 biased items, recognizing the role of guessing in test performance. Finally, the study then concludes that the findings undermine the importance of fairness and validity of the Agricultural Science examination.

This study was conducted following professional ethical guidelines. The guidelines include obtaining informed consent from participants, maintaining ethical treatment, and respecting the fundamental human rights of the participants. Subsequently, we ensure that participants' identities and data are kept private. These guidelines ensure that individual participants are not identified in published or archival results.

The first limitation of this study is that it focuses on the participants' school location and does not consider other factors that may influence their test performance. Future research should consider other demographic variables such as gender and school type. By expanding the scope of the study to include these factors, researchers can gain a more comprehensive understanding of the various elements that contribute to test performance. Additionally, exploring how these different variables interact could provide valuable insights

for developing more effective testing practices and strategies. Second, the study utilizes the Mantel-Haenzel method to determine the magnitude of DIF.

Further studies should be conducted using other methods of DIF detection, including Rasch analysis, Simultaneous Item Bias Test (SIBTEST), and logistic regression analysis, to validate the results obtained from the Mantel-Haenzel method. Lastly, dichotomous data was used for the analysis in this research, but further studies could benefit from incorporating polytomous data to provide a more comprehensive understanding of DIF. By combining different data types, researchers can better understand how variables impact test performance and develop more effective testing practices. Including various analyses and data types will help to ensure the validity and reliability of the results obtained in studies of this nature.

**Acknowledgment**

We sincerely appreciate all our participants who made themselves available for data gathering in this research.

**References**

Adedoyin, O. O. (2010). Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories. *International Journal of Educational Sciences, 2*(2), 107-113. doi:10.1080/09751122.2010.11889987

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.

Akindele, B. (2003). The development of an item bank for selection tests into Nigerian universities: An exploratory study. Unpublished Doctoral Dissertation). Nigeria: University of Ibadan.

Awopeju, O., Afolabi, E., & Opesemowo, O. (2017). An Investigation of Invariance Properties of One, Two and Three Parameter Logistic Item Response Theory Models. *Bulgarian Journal of Science and Education Policy, 11*(2), 197-219.

Aybek, E. C. (2023). The Relation of Item Difficulty Between Classical Test Theory and Item Response Theory: Computerized Adaptive Test Perspective. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 118-127. doi:10.21031/epod.1209284

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *British Journal of Mathematical and Statistical Psychology, 76*(3), 435-461. doi:10.1111/bmsp.12316

Bundsgaard, J. (2019). DIF as a pedagogical tool: analysis of item characteristics in ICILS to understand what students are struggling with. *Large-scale Assessments in Education, 7*(1), 9. doi:10.1186/s40536-019-0077-2

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It Might Not Make a Big DIF:Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement, 76*(1), 114-140. doi:10.1177/0013164415584576

Chankseliani, M., Gorgodze, S., Janashia, S., & Kurakbayev, K. (2020). Rural disadvantage in the context of centralized university admissions: a multiple case study of Georgia and Kazakhstan. *Compare: A Journal of Comparative and International Education, 50*(7), 995-1013. doi:10.1080/03057925.2020.1761294

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. doi:10.4314/gjedr.v20i1.6

Española, R. P. (2022, 5th-7th March 2022). Examining Gender and Urban/Rural School Differences in Empirically-derived Achievement Profiles. Paper presented at the 17th Education and Development Conference

Faleye, B. A., & Dibu-Ojerinde, O. O. (2006). A Review of the Enrolment and Performance of Male and Female Students in Education / Economics Programme of Obafemi Awolowo University, Ile-Ife, Nigeria. *Journal of Social Sciences, 12*(2), 143-146. doi:10.1080/09718923.2006.11978383

Faremi, Y. A., & Jimoh, K. (2022). Differential Item Functioning and Implications for Testing in Nigeria Education System. *Indonesian Journal of Learning Education and Counseling, 5*(1), 1-10. doi:10.31960/ijolec.v5i1.1689

Jang, Y., Pashler, H., & Huber, D. E. (2014). Manipulations of choice familiarity in multiple-choice testing support a retrieval practice account of the testing effect. *Journal of Educational Psychology, 106*(2), 435-447. doi:10.1037/a0035715

Jerome, A. (2023). Public Examinations in Nigeria: Challenges and Prospects in the 21st Century. *Glob Acad J Humanit Soc Sci, 5*(1), 10-15. doi:10.36348/gajhss.2023.v05i01.003

Jimoh, K., Opesemowo, O. A., & Faremi, Y. A. (2022). Psychometric Analysis of Senior Secondary School Certificate Examination (SSCE) 2017 Neco English Language Multiple Choice Test Items in Kwara State Using Item Response Theory. *Journal of Applied Research and Multidisciplinary Studies, 3*(2), 1-19. https://doi.org/10.32350/jarms.32.01

Jodoin, M. G., & 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. doi:10.1207/S15324818AME1404_2

Joo, S., Ali, U., Robin, F., & Shin, H. J. (2022). Impact of differential item functioning on group score reporting in the context of large-scale assessments. *Large-scale Assessments in Education, 10*(1), 18. doi:10.1186/s40536-022-00135-7

Jumadi, J., Sukarelawan, M. I., & Kuswanto, H. (2023). An investigation of item bias in the four-tier diagnostic test using Rasch model. *International Journal of Evaluation and Research in Education (IJERE), 12*(2), 622-629. doi:10.11591/ijere.v12i2.22845

Kaya, Y., Leite, W. L., & Miller, M. D. (2016). A comparison of logistic regression models for DIF detection in polytomous items: the effect of small sample sizes and non-normality of ability distributions. *International Journal of Assessment Tools in Education, 2*(1), 22-39. doi:10.21449/ijate.239563

Khoeruroh, U., & Retnawati, H. (2020). Comparison sensitivity of the differential item function (DIF) detection method. *Journal of Physics: Conference Series, 1511*(1), 012042. doi:10.1088/1742-6596/1511/1/012042

Kishida, W., Fuchimoto, K., Miyazawa, Y., & Ueno, M. (2023). Item Difficulty Constrained Uniform Adaptive Testing, Cham.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A Comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items.

*Educational and Psychological Measurement, 65*(6), 935-953. doi:10.1177/0013164405275668

Kuzu, Y., & Gelbal, S. (2023). Investigation of differential item and step functioning procedures in polytomus items. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 200-221. doi:10.21031/epod.1221823

Lu, L., Phua, Q. S., Bacchi, S., Goh, R., Gupta, A. K., Kovoor, J. G., . . . To, M.-S. (2022). Small Study Effects in Diagnostic Imaging Accuracy: A Meta-Analysis. *JAMA Network Open, 5*(8), e2228776-e2228776. doi:10.1001/jamanetworkopen.2022.28776

Martin, J. A., Tarantino, D. M., & Levy, K. N. (2023). Investigating gender-based differential item functioning on the McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD): An item response theory analysis. *Psychological Assessment, 35*(5), 462-468. doi:10.1037/pas0001229

Metsämuuronen, J. (2023). Seeking the real item difficulty: bias-corrected item difficulty and some consequences in Rasch and IRT modeling. *Behaviormetrika, 50*(1), 121-154. doi:10.1007/s41237-022-00169-9

Mitana, J. M. V., Muwagga, A. M., & Ssempala, C. (2019). The Influence National Examinations on Classroom Practice in Primary Schools in Uganda: Case of Kampala and Kabale Districts. *International Journal of Educational Research Review, 4*(3), 472-480. doi:10.24331/ijere.573954

Ojerinde, D., & Ajeigbe, T. O. (2021). Post-Test Estimates of Item Parameters of Mathematics Multiple-Choice Test of a Public Examination in Nigeria. *Journal of Emerging Trends in Educational Research and Policy Studies, 12*(6), 253-259. doi:10.10520/ejc-sl_jeteraps_v12_n6_a5

Ojerinde, D., Popoola, O., Onyeneho, P., & Egberongbe, A. (2016). A comparative analysis of pre-equating and post-equating in a large-scale assessment, high stakes examination. *Perspectives in Education, 34*(4), 79-98. doi:doi:10.18820/2519593X/pie.v34i4.6

Oluwatimilehin, T., & Opesemowo, O. (2019). *Investigating the prevalence of cheating*: LAP LAMBERT Academic Publishing.

Omale, O., Dike, G. A., & Chibundum, C. (2023). Assessment of Ability Estimate and Model Fit of Students in 2020 NECO Mathematics in Benue State, Nigeria. *Journal of Educational Research and Development, 10*, 43-52.

Opesemowo, O. A. G., Ayanwale, M. A., Opesemowo, T. R., & Afolabi, E. R. I. (2023). Differential Bundle Functioning of National Examinations Council Mathematics Test Items: An Exploratory Structural Equation Modelling Approach. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 1-18. doi:10.21031/epod.1142713

Perez, A. L., & Loken, E. (2023). Person Specific Parameter Heterogeneity in the 2PL IRT Model. *Multivariate Behavioral Research*, 1-7. doi:10.1080/00273171.2023.2224312

Roussos, L. A., & Stout, W. F. (1996). Simulation Studies of the Effects of Small Sample Size and Studied Item Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement, 33*(2), 215-230. doi:10.1111/j.1745-3984.1996.tb00490.x

Song, C., Gadermann, A., Zumbo, B., & Richardson, C. (2022). Differential Item Functioning of the Center for Epidemiologic Studies Depression Scale Among Chinese Adolescents. *Journal of Immigrant and Minority Health, 24*(3), 790-793. doi:10.1007/s10903-021-01275-8

Sweeney, S. M., Sinharay, S., Johnson, M. S., & Steinhauer, E. W. (2022). An Investigation of the Nature and Consequence of the Relationship between IRT Difficulty and

Discrimination. *Educational Measurement: Issues and Practice, 41*(4), 50-67. doi:10.1111/emip.12522

Thissen, D., Steinberg, L., & Wainer, H. (2013). Use of item response theory in the study of group differences in trace lines. In Test validity (pp. 147-169): Routledge.

Wallin, G., Chen, Y., & Moustaki, I. (2024). DIF Analysis with Unknown Groups and Anchor Items. *Psychometrika, 89*(1), 267-295. doi:10.1007/s11336-024-09948-7

Warne, R. T., Yoon, M., & Price, C. J. (2014). Exploring the various interpretations of "test bias". *Cultural Diversity and Ethnic Minority Psychology, 20*(4), 570-582. doi:10.1037/a0036503

Wiggins, B. L., Lily, L. S., Busch, C. A., Landys, M. M., Shlichta, J. G., Shi, T., & Ngwenyama, T. R. (2023). Public exams may decrease anxiety and facilitate deeper conceptual thinking. *Journal of STEM Education: Innovations and Research, 24*(2), 36-48.

Yavuz Temel, G. (2023). A Simulation and Empirical Study of Differential Test Functioning (DTF). *Psych, 5*(2), 478-496.

Yeng, E., Ali, C. A., & Adzifome, N. S. (2023). Relationship between familiarity and competency in integrating information and communication technology into mathematics instruction. *Contemporary Mathematics and Science Education, 4*(2), ep23022. doi:10.30935/conmaths/13405

Zhou, S., & Shen, C. (2022). Avoiding Definitive Conclusions in Meta-analysis of Heterogeneous Studies With Small Sample Sizes. *JAMA Otolaryngology–Head & Neck Surgery, 148*(11), 1003-1004. doi:10.1001/jamaoto.2022.2847

Zhou, Y., & Jia, N. (2023). The Impact of Item Difficulty on Judgment of Confidence—A Cross-Level Moderated Mediation Model. *Journal of Intelligence, 11*(6), 113. doi.org/10.3390/jintelligence11060113