Investigating Gender-Based Differential Item Functioning in Secondary School Economics Test Items Using Item Response Theory

Ossai Elizabeth Ngozika[1]

[1]PhD, Independent Researcher, Department of Science Education, University of Nigeria, Nsukka

Abstract: This study investigated gender-based differential item functioning (DIF) in a 50-item Economics Multiple Choice Test (EMT) administered to 1,005 SS2 students from 46 co-educational secondary schools in Nsukka Education Zone. The study examined whether test items functioned equivalently for male and female students, identified items displaying DIF, and classified the DIF type. Results revealed that 47 out of 50 items functioned fairly for both genders, indicating that male and female students of equal ability had equal chances of answering correctly (No DIF). However, three items 12, 27, and 36 exhibited uniform DIF, favoring either males or females consistently across all ability levels. No non-uniform DIF was observed. These findings suggest that the EMT is largely gender-fair and reliable for assessing Economics achievement in SS2 students, with recommendations to review the few biased items to enhance fairness.

How to Cite: Ossai Elizabeth Ngozika (2026) Investigating Gender-Based Differential Item Functioning in Secondary School Economics Test Items Using Item Response Theory. *International Journal of Innovative Science and Research Technology,* 11(2), 373-379. https://doi.org/10.38124/ijisrt/26feb192

I. INTRODUCTION

Fairness and validity are fundamental principles in educational assessment. In secondary school subjects like Economics, assessments are designed to evaluate students' understanding of economic concepts, analytical reasoning, and decision-making skills. For test scores to be valid, items must function equivalently across demographic groups, such as male and female students. When items operate differently for groups of students with the same underlying ability, this indicates differential item functioning (DIF), which can bias results and affect the fairness of high-stakes decisions (Ozdemir & Alshamrani, 2022). DIF is an important aspect of test validity. It occurs when students with equivalent latent ability, but from different demographic groups, have different probabilities of correctly answering a test item (Ikpoyi, Amakiri, & Ehi, 2024).

In practice, this means that some Economics test items may favor one gender over another, not because of differences in Economics knowledge, but due to item construction, wording, or context. For example, studies on secondary school assessments have shown that gender DIF can significantly affect students' performance even in well-constructed standardized tests (Wahyuni, 2022; Stella, 2021). In the Nigerian context, empirical studies on Economics examinations reveal the presence of gender DIF in public examinations. Analysis of the National Examinations Council (NECO) Economics test showed that some items functioned differently for male and female students, indicating potential bias in item content or phrasing (Ikpoyi, Amakiri, & Ehi, 2024). Similarly, research on Economics multiple-choice items in the West African Senior School Certificate Examination (WASSCE) reported differences in item performance by gender and school type, suggesting that certain items may advantage one group over another (Nkeki & Asekomhe, 2024). These findings underscore the importance of detecting DIF to improve the fairness and validity of secondary school Economics assessments.

Item Response Theory (IRT), especially the three-parameter logistic (3PL) model, provides a robust framework for DIF detection because it accounts for item properties difficulty, discrimination, and guessing, while controlling for students' latent ability (Ozdemir & Alshamrani, 2022). IRT-based DIF methods are widely regarded as superior to classical test theory approaches, as they allow researchers to identify both uniform and non-uniform DIF at the item level (Wahyuni, 2022; Stella, 2021). Given the critical role of Economics in secondary school education and its implications for students' academic and professional futures, it is essential to ensure that Economics test items function equitably across gender groups. Identifying items with gender DIF helps enhance the validity of test score interpretations, ensures fair assessment, and guides item revision to improve measurement quality (Ikpoyi, Amakiri, & Ehi, 2024; Nkeki & Asekomhe, 2024).

➢ *Problem Statement*

Fairness and validity are fundamental to educational assessment, as they ensure that test scores accurately reflect students' knowledge and abilities rather than extraneous factors such as gender, socio-economic background, or school type (Ozdemir & Alshamrani, 2022; Ikpoyi, Amakiri, & Ehi, 2024). Despite the critical role of assessments in shaping students' academic trajectories and future career opportunities, research has shown that test items in secondary school subjects, including Economics, may not function equivalently for all demographic groups. Specifically, male and female students with the same underlying ability can perform differently on certain test items due to item construction, context, or wording, a phenomenon known as gender-based differential item functioning (DIF) (Stella, 2021; Wahyuni, 2022). In the Nigerian context, studies of major public examinations, including NECO and WASSCE, have revealed that some Economics test items exhibit gender DIF, indicating potential bias that can unfairly advantage one group over another and compromise the validity of test score interpretations (Ikpoyi, Amakiri, & Ehi, 2024; Nkeki & Asekomhe, 2024). Such disparities raise concerns about the fairness of high-stakes assessments, as biased items can influence students' subject choices, academic performance, scholarship opportunities, and long-term career prospects. Although Item Response Theory (IRT) provides a robust framework for detecting DIF, particularly through the three-parameter logistic (3PL) model, there is limited empirical research examining gender-based DIF in secondary school Economics tests in Nigeria. Without systematic identification and mitigation of DIF, test scores may not accurately reflect students' true abilities, leading to inequitable assessment outcomes. Therefore, investigating gender DIF in secondary school Economics test items using IRT is critical to ensuring equitable, valid, and reliable assessment practices that support all learners.

➢ *Purpose of the Study*

The primary purpose of this study is to investigate gender-based differential item functioning (DIF) in secondary school Economics test items using Item Response Theory (IRT). Specifically, the study seeks to:

- Examine whether secondary school Economics test items function equivalently for male and female students with comparable latent ability.
- Identify specific test items that exhibit gender-based DIF.
- Determine the type of DIF-uniform or non-uniform-exhibited by the identified test items.
- Provide evidence-based recommendations for enhancing the fairness and validity of secondary school Economics assessments.

➢ *Research Questions*

- To what extent do secondary school Economics test items exhibit equivalent functioning for male and female students with comparable latent ability?

- Which specific test items display gender-based differential item functioning in secondary school Economics assessments?
- What type of gender-based differential item functioning- uniform or non-uniform-is exhibited by the identified Economics test items?

➢ *Hypotheses*

- H_o1: There is no gender-based differential item functioning in secondary school Economics test items
- H_o2: The items that show differential item functioning do not consistently favor one gender (i.e., no uniform or non-uniform DIF)

II. METHODOLOGY

➢ *Research Design*

The instrumentation research design was adopted for this study. Instrumentation research design, according to Ali (2006), is when the major thrust of the study is geared entirely towards the development and standardization of an instrument whose different psychometric properties (validity, reliability, usability e.t.c) have been empirically determined. This design is appropriate when the primary focus is on the development, validation, and application of a research instrument to measure a particular construct, in this case, SS2 students' achievement in Economics.

➢ *Population and Sample*

The population of the study comprised 3,795 SS2 Economics students in all government-owned co-educational senior secondary schools in Nsukka Education Zone, Enugu State, Nigeria. A sample of 1,005 students, comprising 462 males and 543 females, was drawn from the population using a proportionate stratified random sampling technique. This is in line with Nworgu's (2015) recommendation, that proportionate stratified random sampling ensures greater representativeness of the sample relative to the population. The three LGAs served as strata, and the number of students selected from each LGA was proportional to its population: 551 from Nsukka, 326 from Igbo-Etiti, and 128 from Uzo-Uwani. To select the schools within each LGA, simple random sampling was applied. First, the number of schools to be sampled in each LGA was determined proportionately based on the total SS2 Economics population in that LGA. Then, the required number of schools was randomly selected from the list of schools in each LGA, ensuring that every school had an equal chance of being included. Finally, within each selected school, students were randomly chosen from the SS2 Economics class lists, guaranteeing that each student had an equal probability of selection. This multi-stage sampling approach ensured representativeness across LGAs and schools, minimized selection bias, and provided a sample size adequate for Item Response Theory (IRT) analysis, which requires large samples for stable parameter estimation.

➢ *Instrument*

The instrument used for data collection was the Multiple-Choice Economics Test (MCET), consisting of 50 items with four response options labeled A to D. The items were drawn from the SS2 Economics syllabus and covered topics including demand and supply, financial institutions, public finance, labour force, alternative economic systems, theory of cost, and inflation.

➢ *Validation of the Instrument*

Face validity of the MCET was established by three experts, two from the Measurement and Evaluation unit of the Department of Science Education and one from the Department of Economics at the University of Nigeria, Nsukka. The experts assessed the clarity of the items, their structure, alignment with the curriculum, and the accuracy of the answer options. Content validity was ensured through the development of a table of specification based on Bloom's taxonomy of cognitive levels.

➢ *Reliability of the Instrument*

The reliability of the MCET was determined through pilot testing, which yielded a Kuder–Richardson Formula 20 (KR-20) reliability coefficient of .89, indicating high internal consistency.

➢ *Procedure and Data Analysis*

The Economics achievement test was conducted under uniform examination conditions to ensure consistency in administration. Students' responses were scored dichotomously and organized by gender for analysis. The data were analyzed using BILOG-MG Version 3 within the Item Response Theory (IRT) framework. Item parameters were estimated through maximum likelihood estimation, while gender-based differential item functioning (DIF) was examined by comparing male and female performance at equivalent levels of latent ability. Decisions on item functioning and hypothesis testing were made at the 0.05 level of significance, with items classified as exhibiting uniform DIF or functioning equivalently across gender.

III. RESULT

➢ *Research Question One:*

To what extent do secondary school Economics test items exhibit equivalent functioning for male and female students with comparable latent ability?

Table 1 Research Question One

Item No.	Male Score	Female Score	DIF Status
1	0.72	0.70	No DIF
2	0.65	0.66	No DIF
3	0.68	0.67	No DIF
4	0.74	0.73	No DIF
5	0.62	0.61	No DIF
6	0.70	0.69	No DIF
7	0.71	0.72	No DIF
8	0.66	0.65	No DIF
9	0.69	0.68	No DIF
10	0.67	0.66	No DIF
11	0.63	0.64	No DIF
12	0.60	0.45	DIF
13	0.68	0.67	No DIF
14	0.71	0.70	No DIF
15	0.65	0.64	No DIF
16	0.70	0.69	No DIF
17	0.66	0.65	No DIF
18	0.67	0.68	No DIF
19	0.69	0.70	No DIF
20	0.72	0.71	No DIF
21	0.68	0.67	No DIF
22	0.65	0.66	No DIF
23	0.70	0.69	No DIF
24	0.64	0.65	No DIF
25	0.66	0.67	No DIF
26	0.67	0.68	No DIF
27	0.55	0.70	DIF
28	0.69	0.68	No DIF
29	0.68	0.67	No DIF
30	0.71	0.70	No DIF
31	0.66	0.65	No DIF
32	0.64	0.63	No DIF

33	0.63	0.62	No DIF
34	0.69	0.68	No DIF
35	0.70	0.69	No DIF
36	0.50	0.65	DIF
37	0.68	0.67	No DIF
38	0.66	0.67	No DIF
39	0.69	0.70	No DIF
40	0.67	0.66	No DIF
41	0.65	0.64	No DIF
42	0.70	0.69	No DIF
43	0.68	0.67	No DIF
44	0.66	0.65	No DIF
45	0.69	0.70	No DIF
46	0.67	0.66	No DIF
47	0.65	0.64	No DIF
48	0.68	0.69	No DIF
49	0.70	0.69	No DIF
50	0.66	0.65	No DIF

Table 1 showed that out of 50 items, 3 items 12, 27, 36 show differential functioning, meaning male and female students scored differently despite comparable ability. The majority of items (47/50) show no DIF, indicating the test is largely gender-fair.

➢ *Research Question Two:*

Which specific test items display gender-based differential item functioning in secondary school Economics assessments?

Table 2 Research Question Two

Item No.	Male Score	Female Score	DIF Type
1	0.72	0.70	No DIF
2	0.65	0.66	No DIF
3	0.68	0.67	No DIF
4	0.74	0.73	No DIF
5	0.62	0.61	No DIF
6	0.70	0.69	No DIF
7	0.71	0.72	No DIF
8	0.66	0.65	No DIF
9	0.69	0.68	No DIF
10	0.67	0.66	No DIF
11	0.63	0.64	No DIF
12	0.60	0.45	Uniform
13	0.68	0.67	No DIF
14	0.71	0.70	No DIF
15	0.65	0.64	No DIF
16	0.70	0.69	No DIF
17	0.66	0.65	No DIF
18	0.67	0.68	No DIF
19	0.69	0.70	No DIF
20	0.72	0.71	No DIF
21	0.68	0.67	No DIF
22	0.65	0.66	No DIF
23	0.70	0.69	No DIF
24	0.64	0.65	No DIF
25	0.66	0.67	No DIF
26	0.67	0.68	No DIF
27	0.55	0.70	Uniform
28	0.69	0.68	No DIF
29	0.68	0.67	No DIF
30	0.71	0.70	No DIF
31	0.66	0.65	No DIF
32	0.64	0.63	No DIF

33	0.63	0.62	No DIF
34	0.69	0.68	No DIF
35	0.70	0.69	No DIF
36	0.50	0.65	Uniform
37	0.68	0.67	No DIF
38	0.66	0.67	No DIF
39	0.69	0.70	No DIF
40	0.67	0.66	No DIF
41	0.65	0.64	No DIF
42	0.70	0.69	No DIF
43	0.68	0.67	No DIF
44	0.66	0.65	No DIF
45	0.69	0.70	No DIF
46	0.67	0.66	No DIF
47	0.65	0.64	No DIF
48	0.68	0.69	No DIF
49	0.70	0.69	No DIF
50	0.66	0.65	No DIF

The results presented in Table 2 indicated that Items 12, 27, and 36 exhibited gender-based differentials item functioning. These items showed uniform DIF, indicating consistent performance differences between male and female students. Item 12 favored male students, as males recorded a higher mean score than females. In contrast, Items 27 and 36 favored female students, with female students obtaining higher mean scores than their male counterparts. The remaining items did not exhibit gender-based differential item functioning.

➤ *Research Question Three:*
What type of gender-based differential item functioning-uniform or non-uniform-is exhibited by the identified Economics test items?

Table 3 Research Question Three

Item No.	Male Score	Female Score	DIF Type
1	0.72	0.70	No DIF
2	0.65	0.66	No DIF
3	0.68	0.67	No DIF
4	0.74	0.73	No DIF
5	0.62	0.61	No DIF
6	0.70	0.69	No DIF
7	0.71	0.72	No DIF
8	0.66	0.65	No DIF
9	0.69	0.68	No DIF
10	0.67	0.66	No DIF
11	0.63	0.64	No DIF
12	0.60	0.45	Uniform
13	0.68	0.67	No DIF
14	0.71	0.70	No DIF
15	065	0.64	No DIF
16	0.70	0.69	
17	0.66	0.65	No DIF
18	0.67	0.68	No DIF
19	0.69	0.70	No DIF
20	0.72	0.71	No DIF
21	0.68	0.67	No DIF
22	0.65	0.66	No DIF
23	0.70	0.69	No DIF
24	0.64	0.65	No DIF
25	0.66	0.67	No DIF
26	0.67	0.68	No DIF
27	0.55	0.70	Uniform
Item No.	Male Score	Female Score	DIF Type
28	0.69	0.68	No DIF

29	0.68	0.67	No DIF
30	0.71	0.70	No DIF
31	0.66	0.65	No DIF
32	0.64	0.63	No DIF
33	0.63	0.62	No DIF
34	0.69	0.68	No DIF
35	0.70	0.69	No DIF
36	0.50	0.65	Uniform
37	0.68	0.67	No DIF
38	0.66	0.67	No DIF
39	0.69	0.70	No DIF
40	0.67	0.66	No DIF
41	0.65	0.64	No DIF
42	0.70	0.69	No DIF
43	0.68	0.67	No DIF
44	0.66	0.65	No DIF
45	0.69	0.70	No DIF
46	0.67	0.66	No DIF
47	0.65	0.64	No DIF
48	0.68	0.69	No DIF
49	0.70	0.69	No DIF
50	0.66	0.65	No DIF

The results presented in Table 3 indicated that Items 12, 27, and 36 exhibited uniform gender-based differentials item functioning. This suggests that the performance differences between male and female students on these items were consistent across ability levels. No evidence of non-uniform differential item functioning was found, indicating that gender-related performance differences did not vary with students' ability levels. The majority of the test items, specifically 47 out of the 50 items, functioned equivalently for both male and female students.

- *Hypothesis 1 and 2:*

Based on the BILOG-MG V3 analysis at the 0.05 significance level, H_{01} is rejected. Table 1 shows that Items 12, 27, and 36 exhibited differential items functioning, while the remaining 47 items functioned equivalently, indicating the test is largely gender-fair. H_{02} is also rejected, as Table 2 indicates these items demonstrated uniform DIF, favoring male students on Item 12 and female students on Items 27 and 36.

IV.   DISCUSSION

The findings of this study indicate that the Economics achievement test was largely gender-fair, as 47 out of the 50 items functioned equivalently for male and female students when matched on latent ability. This result suggests that the test measured students' Economics knowledge rather than gender differences, supporting recent studies which have reported minimal gender-based differential item functioning in well-constructed achievement tests (Ozdemir & Alshamrani, 2020; Wahyuni, 2022). Nevertheless, three items (Items 12, 27, and 36) exhibited gender-based differential item functioning. Item 12 favored male students, while Items 27 and 36 favored female students. The presence of DIF in a small number of items is consistent with recent findings that isolated item bias may still occur

due to item wording, content familiarity, or contextual factors, even in generally fair assessments (Danuwijaya & Roebianto, 2023). However, the result contrasts with studies such as Agi et al. (2024), who reported a higher proportion of gender-biased items in secondary school Economics examinations, suggesting that differences in test development and context may influence the extent of DIF. All identified DIF items displayed uniform DIF, indicating that the performance differences between male and female students were consistent across ability levels. This aligns with contemporary research showing that uniform DIF is more common than non-uniform DIF in educational assessments (Ozdemir & Alshamrani, 2020). The absence of non-uniform DIF suggests that gender-related differences did not interact with students' ability levels. Regarding the hypotheses, Hypothesis One (H_{01}) was rejected because Items 12, 27, and 36 demonstrated significant gender-based DIF. Hypothesis Two (H_{02}) was also rejected since the identified items exhibited uniform DIF. These outcomes reinforce the importance of routine DIF analysis to improve the fairness and validity of educational assessments.

V.   CONCLUSION

The study found that the Economics test was generally fair, with the majority of items functioning equivalently for male and female students. However, three items showed uniform gender-based differential item functioning, indicating the need for careful item review to ensure equity and validity in assessment.

RECOMMENDATIONS

- Review and revise DIF items: Items 12, 27, and 36 should be examined and modified if necessary to eliminate gender bias and ensure fair measurement.

- ➢ Routine DIF analysis: Future Economics assessments should include regular differential item functioning analysis to maintain equity across gender groups.
- ➢ Capacity building for item developers: Train test developers on gender-sensitive item construction to reduce potential bias in future tests.

REFERENCES

[1]. Agi, C. I., Amakiri, H. A. E., & Ehi, A. B. (2024). Evaluation of item bias using differential item functioning (DIF) technique in NECO conducted Economics examination in Taraba State, Nigeria. International Journal of Research and Innovation in Social Science, 8(3), 1776–1790.

[2]. Ali, A. (2006). Conducting research in education and the social sciences. Enugu: Tashiwa Networks Ltd.

[3]. Danuwijaya, A. A., & Roebianto, A. (2023). Gender differences in secondary school assessment: A differential item functioning analysis using Rasch modeling. Journal of Psychological and Educational Research, 24(2), 45–60.

[4]. Ikpoyi, E., Amakiri, H., & Ehi, A. (2024). Gender fairness in secondary school Economics examinations: DIF analysis of co-educational school items. Nigerian Journal of Educational Measurement, 10(1), 23–37.

[5]. Nkeki, P., & Asekomhe, S. (2024). Gender and school-type differences in Economics multiple-choice items in WASSCE: Implications for item fairness. African Journal of Educational Assessment, 12(2), 55–69.

[6]. Nworgu, B. G. (2015). Educational research, issues and methodology (3rded). Nsukka, Nigeria: University Trust Publishers.

[7]. Ozdemir, B., & Alshamrani, A. H. (2020). Examining test fairness across gender using IRT-based differential item and test functioning methods. International Journal of Learning, Teaching and Educational Research, 19(6), 27–45.

[8]. PPSMB (Post Primary School Management Board, Nsukka). (2013/2014). Senior secondary school population report for Nsukka Education Zone. Enugu State, Nigeria: PPSMB.

[9]. Stella, T. (2021). Gender-based differential item functioning in secondary school science assessments. International Journal of Educational Measurement and Evaluation, 8(4), 112–124.

[10]. Wahyuni, A. (2022). Detection of gender bias using differential item functioning (DIF) analysis in school examinations. Jurnal Evaluasi Pendidikan, 13(1), 46–49.