African Educational Research Journal
Vol. 14(1), pp. 63-70, January 2026
ISSN: 2354-2160
Full Length Research Paper

Application of item response theory in the development and validation of a multiple-choice economics test for senior secondary school students

Ani Elizabeth Ngozika

Department of Science Education, University of Nigeria, Nsukka, Nigeria.

Accepted 20 January, 2026

ABSTRACT

This study applied Item Response Theory (IRT) to develop and validate a multiple-choice Economics test for senior secondary school students. An instrumentation research design was adopted. A total of 1,005 SS2 Economics students were randomly selected from 46 government co-educational secondary schools in Nsukka Education Zone, Enugu State, Nigeria. A 50-item Multiple Choice Economics Test (MCET) was developed, validated by experts, and pilot-tested, yielding a reliability coefficient of 0.89 (KR-20). Data were analyzed using the three-parameter logistic (3PL) IRT model with BILOG-MG software. Results showed that 49 items demonstrated low standard errors of measurement, indicating high precision, while one item showed low reliability. Item fit analysis revealed that 21 items fit the 3PL model, whereas 29 items misfit. Differential Item Functioning (DIF) analysis indicated that most items functioned differently across gender. The findings demonstrate IRT's usefulness in identifying reliable, misfitting, and biased items, supporting evidence-based item revision. The study recommends adopting IRT principles in Economics test development to enhance reliability, validity, and fairness.

Keywords: Economics, item response theory, 3PL model, secondary education, test validation.

E-mail: elizabethngozika027@gmail.com.

INTRODUCTION

Economics is a core subject in senior secondary school that equips students with the knowledge and skills necessary to understand economic concepts, solve societal problems, and make informed decisions. Since its introduction into the Nigerian secondary school curriculum in 1966 (Obemeata, 1991), Economics has aimed to promote rational thinking, efficient resource management, an appreciation of labor, and responsible citizenship (Asadu, 2001). Achieving these objectives largely depends on the quality of assessment instruments used to measure students' achievement. Multiple-choice tests (MCQs) are widely used in Nigerian secondary schools because of their objectivity, efficiency, and broad content coverage (Okoro, 2006; Onunkwo, 2002). However, most MCQs are constructed and evaluated using Classical Test Theory (CTT), which has several limitations. Under CTT, item difficulty and discrimination indices are sample-dependent, and test scores are interpreted primarily at the total-score level, providing limited information on individual item functioning (Crocker and Algina, 2008; Adedoyin, 2010). These limitations may compromise the fairness of the test and the comparability of scores across different student groups. Item Response Theory (IRT) offers a modern alternative, focusing on item-level characteristics and modeling the relationship between students' latent abilities and their responses to test items (Hambleton, Swaminathan and Rogers, 1991). IRT estimates item parameters such as difficulty, discrimination, and guessing, enabling a more precise measurement of student ability. Importantly, IRT allows for the detection of biased items through Differential Item Functioning (DIF) analysis, thereby enhancing the fairness of assessments (Meredith, Joyce and Walter, 2007; Chen, Li and Liu, 2021). Recent studies in Nigeria demonstrate the effectiveness of IRT in

improving test quality across various subjects, including Basic Science, Geography, and Mathematics. These studies showed that IRT provides stable item statistics and identifies problematic items that may be overlooked under CTT.

Despite these advances, the application of IRT to Economics achievement tests at the secondary school level remains limited. Given the importance of Economics for academic progression and career choices, there is a need for reliable, valid, and fair assessment instruments. The IRT model assumes that an examinee's performance can be fully predicted or explained by one or more latent abilities. IRT models the probability of a correct response using three logistic functions. The one-parameter logistic (1PL) model accounts for the probability of a correct response by assigning each question an independent difficulty parameter. For example, the 1PL model allows each question on an achievement test to have a distinct difficulty parameter. The two-parameter logistic (2PL) model accounts for each item's ability to discriminate between high- and low-ability students, while the three-parameter logistic (3PL) model adds a third parameter, called the pseudo-guessing parameter, which reflects the probability that an examinee with very low ability may answer an item correctly by guessing. This implies that students with low ability may answer items correctly by guessing. This study, however, applied the three-parameter logistic (3PL) IRT model to develop and validate a multiple-choice Economics test for Senior Secondary School II (SS2) students. Specifically, the study examined the standard errors of measurement, item fit to the 3PL model, and differential item functioning (DIF) by gender.

Purpose of the study

The main purpose of this study was to apply Item Response Theory (IRT) in the development and validation of a multiple-choice Economics test. Specifically, the study sought to:

1. Determine the standard errors of measurement of the test items.
2. Evaluate the fit of the test items using the three-parameter logistic (3PL) model.
3. Investigate differential item functioning with respect to gender.

Research questions

1. What are the standard errors of measurement of the items of the multiple-choice Economics test?
2. How do the items of the multiple-choice Economics test fit the three-parameter logistic 3PL model?
3. What differential item functioning is exhibited by the test items with respect to gender?

Research hypotheses

H_{01}: There is no significant fit between the items of the multiple-choice Economics test and the 3PL model.
H_{02}: The items of the multiple-choice Economics test do not function differentially between male and female students.

METHODOLOGY

Design and sample

The instrumentation research design was adopted. The population consisted of 3,795 SS2 Economics students in Nsukka Education Zone, Enugu State, Nigeria. The sample comprised 1,005 SS2 Economics students (462 males and 543 females) drawn from government-owned co-educational senior secondary schools in Nsukka Education Zone, Enugu State. The population of schools consisted of all government-owned co-educational schools offering Economics at the SS2 level. A proportionate stratified random sampling technique was employed, with the three Local Government Areas (Nsukka, Igbo-Etiti, and Uzo-Uwani) serving as strata. Students were selected concurrently from all strata in proportion to their population sizes: 551 from Nsukka LGA, 326 from Igbo-Etiti LGA, and 128 from Uzo-Uwani LGA. The number of students selected from each school was determined proportionately based on SS2 Economics enrolment, after which simple random sampling was used to select the students, ensuring equal chances of selection. The sample size was considered adequate for Item Response Theory (IRT) analysis, which requires large samples for stable parameter estimation.

Instrument

The Multiple Choice Economics Test (MCET) consisted of 50 items, each with four response options. Content and face validity were established by three experts, two lecturers in Measurement and Evaluation, Department of Science Education, and one from the Department of Economics, all in University of Nigeria, Nsukka. The experts were asked to examine the instrument with respect to:

• Whether the questions correspond to the table of specifications
• The structure and clarity of the questions;

Whether the answers to the questions tally with the ones in the marking scheme. The corrections and suggestions of these experts helped in modifying the items in MCET. Content validation of the test was carried out by preparing the table of specifications based on the six levels of cognitive domain of Bloom's

taxonomy. Pilot testing yielded a KR-20 reliability coefficient of 0.89, indicating high internal consistency.

Procedure and data analysis

The MCET was administered under standardized testing conditions. Data were analyzed using BILOG-MG software based on the 3PL IRT model. Maximum likelihood estimation was used to estimate item parameters, while the chi-square goodness-of-fit procedure was used to assess item fit and DIF with respect to gender.

RESULTS

Standard errors of measurement

Forty-nine items had standard errors ranging from 0.05 to 0.44, indicating high measurement precision. One item recorded a standard error of 0.58, suggesting lower reliability.

Table 1. Standard errors of measurement of the test items of the multiple-choice test in Economics based on the three-parameter logistic (3PL) model.

Item	S.E
1	0.44
2	0.27
3	0.12
4	0.09
5	0.10
6	0.10
7	0.16
8	0.06
9	0.09
10	0.10
11	0.22
12	0.14
13	0.05
14	0.11
15	0.09
16	0.36
17	0.58
18	0.06
19	0.10
20	0.08
21	0.08
22	0.09
23	0.13
24	0.07
25	0.08
26	0.15
27	0.33
28	0.08
29	0.07
30	0.08
31	0.06
32	0.24
33	0.08
34	0.09
35	0.09
36	0.07

Table 1. Continued.

37	0.06
38	0.14
39	0.05
40	0.12
41	0.07
42	0.05
43	0.07
44	0.16
45	0.05
46	0.15
47	0.09
48	0.07
49	0.16
50	0.20

Standard Error (SE): Precision of each item's ability estimate; lower values indicate higher precision.

Item fit to the 3PL model

Chi-square analysis revealed that 21 items (42%) fit the 3PL model ($p > 0.05$), while 29 items (58%) misfit ($p < 0.05$), indicating a need for item revision.

Table 2. Fits statistics of Economics 3 multiple choice test based on three parameter logistic (3PL) model.

Item	Chi.sq.	Prob
1	51.2	0. 10
2	37.6	0.00*
3	67.3	0.12
4	48.5	0.00*
5	51.9	0.20
6	47.6	0.00*
7	96.6	0.15
8	30.5	0.00*
9	90.9	0.14
10	46.7	0.05
11	79.4	0.00*
12	57.1	0.16
13	31.5	0.03*
14	18.2	0.01*
15	55.0	0.08
16	35.2	0.00*
17	31.4	0.07
18	84.2	0.00*
19	76.0	0.00*
20	43.9	0.03*
21	31.7	0.09
22	44.2	0.18
23	77.4	0.00*
24	13.7	0.06
25	40.0	0.00*
26	84.2	0.13

Table 2. Continued.

27	18.0	0.02*
28	79.0	0.06
29	46.0	0.07
30	43.7	0.00*
31	21.3	0.00*
32	103.4	0.08
33	48.7	0.00*
34	45.4	0.01*
35	92.6	0.24
36	55.2	0.00*
37	52.1	0.00*
38	29.3	0.00*
39	23.2	0.26
40	179.9	0.00*
41	77.8	0.04*
42	23.8	0.00*
43	70.4	0.00*
44	116.5	0.00*
45	26.1	0.09
46	41.0	0.00*
47	138.4	0.13
48	33.5	0.00*
49	45.5	0.02*
50	94.3	0.09

Chi-sq.: Chi-square statistic assessing how well each item fits the 3PL model.
Prob**:** Probability value; $p < 0.05$ indicates significant misfit of the item.

Differential item functioning

DIF analysis showed that 46 items (92%) functioned differently across gender groups, while 4 items (8%) functioned similarly, indicating gender-related bias in many items.

Table 3. Model for group differential item functioning of the test items of the multiple-choice test in Economics.

Item	Group	P	Chi.Sq
1	Male	0.00	120.2*
	Female	0.00	266.8*
2	Male	0.00	68.2*
	Female	0.00	113.5*
3	Male	0.72	5.3*
	Female	0.00	22.2*
4	Male	0.49	7.4*
	Female	0.48	7.5*
5	Male	0.89	3.5*
	Female	0.03	16.8*
6	Male	0.15	11.9*
	Female	0.00	40.8*
7	Male	0.00	26.3*
	Female	0.00	36.9*

Table 3. Continued.

8	Male	0.00	23.7*
	Female	0.00	37.6*
9	Male	0.6	6.2*
	Female	0.59	6.5*
10	Male	0.9	3.0*
	Female	0.92	3.2*
11	Male	0.76	5.0*
	Female	6.0	6.4*
12	Male	0.58	6.6*
	Female	0.01	19.8*
13	Male	0.00	40.0*
	Female	0.00	105.6*
14	Male	0.72	10.7
	Female	0.70	10.7
15	Male	0.61	6.3*
	Female	0.00	22.3*
16	Male	0.00	49.4*
	Female	0.00	109.2*
17	Male	0.00	92.8*
	Female	0.00	242.6*
18	Male	0.00	89.8*
	Female	0.00	90.4*
19	Male	0.00	30.1*
	Female	0.00	30.0*
20	Male	0.10	13.2*
	Female	0.00	20.3*
21	Male	0.29	9.5
	Female	0.01	9.5
22	Male	0.00	27.5*
	Female	0.00	147.4*
23	Male	0.97	2.1*
	Female	0.83	4.2*
24	Male	0.00	80.2*
	Female	0.00	134.8*
25	Male	0.00	20.7*
	Female	0.81	4.5*
26	Male	0.04	16.1*
	Female	0.00	71.9*
27	Male	0.00	107.2
	Female	0.00	107.2
28	Male	0.85	4.0*
	Female	0.57	6.7*
29	Male	0.55	15.2*
	Female	0.00	200.0*
30	Male	0.00	23.8*
	Female	0.00	45.0*
31	Male	0.24	10.4*
	Female	0.00	101.8*
32	Male	0.32	9.2*
	Female	0.00	61.0*
33	Male	0.54	6.9*
	Female	0.65	5.9*

Table 3. Continued.

34	Male	0.32	9.0*
	Female	0.00	21.0*
35	Male	0.99	0.9*
	Female	0.00	68.6*
36	Male	0.04	15.7*
	Female	0.13	12.4*
37	Male	0.23	10.4*
	Female	0.30	9.4*
38	Male	0.25	10.1*
	Female	0.83	4.2*
39	Male	0.00	44.9*
	Female	0.00	78.9*
40	Male	0.24	10.4*
	Female	0.07	14.5*
41	Male	0.00	31.1*
	Female	0.19	11.1*
42	Male	0.00	31.1*
	Female	0.00	68.4*
43	Male	0.00	22.8*
	Female	0.19	11.2*
44	Male	0.72	5.3*
	Female	0.00	24.5*
45	Male	0.00	99.8*
	Female	0.00	83.8*
46	Male	0.00	13.3
	Female	0.46	13.3
47	Male	0.79	4.7*
	Female	0.98	2.0*
48	Male	0.02	18.1*
	Female	0.02	18.0*
49	Male	0.02	17.9*
	Female	0.00	76.2*
50	Male	0.00	141.6*
	Female	0.00	228.0*

Key indicator: Most items have significant Chi-square (χ^2) values with $p < .05$**,** showing evidence of gender-based DIF.

DISCUSSION

IRT provides detailed item-level diagnostics beyond CTT. Most items of the multiple-choice Economics test showed high reliability, supporting prior studies (Mawak, Efomo and Mustapha, 2024; Kaigama, Dadughun and Mustapha, 2025). However, many items misfit the IRT model and exhibited widespread differential item functioning (DIF), contradicting studies reporting minimal misfit and little DIF (van Rijn, Sinharay and Haberman, 2016; Fährmann and Hambleton, 2022), likely due to differences in sample characteristics, instructional context, or item content (Sinharay and Haberman, 2014; Eteng-Uket, 2017).

Regarding the hypotheses, 21 items fit the three-parameter logistic (3PL) model while 29 did not, partially supporting Hypothesis One, and 46 items showed DIF, rejecting Hypothesis Two. These findings highlight the need for iterative item review to ensure valid, reliable, and fair assessment instruments.

Conclusion

IRT is a robust framework for developing and evaluating Economics achievement tests. While most items were reliable, many misfit the 3PL model or exhibited gender-based DIF. IRT-based diagnostics can guide item revision to improve reliability, validity, and fairness.

RECOMMENDATIONS

1. Train Economics teachers and examination authorities in IRT applications.
2. Adopt IRT in the development and validation of secondary school Economics tests.
3. Revise or replace misfitting or biased items.
4. Conduct DIF analyses routinely to ensure gender fairness.

REFERENCES

Adedoyin, O. (2010). *Assessment practices in Nigerian secondary schools*. Lagos: Educational Press.

Asadu, J. I. (2001). *Economics education in Nigeria: Trends and challenges*. Enugu: Academic Press.

Chen, H., Li, Z., & Liu, Y. (2021). Differential item functioning in secondary school assessments. *Journal of Educational Measurement, 58*(3), 245–263.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory* (2nd ed.). Wadsworth.

Eteng-Uket, F. (2017). Evaluation of item bias using differential item functioning (DIF) technique in NECO-conducted examinations in Taraba State, Nigeria. *International Journal of Research and Innovation in Social Science, 1*(1), 12–19.

Fährmann, K., & Hambleton, R. K. (2022). *Practical significance of item misfit in educational assessments.* Large-scale Assessments in Education, 10, Article 7.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Kaigama, D., Dadughun, M., & Mustapha, B. (2025). Application of IRT in Nigerian education. *International Journal of Educational Research, 110*, 100–112.

Meredith, W., Joyce, C., & Walter, P. (2007). *Modern approaches to test development*. Springer.

Mawak, L., Efomo, A., & Mustapha, B. (2024). IRT for improving test reliability. *African Journal of Measurement and Evaluation, 12*(2), 55–68.

Obemeata, J. A. (1991). *History and development of Economics education in Nigeria*. Ibadan: University Press.

Okoro, J. C. (2006). *Secondary school assessment in Nigeria*. Enugu: Academic Publishers.

Onunkwo, I. (2002). *Evaluation of multiple-choice testing in Nigerian schools*. Lagos: Educational Research Press.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*(1), 23–35.

van Rijn, P. W., Sinharay, S., & Haberman, S. J. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education, 4*(10), 1–23.

Citation: Ani, E. N. (2026). Application of item response theory in the development and validation of a multiple-choice economics test for senior secondary school students. *African Educational Research Journal, 14*(1), 63-70.