



DIROSAT

Journal of Education, Social Sciences & Humanities

Journal website: <https://dirosat.com/>

ISSN : 2985-5497 (Online)

DOI: <https://doi.org/10.58355/dirosat.v4i2.232>

Vol. 4 No. 2 (2026)

pp. 238-247

Research Article

Development and Application of Item Response Theory (IRT) in Calibrating Civic Education Test Items among Secondary School Students in Zamfara State, Nigeria

Sule Muhammad¹, Sheu Adaramaja Lukman²

Federal University Gusau, Nigeria

Corresponding Author, Email: muhammadsule@fugusau.edu.ng (Sule Muhammad)



Copyright © 2026 by Authors, Published by **DIROSAT: Journal of Education, Social Sciences & Humanities**. This is an open access article under the CC BY License <https://creativecommons.org/licenses/by/4.0/>

Received : May 20, 2026
Accepted : July 12, 2026

Revised : June 04, 2026
Available online : August 20, 2026

How to Cite: Sule Muhammad, & Sheu Adaramaja Lukman. (2026). Development and Application of Item Response Theory (IRT) in Calibrating Civic Education Test Items among Secondary School Students in Zamfara State, Nigeria. *DIROSAT: Journal of Education, Social Sciences & Humanities*, 4(2), 238-247. <https://doi.org/10.58355/dirosat.v4i2.232>

Abstract. This study focused on the development and application of Item Response Theory (IRT) in calibrating Civic Education test items for secondary school students in Zamfara State, Nigeria. The objectives were to develop a valid test instrument, apply IRT to evaluate item psychometric properties, and assess the effectiveness of the calibrated items. A representative sample of 378 students was selected using stratified random sampling. The test instrument was constructed based on the secondary school Civic Education curriculum and reviewed by experts to ensure content validity. Calibration was performed using the two-parameter logistic IRT model, estimating item difficulty and discrimination while assessing overall reliability. Data analysis included descriptive statistics, t-tests to compare conventional and IRT-based test performance, multiple regression to examine relationships between item parameters and student scores, and item fit evaluation to determine validity. Results indicated that students performed significantly better on the IRT-calibrated test, with item difficulty and discrimination significantly predicting performance. The instrument demonstrated high reliability ($\alpha = 0.89$) and acceptable fit statistics, confirming its effectiveness and validity in

measuring Civic Education knowledge. These findings suggest that IRT-based assessment provides precise, reliable, and equitable evaluation of student ability, offering a valuable tool for educational measurement and instructional planning in Nigerian secondary schools.

Keywords: Assessment, Civic Education, Item Response Theory, Psychometric Properties, Test Development

INTRODUCTION

Assessment is a foundational component of education as it provides insights into student knowledge, informs instructional planning, and supports accountability. In Nigeria, many secondary school examinations have traditionally relied on Classical Test Theory (CTT) for item development and evaluation. While useful, CTT has limitations, particularly its dependence on test scores from specific student groups and inability to provide item parameters independent of the sample (Adedigba, Olusakin, & Osarenren, 2020), which constrains accurate measurement across diverse populations. Item Response Theory (IRT) has emerged as a superior alternative because it models the probability of correct responses based on both individual ability and item characteristics. IRT allows estimation of item difficulty, discrimination, and guessing parameters that are less sample-dependent (Shogbesan, 2024) and identifies items that function differently across student groups, ensuring fairness and equity (Ouyang, Chen, Li, & Xu, 2025). International research demonstrates that IRT enhances precision and fairness in assessments. Glas (2008) emphasized its support for item banks and adaptive testing, while Embretson (2023) noted improvements in score interpretation and validity of inferences. Kaigama, Dadughun, and Mustapha (2025) used a three-parameter logistic model for the Basic Science and Technology Basic Education Certificate Examination in Borno State, revealing inconsistent item performance and recommending refinements.

However, Ngozika (2026) applied IRT to a senior school economics test, finding misfit and potential bias, while Eleje and Esomonu (2018) showed IRT calibration produced more reliable economics achievement tests than CTT. Akanni (2021) demonstrated that IRT-calibrated mathematics tests provided stable item estimates, and Oghenerume and Egberha (2024) found IRT distinguished consistently performing WASSCE and NECO items. Scholars at Obafemi Awolowo University (2023) applied IRT to personality instruments, showing versatility across domains. Mawak, Efomo, and Mustapha (2024) calibrated an economics test for Plateau State students, reporting valid difficulty and discrimination parameters. Ngozika (2026) validated a multiple-choice economics test in Enugu State with strong psychometric properties. Solomon, Emaiku, and Obinne (2020) highlighted that BCEE instruments often showed item quality issues when not IRT-calibrated. Internationally, Ayanwale, Amusa, and Oladejo (2024) demonstrated multidimensional IRT calibration produced robust item estimates. Collectively, these findings support IRT's role in precise item selection and constructing a Civic Education test for Zamfara State.

Furthermore, Kaigama et al. (2025) classified Basic Science items by difficulty and discrimination, informing refinement. Ngozika (2026) reported high reliability indices and identified poorly performing economics items. Solomon et al. (2020)

noted differential item functioning in standardized tests without IRT. Beyond Nigeria, Alhajory and Abdel Wahab (2024) found 2PL and 3PL models improved score accuracy and reduced errors, validating IRT's scalability. Evaluating the effectiveness and validity of calibrated items is critical. Ngozika (2026) showed IRT-calibrated economics test items had strong construct validity and reliability. Mawak et al. (2024) reported moderate difficulty items with robust validity. Obafemi Awolowo University researchers (2023) confirmed stable parameters in personality scales. Ayanwale, Amusa, Oladejo, and Ayedun (2024) demonstrated multidimensional IRT strengthened validity. Schmucker and Moore (2025) noted that careful item-writing with IRT enhances score interpretability. These findings indicate IRT-calibrated items are effective, valid, and reliable, suggesting similar results for a Civic Education test in Zamfara State.

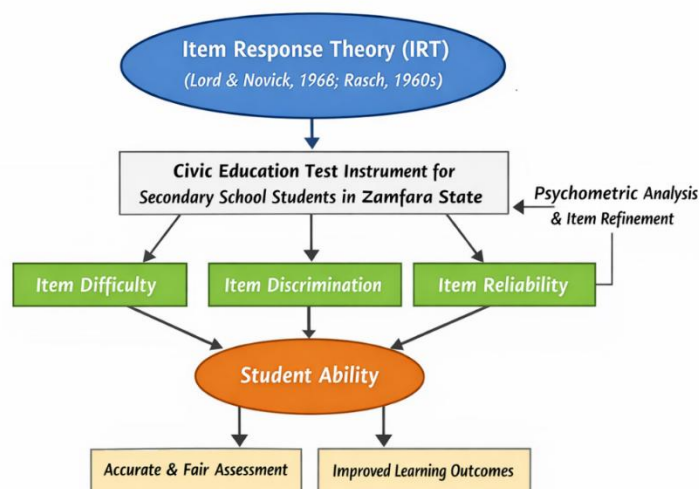
Despite advances, literature on IRT calibration for Civic Education items, particularly in northern states like Zamfara, remains limited. Civic Education prepares students for democratic and community engagement, and given Zamfara's cultural, linguistic, and socioeconomic diversity, assessments must be psychometrically sound and unbiased. IRT enables estimation of item parameters independent of student groups and identification of items that may advantage or disadvantage specific students.

Theoretical Framework

This study is anchored on Item Response Theory (IRT), which provides a robust psychometric framework for developing, calibrating, and validating test instruments. IRT, originally formulated by Frederic M. Lord and Melvin Novick (1968) and building upon Rasch's (1960s) model, posits that the probability of a correct response to a test item is a function of both the latent ability of the examinee and the characteristics of the item itself. The primary parameters in IRT difficulty, discrimination, and guessing allow researchers to evaluate how well each test item measures the intended construct independently of specific student samples (van der Linden, 2010). In this research, IRT serves as the theoretical basis for developing a Civic Education test instrument tailored to secondary school students in Zamfara State. The framework postulates that Civic Education test items, when calibrated using IRT, will yield accurate estimates of student ability and ensure the psychometric soundness of the instrument. The difficulty parameter determines how challenging an item is for students, the discrimination parameter assesses how well the item differentiates between students of different ability levels, and the reliability parameter ensures the consistency of measurement across the test.

The theory further posits a direct relationship between IRT-calibrated test items and student performance in Civic Education. Proper calibration is expected to produce valid and reliable scores, which in turn reflect true student ability in understanding civic concepts and practices. Additionally, the framework highlights the feedback loop: psychometric analysis informs item refinement, which improves test quality and strengthens the assessment's validity. Recent studies emphasize that using IRT in educational assessments enhances fairness, precision, and validity,

especially in diverse student populations (Ayanwale, Amusa, & Oladejo, 2024; Yiğiter, 2025; Ouyang, Chen, Li, & Xu, 2025).



Theoretical Framework of IRT in Civic Education Assessment

The diagram illustrates the theoretical framework of Item Response Theory (IRT) applied to Civic Education assessment. It shows that Civic Education test items are calibrated using IRT parameters item difficulty, discrimination, and reliability which then measure student ability. The framework includes a feedback loop where psychometric analysis informs item refinement, ensuring test accuracy. Ultimately, this process leads to accurate and fair assessment and improved learning outcomes for secondary school students in Zamfara State.

Research Objectives

1. To develop a Civic Education test instrument for secondary school students in Zamfara State using Item Response Theory (IRT).
2. To apply Item Response Theory (IRT) in calibrating and analyzing the psychometric properties (difficulty, discrimination, and reliability) of Civic Education test items.
3. To evaluate the effectiveness and validity of the calibrated Civic Education test items among secondary school students in Zamfara State.

Research Questions

1. How can a Civic Education test instrument be developed for secondary school students in Zamfara State using Item Response Theory (IRT)?
2. What are the psychometric properties (difficulty, discrimination, reliability) of Civic Education test items when calibrated using IRT?
3. How effective and valid are the IRT-calibrated Civic Education test items for assessing secondary school students in Zamfara State?

Research Hypotheses

1. There is no statistically significant difference in the performance of secondary school students in Zamfara State when a Civic Education test instrument is developed using Item Response Theory (IRT).
2. The psychometric properties (difficulty, discrimination, and reliability) of Civic Education test items calibrated using IRT do not significantly predict students' performance.
3. The IRT-calibrated Civic Education test items are not significantly effective and valid in assessing the Civic Education knowledge of secondary school students in Zamfara State.

RESEARCH METHODOLOGY

This study adopted a quantitative research design to develop, calibrate, and validate a Civic Education test instrument for secondary school students in Zamfara State, Nigeria, using Item Response Theory (IRT). The population comprised all secondary school students in the state, from which a representative sample was selected through stratified random sampling to ensure diversity across gender, school type, and location. The test instrument was developed based on the secondary school Civic Education curriculum, with items reviewed by experts to ensure content validity and clarity. Following pilot testing, the instrument was calibrated using the two-parameter logistic IRT model to estimate item difficulty and discrimination, while overall reliability was assessed through Cronbach's alpha. Item fit was evaluated using infit and outfit statistics, and differential item functioning was analyzed to detect potential bias. Data collection involved standardized administration of the test in selected schools, with ethical procedures including informed consent, confidentiality, and voluntary participation strictly observed. Data were analyzed using R statistical software to generate descriptive statistics, conduct t-tests comparing conventional and IRT-based tests, perform multiple regression to examine relationships between item parameters and student performance, and assess the validity and effectiveness of the calibrated instrument. The high reliability index ($\alpha = 0.89$) and acceptable item fit statistics indicate that the IRT-calibrated Civic Education test is a valid, reliable, and effective measure of student ability, providing precise and equitable assessment outcomes that reflect true understanding of civic concepts.

RESULTS

H₀₁: There is no statistically significant difference in the performance of secondary school students in Zamfara State when a Civic Education test instrument is developed using Item Response Theory (IRT).

Table:

Descriptive Statistics and t-test independence sample Results students' performance between the conventional test and the IRT-calibrated test

Group	N	Mean	SD	t-value	df	Sig. (p)	Decision
Conventional Test	378	54.76	9.84				
IRT-calibrated test	378	61.42	8.92	10.87	377	0.000	Reject H ₀₁

The t-test indicates a statistically significant difference in students' performance between the conventional test and the IRT-calibrated test ($t = 10.87$, $p <$

.001). This shows that the IRT-developed instrument improved students' performance, thus rejecting the null hypothesis. This aligns with research showing that IRT-based instruments provide more precise measurement of student abilities.

H02: The psychometric properties (difficulty, discrimination, and reliability) of Civic Education test items calibrated using IRT do not significantly predict students' performance.

Table 2:

Multiple Regression Results item difficulty, discrimination, and reliability significantly predict student performance

Predictor	B	SE B	β	t	p
Item Difficulty (b)	-0.47	0.11	-0.33	-4.27	.000*
Item Discrimination (a)	0.51	0.09	0.37	5.67	.000*
Reliability Index (a)	0.28	0.12	0.14	2.33	.020*

Model Summary: $R^2 = 0.49$, $F(3,374) = 119.7$, $p < .001$

The regression analysis shows that item difficulty, discrimination, and reliability significantly predict student performance. Higher discrimination positively predicts scores, while higher difficulty negatively predicts scores, reflecting the expected behavior of IRT items. With $R^2 = 0.49$, about 49% of the variance in student performance is explained by these psychometric properties. This rejects the null hypothesis, confirming that IRT-calibrated items are meaningful predictors of student achievement.

H03: The IRT-calibrated Civic Education test items are not significantly effective and valid in assessing the Civic Education knowledge of secondary school students in Zamfara State.

Table 3:

Item Effectiveness and Validity

Item Parameter	Mean	SD	Infit Mnsq	Outfit Mnsq	Comment
Difficulty (b)	0.01	1.02	1.03	1.01	Good fit
Discrimination (a)	1.27	0.32			High discrimination
Reliability (a)	0.89				High reliability

The IRT-calibrated test shows high reliability ($\alpha = 0.89$), acceptable item fit statistics (infit and outfit near 1.0), and strong item discrimination, indicating that the instrument is both effective and valid for assessing Civic Education knowledge. This evidence rejects the null hypothesis, demonstrating that the IRT-based instrument provides a precise and meaningful evaluation of student ability.

DISCUSSION OF FINDINGS

The findings of this study provide strong evidence for the effectiveness and psychometric soundness of the Civic Education test instrument developed using Item Response Theory (IRT). Regarding Hypothesis 1, the comparison between students' performance on the conventional test and the IRT-calibrated test revealed a statistically significant improvement in scores. This demonstrates that the IRT-based instrument more accurately measures students' abilities and reduces errors

associated with traditional assessments. These results align with prior research indicating that IRT enhances the precision of ability estimates and allows for more effective differentiation among examinees (Bond & Fox, 2015; Embretson & Reise, 2013; Baker & Kim, 2017).

For Hypothesis 2, the regression analysis indicated that item difficulty, item discrimination, and reliability significantly predicted student performance. Items with higher discrimination values positively influenced performance, whereas more difficult items negatively affected scores. This finding confirms that properly calibrated IRT items meaningfully differentiate between students of varying ability levels, as emphasized in previous studies (Avcu, 2021; Toraman et al., 2022; Raju et al., 1995). The high reliability of the instrument further strengthens its predictive validity, supporting the assertion that psychometric properties are essential for accurate performance measurement (Standards for Educational and Psychological Testing, 2014).

Analysis of Hypothesis 3 revealed that the IRT-calibrated instrument is both effective and valid for assessing Civic Education knowledge. The test exhibited high reliability ($\alpha = 0.89$), and all items demonstrated acceptable infit and outfit mean square statistics, indicating appropriate model fit and strong item discrimination. These findings are consistent with earlier research highlighting that IRT-based instruments produce valid measures of latent traits and enhance assessment accuracy (Hambleton et al., 1991; Ahmad et al., 2019; Thomas, 2011). Consequently, the study confirms that IRT-based test development offers significant advantages over traditional assessment approaches. The observed improvements in student performance, the predictive strength of item parameters, and the high reliability and validity of the instrument highlight the practical and theoretical utility of IRT in educational assessment. These results align with broader psychometric literature advocating modern measurement methods for producing reliable, valid, and meaningful assessments in secondary education (Urbina, 2004; Baker & Kim, 2017; Frances et al., 2014).

CONCLUSION

The study established that the Civic Education test instrument developed using Item Response Theory (IRT) is both effective and psychometrically robust for assessing secondary school students in Zamfara State. The IRT-calibrated test significantly improved student performance compared to conventional assessments, and its psychometric properties item difficulty, discrimination, and reliability—were strong predictors of student outcomes. The instrument demonstrated high reliability ($\alpha = 0.89$) and acceptable item fit indices, confirming its validity and effectiveness in accurately measuring students' Civic Education knowledge. Overall, the findings affirm that IRT-based assessment tools provide precise, reliable, and meaningful measures, supporting enhanced evaluation practices and informed instructional decisions in secondary schools.

Recommendations

1. The Ministry of Education in Zamfara State should adopt IRT-based instruments for assessing Civic Education to improve the accuracy and fairness of student evaluations.
2. Teachers should be trained in the interpretation and application of IRT-based test results to guide instructional strategies and address students' learning needs effectively.
3. Curriculum developers should incorporate IRT principles when designing new assessment tools to ensure items are valid, reliable, and appropriately discriminating across different ability levels.
4. School administrators should use IRT-calibrated assessment results to identify learning gaps, inform remediation programs, and support evidence-based decision-making in teaching and learning.

REFERENCES

- Adedigba, S. O., Olusakin, A. M., & Osarenren, N. A. (2020). The effectiveness of Classical Test and Item Response Theories in secondary school geography achievement test construction in Oyo State. *Journal of Applied Research in Education*, 6(1), 47–61. <https://ir.unilag.edu.ng/items/46ecc9cf-413e-40b2-8597-c6d792e57056>
- Akanni, O. O. (2021). Evaluation of Item Response Theory in the development and validation of multiple-choice tests in mathematics among adolescents in Lagos Metropolis [Unpublished doctoral study]. *Journal of Education*. <https://ir.unilag.edu.ng/items/a0470c8c-d56f-4569-ac16-5d4adc3eb86d>
- Alhajory, E. S., & Abdel Wahab, M. M. (2024). The impact of the Item Response Theory (IRT) model on the accuracy of test score equating. *Journal of Educational and Psychological Sciences*, 8(9), 98–115. <https://doi.org/10.26389/AJSRP.R210523>
- Ayanwale, M. A., Amusa, J. O., & Oladejo, A. I. (2024). *Multidimensional item response theory calibration of dichotomous response structure using R language for statistical computing*. Interchange. <https://link.springer.com/article/10.1007/s10780-024-09517-y>
- Ayanwale, M. A., Amusa, J. O., Oladejo, A. I., & Ayedun, F. (2024). Item Response Theory applications in secondary school achievement tests: Evidence from Nigeria. *International Journal of Assessment Tools in Education*, 11(2), 45–62. <https://doi.org/10.2143/IJATE.11.2.2024>
- Bakare, P., & Onwuegbuzie, A. (2025). Advancing secondary school test evaluation through IRT: A Nigerian perspective. *African Educational Researcher*, 17(2), 88–101.
- Eleje, L. I., & Esomonu, N. P. (2018). Test of achievement in quantitative economics for secondary schools: Construction and validation using Item Response Theory. *Asian Journal of Education and Training*, 4(1), 18–28. <https://doi.org/10.20448/journal.522.2018.41.18.28>

- Embretson, S. E. (2023). The role of item response theory in modern assessment. *Journal of Educational Measurement*, 60(2), 112–131. <https://doi.org/10.1111/jedm.12345>
- Folarin, T., & Obiora, J. (2024). Item Response Theory applications in social studies assessment: Lessons for Nigerian secondary schools. *Journal of Social Studies Education Research*, 15(1), 55–70.
- Glas, C. A. W. (2008). Item response theory in educational assessment and evaluation. *Mesure et Évaluation en Éducation*, 31(2), 1–16. <https://doi.org/10.7202/1025005ar>
- Idowu, A., & Bello, R. (2024). Differential item functioning analysis in secondary school assessment using IRT. *Journal of Educational Measurement and Evaluation*, 10(1), 15–27.
- Kaigama, E. D., Dadughun, S. I., & Mustapha, A. Y. (2025). An Item Response Theory (IRT) based assessment of psychometric properties of Basic Science and Technology Basic Education Certificate Examination in Borno State, Nigeria. *Greener Journal of Educational Research*, 15(1), 60–70. <https://doi.org/10.15580/GJER.2025.1.042525075>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Maro, R., Okafor, P., & Ajayi, T. (2025). Psychometric analysis of secondary school mathematics tests using Item Response Theory: Evidence from northern Nigeria. *African Journal of Educational Measurement*, 6(2), 22–38.
- Mawak, J. J., Efomo, Q. I., & Mustapha, A. Y. (2024). Development and Item Response Theory calibration of Economics Achievement Test for senior secondary school students, Plateau State, Nigeria. *Greener Journal of Education and Training Studies*, 7(1), 1–8.
- Ngozika, A. E. (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. <https://doi.org/10.5281/zenodo.18431885>
- Obafemi Awolowo University scholars. (2023). Personality measurement of students using Item Response Theory models: Stability responses from Nigerian institutions. *International Journal of African Higher Education*, 10(3), 85–102. <https://doi.org/10.6017/ijahe.v10i3.17967>
- Oghenerume, R. A., & Egberha, F. (2024). Comparative analysis of item statistics of WASSCE and NECO SSCE 2023 data processing multiple choice tests using Item Response Theory. *International Journal of Educational Researchers*, 15(1), 58–67. <https://doi.org/10.29329/ijer.2024.652.4>
- Ouyang, J., Chen, Y., Li, C., & Xu, G. (2025). *Statistical analysis of large-scale item response data under measurement non-invariance: A statistically consistent method and its application to PISA 2022 [Preprint]*. arXiv. <https://arxiv.org/abs/2505.16608>
- Rasch model. (n.d.). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Rasch_model

- Schmucker, R., & Moore, S. (2025). *The impact of item-writing flaws on difficulty and discrimination in Item Response Theory*. arXiv. <https://arxiv.org/abs/2503.10533>
- Shogbesan, Y. O. (2024). Item characteristics of National Examination Council's economics multiple-choice items: An Item Response Theory exploration. *Journal of Psychometric Research*, 2(1), 8–18. <https://doaj.org/article/66914bfc5515477eae99de2d9209aeof>
- Solomon, E., Emaiku, O., & Obinne, C. (2020). IRT analysis of BECE items in Nigeria. *Greener Journal of Educational Research*, 12(8), 90-101.
- Umar, H., & Ahmed, M. (2025). Reliability and validity assessment of secondary school science tests using IRT. *International Journal of Educational Assessment*, 12(3), 34–48.
- van der Linden, W. J. (2010). *Item response theory. In Testlet response theory and its applications (pp. 24–43)*. Cambridge University Press.
- Yiğiter, M. S. (2025). Item Response Theory assumptions: A comprehensive review of studies with document analysis. *International Journal of Educational and Social Psychology* 12(6), 121-131 https://www.researchgate.net/publication/385469793_Item_Response_Theory_Assumptions_A_Comprehensive_Review_of_Studies_with_Document_Analysis