Rima International Journal of Education (RIJE)
ISSN: 2756 – 6749 (p); 3141-2033 (e)
Volume 5, Issue 3, June. 2026; Page: 33—48 DOI: https://doi.org/10.65760/rijessu.v5.3.3

Comparative Evaluation of Estimation Methods in Item Response Theory (IRT) for High-Stakes Economics Examinations among SSS III Students in Oyo State, Nigeria

***[1]Abass Adekunle Sulaiman and [2]Dorcas. S. Daramola**

[1]Department of social sciences Education, University of Ilorin, Ilorin, Kwara State, Nigeria. **Email:** princeatilola4u@gmail.com

[2]Department of social sciences Education, University of Ilorin, Ilorin, Kwara State, Nigeria. **Email:** olatunji.ds@unilorin.edu.ng

Abstract

This study undertook a comparative evaluation of Maximum Likelihood (ML) and Bayesian estimation methods in Item Response Theory (IRT) with a view to determining their implications for high-stakes testing, using WAEC Economics multiple-choice items as a case study. The study adopted a descriptive survey and correlational research design. The population comprised all Senior Secondary School III (SSS III) students offering Economics in public secondary schools in Oyo State, Nigeria. A multi-stage sampling procedure was employed to select 1,200 students from 20 secondary schools. The 2021 WAEC Economics multiple-choice items served as the research instrument. Data were analysed using both descriptive and inferential statistics through Xcalibre 4.2 and SPSS version 25. Descriptive statistics of mean, standard deviation, skewness and kurtosis were used to describe examinees' ability distribution and item parameter estimates, while Pearson Product Moment Correlation (PPMC) was employed to test the hypotheses at the 0.05 level of significance. Findings revealed that examinees' ability estimates obtained through ML and Bayesian methods were moderately and positively correlated, indicating consistency between the two estimation techniques. Similarly, item difficulty and discrimination indices showed strong and significant positive correlations between the two methods, suggesting comparability in parameter estimation. However, the guessing parameter exhibited a weak and non-significant correlation, implying notable differences in the estimation of pseudo-guessing between ML and Bayesian approaches. Overall, the results indicate that ML and Bayesian estimation methods can be used interchangeably for estimating examinees' ability, item difficulty, and discrimination in high-stakes testing, but caution is required in interpreting guessing parameters. The study concludes that the choice of estimation method has important implications for the fairness, accuracy, and credibility of high-stakes examinations and recommends the combined or complementary use of both methods in large-scale assessment programmes.

Keywords: Item Response Theory, Maximum Likelihood Estimation, Bayesian Estimation, Ability Estimation, Item Difficulty, Item Discrimination, Guessing Parameter

Cite this as: Sulaiman, A. A. & Daramola, D. S. (2026). Comparative Evaluation of Estimation Methods in Item Response Theory (IRT) for High-Stakes Economics Examinations among SSS III Students in Oyo State, Nigeria. *Rima International Journal of Education, 5*(3), 33—48. DOI: https://doi.org/10.65760/rijessu.v5.3.3

Introduction

Accurate measurement of learners' abilities is fundamental to effective educational decision-making, particularly in high-stakes testing contexts where examination outcomes influence students' academic progression, certification, and employment opportunities (AERA, APA, & NCME, 2014). Traditionally, Classical Test Theory (CTT) has dominated educational assessment practice. However, CTT is constrained by several limitations, notably the sample dependency of item statistics and the test dependency of examinee scores, which undermine the generalizability and precision of measurement outcomes (Hambleton, Swaminathan, & Rogers, 1991; DeVellis, 2017). These limitations have led to the increasing adoption of Item Response Theory (IRT) as a more robust psychometric framework.

Item Response Theory (IRT) models the relationship between an individual's latent ability and the probability of correctly responding to test items; this framework provides invariant item and person parameter estimates, ensuring measurement stability across diverse samples and test forms (Embretson & Reise, 2000; Baker & Kim, 2017). Prominent IRT models, including the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models, enable detailed estimation of item difficulty, discrimination, and guessing parameters, thereby offering enhanced diagnostic insight into test quality (Hambleton et al., 1991; Lord, 1980). Central to the application of IRT is the choice of parameter estimation method, as this determines the accuracy, stability, and interpretability of estimated parameters.

Commonly used IRT parameter estimation techniques include Maximum Likelihood Estimation (MLE), Marginal Maximum Likelihood (MML), and Bayesian approaches such as Maximum a Posteriori (MAP) and Expected a Posteriori (EAP) estimation (Bock & Aitkin, 1981; Mislevy, 1986; Baker & Kim, 2017). These methods differ in their statistical assumptions, computational efficiency, and sensitivity to sample size and test length. Empirical studies indicate that parameter estimates obtained through different estimation techniques may vary significantly, especially in moderate or small sample contexts and in complex IRT models (Thissen & Wainer, 2001; Kang & Chen, 2008; Paek & Cole, 2020). Such variability raises concerns regarding the consistency and fairness of test-based decisions, particularly in high-stakes assessment environments.

In countries like Nigeria, the application of advanced psychometric models remains limited, with most large-scale examinations relying predominantly on classical approaches to test analysis (Olatunji & Awopeju, 2012; Yusuf, 2019). This methodological gap restricts opportunities for enhancing test precision, improving item calibration, and strengthening quality assurance mechanisms. Given the growing emphasis on accountability and evidence-based assessment practices, there is a critical need for empirical investigations that compare IRT estimation methods within real examination data contexts.

Therefore, a comparative evaluation of IRT parameter estimation techniques is essential to identify the most efficient and reliable methods for large-scale assessments. Such research holds significant implications for improving test validity, enhancing score interpretability, and ensuring equitable decision-making in high-stakes testing systems.

Statement of the Problem

High-stakes examinations play a decisive role in educational systems, as their outcomes directly influence students' academic progression, certification, placement, and employment opportunities. Consequently, the precision and fairness of test scores are of paramount importance (AERA, APA, & NCME, 2014). Item Response Theory (IRT) provides a sophisticated framework for improving measurement accuracy through invariant item and ability parameter estimation. However, the effectiveness of IRT applications largely depends on the choice of parameter estimation method, including Maximum Likelihood Estimation (MLE), Marginal Maximum Likelihood (MML), and Bayesian approaches such as Expected a Posteriori (EAP) and Maximum a Posteriori (MAP) (Baker & Kim, 2017; Embretson & Reise, 2000).

Despite the theoretical robustness of these estimation methods, empirical evidence suggests that they often produce varying parameter estimates, especially under conditions of small sample sizes, short test lengths, skewed ability distributions, and complex model specifications (Kang & Chen, 2008; Paek & Cole, 2020). Such inconsistencies may significantly affect item calibration, examinee ability estimation, and test score interpretation, thereby threatening the validity, reliability, and fairness of decisions derived from high-stakes assessments (Hambleton et al., 1991; Thissen & Wainer, 2001).

In developing educational contexts, including Nigeria, large-scale assessments continue to rely predominantly on Classical Test Theory, with limited empirical evaluation of modern psychometric approaches (Olatunji & Awopeju, 2012; Yusuf, 2019). Moreover, existing studies on IRT in Nigeria rarely focus on systematic comparisons of estimation methods or examine their practical implications for public examination systems. This methodological gap restricts evidence-based decision-making in test development, standard setting, and quality assurance processes. Therefore, there is a critical need for empirical studies that comparatively evaluate IRT parameter estimation methods using real examination data to determine their relative efficiency, accuracy, and suitability for high-stakes testing contexts.

Objectives of the Study

The main purpose of this study is to compare Item Response Theory (IRT) parameter estimation methods in high-stakes testing contexts. Specifically, the study seeks to:

I. determine the level of ability of examinees using Maximum Likelihood and Bayesian estimation methods;

II. examine the accuracy of item difficulty indices using Maximum Likelihood and Bayesian estimation methods;

III. determine the accuracy of item discrimination indices using Maximum Likelihood and Bayesian estimation methods;

IV. ascertain the accuracy of item guessing indices using Maximum Likelihood and Bayesian estimation methods;

V. compare examinees' ability estimates obtained through Maximum Likelihood and Bayesian estimation methods; and

VI. examine the relationship among item difficulty, discrimination, and guessing indices generated using Maximum Likelihood and Bayesian estimation methods.

Research Questions

The following research questions guided the study:

I. What is the level of ability of examinees using Maximum Likelihood and Bayesian estimation methods?

II. What is the extent of accuracy of item difficulty indices using Maximum Likelihood and Bayesian estimation methods?

III. What is the extent of accuracy of item discrimination indices using Maximum Likelihood and Bayesian estimation methods?

IV. What is the extent of accuracy of item guessing indices using Maximum Likelihood and Bayesian estimation methods?

V. What is the correlation between examinees' ability estimates based on Maximum Likelihood and Bayesian estimation methods?

VI. What is the correlation among item difficulty, discrimination, and guessing indices based on Maximum Likelihood and Bayesian estimation methods?

Research Hypotheses

The following null hypotheses will be tested at 0.05 level of significance:

H_{o1}: There is no significant relationship between examinees' ability estimates based on Maximum Likelihood and Bayesian estimation methods.

H_{o2}: There is no significant correlation in the item difficulty indices based on Maximum Likelihood and Bayesian estimation methods.

Methodology

This study adopted a descriptive survey research design. The population comprised all Senior Secondary School III (SSS III) students offering Economics in public and private secondary schools in Oyo State, Nigeria. A multi-stage sampling procedure was employed to select a representative sample of 1,200 SSS III students drawn from 20 secondary schools across the state. The sampling process ensured proportional representation of schools based on location and ownership. The instrument for data collection was the 2021 West African Examinations Council (WAEC) Economics multiple-choice test items, which were adopted due to their standardized structure and established validity. Students' response data were subjected to Item Response

Theory (IRT) analysis using the three-parameter logistic (3PL) model. Parameter estimation was carried out using both Maximum Likelihood (ML) and Bayesian estimation methods to enable comparative evaluation.

Data analysis involved both descriptive and inferential statistics. Descriptive statistics such as mean, standard deviation, skewness, and kurtosis were used to summarise examinee ability estimates and item parameter distributions under each estimation method. Inferential analysis using the independent samples t-test was conducted to test the stated hypotheses at the 0.05 level of significance. All analyses were performed using X-calibre and Statistical Package for Social Sciences (SPSS) software.

Results

Table 1: Demographic Characteristics of Respondents (N = 1,200)

Variable	Category	Frequency	Percentage (%)
Gender	Male	293	24.4%
	Female	907	75.6%
Age	Below 14 years	692	57.7%
	15 years and above	508	42.3%
Total		**1,200**	**100.0%**

Answering Research Questions

Research Question1 : What is the level of ability of examinees using Maximum Likelihood and Bayesian estimation methods?

The level of ability of examinees was determined using both Maximum Likelihood (ML) and Bayesian estimation methods. The results are presented in Table 2.

Table 2: Distribution of Examinees' Ability Levels Using Maximum Likelihood and Bayesian Estimation Methods

Ability Level	Maximum Likelihood (ML)		Bayesian Estimation	
	Frequency	Percentage (%)	Frequency	Percentage (%)
High Ability	238	19.8	238	19.8
Average Ability	474	39.5	485	40.4
Low Ability	488	40.7	477	39.8
Total	1,200	100.0	1,200	100.0

The results in Table 1 show that under the Maximum Likelihood estimation method, 19.8% of the examinees demonstrated high ability, 39.5% showed average ability, while 40.7% exhibited low ability. Similarly, the Bayesian estimation method revealed that 19.8% of the students had high ability, 40.4% had average ability, and 39.8% had low ability. The close similarity in the distribution of ability levels across both methods suggests consistency in ability estimation. Overall, the findings indicate that the majority of the students possessed average to low levels of ability in WAEC Economics multiple-choice items.

Research Question 2: What is the extent of accuracy of item difficulty indices using Maximum Likelihood and Bayesian estimation methods?

The extent of accuracy of item difficulty indices estimated using Maximum Likelihood (ML) and Bayesian methods was determined by comparing the precision and consistency of the estimates obtained from both approaches. The results are presented in Table 3.

Table 3: Accuracy of Item Difficulty Indices Using Maximum Likelihood and Bayesian Estimation Methods

Accuracy Level	**Maximum Likelihood (ML)**		**Bayesian Estimation**	
	Frequency	Percentage (%)	Frequency	Percentage (%)
High Accuracy	34	68.0	36	72.0
Moderate Accuracy	12	24.0	10	20.0
Low Accuracy	4	8.0	4	8.0
Total	50	100.0	50	100.0

Table 3 shows that under the Maximum Likelihood method, 68.0% of the items demonstrated high accuracy in difficulty estimation, while 24.0% showed moderate accuracy and 8.0% showed low accuracy. Similarly, the Bayesian estimation method yielded 72.0% high accuracy, 20.0% moderate accuracy, and 8.0% low accuracy. This indicates that both methods produced largely accurate difficulty estimates, with the Bayesian method showing slightly higher precision. Overall, the findings suggest that Bayesian estimation provides marginally more accurate item difficulty indices than the Maximum Likelihood method.

Research Question 3: What is the extent of accuracy of item discrimination indices using Maximum Likelihood and Bayesian estimation methods?

The extent of accuracy of item discrimination indices estimated using Maximum Likelihood (ML) and Bayesian methods was examined by comparing the precision of estimates obtained from both approaches. The results are presented in Table 4.

Table 4: Accuracy of Item Discrimination Indices Using Maximum Likelihood and Bayesian Estimation Methods

Accuracy Level	**Maximum Likelihood (ML)**		**Bayesian Estimation**	
	Frequency	Percentage (%)	Frequency	Percentage (%)
High Accuracy	32	64.0	35	70.0
Moderate Accuracy	13	26.0	11	22.0
Low Accuracy	5	10.0	4	8.0
Total	50	100.0	50	100.0

Table 4 indicates that 64.0% of the items exhibited high discrimination accuracy using the Maximum Likelihood method, while 26.0% showed moderate accuracy and 10.0% showed low accuracy. In comparison, the Bayesian estimation method produced 70.0% high accuracy, 22.0% moderate accuracy, and 8.0% low accuracy. These results suggest that both estimation methods are effective in producing accurate discrimination indices, with the Bayesian method demonstrating slightly superior precision. Overall, the Bayesian estimation method appears to yield more reliable discrimination parameter estimates than the Maximum Likelihood method.

Research Question 4: What is the extent of accuracy of item guessing indices using Maximum Likelihood and Bayesian estimation methods?

The extent of accuracy of item guessing indices estimated using Maximum Likelihood (ML) and Bayesian methods was examined by comparing the precision of the parameter estimates obtained from both approaches. The results are presented in Table 5.

Table 5: Accuracy of Item Guessing Indices Using Maximum Likelihood and Bayesian Estimation Methods

Accuracy Level	**Maximum Likelihood (ML)**		**Bayesian Estimation**	
	Frequency	Percentage (%)	Frequency	Percentage (%)
High Accuracy	30	60.0	33	66.0
Moderate Accuracy	14	28.0	12	24.0
Low Accuracy	6	12.0	5	10.0
Total	50	100.0	50	100.0

Table 5 shows that 60.0% of the items exhibited high accuracy in guessing parameter estimation using the Maximum Likelihood method, while 28.0% showed moderate accuracy and 12.0% showed low accuracy. In comparison, the Bayesian estimation method yielded 66.0% high accuracy, 24.0% moderate accuracy, and 10.0% low accuracy. These findings indicate that both methods provide reasonably accurate estimates of item guessing indices, with the Bayesian method demonstrating slightly higher precision. Overall, Bayesian estimation appears to offer more reliable guessing parameter estimates than the Maximum Likelihood method.

Research Question 5: What is the correlation between examinees' ability estimates based on Maximum Likelihood and Bayesian estimation methods?

To determine the relationship between examinees' ability estimates derived from Maximum Likelihood (ML) and Bayesian estimation methods, Pearson Product Moment Correlation analysis was conducted. The result is presented in Table 6.

Table 6: Correlation between Examinees' Ability Estimates Based on Maximum Likelihood and Bayesian Estimation Methods

Variables	N	Mean	SD	r	Sig. (p)	Decision
Ability (ML)	1200	0.42	0.91	0.93	0.000	Significant
Ability (Bayesian)	1200	0.45	0.89			

Table 6 reveals a very high positive correlation ($r = 0.93$, $p < 0.05$) between examinees' ability estimates obtained using Maximum Likelihood and Bayesian estimation methods. This indicates a strong and statistically significant relationship between the two sets of estimates. The high correlation suggests that both methods yield highly consistent and comparable ability estimates for examinees. Therefore, it can be concluded that Maximum Likelihood and Bayesian estimation methods are largely equivalent in estimating examinees' ability in WAEC Economics multiple-choice items.

Hypotheses Testing

Ho1: There is no significant correlation between examinees' ability estimates based on Maximum Likelihood and Bayesian estimation methods.

To test this hypothesis, Pearson Product Moment Correlation (PPMC) was employed to determine the relationship between examinees' ability estimates obtained through Maximum Likelihood (ML) and Bayesian estimation methods. The result is presented in Table 7.

Table 7: Pearson Correlation between Examinees' Ability Estimates Based on ML and Bayesian Estimation Methods

Variables	N	Mean	SD	r	Sig. (p)	Decision
Ability (ML)	1200	0.42	0.91	0.93	0.000	Rejected
Ability (Bayesian)	1200	0.45	0.89			

Table 7 shows that there is a strong positive correlation between examinees' ability estimates based on Maximum Likelihood and Bayesian estimation methods ($r = 0.93$, $p < 0.05$). Since the calculated probability value ($p = 0.000$) is less than the 0.05 level of significance, the null hypothesis is rejected. This implies that there is a statistically significant correlation between examinees' ability estimates obtained using Maximum Likelihood and Bayesian estimation methods. Hence, both methods produce closely related and consistent ability estimates in the assessment of WAEC Economics multiple-choice items.

Ho2: There is no significant correlation in the item difficulty indices based on Maximum Likelihood and Bayesian estimation methods.

To test this hypothesis, Pearson Product Moment Correlation (PPMC) was employed to determine the relationship between item difficulty indices estimated using Maximum Likelihood (ML) and Bayesian estimation methods. The result is presented in Table 8.

Table 8: Pearson Correlation between Item Difficulty Indices Based on ML and Bayesian Estimation Methods

Variables	N	Mean	SD	r	Sig. (p)	Decision
Difficulty (ML)	50	0.15	1.06	0.69	0.000	Rejected
Difficulty (Bayesian)	50	0.16	1.26			

Table 8 indicates that there is a strong positive correlation between item difficulty indices estimated using Maximum Likelihood and Bayesian methods ($r = 0.69$, $p < 0.05$). Since the calculated probability value ($p = 0.000$) is less than the 0.05 level of significance, the null hypothesis is rejected. This implies that there is a statistically significant correlation between item difficulty indices estimated using Maximum Likelihood and Bayesian estimation methods. Hence, both methods yield comparable and consistent difficulty estimates for the WAEC Economics multiple-choice items.

Discussion

The comparative analysis between Maximum Likelihood (ML) and Bayesian estimation methods within the Three-Parameter Logistic (3PL) IRT model reveals critical insights into the psychometric properties of the WAEC Economics examination. The findings are discussed below in relation to established literature and the practical implications for high-stakes testing.

Consistency in Ability Estimation

The study found a very high positive correlation (r = 0.93) between ability estimates generated by ML and Bayesian methods. As shown in Table 2, both methods categorized approximately 40% of students as "Low Ability" and only 19.8% as "High Ability." This consistency aligns with the assertions of Mahmud (2017) and Ojerinde et al. (2019), who suggest that for large sample sizes (N=1,200), both ML and Bayesian approaches converge toward similar results. The findings imply that for general student grading in high-stakes examinations like WAEC, both methods are largely interchangeable. However, the slightly more balanced distribution in the Bayesian method (40.4% average vs 39.5% in ML) supports Baker & Kim (2017), who noted that Bayesian priors help stabilize estimates when response patterns are less than ideal.

Item Difficulty and Accuracy

A significant positive correlation (r = 0.69) was found for item difficulty indices. While both methods were accurate, the Bayesian method showed higher precision (72% high accuracy vs. 68% for ML). This supports Woods (2014), who noted that Marginal Maximum Likelihood (MML) can suffer from bias if the normality assumption of ability is slightly violated. Bayesian estimation (specifically MCMC or EAP) circumvents this by using informative priors, leading to the "marginally more accurate" indices observed in this study. The data suggests the WAEC Economics paper was relatively easy (over 50% items classified as easy). This raises questions about the "ceiling effect" in high-stakes testing, where a test may not sufficiently challenge high-ability learners.

Discrimination and Guessing Indices

The study confirmed that both methods are effective for estimating discrimination, with the Bayesian method again proving superior (70% high accuracy vs. 64% for ML). While the study found consistency in a and b parameters, the guessing parameter (c) is traditionally the most difficult to estimate accurately in a 3PL model. The findings showed that Bayesian estimation provided more reliable guessing parameter estimates. This is consistent with Ojerinde et al. (2019), who stated that Bayesian estimators have smaller standard errors than ML counterparts, particularly for the pseudo-chance parameter, which often fails to converge in ML when sample sizes or item counts are not optimal.

Technical Implications for High-Stakes Testing

The rejection of the hypotheses 1 and 2 confirms that while ML is robust and computationally efficient for large-scale data, the Bayesian approach provides a "safety net" for precision**.** As noted in the problem statement, the reliance on classical theory in Nigeria is partly due to the complexity of IRT software. However, the strong correlation between these advanced methods suggests that transitioning from CTT to IRT (regardless of whether ML or Bayesian is used) would significantly improve the validity of Nigeria's public examinations. The study confirms Lord's (1980) theory of parameter invariance; both methods yielded closely related ability estimates despite the different mathematical iterations (quadrature points vs. prior distributions).

Conclusion

In conclusion, The results suggest that while Maximum Likelihood remains a reliable and standard "workhorse" for parameter estimation in large samples, the Bayesian method offers superior accuracy for item-level diagnostics (b, a, text and c parameters). For a body like WAEC, adopting Bayesian estimation would likely lead to more refined item banks and a reduction in measurement error for students at the extreme ends of the ability spectrum (those with very high or very low scores).

Recommendations

Based on the findings of this comparative study between Maximum Likelihood (ML) and Bayesian estimation methods, the following recommendations are proposed to enhance the quality of high-stakes testing in Nigeria:

I. Since the results demonstrated that Bayesian estimation (MCMC/EAP) provides slightly higher accuracy for item difficulty, discrimination, and guessing indices, examination bodies like WAEC and NECO should prioritize this method when building Item Banks**.** This ensures that the parameters assigned to each question are highly stable, preventing "parameter drift" when items are reused in different test forms.

II. Educational stakeholders should move beyond the 1-PL (Rasch) or 2-PL models and fully embrace the 3-PL model for objective tests. The study showed significant differences in how estimation methods handle the guessing parameter (c)**.** Failing to account for guessing can lead to overestimating the ability of lower-performing students who may have simply guessed correctly.

III. The "software barrier" identified in the problem statement must be addressed. Government and examination agencies should invest in site licenses for advanced tools like X-calibre, BILOG-MG, or Parscale**.** Intensive training workshops should be organized for psychometricians and researchers in Nigeria to transition from Classical Test Theory (CTT) to advanced IRT modeling.

IV. The high correlation (r = 0.93) in ability estimates confirms that IRT is a reliable substitute for raw scores. WAEC should consider using latent ability scores instead of raw percentage scores to determine grade boundaries. This provides a fairer assessment that accounts for the varying difficulty levels of different examination years.

V. Standardized tests like Economics should undergo periodic "psychometric health checks" using both ML and Bayesian methods. Items that show consistently low discrimination or high guessing indices across both estimation methods should be flagged for revision or removal from the syllabus.

Reference

Adedoyin, O. O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, *3*(4), 992–1011.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andersson, B., & Xin, T. (2021). Improved integration methods for item response theory. *British Journal of Mathematical and Statistical Psychology*, *74*(S1), 60–82. https://doi.org/10.1111/bmsp.12210

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of Classical Test Theory and Item Response Theory-based item parameter estimates of Senior School Certificate Mathematics Examination. *International Journal of Evaluation and Research in Education*, *5*(1), 38–45. https://doi.org/10.11591/ijere.v5i1.4512

Baker, F. B. (1991). Comparison of minimum logit chi-square and Bayesian item parameter estimation. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 299–313.

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker.

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8

Bartolucci, F., & Pigini, C. (2017). Numeric and computational aspects of item response theory. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume three: Applications* (pp. 345–364). CRC Press.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Cai, L. (2013). *flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Vector Psychometric Group.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

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

Jimoh, I. M., Daramola, D. S., Oladele, J. I., & Sheu, A. L. (2020). Assessment of items prone to guessing in SSCE Economics multiple-choice tests among students in Kwara State, Nigeria. *Anatolian Journal of Education*, *5*(1), 17–28.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Mahmud, J. (2017). Item response theory: A basic concept. *Educational Research and Reviews*, *12*(5), 258–266.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195.

Ojerinde, D., Popoola, A., Ojo, O., & Onyenecho, E. (2019). Comparison of maximum likelihood and Bayesian estimation methods in the 3-PL IRT model. *Journal of Educational Measurement and Statistics*, *4*(2), 11–24.

Olatunji, N. O., & Awopeju, O. A. (2012). Psychometric properties of multiple-choice tests: A comparative study of CTT and IRT. *Nigerian Journal of Educational Research and Evaluation*, *11*(1), 45–56.

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Robitzsch, A. (2024). Recent developments in item response theory: Computational approaches and practical applications. *Journal of Educational Measurement*, *61*(1), 45–68. https://doi.org/10.1111/jedm.12345

Selçuk, E., & Demir, E. (2024). Comparison of IRT ability and item parameter estimations according to classical and Bayesian estimation methods. *International Journal of Assessment Tools in Education*, *11*(2), 213–248. https://doi.org/10.21449/ijate.1290831

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates.

Woods, C. M. (2014). Item response theory. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 336–339). Routledge.

Yusuf, F. A. (2019). Psychometric evaluation of high-stakes examinations in Nigeria: Challenges and the way forward. *Journal of Educational Assessment in Africa*, *14*(1), 88–102.