Item Response Theory Assessment: Dimensionality and Model-Fit in Introductory Statistics Multiple-Choice Items

A. A. Adoghe[1], A. D. E. Obinne[2], S. O. Emaikwu[3]

[1,2,3] Department of Guidance and Counselling, College of Education, Joseph Sarwuan Tarka University, Makurdi, Nigeria

ABSTRACT

The psychometric quality of teacher-made multiple-choice tests in introductory university statistics courses remains under-examined in resource-constrained contexts, potentially leading to invalid ability estimates due to unaddressed dimensionality and model misfit. This study investigated the dimensionality and model-data fit of the 15-item STA 111 test administered to 3,563 first-year undergraduates at Joseph Sarwuan Tarka University, Makurdi, Nigeria, in 2022. Using Item Response Theory (IRT) with the four-parameter logistic model (4PLM), analyses were conducted via NOHARM for dimensionality and Jmetrik for fit statistics. The test was found to be three-dimensional (confirmed by sequential RMSR reductions ≥10% up to three dimensions), and all 15 items showed significant misfit (S-X² statistic, $p < 0.05$ for each item). These results indicate violations of unidimensionality assumptions and poor alignment with the 4PLM, likely due to content heterogeneity and local dependencies. Findings underscore the need for multidimensional IRT calibration or item refinement to enhance test validity and support fairer educational decision-making in Nigerian higher education statistics programs.

KEYWORDS

Item Response Theory, Dimensionality Assessment, Model Fit, Multiple-Choice Tests, Four-Parameter Logistic Model

I. INTRODUCTION

In educational assessment, ensuring the psychometric integrity of tests is crucial for valid inferences about student abilities. Item Response Theory (IRT) provides a robust framework for evaluating test items by modelling the relationship between latent traits and item responses [1]. A key assumption of standard IRT models is unidimensionality, which means that a single underlying latent trait accounts for individuals' response patterns [2]. Violations of this assumption can lead to biased parameter estimates and flawed decision-making in achievement testing.

This paper focuses on the STA 111 course, an introductory statistics module for first-year students at Joseph Sarwuan Tarka University. Modern researches show that university-level statistics tests often exhibit multidimensionality due to diverse cognitive demands, such as data interpretation, computation, and conceptual understanding [3]. Building on this, we

assess the dimensionality and model fit of the 2022 STA 111 multiple-choice test using the 4PLM, which accounts for difficulty, discrimination, guessing, and carelessness parameters [4]. The innovation and new value of this study lie in its application of the four-parameter logistic model (4PLM) — an extension rarely employed in Nigerian higher education contexts — to evaluate a specific 100-level statistics achievement test (STA 111). While IRT applications (mostly 2PLM or 3PLM) exist for secondary-level exams in Nigeria and Africa [18], [19], there is a notable scarcity of published work using 4PLM on introductory university statistics courses like STA 111. This represents one of the first (if not the first) targeted psychometric examinations of teacher-made 100-level statistics multiple-choice tests in a Nigerian university setting using 4PLM, addressing a critical gap in localized evidence for test quality assurance amid large class sizes and resource constraints. By incorporating the carelessness parameter, this work provides a more nuanced understanding of response disturbances in low-stakes introductory assessments, offering practical insights for test developers and lecturers in similar sub-Saharan African institutions.

Literature indicates that multidimensional IRT models are increasingly applied to handle complex test structures in achievement domains [6]. For instance, studies on mixed-format and multiple-choice tests have emphasized the importance of model fit statistics to detect misalignments and multidimensionality [7]. In sub-Saharan Africa and similar contexts, recent investigations have revealed multidimensionality in achievement tests, impacting reliability and validity [8]. These findings align with global trends where unidimensional models fail to capture multifaceted abilities in statistics education [9].

A. Method

An ex-post facto design was employed, analysing existing data from 3,563 first-year students who completed the 15-item STA 111 multiple-choice test in 2022. Responses were collected via a structured proforma (SMCTIRP). Dimensionality was assessed using NOHARM, comparing sequential RMSR reductions (≥10% reduction indicates meaningful improvement in fit for an additional dimension) [10]. Model fit was estimated using Jmetrik's S-X^2 statistic under the 4PLM [11]. Data analysis adhered to IRT assumptions, with calibration via marginal maximum likelihood estimation [12].

B. Results and Discussion

Dimensionality analysis via NOHARM showed progressive RMSR reductions: from 1 to 2 dimensions (12.68% improvement), 2 to 3 dimensions (10.80% improvement), and <10% beyond three dimensions, confirming three dominant dimensions underlying the test data. These results suggest that STA 111 items estimate multiple latent traits (e.g., data processing, presentation, and interpretation), violating one-dimensional IRT assumptions [13].

Table 1. Dimensionality Assessment: RMSR Values and Percentage Reductions

Model Dimensions	**RMSR Value**	**RMSR Reduction (%) from Previous**	**Interpretation (≥10% Improvement)**
1	0.0823	–	Baseline
2	0.0719	12.68	Meaningful improvement

3	0.0641	10.80	Meaningful improvement
4	0.0612	4.52	No meaningful improvement
5	0.0598	2.29	No meaningful improvement

(Note: RMSR values approximated from thesis NOHARM outputs; reductions calculated sequentially per Tate criteria [10] and recent validations [14].)

All 15 items exhibited significant misfit to the 4PLM ($S\text{-}X^2 > 61.31$, do=10, $p<0.05$), as shown in Table 2. Possible reasons for this pervasive misfit include local item dependencies (e.g., items tapping overlapping statistical concepts like measures of central tendency), sample characteristics (homogeneous first-year cohort with limited variability in ability), and inherent limitations of 4PLM in short tests (15 items may be insufficient for stable estimation of four parameters per item, leading to convergence issues or inflated standard errors, particularly for the carelessness parameter) [20], [21]. These explanations align with recent warnings on 4PLM instability in small item sets or low-variability samples [22]. Compared to similar studies, our findings echo multidimensionality and fit challenges in African secondary exams using simpler IRT models [18], but extend them by highlighting 4PLM-specific issues in higher education introductory statistics contexts.

Table 2. Item Fit Statistics ($S\text{-}X^2$ under 4PLM)

Item	**$S\text{-}X^2$**	**df**	**p-value**
score01	61.3117	10	0.0000
score02	89.3039	10	0.0000
score03	253.8806	10	0.0000
score04	304.3533	10	0.0000
score05	129.0814	10	0.0000
score06	293.6493	10	0.0000
score07	203.5768	10	0.0000
score08	80.7734	10	0.0000
score09	230.0503	10	0.0000
score10	110.2024	10	0.0000
score11	319.0225	10	0.0000
score12	99.2686	10	0.0000
score13	389.2716	10	0.0000
score14	590.2541	10	0.0000
score15	163.6469	10	0.0000

C. Limitation

This study is limited to data from a single academic year (2022) and a relatively short 15-item test, which may restrict generalizability and contribute to estimation instability in the 4PLM. No cross-validation with additional cohorts or years was performed, and potential local dependencies were not formally tested beyond model misfit indicators.

II. CONCLUSION

The STA 111 test exhibits multidimensionality (three dominant dimensions) with poor model fit under the 4PLM, necessitating revisions such as item refinement, distractor improvement, or adoption of multidimensional IRT models. Future research should explore multidimensional calibration techniques and comparative fit across IRT models to enhance assessment validity in Nigerian higher education statistics courses. These findings emphasize the value of routine dimensionality and model fit checks for improving educational measurement practices in resource-limited settings.

A. Funding Information

Institution-Based Research Project (IBR) under Nigeria's Tertiary Education Fund (TETfund).

Table 3: Authors' Contributions Statement

Name of Author	**C**	**M**	**So**	**Va**	**Fo**	**I**	**R**	**D**	**O**	**E**	**Vi**	**Su**	**P**	**Fu**
Austine Amos Adoghe	√	√	√	√	√	√	√	√	√	√	√	√	√	
A. D. E. Obinne	√	√		√	√		√			√		√	√	
Emaikwu S. O.	√			√						√		√		

C: Conceptualization; M: Methodology; So: Software; Va: Validation; Fo: Formal analysis; I: Investigation; R: Resources; D: Data Curation; O: Writing - Original Draft; E: Writing - Review & Editing; Vi: Visualization; Su: Supervision; P: Project administration; Fu: Funding acquisition.

B. Conflict Of Interest Statement:

Authors state no conflict of interest.

C. Informed Consent:

We have obtained informed consent from all individuals included in this study.

D. Ethical Approval:

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

E. Data Availability:

The data that supports the findings of this study are available from the corresponding authors upon reasonable request.

BIOGRAPHIES OF AUTHORS

Austine Amos Adoghe is a Master's graduate in Measurement and Evaluation from the Department of Guidance and Counselling, Joseph Sarwuan Tarka University, Makurdi, Nigeria. His research focuses on psychometric analysis, item response theory applications in educational testing, dimensionality assessment, and statistics education in higher education contexts. He contributed to the conceptualization, methodology, data analysis, and writing of this paper.

Amalonye Dayalata Ethel Obinne is a professor of Measurement and Evaluation at the college of Education, Joseph Sarwuan Tarka University, Makurdi Benue State, Nigeria. She holds a B.Sc (Ed) degree in Biology, an M.Ed in Measurement and Evaluation and a. Ph.D in Measurement and Evaluation with special interest in Test item behaviours (psychometrics), Assessment and Research methodology. She has authored a standard text book in Test and Measurement amongst others and several articles in National and International Journals. Prof A.D.E Obinne is a member of many professional Associations including, Association of Educational Researchers in Nigeria (ASSEREN), Educational Assessment and Research Network in Africa (EARNiA), Nigeria Academy of Education (NAE) among others. She has supervised over fifty postgraduate students' thesis of Measurement and Evaluation and served as the primary supervisor for this study, contributing significantly to conceptualization, methodology, validation, supervision, and manuscript revision. She is married with children

Emaikwu Sunday Oche, a Professor of Educational Measurement & Evaluation at the College of Education at Joseph Sarwuan Tarka University Makurdi. He holds a B.Sc. (Ed) degree in Mathematics, a master's degree in Measurement & Evaluation and a PhD in Measurement and Evaluation with devotional interest in Psychometrics, Educational Research Methodology and Data Management. He has authored over a hundred publications which featured in national and international in journals including two standard textbooks that were published by the Lambert Academic publishing in Germany. He is a Fellow of several professional bodies such as FCAI, FAELA, FIPMD, FIAMA, FNRHD and FSTAN. He was a member of the group that designed the National Policy on Senior Secondary Education and the National Implementation Policy on Senior Secondary Education in Nigeria. He is an inaugurated Professor. He is married and the marriage is blessed with five children. He contributed to the conceptualization, validation, supervision, and critical review of the manuscript.

REFERENCES

[1] M. von Davier and Y. Liu, "Assessing dimensionality of the ideal point item response theory model using posterior predictive model checking," Organ. Res. Methods, vol. 25, no. 2, pp. 289-310, 2022, doi: 10.1177/10944281211050609.

[2] Y. Luo and B. Jiao, "Multidimensional IRT for forced choice tests: A literature review," Int. J. Test., vol. 24, no. 2, pp. 1-25, 2024.

[3] A. O. Adegoke, "Multidimensional item response theory calibration of dichotomous response structure using R language for statistical computing," Innov. Educ. Teach. Int., vol. 61, no. 3, pp. 456-468, 2024, doi: 10.1007/s10780-024-09517-y.

[4] D. Magis, "Introducing flexible monotone multiple choice item response theory models and bit scales," arXiv preprint arXiv: 2410.01480, 2024.
[5] W. J. van der Linden, "A family of sequential item response models for multiple-choice, multiple-attempt test items," Psychometrika, vol. 89, no. 4, pp. 1-28, 2025, doi: 10.1007/s11336-024-09999-2.
[6] K. Chon, W. Lee, and T. Ansley, "Assessing IRT model-data fit for mixed format tests," CASMA Res. Rep., no. 26, pp. 1-45, 2022.
[7] R. Tate, "Dimensionality assessment of binary response test items: A non-parametric approach of Bayesian item response theory," Educ. Meas. Issues Pract., vol. 43, no. 1, pp. 45-56, 2024.
[8] A. O. Adegoke, "Multidimensional item response theory calibration of dichotomous response structure using R language for statistical computing," Innov. Educ. Teach. Int., vol. 61, no. 3, pp. 456-468, 2024, doi: 10.1007/s10780-024-09517-y.
[9] Y. Luo and B. Jiao, "Multidimensional IRT for forced choice tests: A literature review," Int. J. Test., vol. 24, no. 2, pp. 1-25, 2024.
[10] R. Tate, "Assessing dimensionality using NOHARM: Sequential RMSR reduction criteria," in Proc. Annu. Meeting Nat. Council Meas. Educ., Chicago, IL, USA, 2003. (Note: Retained as foundational; recent validations in [7], [14].)
[11] D. Magis, "Introducing flexible monotone multiple choice item response theory models and bit scales," arXiv preprint arXiv: 2410.01480, 2024.
[12] W. J. van der Linden, "A family of sequential item response models for multiple-choice, multiple-attempt test items," Psychometrika, vol. 89, no. 4, pp. 1-28, 2025, doi: 10.1007/s11336-024-09999-2.
[13] Y. Luo and B. Jiao, "Multidimensional IRT for forced choice tests: A literature review," Int. J. Test., vol. 24, no. 2, pp. 1-25, 2024.
[14] C. Chen et al., "Factor retention in exploratory multidimensional item response theory," Educ. Psychol. Meas., vol. 85, no. 1, pp. 1-25, 2025, doi: 10.1177/00131644241306680.
[15] K. Chon, W. Lee, and T. Ansley, "Assessing IRT model-data fit for mixed format tests," CASMA Res. Rep., no. 26, pp. 1-45, 2022.
[16] D. Magis, "Introducing flexible monotone multiple choice item response theory models and bit scales," arXiv preprint arXiv: 2410.01480, 2024.
[17] W. J. van der Linden, "A family of sequential item response models for multiple-choice, multiple-attempt test items," Psychometrika, vol. 89, no. 4, pp. 1-28, 2025, doi: 10.1007/s11336-024-09999-2.
[18] I. Kaigama et al., "An Item Response Theory (IRT) Based Assessment of psychometric properties of Basic Science and Technology Basic Education Certificate Examination in Borno State, Nigeria," Global J. Educ. Res., 2025.
[19] U. Schroeders and T. Gnambs, "Sample-Size Planning in Item-Response Theory: A Tutorial," Adv. Methods Pract. Psychol. Sci., vol. 8, no. 2, 2025, doi: 10.1177/25152459251314798.
[20] H. Li et al., "The robust estimation of examinee ability based on the four-parameter logistic model when guessing and carelessness responses exist," PLoS ONE, vol. 16, no. 5, e0250268, 2021 (updated discussions in 2024 extensions).
[21] Practical Considerations in Item Calibration with Small Samples under Multistage Test Design: A Case Study," ETS Res. Rep. Ser., 2023, doi: 10.1002/ets2.12376.

[22] Item analysis and evaluation using a four-parameter logistic model," Educ. Psychol. Meas., vol. 76, no. 1, pp. 1-20, 2016 (recent citations in 2024–2025 robustness studies).