British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 2, 2025 (pp. 75-85)
www.abjournals.org

ASSESSING ITEM DIFFICULTY, DISCRIMINATION, GUESSING, AND CARELESSNESS PARAMETERS OF THE MATHEMATICS ACHIEVEMENT TEST FOR SECONDARY SCHOOL STUDENTS IN EDO STATE, NIGERIA

Omaze Anthony Afemikhe[1*] and Kennedy Imasuen[2]

[1]Institute of Education, University of Benin, Benin City, Nigeria.
Email: tonyafemikhe@yahoo.co.uk

[2]Institute of Education, University of Benin, Benin City, Nigeria.
Email: kennedy.imasuen@uniben.edu

Cite this article:

Afemikhe, O. A., Imasuen, K. (2025), Assessing Item Difficulty, Discrimination, Guessing, and Carelessness Parameters of the Mathematics Achievement test for Secondary School Students in Edo State, Nigeria. British Journal of Education, Learning and Development Psychology 8(2), 75-85. DOI: 10.52589/BJELDP-4SKVBGUA

Manuscript History

Received: 20 Jun 2025

Accepted: 22 Jul 2025

Published: 30 Jul 2025

ABSTRACT: *This study assessed the psychometric properties of the Mathematics Achievement test for Secondary School Students in Edo State, Nigeria, using the four-parameter logistic model (4PLM) of Item Response Theory (IRT). The study adopted a descriptive survey design. The population comprised students from 312 public junior secondary schools in Edo State, while the sample consisted of 2,204 students selected from this population. The research instrument was a 40-item multiple-choice Mathematics Achievement developed by Afemikhe and Imasuen (2024). The instrument, previously validated and standardized, had a reliability coefficient of 0.89 using the Kuder-Richardson Formula 20 (KR-20). Unidimensionality of the data was verified through Principal Component Analysis using SPSS, while item calibration was conducted with Jmetrik IRT software to estimate item difficulty, discrimination, guessing, and carelessness parameters. The results revealed that most items demonstrated very high discrimination, indicating a strong capacity to differentiate between students with high and low levels of achievement in mathematics. Most items were difficult, suggesting that the test provided sufficient challenge for students. However, a high proportion of items displayed elevated guessing parameters, reflecting issues with distractor quality. On the positive side, carelessness was generally low, suggesting that students responded attentively. Based on the findings, it was recommended that the distractors of test items of the test be reviewed and improved to reduce guessing and that IRT frameworks be more widely adopted in the evaluation of educational assessments.*

KEYWORDS: Achievement Test, Discrimination, Difficulty, Guessing, Carelessness.

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 2, 2025 (pp. 75-85)

www.abjournals.org

INTRODUCTION

Mathematics, often perceived as abstract and demanding, necessitates consistent learner attention, effort, and motivation. The degree to which students involve themselves in mathematics-related learning activities is referred to as learners' engagement, a construct vital for educational success. Engagement significantly influences not only academic achievement but also student retention and interest in Science, Technology, Engineering, and Mathematics (STEM) fields (Fredricks et al., 2004). Therefore, understanding and accurately measuring student engagement in mathematics is critical for improving academic performance, particularly at the secondary school level.

Learners' engagement is typically examined across three interconnected dimensions: behavioural, emotional, and cognitive. Behavioural engagement encompasses observable actions such as attendance, active participation, and sustained effort in learning activities. Emotional engagement reflects students' interest, enjoyment, or sense of belonging within mathematics classrooms, while cognitive engagement entails the willingness to exert mental effort, employ effective learning strategies, and reflect critically on mathematical concepts (Fredricks et al., 2004; Adodo & Ojerinde, 2022). Accurately measuring these multidimensional components of engagement requires carefully designed instruments, often comprising items or statements that students respond to using rating scales.

For these engagement items to yield meaningful and valid results, their psychometric properties must be rigorously examined. Traditional psychometric approaches, particularly Classical Test Theory (CTT), predominantly focus on scale-level reliability indicators like Cronbach's alpha. However, a significant limitation of CTT is its inability to account for how individual items function across varying levels of the latent trait (e.g., engagement). Furthermore, CTT's item and person parameters are inherently dependent on the specific sample and examinee cohort under investigation (Hambleton et al., cited in Pardede et al., 2023; Zanon et al., 2016). This dependency implies that an item's perceived difficulty is contingent upon the ability distribution of the test-takers, leading to compromised discernment of examinee ability and item characteristics that fluctuate with changes in the examinee population and test context (Pardede et al., 2023). Consequently, relying solely on CTT may obscure flawed items that could diminish the overall validity of a scale (Ferrando & Lorenzo-Seva, 2018), raising concerns about measurement precision and fairness, especially in diverse populations.

To surmount the limitations inherent in CTT, the field of psychometrics has progressively embraced Item Response Theory (IRT), also referred to as modern trait theory or latent trait theory. IRT offers a probabilistic framework that elucidates the relationship between unobservable latent traits (e.g., mathematics engagement) and observed item responses (Gyamfi & Wren, 2022; Butakor, 2022). Unlike CTT, IRT employs statistical models where both person and item parameters serve as predictors of observed performance (Gyamfi, 2023), offering deeper insights into the quality of measurement instruments (Embretson & Reise, 2018; DeMars, 2021). IRT enables the estimation of several critical item parameters: Item Difficulty (b-parameter), Item Discrimination (a-parameter), Guessing Parameter (c-parameter), and Carelessness Parameter (d-parameter). Understanding these parameters is crucial for evaluating and improving the validity and reliability of engagement scales.

Prior to the application of IRT models, certain foundational assumptions must be verified. For polytomous response formats, two principal assumptions are paramount: unidimensionality

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

and local independence (Bulut, 2015). A third assumption, parameter invariance, is particularly salient for dichotomous items. Several logistic models are employed within the IRT paradigm, each incorporating different combinations of these parameters: One-Parameter Logistic Model (1PLM), Two-Parameter Logistic Model (2PLM), Three-Parameter Logistic Model (3PLM), and Four-Parameter Logistic Model (4PLM). While the 4PLM offers a more nuanced representation, its practical application has historically been limited by challenges in parameter estimation and software availability. Nonetheless, recent advancements in computerized adaptive testing (CAT) have renewed scholarly interest in the 4PLM, especially for efficiently estimating abilities in the presence of careless errors (Dogruoz & Arikan, 2020; Kalkan, 2022; Liao et al., 2012; Loken & Rulison, 2010; Ogasuwar, 2017; Primi et al., 2018; Robitzsch, 2022; Pardede et al., 2023). Despite these developments, empirical investigations of carelessness within traditional paper-and-pencil testing contexts remain scarce (Pardede et al., 2023).

Numerous studies have explored learners' engagement and its influence on academic outcomes. Fredricks et al. (2004) provided the theoretical foundation for understanding engagement as a multifaceted construct, informing a wide array of global measurement tools. In the Nigerian context, Adodo and Ojerinde (2022) found that emotionally and cognitively engaged mathematics students performed significantly better, while Ogunleye (2023) emphasized high cognitive engagement as a strong predictor of mathematics achievement in senior secondary schools.

From a measurement perspective, many instruments used in engagement research rely on CTT, which typically assesses internal consistency but fails to examine how well individual items function. This reliance on CTT has led to concerns about measurement precision and fairness, particularly in diverse populations. In response to these limitations, researchers have increasingly turned to IRT. Baker and Kim (2017) highlighted IRT's superiority in analyzing item-level properties and ensuring more accurate measurement. IRT has been successfully applied in various educational settings to identify poor-performing items and improve assessment instruments. For instance, Chen and Thissen (2020) demonstrated IRT's ability to detect items affected by guessing and carelessness, factors that often compromise the reliability of rating scales.

Despite these advancements and the recognized importance of psychometrically sound instruments, there is a noticeable gap in research that applies IRT to the validation of engagement scales in mathematics, especially within the Nigerian secondary school context. Specifically in Edo State, little has been done to evaluate how well engagement items perform based on IRT parameters, including the often-overlooked carelessness parameter. Given the critical importance of mathematics to national development and the persistent challenges students face in engaging with the subject, it is imperative to comprehensively assess the difficulty, discrimination, guessing, and carelessness parameters of achievement test items used in secondary schools in Edo State. This study addresses this critical gap by employing IRT to provide a nuanced understanding of the psychometric properties of the Mathematics Achievement Test, contributing to the development of more valid and reliable assessment tools.

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

Research Question

1. What are the item properties of the Mathematics Achievement Test for secondary school students in Edo State, Nigeria, based on the discrimination, difficulty, guessing, and carelessness parameters?

MATERIALS AND METHODS

The study was a descriptive survey. The population of the study consists of the students in the 312 public junior secondary schools in Edo State. The sample size of the study was 2,204 students drawn from the schools. The research instrument for the study was the Mathematics Achievement Test for Secondary School Students developed by Afemikhe and Imasuen (2024). The instrument consists of 40 items with four (4) options lettered A-D. The candidates were required to select one correct answer from the four options. The response to each item was coded as 1 for a correct response and 0 for an incorrect response. The instrument was already subjected to the process of validation and standardization; hence, they are presumed to be valid. However, the reliability of the scores was ascertained using the Kuder-Richardson formula 20 (KR-20), and it gave an index of 0.89.

In analyzing the data, the principal component analysis was used to determine the unidimensionality of the data for the Mathematics Achievement multiple choice test items, using the Statistical Package for Social Sciences (SPSS). Thereafter, item calibration was done using Jmetrik IRT software, which can handle the four-parameter model. The criteria for grouping the difficulty parameter as proposed by Georgiev (2008) is b < -1.00 (very easy), - 1 ≤ b <0.00 (easy), 0.00 ≤ b <1 (hard), and b >1 (very hard). The criteria for grouping the discrimination parameter, as proposed by Baker and Kim (2017), 0 ≤ a <0.34 (very low), $0.35 \leq a \leq 0.64$ (low), $0.65 \leq a \leq 1.34$ is regarded as moderate, $1.35 \leq a \leq 1.69$ is seen as high, and $a \geq 1.70$ *as* very high). The criteria for grouping the guessing parameter is $\frac{1}{k}$ where k is the number of options in the test. In this study, the test used has 4 options. Therefore, the guessing parameter c is grouped as $0 \leq c \leq 0.25$ for low guessing, and c >0.25 for high guessing. The criteria for grouping the carelessness parameter d, as proposed by Guyer and Thompson, cited in Pardede et al. (2023), is 0 ≤ d <1. Less than 0.90 will be seen as high carelessness, and 0.90 and above as low carelessness. The research question was answered using frequency counts, mean, and standard deviation.

RESULTS

Before checking for the assumption of unidimensionality using the factor analysis, we needed to ensure the suitability of the response of the examinees for factor analysis and principal components. This was achieved by considering the correlation matrix and sample adequacy. Bartlett's test of sphericity ($\chi^2_{(780)}$ = 2069.047, p < 0.05) suggests that there is sufficient evidence not to accept that the correlation matrix formed is an identity matrix. The Kaiser-Meyer Olkin (KMO) factor adequacy (overall MSA = 0.928) demonstrated that the sample of responses on the test items was sufficient for each variable in the model and the complete model. Based on the two results, we proceeded with the analysis to check for the assumption

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

of unidimensionality through the factor theory analysis and principal components using the Statistical Package for Social Science (SPSS) version 27. This was to ascertain if the Mathematics Achievement multiple-choice test items measure only one dominant factor or latent trait.

Unidimensionality was established in the Mathematics Achievement multiple-choice test items; there was a presence of a dominant factor of the first Eigenvalue of 8.04, which accounted for 20.11% and is larger than the second eigenvalue of 2.77, which accounted for 6.92%. The spree plot also shows that the instrument is unidimensional, hence necessitating the analysis to be done using the IRT model.

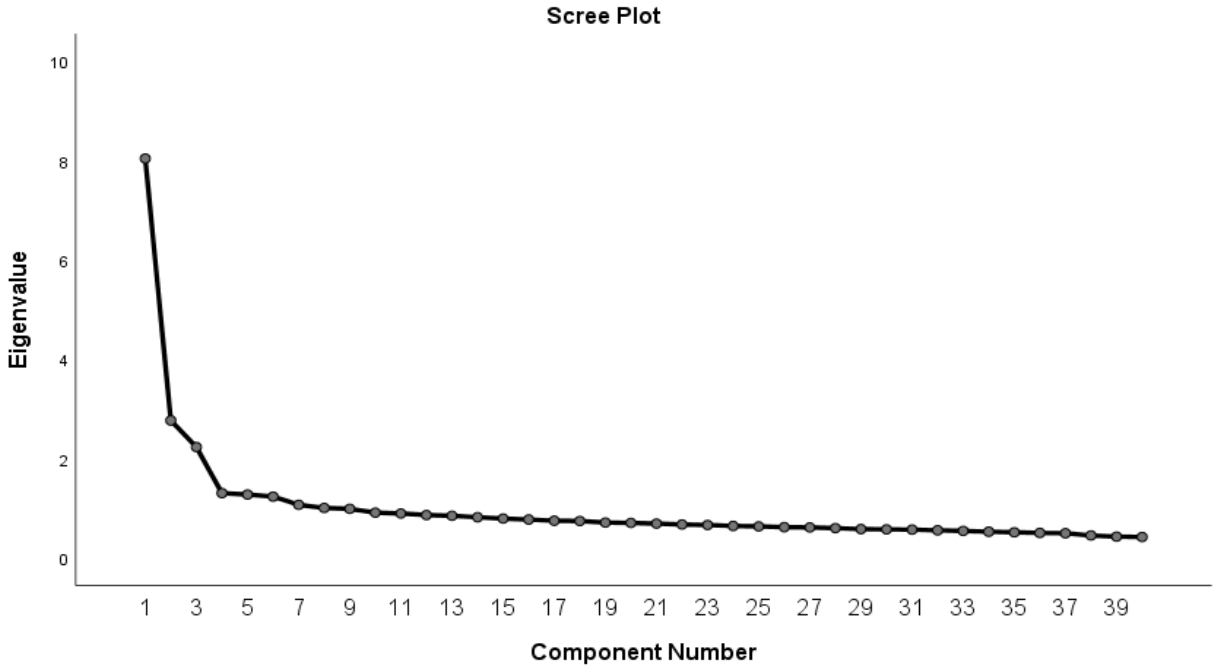

Figure 1: Spree Plot of the Mathematics Achievement Test

Table 1: The 4 Item Parameters of the Mathematics Achievement Test

Items	a (Discrimination)	b (Difficulty)	c (Guessing)	d (Carelessness)
1.	2.79	-0.89	0.04	0.86
2.	2.71	-0.93	0.06	0.88
3.	2.42	-0.80	0.15	0.84
4.	2.64	-0.81	0.05	0.85
5.	2.47	-0.82	0.07	0.84
6.	2.49	-0.86	0.09	0.85
7.	1.83	-0.04	0.50	0.97
8.	1.89	0.01	0.40	0.96
9.	1.95	-0.04	0.40	0.96
10.	2.53	1.10	0.36	0.98
11.	2.28	0.15	0.43	0.98
12.	2.44	0.44	0.38	0.98
13.	2.57	0.62	0.37	0.97

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

14.	2.59	0.29	0.41	0.98
15.	2.27	0.41	0.36	0.86
16.	2.60	0.53	0.39	0.99
17.	2.57	0.53	0.40	0.99
18.	2.50	0.83	0.35	0.98
19.	2.63	0.63	0.33	0.98
20.	2.66	0.33	0.38	0.99
21.	2.53	0.28	0.40	0.98
22.	2.11	0.21	0.41	0.99
23.	2.58	0.20	0.39	0.87
24.	2.75	0.24	0.34	0.99
25.	2.36	0.48	0.34	0.99
26.	2.72	0.40	0.37	0.99
27.	2.40	0.55	0.31	0.98
28.	2.75	0.90	0.32	0.99
29.	2.72	0.68	0.53	0.99
30.	2.68	0.45	0.39	0.99
31.	2.83	0.54	0.32	0.99
32.	2.50	0.54	0.28	0.88
33.	2.63	0.81	0.34	0.98
34.	2.71	0.65	0.33	0.99
35.	2.36	0.15	0.27	0.98
36.	2.67	0.14	0.30	0.98
37.	2.38	0.07	0.31	0.99
38.	2.45	0.16	0.27	0.98
39.	2.34	0.59	0.31	0.98
40.	2.49	0.75	0.34	0.98

From the data in Table 1, almost all the items (92.5%) demonstrated very high discrimination, indicating a strong ability to differentiate between high- and low-performing students. The majority of items (82.5%) were hard, with only one item classified as very hard. A large majority (92.5%) of the items had high guessing parameters, suggesting potential issues with distractor quality that make guessing more likely, and most items (85%) showed low carelessness, indicating that students generally responded attentively and that incorrect responses likely reflected misunderstanding rather than random errors.

Table 2 shows that the mean discrimination parameter was 2.49, with a standard deviation of 0.24 and a standard error of 0.04. This high average discrimination value suggests that the items were very effective in differentiating between high- and low-ability students. The relatively low standard deviation and standard error indicate that this was a consistent pattern across the test, reinforcing the test's strength in measuring students' true ability levels. For the difficulty parameter, the mean was 0.24, with a standard deviation of 0.53 and a standard error of 0.08. This indicates that, on average, the test items were of moderate difficulty, tending slightly toward the easier end of the spectrum. The variability in difficulty, as shown by the standard deviation, suggests that while many items were moderately challenging, some were notably easier or harder, providing a good spread across the ability continuum.

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

The guessing parameter had a mean of 0.32, with a standard deviation of 0.12 and a standard error of 0.02. This average is somewhat higher than the ideal value expected for a four-option multiple-choice test, where a guessing value closer to 0.25 is preferable. The elevated guessing parameter implies that some items may have had distractors that were not sufficiently misleading, making it easier for students to guess the correct answers. The standard deviation reflects a moderate range of guessing probabilities among items. Lastly, the carelessness parameter, interpreted here as the upper asymptote (U), had a mean of 0.95, a standard deviation of 0.05, and a standard error of 0.01. A value close to 1.0 signifies minimal carelessness, indicating that most students who had the requisite ability answered the items correctly. This high average and low variability confirm that students were generally attentive and deliberate in their responses, and that errors were more likely due to actual lack of knowledge rather than random guessing or inattention.

Table 2: Mean and Standard Deviation of the Item Parameters of the Mathematics Achievement Test

Parameter	Mean	SD	SEM	Remarks
Discrimination	2.4902	0.2363	0.0374	Very high
Difficulty	0.2368	0.5306	0.0839	Moderately difficult
Guessing	0.3195	0.1168	0.0185	High guessing
Carelessness	0.9545	0.0532	0.0084	Low carelessness

DISCUSSION OF FINDINGS

The findings from this study revealed that an overwhelming majority of items (92.5%) demonstrated very high discrimination, indicating that these items were exceptionally effective at differentiating between students with high and low levels of achievement. This is a highly desirable property in educational assessments, particularly in mathematics, as it enhances the ability to detect subtle differences in learner traits across the mathematic achievement spectrum (Embretson & Reise, 2018; DeMars, 2021). According to Baker and Kim (2017), high discrimination parameters ensure that item responses accurately reflect the underlying latent trait, in this case, students' mathematics achievement. This result aligns with the conceptualization of achievement as a nuanced, multidimensional construct that requires instruments capable of distinguishing learners across a continuum of behavioural, emotional, and cognitive involvement (Fredricks et al., 2004; Adodo & Ojerinde, 2022).

Furthermore, the majority of items (82.5%) were classified as hard, with only one item (2.5%) considered very hard. While these values suggest that the items were appropriately challenging, they may slightly exceed the average engagement or ability level of the target student population. However, such difficulty levels may be defensible, given that achievement in mathematics necessitates sustained cognitive effort, motivation, and persistence, traits often underdeveloped among disengaged or low-performing students (Skinner et al., 2009). Still, the dominance of hard items underscores the need for a more balanced representation of item difficulty to accommodate a wider range of learner abilities, a point echoed by Ferrando and Lorenzo-Seva (2018), who emphasized that assessments should include both lower- and higher-difficulty items to support measurement precision and fairness.

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

A notable concern emerged from the analysis of the guessing parameter, where 92.5% of items had values exceeding the ideal threshold of 0.25. This suggests potential problems with distractor quality, where incorrect response options may not have been sufficiently plausible, allowing students to answer correctly by guessing rather than demonstrating actual engagement or knowledge. Chen and Thissen (2020) observed that high guessing parameters are often symptomatic of poorly constructed distractors, which can significantly compromise the validity of test scores. Similarly, Haladyna and Downing (2004) emphasized that the effectiveness of multiple-choice items largely depends on the quality of their distractors. Items susceptible to guessing undermine the interpretive accuracy of engagement scores by inflating responses that do not correspond to genuine learner traits (Crocker & Algina, 2008). On a more encouraging note, the carelessness parameter (represented by the upper asymptote, *d*) showed that 85% of the items had values above 0.90, indicating low levels of carelessness. This suggests that students engaged thoughtfully with the test items, and their responses were indicative of authentic levels of engagement rather than inattentive or random behaviour. This result supports the notion that when learners are motivated or perceive an assessment as meaningful, they are more likely to respond with intentionality (Primi et al., 2018). According to Pardede et al. (2023), accounting for carelessness using models like the 4PLM enhances the reliability and interpretability of assessment scores, especially in high-stakes or diagnostic settings.

The summary statistics further reinforced these findings. The mean discrimination value of 2.49, coupled with a standard deviation of 0.24 and standard error of 0.04, confirms that the items consistently possessed strong discriminative power across the scale. This level of precision is in line with the standards proposed by Hambleton et al. (1991), who asserted that high discrimination values are essential for effective latent trait measurement. It also affirms the theoretical expectation that engagement instruments should be sensitive enough to differentiate varying levels of motivation, attention, and cognitive investment (Gyamfi, 2023). Similarly, the mean difficulty index of 0.24 falls within the acceptable range for educational testing, although it leans toward moderately easy items. The standard deviation of 0.53 indicates a fairly wide spread of item difficulty levels, which is beneficial for capturing diverse learner profiles (Butakor, 2022). A varied difficulty distribution enhances the scale's ability to measure engagement across different levels of mathematical understanding and academic readiness.

However, the mean guessing parameter of 0.32, higher than the recommended value of 0.25 for four-option multiple-choice items, raises persistent concerns about distractor efficacy. Fredricks et al. (2004) and Ogunleye (2023) emphasized that valid engagement assessment must capture students' genuine behavioural, emotional, and cognitive involvement, not artefacts of test-taking strategies or lucky guesses. Elevated guessing undermines the interpretive validity of such assessments by distorting the relationship between item responses and underlying traits.

Finally, the carelessness parameter mean of 0.95, with a standard deviation of 0.05, further confirmed that the vast majority of items were not significantly affected by inattentive responding. This result aligns with prior studies by Primi et al. (2018) and Loken and Rulison (2010), which advocate for incorporating the upper asymptote parameter in IRT models to capture and control for careless errors, especially in paper-based testing environments.

Taken together, these findings provide strong empirical support for the psychometric quality of the Mathematics Achievement Test, particularly regarding item discrimination and

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

respondent attentiveness. Nonetheless, the elevated guessing parameter and prevalence of hard items suggest areas for improvement, particularly in ensuring that the scale is well-targeted and the distractors are plausible. Addressing these issues will enhance the validity and fairness of the scale, ensuring that it more accurately reflects students' true levels of mathematics engagement. These findings also corroborate the broader theoretical understanding of student engagement as a complex, multidimensional construct that demands precise, nuanced measurement tools (Fredricks et al., 2004; Skinner et al., 2009). The application of Item Response Theory, especially the 4PLM, offers a rigorous and interpretable framework for refining such tools. Unlike Classical Test Theory, which assumes sample-dependent parameters, IRT allows item characteristics to remain invariant across populations, thereby supporting greater generalizability, fairness, and diagnostic utility (Embretson & Reise, 2018; Baker & Kim, 2017; Zanon et al., 2016).

CONCLUSION

This study employed the four-parameter logistic model (4PLM) of Item Response Theory (IRT) to evaluate the psychometric properties of the Mathematics Achievement Test among secondary school students in Edo State, Nigeria. The findings revealed that the majority of items exhibited very high discrimination, indicating a strong ability to differentiate between students with varying levels of engagement. Most items were categorized as hard, suggesting that the scale appropriately challenges students, though some may be overly difficult for the average learner. The analysis also uncovered a high rate of guessing behaviour, highlighting weaknesses in distractor design that may affect the validity of some items. Nonetheless, carelessness was generally low, signifying that students approached the items attentively and that responses likely reflected genuine levels of engagement.

RECOMMENDATIONS

1. Given the high guessing parameter observed in most items, test developers should critically examine and revise the distractors (incorrect options) in the scale.
2. Schools and examination bodies should adopt IRT frameworks for evaluating test instruments, especially when measuring latent traits like engagement.
3. Training programs should be organized to equip teachers and item writers with knowledge of IRT principles, including how to construct high-quality multiple-choice items, develop effective distractors, and interpret item parameters, such as discrimination and guessing.
4. The validated engagement scale can be used by school administrators, counsellors, and classroom teachers to identify students with low levels of behavioural, emotional, or cognitive engagement.

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

Acknowledgement

This work was financially supported by Tertiary Education Trust Fund (TETFUND) Nigeria (2020/2023) Research Project (RP) Intervention for the University of Benin, Benin City.

Ethics Statements

Consent was obtained from the school principals who act as parents to these students while in school.

Conflict of Interest

There is no conflict of interest among the authors.

REFERENCES

Adodo, S. O., & Ojerinde, D. (2022). Students' engagement and academic achievement in mathematics: The mediating role of emotional and cognitive participation. *Journal of Educational Research and Development*, 13(2), 84–98.

Afemikhe, O.A. & Imasuen, K. (2024). Student engagement in learning secondary school mathematics in Edo state. An unpublished research report; financially supported by Tertiary Education Trust Fund (TETFUND) Nigeria (2020/2023) Research Project (RP) Intervention for the University of Benin, Benin City.

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

Bulut, O. (2015). Applying item response theory models to entrance examination for graduate studies: Practical issues and insights. *Journal of Measurement and Evaluation in Education and Psychology*, 6(2), 313–330. https://doi.org/10.21031/epod.17523

Butakor P. K. (2022). Using classical test and item response theories to evaluate the psychometric quality of the teacher-made test in Ghana. *European Scientific Journal, ESJ*, 18(1), 139. https://doi.org/10.19044/esj.2022.v18n1p139

Chen, S. Y., & Thissen, D. (2020). Detecting problematic items in educational assessments using IRT-based indices. Psychological Test and Assessment Modelling, 62(1), 23–45.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

DeMars, C. E. (2021). *Item response theory*. Oxford University Press.

DeMars, C. (2010). *Item response theory: Understanding statistical measurement*. Oxford University Press.

Doğruöz, E., & Arikan, Ç. A. (2020). Comparison of different ability estimation methods based on 3 and 4PL item response theory. *Pamukkale University Journal of Education, 50(1)*, 50– 69. https://doi.org/10.9779/pauefd.585774

Embretson, S. E., & Reise, S. P. (2018). *Item response theory for psychologists*. Psychology Press.

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA

Ferrando, P. J., & Lorenzo-Seva, U. (2018). *Assessing the quality and functioning of test items: A review and practical guide using IRT.* Applied Measurement in Education, 31(2), 119–134.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). *School engagement: Potential of the concept, state of the evidence.* Review of Educational Research, 74(1), 59–109. https://doi.org/10.3102/00346543074001059

Georgiev, N. (2008). Item analysis of C, D and E series from Raven's standard progressive matrices with item response theory two-parameter logistic model. *Europe's Journal of Psychology*, 4(3). https://doi.org/10.5964/ejop.v4i3.431

Gyamfi, A. & Wren, D. (2022). Determining the difficulty and discrimination parameters of a Mathematics performance-based assessment. *Creative Education*, 13(11), 3483-3489.

Gyamfi, A. (2023). Differential item functioning of performance-based assessment in mathematics for senior high schools. *Journal of Evaluation and Learning*, 5(1), 20-34.

Haladyna, T. M., & Downing, S. M. (2004). *Construct-irrelevant variance in high-stakes testing.* Educational Measurement: Issues and Practice, 23(1), 17–27.

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

Kalkan, O. K. (2020) The comparison of estimation methods for the four-parameter logistic item response theory model. *Measurement Interdisciplinary Research and Perspectives*, 20(2):73-90

Liao, W. W., Ho, R. G., Yen, Y. C. and Cheng, H. C. (2012). The four-parameter logistic item response theory model is a robust method of estimating ability despite aberrant responses. *Social Behaviour and Personality: An International Journal*, 40(10), 1679–1694.

Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63(3), 509–525

Ogasawara, H. (2017). Identified and unidentified cases of the fixed-effects 3- and 4-parameter models in item response theory. *Behaviormetrika*, 44(2), 405–423. https://doi.org/10.1007/s41237-017-0032-x

Ogunleye, A. A. (2023). Cognitive engagement and achievement in mathematics among senior secondary students in Lagos State. *Nigerian Journal of Mathematics Education*, 17(1), 33–49.

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL I RT model and strategies for estimating it. REID (*Research and Evaluation in Education*), 9(1), 86-117. https://doi.org/10.21831/reid.v9i1.63230

Primi, R., Nakano, T. D. C., & Wechsler, S. M. (2018). Using four-parameter item response theory to model human figure drawings. *Journal of Psychological Assessment*, 17(4), 473–483. https://doi.org/10.15689/ap.2018.1704.7.07

Robitzsch, A. (2022). Four-parameter guessing model and related item response models. *Mathematical and Computational Applications*, 27(6), 1–16. https://doi.org/10.3390/mca27060095

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioural and emotional participation in academic activities. *Educational and Psychological Measurement*, 69(3), 493–525.

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psychology: Reflection and Criticism*, 29(1), 1– 10. https://doi.org/10.1186/s41155-016-0040-x

Article DOI: 10.52589/BJELDP-4SKVBGUA
DOI URL: https://doi.org/10.52589/BJELDP-4SKVBGUA