Full research paper

EFFECT OF 2-PL AND 3-PL MODELS ON THE ABILITY ESTIMATE IN MATHEMATICS BINARY ITEMS

Rukayat Oyebola Iwintolu[1]
Oluwaseyi Aina Gbolade Opesemowo[2]✉
Phebean Oluwaseyi Adetutu[3]

[1]Osun State University, Nigeria

[2]University of Johannesburg, South Africa

[3]Obafemi Awolowo University, Nigeria

✉ oopesemowo@uj.ac.za

ABSTRACT

The investigation delves into examining the influence of 2-parameter logistic (PL) and 3-parameter logistic models on the ability estimates of students in mathematical binary items. It ascertained the parameters of the items in the 2-PL and 3-PL models. We employed Item Response Theory (IRT) in the design of this research survey, with a sample comprising 1015 senior secondary (SS) students in SS III classes who were analyzed using both models in the investigation. The Mathematics Achievement Test instrument was adapted from the General Mathematics Paper 1 of the Senior School Certificate Examination administered by the West Africa Examinations Council (WAEC). Results indicated that the 2-PL model shows lower difficulty levels but higher discriminatory indices. Statistical analysis revealed a significant ($F = 19.52$, $p < 0.05$ and $F = 18.52$, $p < 0.05$) effect of both models, respectively, on ability estimates in mathematics binary items among Nigerian secondary school students. We established that item parameters in the 2-PL and 3-PL models significantly affected the ability estimate of Nigeria secondary school students in binary mathematics items, while the 3-PL model provided a better ability estimate than the 2-PL model.

KEYWORDS

Items, binary items, models, item response theory, item parameters

HOW TO CITE

Iwintolu R. O., Opesemowo O. A. G., Adetutu P. O. (2024) 'Effect of 2-PL and 3-PL Models on the Ability Estimate in Mathematics Binary Items', *Journal on Efficiency and Responsibility in Education and Science*, vol. 17, no. 3, pp. 257–272. http://dx.doi.org/10.7160/eriesj.2024.170308

Article history
Received
April 13, 2024
Received in revised form
July 10, 2024
Accepted
September 3, 2024
Available on-line
September 30, 2024

Highlights

- *The efficacy of 2-PL and 3-PL models were evaluated in estimating students' ability in Mathematics binary items.*
- *The models had a significant effect in estimating students' ability*
- *The 3-PL model estimated ability better than the 2-PL model.*
- *Continued research into the effects of item parameters on ability estimation across different subjects and grade levels will be crucial for advancing assessment practices and promoting academic success.*

INTRODUCTION

Tests are standardized instruments used to obtain a sample of an examinee's best attempt at aptitude/achievement test, which gives an estimate of their performance/ability (Adetutu and Iwintolu, 2017; Breuer et al., 2023; Gates, 2023; Opesemowo et al., 2018) or a representation of an individual's standard performance on surveys or assessments where they reveal their typical emotions, beliefs, preferences, or responses to situations (O'Connor et al., 2019; Powers, 2019). Different peculiarities, strengths, and weaknesses characterize the aptitude/achievement tests, including the essay and objective tests. There are various objective tests: the short-answered test, the completion test, multiple choice, matching, cloze tests, and binary choice tests. The multiple-choice tests (binary scored) have gained significant acceptance among item-generation experts, even in Nigeria's standardized tests (Opesemowo et al., 2023). Among the objective test types, the multiple-choice test is generally known as the most commonly relevant, valuable, and used (Danh et al., 2020). It is fit for measuring complex outcomes in knowledge, understanding, application, and problem-solving skills. In Nigeria, like other countries, multiple-choice items are popular test types among examination organizations, as Douglas et al. (2023), Kalhori and Abbasi (2017), and Rios and Soland (2022) alluded. The organizations include the West Africa Examinations Council (WAEC), National Examinations Council (NECO), National Teachers Institution (NTI) Examination, and Joint Admission and Matriculation Board (JAMB) are some of the organizations involved.

Essentially, users of multiple-choice tests typically use binary scoring (i.e., assigning a value of one for a correct response and zero for an incorrect response), commonly analyzed using Classical Test Theory (CTT) techniques due to its ease of interpretation. Using CTT, examinees' raw scores are summed;

Electronic ISSN 1803-1617 Printed ISSN 2336-2375

therefore, all tests and examinees are considered together in this model. Despite its widespread use, CTT has been critiqued based on its inability to capture the essence of a test taker's ability, as the actual score is not an inherent trait. Moreover, the difficulty of individual test items may fluctuate based on the composition of the test-takers, making comparison of results difficult across different tests. In contrast, the Item Response Theory (IRT) technique is garnering recognition in the fields of psychological and educational testing owing to its provision of more flexible and efficient approaches to test development, evaluation, and scoring compared to those stemming from CTT (Adetutu and Iwintolu, 2017; Olagunju and Iwintolu, 2023). In IRT, individual items and individual test takers are the objects of analysis (Awopeju and Afolabi, 2016; Setiawati et al., 2023). As a result, IRT is contingent upon the individual items within a test instead of a collective measure of item responses like test scores for its basic concepts (Alordiah, 2015; Baker, 2001).

Furthermore, among the ultimate assumptions of IRT is that the respective test examinee responding to a test item has some level of the underlying ability that the item is intended to measure. This underlying ability is called a latent trait, and IRT models seek to estimate the latent based on the examinee's responses to various test items. By accurately estimating an examinee's latent trait, IRT can provide more precise and reliable measures of ability than traditional test-scoring methods. However, in practical terms, we cannot directly measure the value of the examinee's ability parameter; thus, the best approach is to estimate it (Ayanwale, 2023; Bichi and Talib, 2018). A numerical score on the ability scale can represent each examinee. At different ability levels (θ), there is a probability that an examinee will answer an item correctly regardless of their ability level. This probability, denoted as P(θ), is low for examinees with lower abilities and high for those with higher abilities. When an examinee faces a set of test items during an examination, they bring their inherent ability (θ) or trait into the testing environment (Rafi et al., 2023; Zanon et al., 2016). Tests are designed to evaluate an examinee's position on the ability scale, enabling a standardized comparison of examinees to ascertain their relative placements (Rudner, 2019; Scheibling-Sève et al., 2020). Obtaining ability measures for everyone taking the test can help achieve two critical objectives. Firstly, it allows for appraising the examinee's underlying ability level. Secondly, it enables comparisons among examinees to determine rates, assign grades, award scholarships, and more.

IRT models, such as the 2 Parameter Logistic and 3 Parameter Logistic (PL) models, have become instrumental in educational assessment, particularly in measuring the student's abilities in mathematics. These models provide an advanced framework for analyzing binary items, where responses are either correct or incorrect, and have been extensively utilized in various educational settings (Jimoh et al., 2022). The influential nature of these models in estimating mathematics abilities is a subject of significant interest and research due to its implications for curriculum design, instructional strategies, and student evaluation. IRT presupposes an examinee can steadily provide correct responses to test items, contingent upon possessing the requisite abilities as demanded by the items. The interaction between the individual's trait and the parameters of the items determines the probability of answering a test item correctly. One of the main objectives of IRT is to establish a relationship between latent variables, such as the examinee's ability, and the likelihood of providing correct responses to test items. The primary models currently utilized are the 1-PL, 2-PL, and 3-PL.

THEORETICAL STRUCTURE

1-Parameter Logistic Model: This model is regarded as the most foundational IRT model. It is presumed that only one item parameter underpins the item response procedure. IRT literature often refers to this item parameter as difficulty, symbolized by *b* in the 1-PL model (Yustiandi and Saepuzaman, 2021). The *b*-parameter, representing a test item, typically aligns with θ, indicating a trait of an individual under consideration. Within this framework, all test items display an identical Item Characteristic Curve (ICC), differing solely in their positioning along the horizontal axis (*θ*). The b-parameter represents the item's or task's cognitive resistance in each cognitive task. The formula is presented below:

$$P_i\left(\theta_j\right)=\frac{1}{1+\exp\left[-\left(\theta_j-b_i\right)\right]} \qquad (1)$$

where $P_i(\theta_j)$ = probability of examinee with ability θ_j answering item *i* correctly

exp = exponential is constant (2.718)

θ_j = ability estimates

b_i = item *i* difficulty parameter

2-Parameter Logistic Model: the parameter uncovers possible flaws in the 1-PL model if all test items have identical shapes in the ICC. In response, the 2-PL model introduces a parameter called discrimination, expressed by *a*, which permits the ICC for diverse items to have distinct slopes (Perez and Loken, 2023). The discrimination parameter will enable us to model items with more significant (or weaker) relationships to the assessed construct (*θ*) than others; a high discrimination index indicates stronger ties between the item and the construct, while a low discrimination index indicates a weaker relationship. The a-parameter is significant in IRT since it directly influences an item's information. This model assumes that the examinee's competence and difficulty level of the question ascertain the chance of responding correctly to an item. The level of simplicity and practicality of the 2PL model has given it a wide application in educational testing. It provided valuable insights into students' abilities based on their responses to binary items. Mathematically, it is expressed as below:

$$P_i\left(\theta_j\right)=\frac{1}{1+\exp\left[-a_i\left(\theta_j-b_i\right)\right]} \qquad (2)$$

where $P_i(\theta_j)$ = probability of examinee with ability θ_j answering item *i* correctly

θ_j = ability estimates

a_i = item *i* discrimination parameter

b_i = item *i* difficulty parameter

3-Parameter Logistic Model: the 2-PL model aims to tackle a specific critique of the Rasch model, which assumes that all test items exhibit uniform discriminating capability. Still, it has been found deficient in addressing another possibly essential factor that may vary among items: the lower asymptote of the ICC refers to the anticipated proportion of correct or key responses exhibited by participants with exceedingly low θ scores. Including the *c*-parameter in the 3-PL model accounts for the likelihood of correctly predicting the item, causing the lower asymptote of the ICC to be potentially non-zero. This contrasts the 1-PL and 2-PL models, where the ICC's lower asymptote is permanently set to zero (Paek et al., 2023). This parameter accounts for the likelihood of guessing correct responses to items despite examinees' lack of the necessary knowledge or skill to be successful at items. In mathematics assessment, the 3-PL model provides a refined method for estimating abilities, wildly when guessing influences IRT, such as in multiple-choice tests that allow partial credit. In assessments comprising multiple-choice questions, examinees lacking the requisite knowledge of the accurate solution are likely to resort to guessing, leading to the need for a non-zero lower asymptote (and on occasion, they may select the correct option). The formula is presented below:

$$P_i\left(\theta_j\right) = c_i + \frac{1-c_i}{1+\exp^{-1.7a_i\left(\theta_j - b_i\right)}} \qquad (3)$$

where $P_i(\theta_j)$ = probability of participant with ability θ_j answering item *i* correctly

θ_j = ability estimates

a_i = item *i* discrimination parameter

b_i = item *i* difficulty parameter

c_i = item *i* guessing parameter

Statement of the Problem

The objective of testing is to attain an accurate measure of what we need to assess, and the accuracy of test measurements is decided mainly by the test data at each examinee's ability level. The fundamental concept of IRT originates from the principles of the item response model, which entails a mathematical function elucidating the likelihood of particular responses to an item based on various quantitative attributes of the respondents (Frick et al., 2024; Jimoh et al., 2022; von Davier, 2019). Three IRT models, namely the 1-PL, 2-PL, and 3-PL models, have been developed and implemented in various studies for item calibration. This process involves determining the properties of items and estimating the examinees' abilities. As a result, the difficulty in calibrating objects stems from deciding which models to use. However, the Rasch model claims to be a reliable measuring criterion. It contends that factors other than the difficulty of the items are likely to influence examinees' responses (Stemler and Naples, 2021).

Nevertheless, other researchers have dismissed it as experimentally meaningless because it does not account for changes in discrimination and guessing (i.e., chance) factors. The Rasch model implies that guessing is irrelevant and that all objects have the same discrimination value. This study did not examine the Rasch model. Conversely, this study aims to ascertain the effect of the 2-PL and 3-PL on examinees' ability estimates in binary mathematics items. Specifically, the study also aims to establish difficulty and discrimination indices in the 2-PL and 3-PL models of the examinees' responses, determine the effect of the 2-PL model on ability estimates in mathematics binary items, and assess the effect of the 3-PL model on ability estimates in mathematics binary items. By analyzing the binary data collected from administering the instrument, we hope to understand how each model affects the accuracy of the ability estimates. This study will provide valuable insight for educators and test developers looking to elevate the consistency and authenticity of their assessment tools in mathematics education.

Research Questions

1. What are the items' difficulty and discrimination indices in the 2-PL?
2. What are the items' difficulty, discrimination, and guessing indices in the 3-PL?

Research Hypotheses

1. The effect of the 2-PL model on the ability estimates in mathematics binary items is deemed significant.
2. The effect of the 3-PL model on the ability estimates in mathematics binary items is deemed significant.

MATERIALS AND METHODS

A descriptive survey design was utilized in this study. This type of research design focuses on describing data without manipulating variables. There were 1015 participants in the study, all from senior secondary school three (SS3). The study sample consisted of 522 (51.4%) male students and 493 (48.6%) female students. Approximately 42% of the participants were enrolled in private educational institutions, whereas 58% were in public institutions, including federal and state-owned schools. We adapted the Mathematics Achievement Test (MAT) instrument from the WAEC General Mathematics Paper 1 of the June/July (2006-2014) SSCE. WAEC has been responsible for conducting standardized examinations across the West African region (Kennedy and Ebuwa, 2022). This exam is crucial in determining students' academic performance and progress in their respective countries. The WAEC ensures that the exams are fair and transparent, allowing all students to showcase their knowledge and ability. The instrument (MAT) consists of a 20-item mathematics multiple-choice test, and we scored the response binarily. Using a stratified random sampling method, the participants were chosen to guarantee inclusivity from private and public educational institutions. The MAT was administered to the students under standard examination conditions to measure their mathematical achievement. The data were analyzed using IRTPRO. The IRTPRO is a statistical analysis software tool used for IRT analysis for binary and polytomous datasets. It can also perform unidimensional and multidimensional IRT analysis and support models, including 1, 2, and 3-PL models.

Analysis

The data went through a preliminary analysis. Descriptive analysis was employed to ascertain the occurrence rate of each item's response, alongside calculating the mean, maximum, minimum, and standard deviation for each item. The test unidimensionality, which corresponds with the IRT assumptions, was established. According to Choi et al. (2023), Kim (2017), and Opesemowo et al. (2023), the assumption of unidimensionality implies that the item examines a single ability and that the response satisfies the local independence principle, which states that item responses depend on a particular ability level independently. Nevertheless, an exploratory factor analysis (EFA) can verify unidimensionality if one of the two conditions is met. First, the unrotated factor matrix should show that the first component explains at least 20% of the variance based on the inter-item correlation matrix. Second, the eigenvalues of the first component must be greater than those of the second factor. In this study, we assessed unidimensionality using exploratory factor analysis. Furthermore, a scree plot was created to see if unidimensionality could be inferred. A scree plot is a valuable diagram for visualizing a principal component analysis (PCA) leading factor. In a scree plot, a dominating factor stands out over the ICC's elbow break.

The item difficulty parameter estimations were analyzed when responding to the research questions. Items with high *b*-values are often tricky (difficult) items under the IRT model; these are the questions that low-ability examinees are unlikely to answer accurately. Items with low *b*-values, on the other hand, are classed as easy (simple) items; these are questions that most examinees, including those with little aptitude, will have at least a moderate chance of answering correctly. Consequently, when interpreting the difficulty values, the following criteria are used: Difficulty values (*b*) that ranged between -3.00 ≤ -2.00 is classified as very easy; *b*-values that ranged between -2.00 ≤ -1.00 is classified as easy; -1.00 ≤ 1.00 is classified as moderately difficult; 1.00 ≤ 2.00 is classified as difficult while ≥ 2.00 is categorized as very difficult (Bichi and Talib, 2018). In addition, the discriminating value reveals how well an item distinguishes between examinees of varied abilities. Discrimination indices for good items typically range from 0.5 to 2.0. In the 3-PL model, item discrimination is proportional to the slope of the item response function at the inflection point (0.25). The c-parameter has a theoretical range of $0 \leq C \leq 1.0$, although values higher than 0.35 are unacceptable (Adedoyin and Adedoyin, 2013; Baker, 2001).

To conduct the first hypothesis test, the effects of the 2-PL model (item difficulty and discrimination) on ability estimates were analyzed using an ANOVA. An ANOVA of the 3-PL model's (item difficulty, discrimination, and guessing) effects on ability estimates was also used to test the second hypothesis.

Ethical Consideration

Before the data collection, the ethical consideration was approved, and the participants were informed about the need to complete the MAT instrument responsibly and honestly, and their participation was completely voluntary. We established a confidentiality agreement to ensure that we would keep the collected data anonymous and use it solely for the research project. All ethical rules and procedures were strictly followed throughout the data collection process to preserve the participants' rights. We also notified participants that they could opt out of the study without repercussions. This made participants feel comfortable and confident that their privacy was respected throughout the research process. The ethical considerations the researchers took were vital in upholding the integrity of the study and respecting the individuals who had chosen to participate. Ultimately, these measures helped establish trust between the researchers and participants, creating a safe and respectful data collection environment.

RESULTS

Table 1 exhibits MAT items' mean, standard deviation, minimum and maximum scores, and response frequency.

Table 1 displays the frequencies of each item answer option and MAT's mean, maximum, minimum, and standard deviation.

Table 2 presents the eigenvalues and total variance explained with evidence that the test is unidimensional.

Table 2 showcases the EFA conducted on the 20-MAT. It produced six eigenvalues that are significantly more than one. The initial eigenvalue, 4.282, was more significant than the successive five eigenvalues (1.279, 1.131, 1.085, 1.010, and 1.004). The initial factor accounted for 21.41% of the total variation in the sample. The subsequent component accounted for 6.397% of the residual variance. However, the other 18 factors accounted for the rest of the variance. A scree plot, demonstrated in Figure 1, further validated the data's unidimensionality.

Research Question 1: What are the items' difficulty and discrimination indices in the 2-PL?

The items were subjected to a 2-PL model in the IRTPRO. Table 4 displays the item parameters, including the difficulty and discrimination indices.

Table 3 indicates that none of the items in the 2-PL were rated tricky. In contrast, only two items (1 and 15) exhibited poor discrimination since their discrimination indices fell below the 0.5 threshold.

Figure 2 illustrates the Total Information Curve (TIC), which compares the test data against the theta (ability) levels with their standard measurement error. The TIC enables researchers to visually analyze the relationship between test data and ability levels, providing valuable insight into the accuracy and precision of the test measurements. By examining the curve, researchers can assess how well the test differentiates between individuals with different ability levels and identify where measurement error is most likely to occur. This information can be used to make informed decisions about test design and interpretation, improving assessments' overall quality and reliability.

The curve shows a normal ability distribution, indicating the highly discriminatory test. The test characteristics curve (TCC) shown in Figure 3 further verified this conclusion. To display the psychometric structure, we also presented each item using an item category curve (ICC) (see Appendix A). Appendix A contains graphical representations of the components, known as ICC. They can show items that discriminate effectively and items that do not distinguish individuals at different levels of the items.

	Statistical Properties				Frequency of Response Options	
Item	M	SD	Min	Max	0	1
1	0.15	0.36	0	1	863	152
2	0.61	0.49	0	1	393	622
3	0.31	0.46	0	1	697	318
4	0.31	0.46	0	1	697	318
5	0.57	0.50	0	1	467	548
6	0.82	0.38	0	1	181	834
7	0.74	0.44	0	1	264	751
8	0.33	0.47	0	1	676	339
9	0.55	0.50	0	1	452	564
10	0.47	0.50	0	1	535	480
11	0.62	0.49	0	1	390	625
12	0.54	0.50	0	1	465	550
13	0.39	0.49	0	1	615	400
14	0.72	0.45	0	1	285	730
15	0.50	0.50	0	1	506	509
16	0.64	0.48	0	1	362	653
17	0.13	0.34	0	1	882	133
18	0.30	0.46	0	1	712	303
19	0.33	0.47	0	1	681	334
20	0.53	0.5	0	1	479	536

Note: Response option frequencies for each item total 1015 responses.

Table 1: Descriptive Statistics of MAT

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.282	21.411	21.411
2	1.279	6.397	27.809
3	1.131	5.656	33.465
4	1.085	5.427	38.892
5	1.010	5.051	43.943
6	1.004	5.018	48.961

Extraction method: PCA

Table 2: Eigenvalues and Total Variance Explained

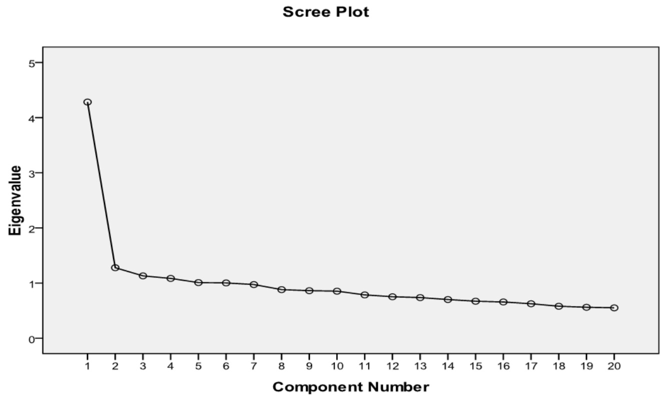


Figure 1: Scree Plot of MAT

Item	*b* (difficulty)	*a* (discrimination)
1	-3.53	-0.52
2	-0.28	1.21
3	-0.47	1.18
4	-0.47	1.03
5	-0.21	0.91
6	-1.71	1.09
7	-1.10	0.79
8	-0.17	1.12
9	-0.21	1.79
10	0.18	0.63
11	-0.36	0.89
12	-0.21	1.02
13	-0.18	1.31
14	-0.77	2.14
15	-0.02	0.46
16	-0.53	1.78
17	-1.62	0.94
18	-0.44	1.97
19	-0.05	0.82
20	-0.12	1.35

Table 3: Item Parameters of 2-PL Model

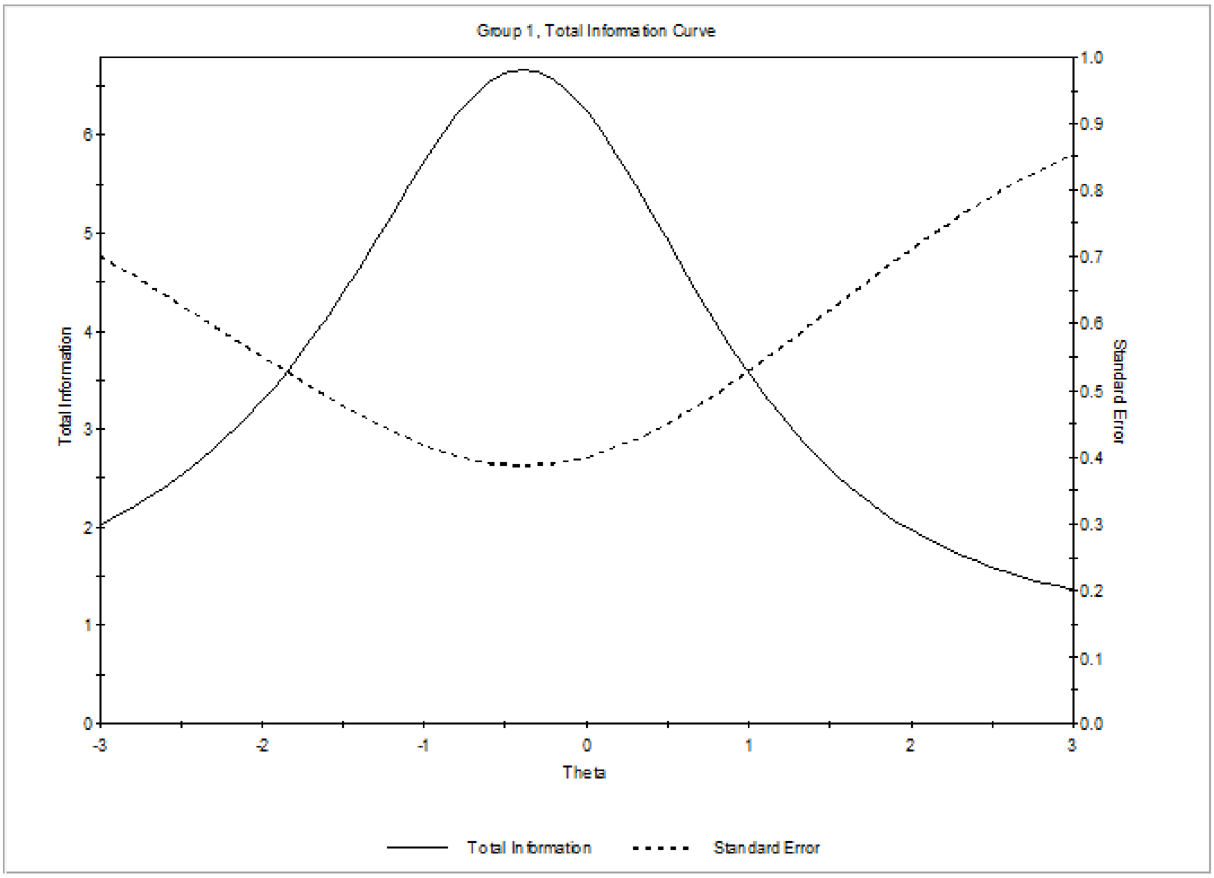


Figure 2: Test Information Curve of MAT

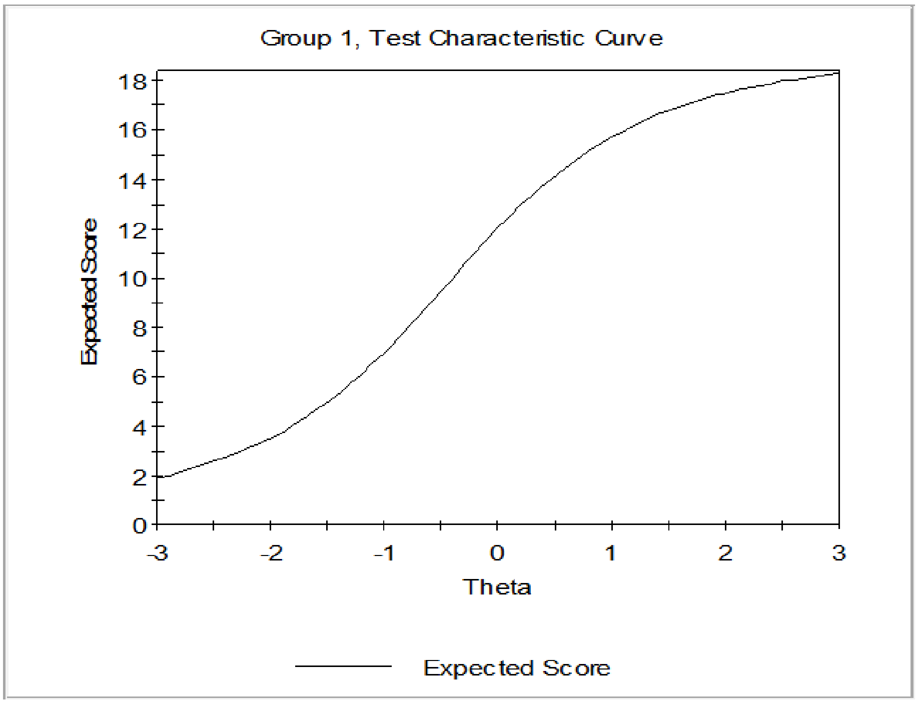


Figure 3: Test Characteristic Curve of MAT

Research Questions 2: What are the items' difficulty, discrimination, and guessing indices in the 3-PL?

The items underwent a 3-PL model within the IRTPRO software. Table 4 displays the item parameters, encompassing difficulty, discrimination, and guessing indices.

Table 4 exhibits the item parameters of the 3-PL model. Only item 1 was adjudged very difficult among the item difficulty parameter estimates; 11 items are moderately difficult, while eight items are easy. Based on the discrimination parameter estimations presented in Table 5, it was observed that only one item from a total of twenty failed to differentiate among the examinees. It is further obverse that the guessing parameter (*c*) highlights the exclusion of six items per the predefined criteria.

Item	*b* (difficulty)	*a* (discrimination)	*c* (guessing)
1	467.72	0.37	-171.99
2	0.12	1.61	-0.19
3	-0.02	1.55	0.03
4	-0.01	1.32	0.01
5	0.17	1.09	-0.19
6	-1.44	1.11	1.60
7	-0.68	0.85	0.58
8	0.36	1.91	-0.69
9	0.10	2.66	-0.27
10	0.93	1.15	-1.07
11	0.01	1.00	-0.01
12	0.20	1.30	-0.26
13	0.13	1.73	-0.23
14	-0.48	2.68	1.30
15	0.79	0.60	-0.47
16	-0.28	2.12	0.59
17	-1.23	0.98	1.21
18	-0.28	2.16	0.60
19	0.55	1.26	-0.69
20	0.14	1.67	-0.24

Table 4: Item Parameters of 3-PL model

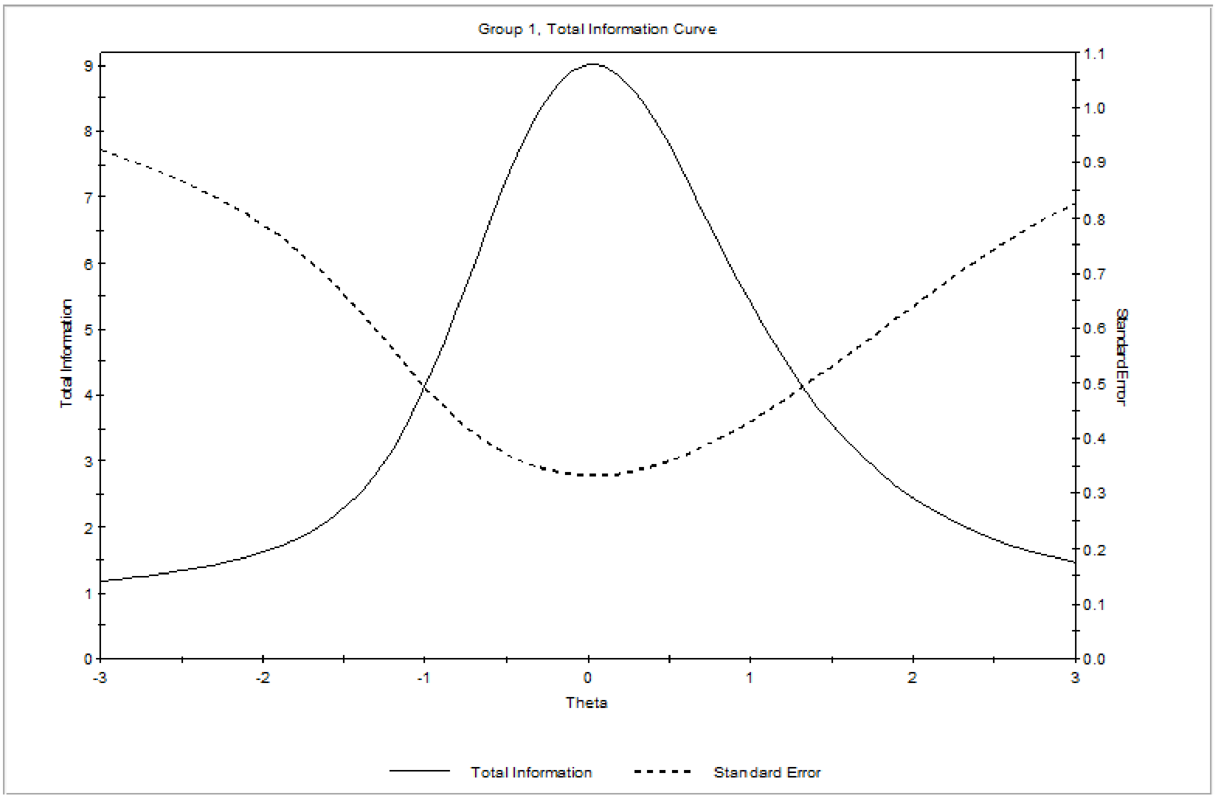


Figure 4: Total Information Curve of MAT

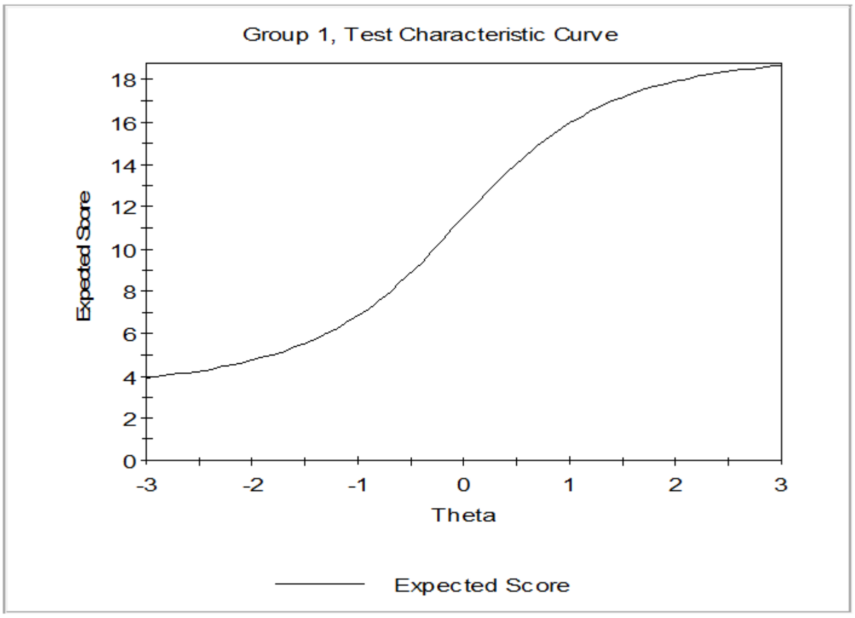


Figure 5: Test Characteristics Curve of MAT

The diagram (Figure 4) illustrates a typical distribution of abilities, demonstrating the high level of discrimination in the test. This allows researchers to identify potential areas for improvement in the test to ensure that it accurately reflects individuals' true abilities. Test developers can also make adjustments to minimize measurement error and increase the test's reliability by analyzing the curve. Figure 5 provided additional evidence supporting the Test Characteristic Curve (TCC). Each item's psychometric structure was characterized using an item category curve (see Appendix B). The item category curves presented in Appendix B serve as visual representations of the items, allowing for the identification of items that effectively discriminate and those that do not differentiate between individuals with different ability levels.

Hypothesis one: the effect of the 2-PL model on the ability estimates in mathematics binary items is deemed significant.

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between groups	31.44	2	15.72	19.52	0.000
Within groups	847.32	1052	0.80		
Total	878.76	1054			

Table 5: ANOVA of the Effect of 2-PL parameters on Ability Estimates

The ANOVA results for the effect of the 2-PL model on ability estimates indicate a significant difference among the groups, as evidenced by a large *F*-value of 19.52 ($p < 0.05$). The between-groups variance (15.72) is substantially higher than the within-groups variance (0.80), suggesting that the variation in mathematics ability estimates of examinees because of item parameters (difficulty and discrimination) is explained mainly by differences between the groups rather than within them. This implies that the 2-PL model notably impacts ability estimates, underscoring the importance of considering these parameters in psychometric modelling.

Hypothesis Two: the effect of the 3-PL model on ability estimates in mathematics binary items is deemed significant.

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	12290.52	3	4096.84	18.52	0.000
Within groups	236861.19	1071	221.16		
Total	249151.71	1074			

Table 6: ANOVA of the Effect of the 3-PL Model on Ability Estimates

The result unveils that the effect of the 3-PL (difficult, discrimination, and guessing) model on ability estimates in mathematics binary items is deemed significant with $F = 18.52$; $p < 0.05$ while the between-groups variance (4096.84) is higher than the within-groups variance (221.16). Therefore, the alternative hypothesis was supported.

Comparing the two ANOVA results in the 2-PL and 3-PL models, it is explicitly uncovered that the effect of the model on ability estimates is higher in the 3-PL model than in the 2-PL model. However, both have a significant effect on ability estimates. The 3-PL model can capture more nuanced variations in ability levels compared to the 2-PL model. This suggests that incorporating the additional parameter (i.e., guessing parameter) in the 3-PL model allows for a more accurate and precise estimation of individuals' abilities. These findings highlight the importance of selecting the appropriate item response theory model to ensure a valid and reliable measurement of abilities in educational and psychological assessments.

DISCUSSION

The 2-PL and 3-PL models utilized in the estimation of binary data raised apprehension due to the inclusion of parameters representing the item's difficulty, discrimination, and lower asymptote. These parameters were crucial in precisely evaluating the correlation between an individual's proficiency level and their reactions to particular items. The 2-PL model was beneficial for measuring discrimination between individuals with different levels of ability, while the 3-PL model also accounted for the guessing behavior of participants. Overall, these models provided a comprehensive framework for understanding and interpreting binary data in assessments and measurements of abilities. The assumption in the mathematics items holds reasonably with the factor analysis results.

The parameters derived from the 2-PL model exhibit low difficulty levels but with high discriminating indices (i.e., only two items exhibit low discrimination values). These findings suggest that the 2-PL model may not be the most suitable for accurately measuring examinees' abilities, as it tends to underestimate the discrimination of items. Additional research is needed to explore alternative models that may provide more precise estimations of item parameters and better reflect the true abilities of individuals. Additionally, considering the potential effect of item discrimination on test validity and reliability, it is crucial for researchers and practitioners to carefully evaluate the appropriateness of the chosen IRT model for their specific assessment needs. This was in tandem with the study of Perez and Loken (2023), who affirmed that item parameters in the 2-PL IRT model demonstrated well-estimated difficulties but noticeably underestimated discriminations, indicating low discrimination values rather than high. The result further aligned with the findings of Setiawati et al. (2023), presenting that the 2-PL model in the study showed low item difficulties but high discrimination indices, with low discrimination values indicating unique characteristics of the items.

The result of the 3-PL model introduced a *c*-parameter that affected the relationship between item difficulty and discrimination indices. Under the 3-PL model, item difficulties tend to be mostly average, leading to a higher discriminating index than in the 2-PL model (Perez and Loken, 2023). This validates the findings of Sweeney et al. (2022), who found that item difficulty and discrimination are mostly positively connected in the 3-PL model, as opposed to the negative correlation reported in the 2-PL. As a result, the 3-PL model's incorporation of the c-parameter alters these relationships, which may contribute to the model's higher discriminating indices (Ferreira-Junior et al., 2023). As a result, integrating the c-parameter in the 3-PL model significantly affects the item's difficulty and discrimination characteristics, distinguishing it from the 2-PL model. This suggests that most examinees, including those of low and medium aptitude, will have a reasonable probability of answering correctly.

Notably, item 1 in the 3-PL had the highest difficulty and the lowest discrimination indices. We traced this to uncertainty in the item's development, which resulted in examinees misinterpreting it.

The findings revealed that both models accounted for the effect of item parameters (item difficulty, discrimination, and guessing) on examinees' ability assessments in mathematics binary items. The 3-PL model introduces a c-parameter, affecting the relationship between item difficulty and discrimination indices. Under the 3-PL model, item difficulties tend to be mostly average, leading to a higher discriminating index than in the 2-PL model (Perez and Loken, 2023). The result supports the findings of Sweeney et al. (2022), revealing that item difficulty and discrimination are mostly positively correlated in the 3-PL model, in contrast to the negative correlation observed in the 2-PL model. Additionally, the 3-PL model's incorporation of the c-parameter influences these relationships, potentially contributing to the higher discriminating indices observed in this model (Ferreira-Junior et al., 2023). As a result, integrating the c-parameter in the 3-PL model significantly impacts the item's difficulty and discrimination indices, separating it from the 2-PL model. This implies that examinee estimates were dependent on item parameters. Although the models' item parameters differed in the index, both impacted ability estimates. The findings of this study correspond with the findings of Setiawati et al. (2023), which suggest that few and possibly non-significant differences exist in the assessment of item parameters in the 1-PL, 2-PL, and 3-PL.

In estimating a person's abilities, some empirical studies have explored the efficacy of IRT compared to CTT. A more accurate estimation of abilities is possible when using IRT due to its sensitivity to item characteristics (Suparman and Juandi, 2022). Some studies have shown slight differences between CTT and IRT estimates of abilities (Mutiawani et al., 2022). These studies emphasize the importance of accurate ability estimation in educational assessments, especially in scenarios like Computer Adaptive Testing (CAT), where precise estimations are crucial (Oladele et al., 2022; Opesemowo and Ndlovu, 2023). Additionally, IRT models, such as many-facet Rasch models (MFRMs), have been purported to enhance accuracy in measuring higher-order abilities, considering factors like rater severity and task difficulty (Sideridis and Alahmadi, 2022). Overall, the research underscores the significance of employing advanced statistical models like IRT for a more precise and reliable estimation of a person's abilities in various assessment contexts.

Despite the valuable results of this study, we should identify some limitations. First, the participants in the research were restricted to senior secondary school students in Nigeria. Hence, it is worth noting that other studies should consider junior and senior secondary school students from different countries to have a broader perspective of the student's ability estimate. Second, the subject of focus was mathematics assessment in Nigeria, which may have restricted the generalizability of the findings. However, it is essential to conduct further research across different disciplines and other countries to understand the ability to estimate better. Third, the data used in this study was binary, which may have impeded the findings.

Further studies can incorporate polytomous data to conduct a broader analysis. Lastly, the study concentrated on quantitative data, neglecting qualitative data that might provide impactful information. Integrating quantitative data into future research could provide more insight into the effects of 2-PL and 3-PL models on ability estimates in mathematics binary items.

CONCLUSION

Based on the study results, we concluded that the item parameters of the 2-PL and 3-PL models affected the ability estimates of examinees in Nigerian secondary schools for mathematics binary items. Furthermore, the study found that the 3-PL model bestowed more precise estimates of examinees' abilities than the 2-PL model. This suggests that using the 3-PL model for mathematics assessments in Nigerian secondary schools may create more accurate and reliable results. In addition, the study recommended further investigation into the effect of item parameters on ability estimation in other subject areas and grade levels to improve assessment practices in the country. In conclusion, the study highlighted the importance of utilizing advanced measurement models, such as the 3-PL model, to enhance the accuracy of ability estimates in mathematics assessments. By implementing this model in secondary school, educators and policymakers can make more informed decisions about students' academic performance and tailor instructional strategies to meet their needs better. Moving forward, continued research into the effects of item parameters on ability estimation across different subjects and grade levels will be crucial for advancing assessment practices and promoting academic success in Nigerian schools.

REFERENCES

Adedoyin, O. O. and Adedoyin, J. (2013) 'Assessing the Comparability between Classical Test Theory (CTT) and Item Response Theory (IRT) Models in Estimating Test Item Parameters', *Herald Journal of Education and General Studies*, Vol. 2, No. 3, pp. 107–114. http://dx.doi.org/10.12691/education-6-3-11

Adetutu, P. O. and Iwintolu, R. O. (2017) 'An Item Response Theory Analysis of the Academic Amotivation Inventory for Secondary School Students in Southwestern Nigeria', *Journal of Research and Method in Education (IOSR-JRME)*, Vol. 7, No. 4, pp. 22–31. https://dx.doi.org/10.9790/7388-0704032231

Alordiah, C. (2015) '*Comparison of Index of Differential Item Functioning under the Methods of Item Response Theory and Classical Test Theory in Mathematics*', An unpublished Ph. D thesis of Delta State University, Abraka, Delta State, Nigeria.

Awopeju, O. and Afolabi, E. (2016) 'Comparative Analysis of Classical Test Theory and Item Response Theory based Item Parameter Estimates of Senior School Certificate Mathematics Examination', *European Scientific Journal*, Vol. 12, No. 28, pp. 263–284. http://dx.doi.org/10.19044/esj.2016.v12n28p263

Ayanwale, M. A. (2023) 'Test Score Equating of Multiple-Choice Mathematics Items: Techniques from Characteristic Curve of Modern Psychometric Theory', *Discover Education*, Vol. 2, No. 1, p. 30. https://doi.org/10.1007/s44217-023-00052-z

Baker, F. B. (2001) *The basics of item response theory*, 2nd Edition, Available: https://files.eric.ed.gov/fulltext/ED458219.pdf [10 January 2024].

Bichi, A. A. and Talib, R. (2018) 'Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development', *International Journal of Evaluation and Research in Education*, Vol. 7, No. 2, pp. 142–151. https://doi.org/10.11591/ijere.v7i2.12900

Breuer, S., Scherndl, T. and Ortner, T. M. (2023) "Effects of Response Format on Achievement and Aptitude Assessment Results: Multi-Level Random Effects Meta-Analyses', *Royal Society Open Science*, Vol. 10, No. 5, p. 220456. https://doi.org/doi:10.1098/rsos.220456

Choi, S., Jang, Y. and Kim, H. (2023) 'Influence of Pedagogical Beliefs and Perceived Trust on Teachers' Acceptance of Educational Artificial Intelligence Tools', *International Journal of Human–Computer Interaction*, Vol. 39, No. 4, pp. 910–922. https://doi.org/10.1080/10447318.2022.20491455

Danh, T., Desiderio, T., Herrmann, V., Lyons, H. M., Patrick, F., Wantuch, G. A. and Dell, K. A. (2020) 'Evaluating the Equality of Multiple-Choice Questions in a NAPLEX Preparation Book', *Currents in Pharmacy Teaching and Learning*, Vol. 12, No. 10, pp. 1188–1193. https://doi.org/10.1016/j.cptl.2020.05.006

Douglas, K. A., Neumann, K. and Oliveri, M. E. (2023) 'Contemporary Approaches to Assessment of Engineering Competencies for Diverse Learners', in Johri, A. (ed.) *International handbook of Engineering Education Research*, New York: Routledge. https://doi.org/10.4324/9781003287483-37

Ferreira-Junior, M., Reinaldo, J. T., Neto, E. A. L. and Prudencio, R. B. (2023) 'β4-IRT: A new β3-IRT with Enhanced Discrimination Estimation', ArXiv Preprint arXiv:2303.17731. https://doi.org/10.48550/arXiv.2303.17731

Frick, S., Krivosija, A. and Munteanu, A., (2024) 'Scalable Learning of Item Response Theory Models', *International Conference on Artificial Intelligence and Statistics (AISTATS),* London, pp. 1234–1242.

Gates, J. A. (2023) '*School Board Member Perceptions of the Race Based Academic Achievement Disparity* [Ed.D., Concordia University Chicago]', ProQuest Dissertations & Theses Global. United States -- Illinois. https://www.proquest.com/dissertations-theses/school-board-member-perceptions-race-based/docview/2910119963/se-2?accountid=13425

Jimoh, K., Opesemowo, O. A. and Faremi, Y. A. (2022) 'Psychometric Analysis of Senior Secondary School Certificate Examination (SSCE) 2017 Neco English Language Multiple Choice Test Items in Kwara State Using Item Response Theory', *Journal of Applied Research and Multidisciplinary Studies*, Vol. 3, No. 2, pp. 1–19. https://doi.org/10.32350/jarms.32.01

Kalhori, R. P. and Abbasi, M. (2017) 'Are Faculty Members of Paramedics Able to Designed Accurate Multiple Choice Questions', *Global Journal of Health Science*, Vol. 9, No. 1, pp. 211–216. http://dx.doi.org/10.5539/gjhs.v9n1p211

Kennedy, I. and Ebuwa, S. O. (2022) 'Assessing Score Dependability of West Africa Examinations Council (WAEC) 2019 Mathematics Objective Test using Generalisability Theory', *British Journal of Contemporary Education*, Vol. 2, No. 1, pp. 64–73. https://doi.org/10.52589/BJCE-OCA9OZJT

Kim, K. Y. (2017) '*IRT Linking Methods for the Bifactor Model: A Special Case of the Two-Tier Item Factor Analysis Model* [Ph.D., The University of Iowa]', ProQuest Dissertations & Theses Global. United States -- Iowa. https://www.proquest.com/dissertations-theses/irt-linking-methods-bifactor-model-special-case/docview/1964934014/se-2?accountid=13425

Mutiawani, V., Athaya, A. M., Saputra, K. and Subianto, M. (2022) 'Implementing Item Response Theory (IRT) Method in Quiz Assessment System', *TEM Journal*, Vol. 11, No. 1, pp. 210–218. https://doi.org/10.18421/TEM111-26

O'Connor, P. J., Hill, A. and Martin, B. (2019) 'The Measurement of Emotional Intelligence: A Critical Review of the Literature and Recommendations for Researchers and Practitioners', *Frontiers in Psychology*, Vol. 10, No. 1116, pp. 1–19. https://doi.org/10.3389/fpsyg.2019.01116

Oladele, J. I., Ndlovu, M. and Spangenberg, E. D. (2022) 'Simulated Computer Adaptive Testing Method Choices for Ability Estimation with Empirical Evidence', *International Journal of Evaluation and Research in Education (IJERE)* Vol. 3, pp. 1392–1399. https://doi.org/10.11591/ijere.v11i3.21986

Olagunju, B. A. and Iwintolu, R. O. (2023) 'Development of Pedagogical Competence Scale for Lecturers in Universities Using Item Response Theory', *Mimbar Sekolah Dasar*, Vol. 10, No.1, pp. 135–148. https://doi.org/10.53400/mimbar-sd.v10i1.51422

Opesemowo, O., Afolabi, E. and Oluwatimilehin, T. (2018) 'Development of a Scale for Measuring Students' Testwiseness in Senior Secondary School Examination in Nigerian', *International Journal of Research*, Vol. 5, No. 19, pp. 464–474.

Opesemowo, O. A. G., Ayanwale, M. A., Opesemowo, T. R. and Afolabi, E. R. I. (2023) 'Differential Bundle Functioning of National Examinations Council Mathematics Test Items: An Exploratory Structural Equation Modelling Approach', *Journal of Measurement and Evaluation in Education and Psychology*, Vol. 14, No. 1, pp. 1–18. https://doi.org/10.21031/epod.1142713

Opesemowo, O. A. G. and Ndlovu, M. (2023) 'Status and Experience of Mathematics Teachers' Perception of Integrating Computer Adaptive Testing into Unified Tertiary Matriculation Examination Mathematics', *Multicultural Education*, Vol. 9, No. 2, pp. 66–78.

Paek, I., Lin, Z., and Chalmers, R. P. (2023) 'Investigating Confidence Intervals of Item Parameters When Some Item Parameters Take Priors in the 2PL and 3PL Models', *Educational and Psychological Measurement*, Vol. 83, No. 2, pp. 375–400. https://doi.org/10.1177/00131644221096431

Perez, A. L. and Loken, E. (2023) 'Person Specific Parameter Heterogeneity in the 2PL IRT Model', *Multivariate Behavioral Research*, pp. 1–7. https://doi.org/10.1080/00273171.2023.2224312

Powers, K. (2019) '*Personality, Attitudes, and Behaviors*', in Workplace Psychology, Chemeketa Community College.

Rafi, I., Retnawati, H., Apino, E., Hadiana, D., Lydiati, I. and Rosyada, M. N. (2023) 'What Might Be Frequently Overlooked Is Actually Still Beneficial: Learning from Post National-Standardized School Examination', *Pedagogical Research*, Vol. 8, No. 1, pp. 1–15. https://doi.org/10.29333/pr/12657

Rios, J. A. and Soland, J. (2022) 'An Investigation of Item, Examinee, and Country Correlates of Rapid Guessing in PISA', *International Journal of Testing*, Vol. 22, No. 2, pp. 154–184. https://doi.org/10.1080/15305058.2022.2036161

Rudner, L. M. (2019) 'Scoring and Classifying Examinees using Measurement Decision Theory', *Practical Assessment, Research, and Evaluation*, Vol. 14, pp. 1–12. https://doi.org/10.7275/vksg-rh07

Scheibling-Sève, C., Pasquinelli, E. and Sander, E. (2020) 'Assessing Conceptual Knowledge through Solving Arithmetic Word Problems', *Educational Studies in Mathematics*, Vol. 10*3,* No. 3, 293–311. https://doi.org/10.1007/s10649-020-09938-3

Setiawati, F. A., Amelia, R. N., Sumintono, B. and Purwanta, E. (2023) 'Study Item Parameters of Classical and Modern Theory of Differential Aptitude Test: is it Comparable?', *European Journal of Educational Research*, Vol. 12, No. 2, pp. 1097–1107. https://doi.org/10.12973/eu-jer.12.2.1097

Sideridis, G. and Alahmadi, M. (2022) 'Estimation of Person Ability under Rapid and Effortful Responding', *Journal of Intelligence*, Vol. 10, No. 3, p. 67. https://www.mdpi.com/2079-3200/10/3/67

Stemler, S. E. and Naples, A. (2021) 'Rasch Measurement v. Item Response Theory: knowing when to cross the line', *Practical Assessment, Research and Evaluation*, Vol. 26, No. 11, pp. 1–16. https://doi.org/10.7275/v2gd-4441

Suparman, S. and Juandi, D. (2022) 'Self-efficacy and Mathematical Ability: A Meta-Analysis of Studies conducted in Indonesia', *Pedagogika/Pedagogy*, Vol. 147, No. 3, pp. 26–57. https://doi.org/10.15823/p.2022.147.2

Sweeney, S. M., Sinharay, S., Johnson, M. S. and Steinhauer, E. W. (2022) 'An Investigation of the Nature and Consequence of the Relationship between IRT Difficulty and Discrimination', *Educational Measurement: Issues and Practice*, Vol. 41, No. 4, pp. 50–67. https://doi.org/10.1111/emip.12522

von Davier, M. (2019) 'TIMSS 2019 Scaling Methodology: Item Response Theory, Population Models, and linking across Modes', *Methods and procedures: TIMSS*, Vol. 11, No. 11, pp. 11–25.

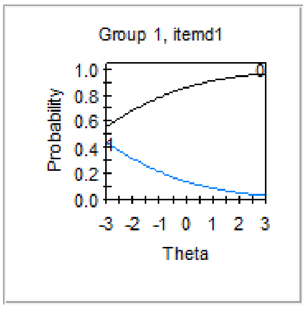
Yustiandi, Y. and Saepuzaman, D. (2021) 'Analysis of Model Fit and Item Parameter of Work and Energy Test using Item Response Theory', *Gravity: Jurnal Ilmiah Penelitian dan Pembelajaran Fisika*, Vol. 7, No, 2, pp. 21–28. https://doi.org/10.30870/gravity.v7i2.10563

Zanon, C., Hutz, C. S., Yoo, H. and Hambleton, R. K. (2016) 'An Application of Item Response Theory to Psychological Test Development', *Psicologia: Reflexão e Crítica*, Vol. 29, No. 18. https://doi.org/10.1186/s41155-016-0040-x

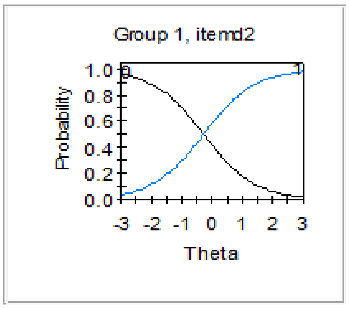
APPENDIX

APPENDIX A

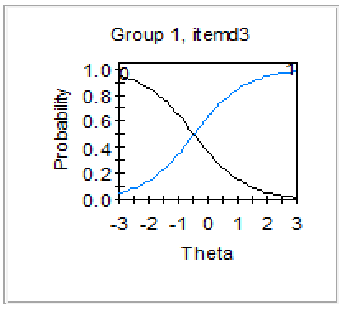
MAT Item Category Curves of 2-PL



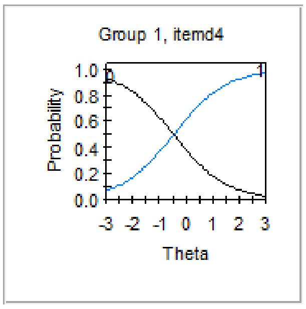
a = -0.52 b = -3.53 *M* = 0.15



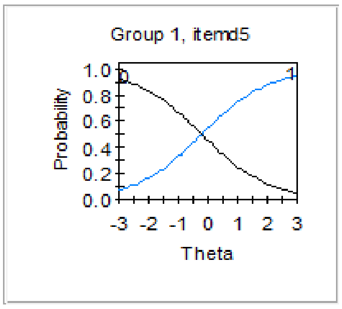
a = 1.21 b = -0.28 *M* = 0.61



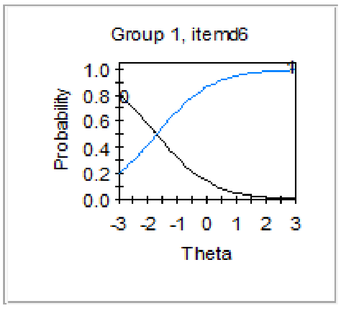
a= 1.18 b = -0.47 *M* = 0.31



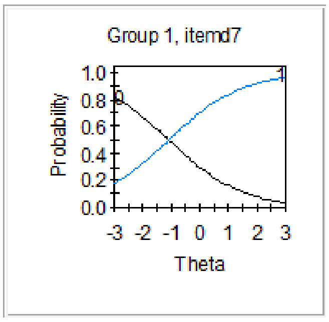
a = 1.03 b = -0.47 *M* = 0.31



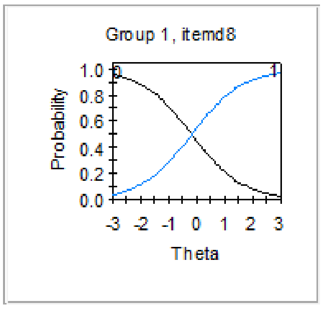
a = 0.91 b = -0.21 *M* = 0.57



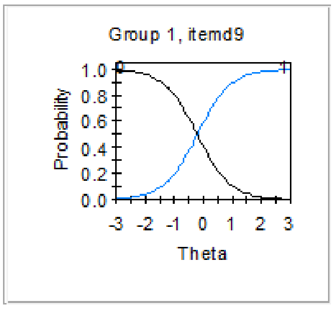
a = 1.09 b = -1.71 *M* = 0.82



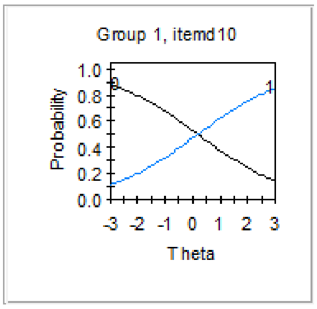
a = 0.79 *b* = -1.10 *M* = 0.74



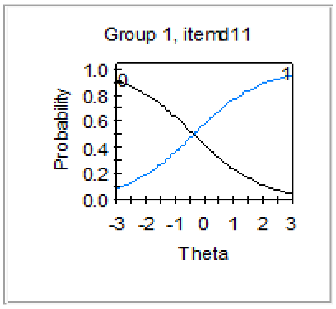
a = 1.12 b = -0.17 *M* = 0.33



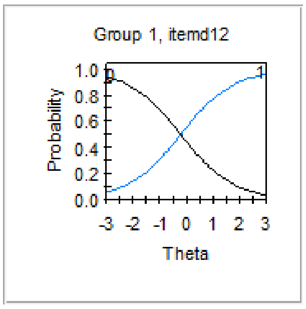
a = 1.79 b =-0.21 M=0.55



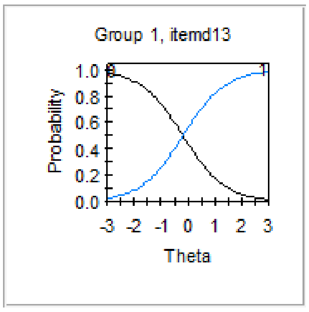
a = 0.63 *b* = 0.18 *M* = 0.47



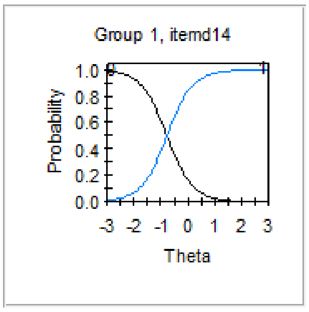
a = 0.89 *b* = -0.36 *M* = 0.62



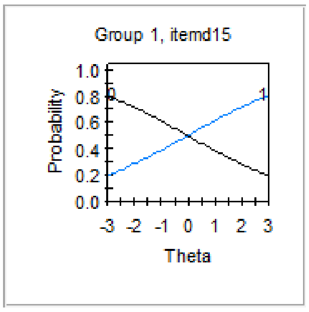
a = 1.02 *b* =-0.21 *M* = 0.54



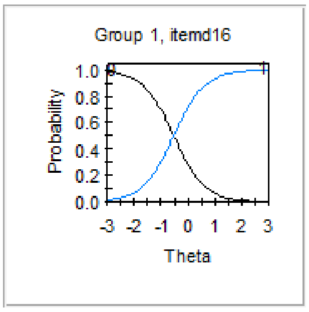
$a = 1.31$ b= -18 $M = 0.39$



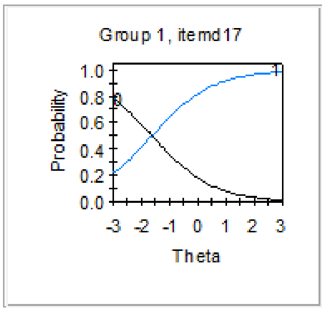
$a = 2.14$ $b = -0.77$ $M = 0.72$



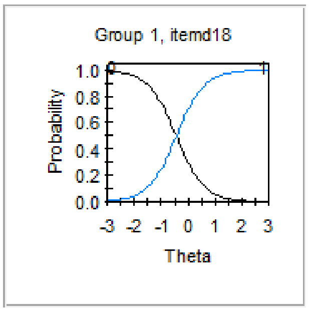
$a =0.46$ $b= -0.02$ $M = 0.50$



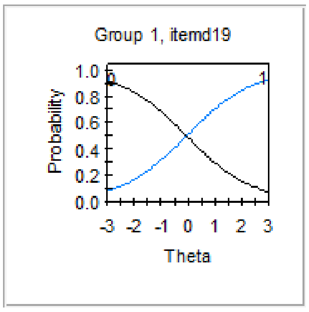
$a = 1.78$ $b = -0.53$ $M = 0.64$



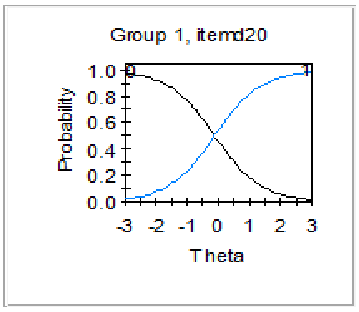
$a = 0.94$ $b = -1.62$ $M = 0.13$



$a = 1.97$ $b = -0.44$ $M =0.30$



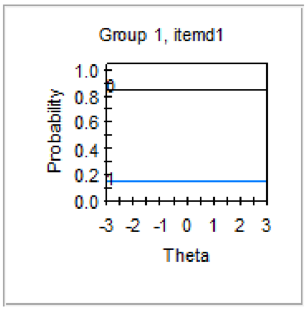
$a = 0.82$ $b = -0.05$ $M = 0.33$



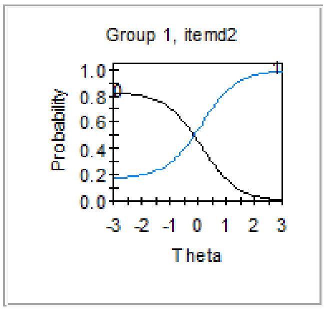
$a = 1.35$ $b = -0.12$ $M = 0.53$

APPENDIX B

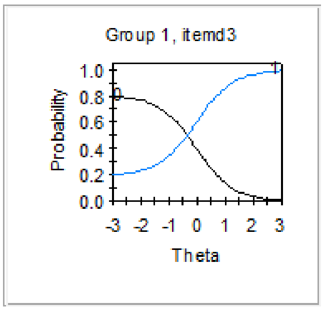
MAT Item Category Curves of 3-PL



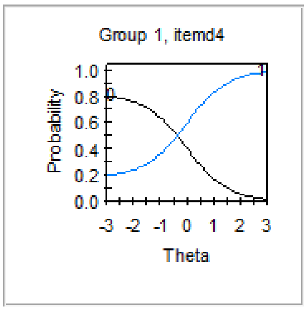
$a = 0.37$ $b = 467.72$ $c = -171.99$



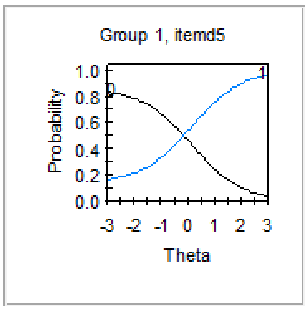
$a = 1.61$ $b = 0.12$ $c = -0.19$



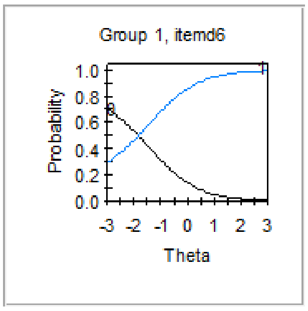
$a = 1.55$ $b = -0.02$ $c = 0.03$



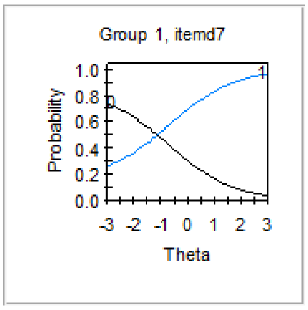
$a = 1.32$ $b = -0.01$ $c = 0.01$



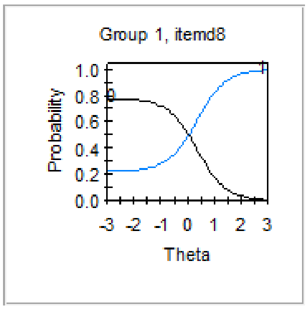
$a = 1.\ 09$ $b = 0.17$ c $= -0.19$



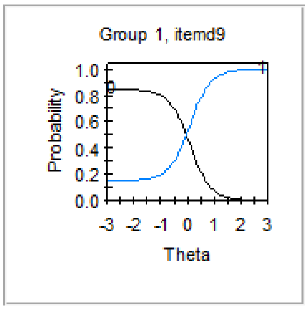
$a = 1.11$ $b = -1.44$ $c = 1.60$



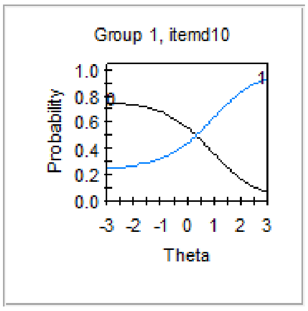
$a = 0.85$ $b = -0.68$ $c = 0.58$



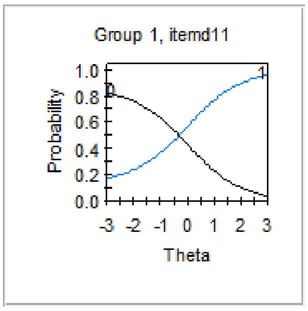
$a = 1.91$ b $= 0.36$ $c = -0.69$



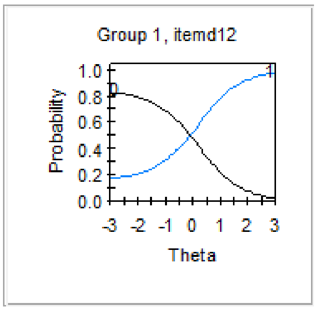
$a = 2.66$ $b =0.10$ c=-0.27



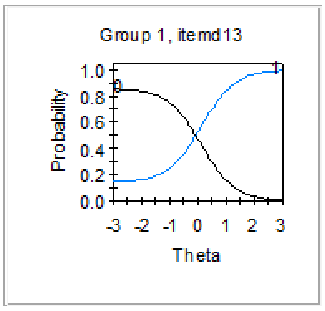
$a =1.15$ $b = 0.93$ $c = -1.02$



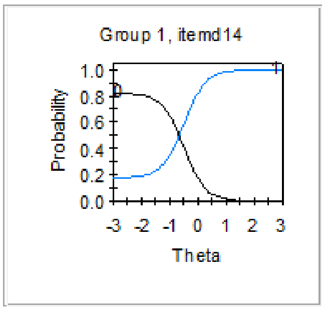
$a = 1.00$ $b = 0.01$ $c =- 0.01$



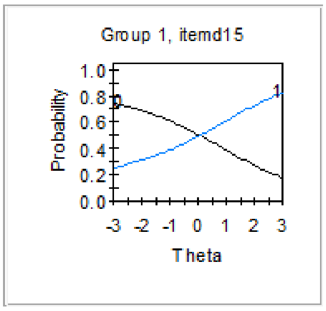
$a = 1.30$ $b =0.20$ $c = -0.26$



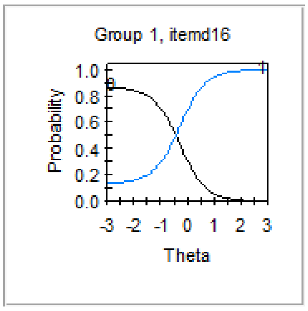
a = 1.73 *b*= 0.13 *c* =-0.23



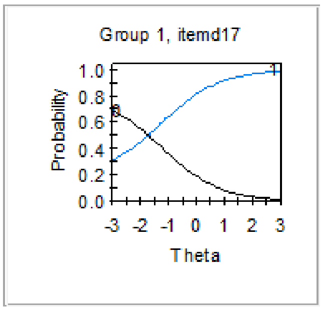
a = 2.68 *b* = -0.48 *c* = 1.30



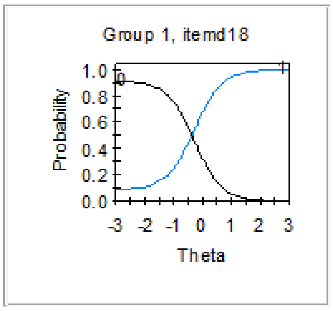
a =0.60 *b*= 0.79 *c* =- 0.47



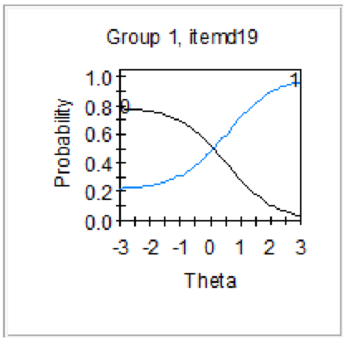
a = 2.12 *b* = -0.28 *c* = 0.59



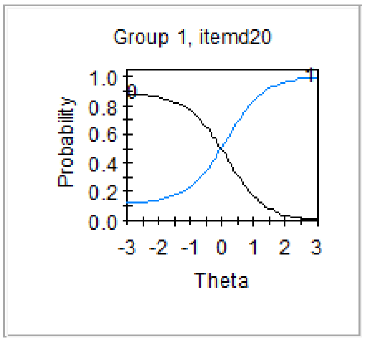
a = 0.98 *b* = -1.23 *c* = 1.21



a = 2.16 *b* = -0.28 *c* =0.60



a =1.26 *b* = 0.55 *c* = -0.69



a = 1.67 *b* = 0.14 *c* = -0.24