Scientific African 31 (2026) e03226

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Item Response Theory for trait assessment in randomized item pool for computer based test

Rhydal Esi Eghan *, Edward Osei-Sarpong, Gaston Edem Awashie, Reindorf Nartey Borkor, Evans Yaokumah, Aditta Abigail N'ganomah

Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

ARTICLE INFO

Editor name: Mohamed Fathy El-Amin Mousa

Keywords:
Computer-based examinations
Randomized item pools
Item response theory
2PL model
Item difficulty
Item discrimination

ABSTRACT

The rapid adoption of computer-based examinations (CBE) has raised concerns regarding fairness and reliability in assessing student performance. This study examines randomized test item fairness to test takers' traits in CBE through the application of Item Response Theory (IRT), with a focus on the two-parameter logistic (2PL) model. Data were collected from 500 level 100 undergraduate students who responded to ten Algebra test items. Raw scores were dichotomized into binary outcomes and analyzed to estimate each item's difficulty level (b), item's discrimination power (a), student ability (latent trait) (θ) and overall test performance. The estimated item discrimination parameter ranged from 0.87 to 2.14, while the item difficulty parameter estimate ranged from 0.26 to 1.91. These estimates indicated moderate to strong discriminatory power across most test items and its effectiveness between low and high ability examinees, as well as a balanced mix of less difficult, moderately difficult, and highly difficult items, respectively. The test-level analyses using the Test Characteristic Curve (TCC) and Test Information Curve (TIC) showed that the test was most reliable for candidates with average ability levels, clustered at latent traits of $0 \leq \theta \leq 2$. The item-fit metric evaluation utilizing the S-X^2 statistic ranging from 1.182 to 5.304, with corresponding p-values between 0.505 and 0.978 above the chosen significance level ($\alpha = 0.05$), and RMSEA values close to 0.000 highlighted eight out of the ten items within the acceptable model fit criteria. This implied that although the majority of examinees receive fair and reliable measurements from the randomized test, some items may need to be revised in order to increase measurement accuracy and equity. The study shows how IRT may be used practically for trait assessment and test item fairness evaluation in computer-based exams.

Introduction

Fairness in assessment has long been a central theme in educational research, as it ensures that students are evaluated on the basis of their true ability rather than external or biased factors [1]. With the rise of Computer-Based Examinations (CBEs), institutions have increasingly adopted randomized item pools to strengthen exam security and reduce opportunities for collusion [2]. While this approach enhances efficiency and minimizes predictability, it raises important concerns about fairness, since students may receive test versions of varying levels of difficulty [3].

* Corresponding author.
E-mail addresses: esi.eghan@knust.edu.gh (R.E. Eghan), oseisarpongedward@gmail.com (E. Osei-Sarpong), gasedlla@gmail.com (G.E. Awashie), reinbork@knust.edu.gh (R.N. Borkor), eyaokumah@st.knust.edu.gh (E. Yaokumah), abigail.addita@gmail.com (A.A. N'ganomah).

https://doi.org/10.1016/j.sciaf.2026.e03226
Received 28 August 2025; Received in revised form 28 January 2026; Accepted 30 January 2026
Available online 31 January 2026

The literature highlights that fairness extends beyond traditional metrics of reliability and validity, encompassing broader principles of equity and comparability across diverse groups of learners [1]. In CBEs, randomization can unintentionally compromise fairness when some examinees encounter disproportionately difficult items. Such inequities undermine confidence in the credibility of examinations and present challenges for institutions that rely on CBEs as a primary assessment tool. Scholars argue that true fairness requires not only secure administration but also consistency in measurement across test forms [4].

These concerns are further magnified in large-scale online learning environments such as Massive Open Online Courses (MOOCs), where learners come from highly diverse cultural, linguistic, and educational backgrounds. Studies in this area show that item properties such as difficulty and discrimination often vary across groups, potentially leading to biased outcomes if left unchecked [4]. To counter such disparities, psychometric frameworks particularly Item Response Theory (IRT) have been employed to ensure that test scores reflect actual ability rather than artifacts of item allocation.

Beyond conceptual discussions, technological innovations have also been proposed as a means to safeguard fairness. For instance, adaptive or personalized assessments dynamically tailor question difficulty to students while maintaining comparability across versions [2]. Similarly, research into vulnerabilities in exam administration emphasizes that fairness is not only about test design but also about protecting the integrity of assessment systems from risks such as question paper leakages [5]. Together, these perspectives highlight fairness as a multidimensional issue involving psychometrics, technology, and security.

The relevance of IRT extends well beyond education. Its applications in health sciences and psychology for patient-reported outcomes, adaptive testing in large scale examinations such as the Graduate Record Examination (GRE) and the Scholastic Assessment Test (SAT), and in test equating and item banking, all demonstrate its robustness as a methodological framework [6]. These diverse applications reinforce IRT's suitability for addressing fairness in CBEs, where the stakes of assessment are high and the risks of inequity are significant [7].

The CBE system has been extended at Kwame Nkrumah University of Science and Technology as part of a larger initiative to update assessments. Nearly half of all exams are now administered using computers since the university has moved tens of thousands of students to take tests and semester exams digitally in addition to its traditional paper-based format. According to KNUST Management, this change improves efficiency, security, fairness, and automatic marking while lowering malpractice through question randomization and biometric logins. Students as well as instructors have received training on how to utilize the KNUST virtual classroom and examroom platforms.

Despite the growing adoption of CBE with randomized item pools at this institution, empirical evidence on whether such randomization yields fair and comparable measurement across examinees remains limited, particularly in university's real assessment settings. Existing functionalities have largely focused on test security, cheating prevention, with fewer investigations examining test item distribution fairness to the student traits using empirical student data within a formal psychometric framework.

Building on this foundation, the specific objectives of the study are to:

1. estimate item discrimination and difficulty parameters for a randomized 10 item Algebra test using the 2PL model
2. assess item level and test level model fit using chi-square based statistics and RMSEA
3. examine whether the distribution of item information supports fair measurement across different ability levels.

The study also addresses the following research questions:

1. What are the item discrimination and difficulty estimates under a 2PL model for the 10-item algebra test?
2. To what extent do items show adequate fit and does the test provide sufficient information at relevant ability levels?
3. Is there evidence of differential items indicating fairness concerns across examinees?

Rationale for test property evaluation

The Algebra computer-based test is built on parameterized items, algorithmic randomization item selection and scoring in its delivery processes. Concerns of fairness, equivalency, and measurement consistency among test forms are raised by the possibility that different examinees may receive different item variants in the randomized computer-based testing environments. This may inadvertently favor or disadvantage particular examinee groups if they are not empirically tested. To ensure that all item variations operate similarly and that examinees are evaluated on the same underlying algebraic ability independent of the specific items administered, evaluating test properties like item difficulty, discrimination, reliability, and item information becomes necessary. Therefore, examining test characteristics using psychometric frameworks like Item Response Theory shows that the test generates accurate ability estimates, upholds administration fairness, and promotes legitimate interpretations of test results.

The remainder of this paper is organized as follows. Section "Mathematical framework" presents the mathematical framework and outlines the key attributes of the Item Response Theory (IRT) model. Section "Model description" describes the model description, focusing on the interpretation of Item Characteristic Curves (ICCs), the Test Characteristic Curve (TCC), Test Information Function (TIF), and Standard Errors (SE). Section "Results and discussion" reports and discusses the results, highlighting both item-level and test-level analyses. Finally, Section "Conclusion" concludes the study and discusses the implications of the findings for fairness in computer-based examinations.

Mathematical framework

The mathematical framework of this study is based on Item Response Theory (IRT), with particular focus on the Two-Parameter Logistic (2PL) model. The framework begins with raw test scores, which are transformed into binary outcomes to fit the IRT structure. The 2PL model then relates the probability of a correct response to a student's latent ability, capturing both item difficulty and discrimination. To ensure validity, the model relies on key assumptions such as unidimensionality, local independence, invariance, and monotonicity. These components provide the theoretical foundation for analyzing fairness in computer-based examinations by connecting observed performance to latent ability through rigorous probabilistic modeling [8].

Raw scores and binary transformation

The data for this study were obtained from a cohort of N students who responded to k items in an Algebra test. Each student's raw score on item j ($j = 1, 2, \ldots, k$) is denoted by X_{ij} for student i ($i = 1, 2, \ldots, N$). These raw scores can be represented as a matrix:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Nk} \end{bmatrix},$$

where X_{ij} is the raw score of student i on item j.

To enable Item Response Theory (IRT) analysis, these raw scores were transformed into dichotomous outcomes (correct/incorrect). Specifically,

$$B_{ij} = \begin{cases} 1 & \text{if } X_{ij} \geq T_j, \\ 0 & \text{if } X_{ij} < T_j, \end{cases}$$

where T_j is a threshold set for item j. This binary matrix $\mathbf{B}$ provides the input for IRT modeling.

IRT models

The IRT has 3 fitting models, One, Two and Three parameter logistic. The One Parameter Logistic (1PL/Rasch) model which assumes that all items discriminate equally well, using only a single parameter (difficulty). This assumption is often unrealistic for educational assessments where items vary in how effectively they differentiate between high and low ability examinees.[9] The Three-Parameter Logistic (3PL) model has both discrimination and difficulty parameters, and adds a guessing parameter, making it useful for multiple-choice tests with high guessing probability. However, the additional parameter increases model complexity and requires larger samples for stable estimation [10] The Two-Parameter Logistic (2PL) model provides an appropriate balance between simplicity and flexibility. By allowing both item difficulty and item discrimination to vary, it accurately captures item behavior more without the estimation challenges of the 3PL model.

For this study, where items differ in their discriminative power but guessing effects are minimal, the 2PL model is the most suitable choice, making it more relevant for constructed algebra tasks with little chance-level guessing. It is worth noting that the 2PL is sensitive to sample size for stable discrimination estimates

The two-Parameter Logistic (2PL) model

All mathematical formulations presented in these subsections, 2.5 through to 3.6 are based on standard Item Response Theory derivations as documented in [2,8–11] . In particular, the 2PL probability model, item and test information functions, and item fit statistics used in this study adopt their conventional forms as defined in these sources.

The 2PL model relates the probability of a correct response to a student's latent ability θ_i through two item-specific parameters: discrimination a_j and difficulty b_j. The probability that student i answers item j correctly is given by:

$$P_{ij}(\theta_i) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}. \tag{1}$$

Here, b_j indicates the ability level at which an examinee has a 50% chance of success, while a_j controls how sharply the probability changes around b_j.

Attributes of the 2PL model

The Two-Parameter Logistic (2PL) model in IRT is based on four fundamental assumptions that ensure validity and interpretability of the results [8]:

- Unidimensionality: A single latent trait θ (algebraic ability) explains performance on all items.

- Local Independence: Given θ, responses to different items are statistically independent:

$$P(X_1, X_2, \ldots, X_k \mid \theta) = \prod_{j=1}^{k} P(X_j \mid \theta).$$

- Invariance: Item parameters (a_j, b_j) and ability estimates remain stable across different samples, unlike in Classical Test Theory.
- Monotonicity: The probability of a correct response increases with ability, i.e.,

$$\frac{\partial P(X_j = 1 \mid \theta)}{\partial \theta} > 0.$$

These assumptions guarantee that item characteristic curves are meaningful, parameter estimates are interpretable, and fairness analysis through the 2PL model is theoretically sound.

Parameter estimation

Item parameters (a_j, b_j) and student abilities (θ_i) are estimated via marginal maximum likelihood and Bayesian procedures [5]. The likelihood for the observed responses is:

$$L(\mathbf{a}, \mathbf{b} \mid \mathbf{B}) = \prod_{j=1}^{k} \int_{-\infty}^{+\infty} \prod_{i=1}^{N} P_{ij}^{B_{ij}} (1 - P_{ij})^{(1-B_{ij})} f(\theta_i)\, d\theta_i. \tag{2}$$

Posterior ability estimates are then obtained, for example, through the Expected A Posteriori (EAP) method:

$$\hat{\theta}_i = \int_{-\infty}^{+\infty} \theta_i f(\theta_i \mid \mathbf{B}_i, \mathbf{a}, \mathbf{b})\, d\theta_i. \tag{3}$$

Model description

The 2PL model is best understood through a set of descriptive curves and functions that illustrate item and test behavior across ability levels. The Item Characteristic Curve (ICC) shows how the probability of success varies with ability, while the Test Characteristic Curve (TCC) aggregates these probabilities to represent expected total scores. Similarly, the Item Information Function (IIF) and Test Information Function (TIF) describe how much information is provided at different points on the ability scale, and the Standard Error of Measurement (SE) reflects the precision of ability estimates. Together, these functions provide a comprehensive view of test performance and fairness, complementing the parameter estimates obtained from the model [2].

Item Characteristic Curve (ICC)

The Item Characteristic Curve (ICC) describes the probability that a student with latent ability θ answers an item correctly. Under the 2PL model, the ICC is defined as:

$$P_{ij}(\theta) = \frac{1}{1 + \exp[-a_j(\theta - b_j)]}. \tag{4}$$

The ICC is an S-shaped logistic curve. Its position along the θ-axis is determined by the difficulty parameter b_j, while its steepness is governed by the discrimination parameter a_j. Items with high a_j values provide sharper discrimination between high- and low-ability students, which is essential for fairness in assessment. Items with low slopes, by contrast, may fail to differentiate effectively.

Test Characteristic Curve (TCC)

The Test Characteristic Curve (TCC) extends the concept of the ICC to the entire test. It represents the expected total score of a student at ability level θ:

$$TCC(\theta) = \sum_{j=1}^{k} P_{ij}(\theta). \tag{5}$$

The TCC provides a smooth function linking latent ability to observed scores. A fair and well-constructed test shows a monotonic increase, ensuring that students with higher ability consistently achieve higher expected scores.

Item Information Function (IIF)

The Item Information Function (IIF) quantifies how much statistical information an item contributes about ability at a specific θ value. In the 2PL model:

$$I_j(\theta) = a_j^2 \cdot P_j(\theta) \cdot (1 - P_j(\theta)). \tag{6}$$

Items provide the most information near their difficulty parameter b_j. Thus, a balanced test should include items targeting a wide range of ability levels to ensure fairness across the spectrum.

Test Information Function (TIF)

The Test Information Function (TIF) is obtained by summing the information across all items:

$$TIF(\theta) = \sum_{j=1}^{k} I_j(\theta). \tag{7}$$

The TIF indicates the precision of the entire test at different ability levels. High TIF values signify that the test provides more reliable measurement. A fair test distributes information evenly across relevant ability levels rather than clustering it around a narrow band.

Standard Error of measurement (SE)

The precision of ability estimates can also be expressed in terms of the Standard Error of Measurement (SE), which is inversely related to test information:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}. \tag{8}$$

Low SE values indicate higher measurement precision, while higher SE values suggest less confidence in the ability estimates. In practice, fairness requires that SE remain acceptably low across a broad range of abilities, ensuring that no group of students is unfairly assessed with greater uncertainty.

Model validation metric

The model-fit assessment was performed at the local item level to verify that the 2PL model adequately represented the observed response patterns of test takers. Item level fit was evaluated using the standardized chi-square statistic $S\text{–}X^2$, which measures the discrepancy between observed and model-predicted response frequencies across groups of examinees with similar ability levels. For each item j, the statistic was computed as

$$\chi_j^2 = \sum_{g=1}^{G} \frac{(O_{jg} - E_{jg})^2}{E_{jg}}, \tag{9}$$

where O_{jg} and E_{jg} denote the observed and model-expected numbers of correct responses in ability group g, with expected values derived from the 2PL model.

$$P_{ij} = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}, \qquad E_{jg} = \sum_{i \in g} P_{ij}. \tag{10}$$

The hypothesis test for each item is stated as

$$H_0 : \text{Item } j \text{ fits the 2PL model}, \qquad H_1 : \text{Item } j \text{ does not fit the 2PL model.}$$

Under H_0, the statistic χ_j^2 follows an approximate chi-square distribution with

$$df = G - 2,$$

reflecting the two estimated item parameters. To supplement the chi-square decision rule, RMSEA was computed as

$$RMSEA = \sqrt{\frac{\max(S\text{–}X^2 - df, 0)}{df \cdot N}},$$

together with the associated p-value. Items with non-significant p-values and low RMSEA were judged to exhibit acceptable fit to the 2PL model.

Item fit criteria

- An item is considered misfitting if its chi-square value χ_j^2 corresponds to a p-value below the chosen significance level ($\alpha = 0.05$), leading to rejection of the null hypothesis that the item fits the model.
- The RMSEA serves as a complementary measure, with values close to 0 indicating good fit and values above 0.05 suggesting potential misfit.

Therefore, items with non-significant p-values ($p > 0.05$) and low RMSEA values are regarded as adequately fitting the 2PL model, while items failing either criterion are flagged as misfitting.

Software usage

The R software, which provides a versatile framework for IRT modeling, was used for all computations. The SX^2 statistic item-fit diagnostics were generated and the item parameter estimates were extracted using the marginal maximum likelihood that was supplied in the *mirt* and *ltm* packages. Additionally, as the typical IRT outputs, these packages offered complete graphical diagnostics, including item characteristic curves and information functions.

Results and discussion

This section presents and interprets the outcomes of the analysis conducted using the Two-Parameter Logistic (2PL) Item Response Theory model [12]. The results are organized to highlight key aspects of the data and model, beginning with the raw and transformed responses, followed by item-level analyses, ability estimates of students, and test-level characteristics. Each subsection discusses not only the numerical results but also their implications for fairness, item performance, and the overall effectiveness of the test in measuring algebraic ability.

Test development

The ten major instructional units of the algebra course, thus, set theory, indices and logarithms, surds, trigonometry, polynomials, rational and partial functions, sequences and series, the binomial theorem, permutations and combinations, and matrices were taken into consideration in creating the ten test items. Every item matched a single unit, guaranteeing comprehensive curricular coverage and fair representation of the course material. The assessments used a three variety of Moodle quiz formats, thus, numerical answer format, multiple choice questions with single correct response, and cloze (embedded items) format. This aligned with the intended learning objective of each unit, and the absence of ambiguous or misleading response options, and also reflected the range of problem solving approaches suitable for the course. Additionally, this variety made sure that various cognitive processes including computing, conceptual thinking, and symbolic manipulation, were recorded in the tests.

Each test item was developed in LaTeX using the `moodle` package and parameterized with Python scripts, allowing the automated generation of several equivalent variants of each item for randomized test assembly. The XML outputs were imported into the virtual learning platform. During test administration, the Moodle engine automatically and randomly selects one item variant for each examinee. Randomization maintained test security and comparability by ensuring that, although all examinees are evaluated on the same learning objectives, no two students necessarily receive the same item combination.

Data collection

The empirical dataset used for this study is made up of test item responses from examinees who, between February 3 and March 20, 2025, finished a 10 item computer-based algebra test on the KNUST Virtual Classroom (Moodle framework). Examinee used their unique institutional credentials (username and password) to access the online tests. Each test had a password that was unique to it, and Moodle's built-in limitations prevented attempts after the deadline, guaranteeing adherence to deadlines. All units used the same testing procedure, which included 20-minute time limitations, restricted access, computerized scoring, and instant feedback following each attempt. Comparability of responses across the ten units sampled for the study was assured by this uniform approach. Test scores were downloaded from the learning platform, exported as spreadsheet format, and subsequently imported into the chosen statistical software for data cleaning and analysis. There were 525 students enrolled in the Level 100 Biochemistry for the Algebra course; however, the analysis only included 500 students who completed all ten quizzes without missing any items

Data preparation

Each test was examined on raw scores of 0 to 10 scale. This provided an initial overview of student performance variation shown in (Table 1). To fit the Item Response Theory (IRT) framework, raw scores were dichotomized: values ≥ 5 were coded as 1 (correct), and scores < 5 as 0 (incorrect).

This binary transformation yielded a dataset suitable for modeling under the Two-Parameter Logistic (2PL) IRT model (Table 2). The binary responses were further summarized into a score distribution (Table 3), showing the number and percentage of students who achieved each possible score (0–10). This distribution provided insight into the general difficulty of the test and the spread of student performance, laying the foundation for subsequent item parameter estimation and ability analysis.

Distribution of raw scores

The overall performance distribution was heavily skewed toward the lower end of the scale. As shown in Table 3 and Fig. 1, nearly one-fifth of students (18.8%) scored zero, while scores of 1–3 accounted for an additional 46.4%. Mid-range scores (5–6) were less common, and fewer than 10% of students achieved scores above 7. Only 0.6% reached the maximum score of 10.

This pattern indicates that the Algebra test was challenging for the majority of students, with only a small fraction demonstrating high mastery of the content.

Table 1
Sample of students' raw scores (0–10 scale) across 10 items.

Student	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Student 1	9	0	1	6	5	3	7	9	0	1
Student 2	1	10	1	4	6	0	7	0	3	0
Student 3	3	4	0	1	10	6	5	2	7	2
					⋮					
Student 500	0	4	1	3	1	10	0	6	2	1

Table 2
Sample of students' binary responses across 10 items (1 = correct, 0 = incorrect).

Student	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Student 1	1	0	0	1	1	0	1	1	0	0
Student 2	0	1	0	0	1	0	1	0	0	0
Student 3	0	0	0	0	1	1	1	0	1	0
					⋮					
Student 500	0	0	0	0	0	1	0	1	0	0

Table 3
Frequency and proportion of students' total raw scores.

Score value	Counts	Proportion (%)
0	94	18.8
1	76	15.2
2	82	16.4
3	74	14.8
4	53	10.6
5	32	6.4
6	41	8.2
7	18	3.6
8	23	4.6
9	4	0.8
10	3	0.6

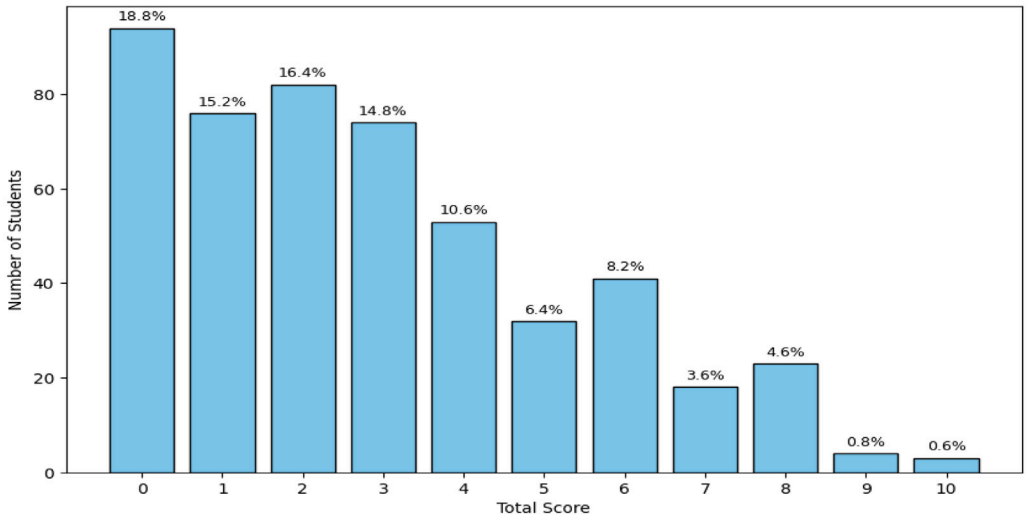


Fig. 1. Histogram of raw score distribution across the 10-item Algebra test.

Item-level distribution of correct responses

Substantial variation in item difficulty was observed across the ten test items. As shown in Table 4 and Fig. 2, Item 5 (43.8%) and Item 7 (41.2%) recorded the highest success rates, indicating they were relatively easy. By contrast, Item 10 was the most difficult, with only 7.0% of students answering correctly. Items 3, 4, 6, 8, and 9 showed moderate success rates (30%–33%), while Items 1 (15.2%) and 2 (26.6%) were more challenging.

Table 4
Distribution of correct responses across test items.

Item	Counts	Percentage (%)
Item 1	76	15.2
Item 2	133	26.6
Item 3	158	31.6
Item 4	148	29.6
Item 5	219	43.8
Item 6	156	31.2
Item 7	206	41.2
Item 8	167	33.4
Item 9	158	31.6
Item 10	35	7.0

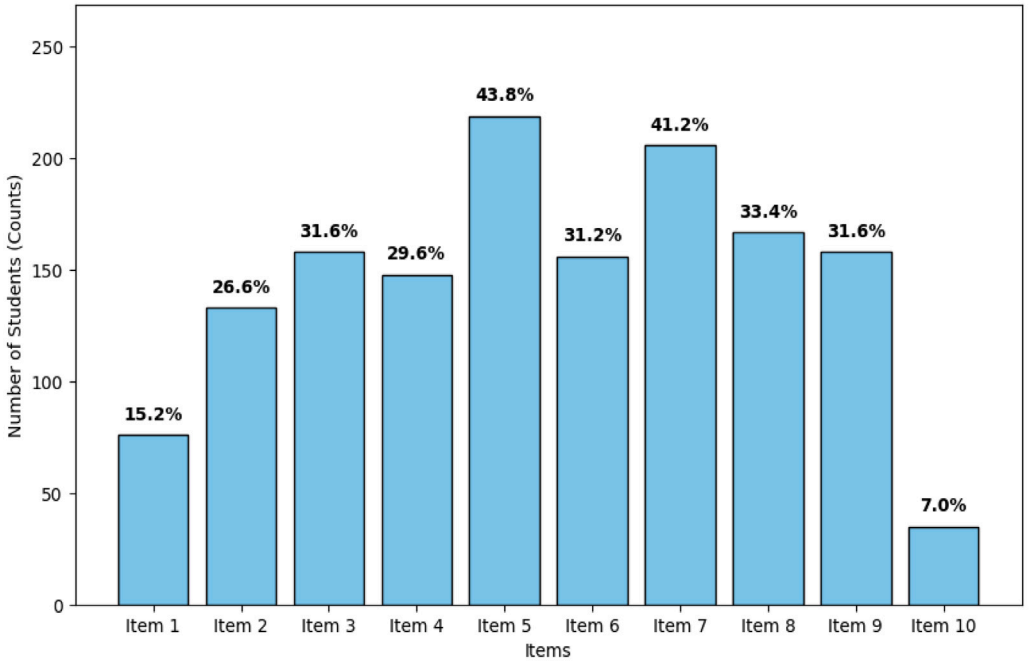


Fig. 2. Proportion of correct responses across test items.

Overall, this spread in item performance reflects an appropriate balance of easy, moderate, and difficult questions, supporting effective discrimination across student ability levels.

Two-Parameter Logistic (2PL) IRT model

After exploring the actual data and its binary transformation, the next stage of the analysis applies the Two-Parameter Logistic (2PL) Item Response Theory (IRT) model. This model captures two essential characteristics of each test item: the discrimination parameter (a) and the difficulty parameter (b). The discrimination parameter (a) measures how well an item differentiates between students of varying ability levels, while the difficulty parameter (b) indicates the ability level at which a student has a 50% chance of answering the item correctly [1].

By estimating these parameters, the 2PL model provides deeper insights into test design, identifying items that are highly effective in differentiating student ability as well as those that may be less difficult or most difficult.

Default parameter ranges

For stability and interpretability, the parameter estimation was constrained within the following ranges:

- Discrimination Parameter (a): 0.01 to 6.0
- Difficulty Parameter (b): -4 to $+4$
- Ability Parameter (θ): $-6 \leq \theta \leq 6$

Table 5
Estimated discrimination parameters (a) for test items.

Item	Discrimination (a)
1	1.4397
2	1.1279
3	1.1748
4	1.6387
5	1.3383
6	1.4792
7	0.8733
8	1.5085
9	1.8472
10	2.1417

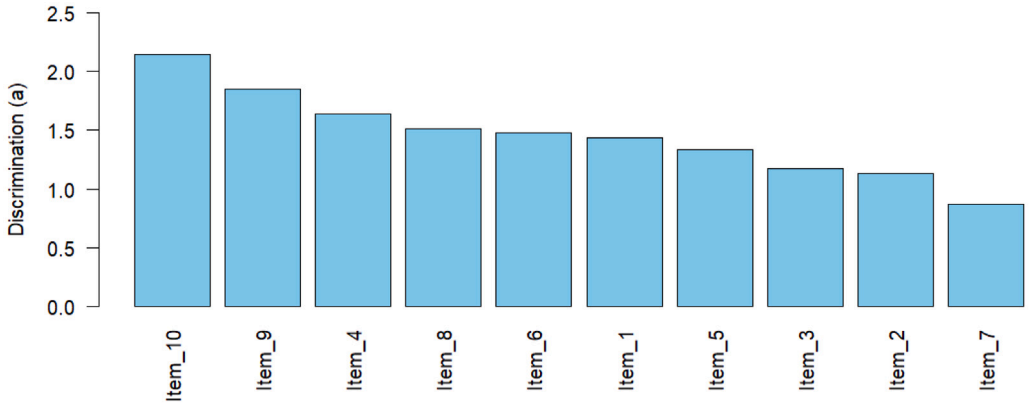


Fig. 3. Discrimination parameter values for the 10 test items.

Table 6
Estimated difficulty parameters (b) for test items.

Item	Difficulty (b)
1	1.6038
2	1.1238
3	0.8380
4	0.7792
5	0.2556
6	0.7507
7	0.4773
8	0.6507
9	0.6601
10	1.9110

Discrimination parameters

The discrimination parameter (a) indicates how well an item distinguishes among students of different ability levels. Higher values of a suggest that an item is more effective in separating high-ability from low-ability students, while lower values indicate weaker differentiation.

Table 5 shows that Items 10 ($a = 2.1417$) and 9 ($a = 1.8472$) were the highest discriminators, while Item 7 ($a = 0.8733$) was the lowest. Most items clustered between 1.1 and 1.6, reflecting moderate discriminatory power.

Fig. 3 visually reinforces the table, showing Items 10 and 9 as clear outliers with the highest discriminatory strength. Item 7 lags behind, while the rest form a middle cluster. This indicates that while most items effectively distinguish ability, a few are particularly high or low contributors.

Difficulty parameters

The difficulty parameter (b) reflects the ability level required for a 50% chance of answering an item correctly. Lower b values indicate less difficult items, while higher values represent most difficult items.

Table 6 shows that Item 5 ($b = 0.2556$) was least difficult, while Item 10 ($b = 1.9110$) was most difficult. Items 3, 4, 6, 8, and 9 fall into the moderate range (0.65–0.85), ensuring coverage across a spectrum of student abilities.

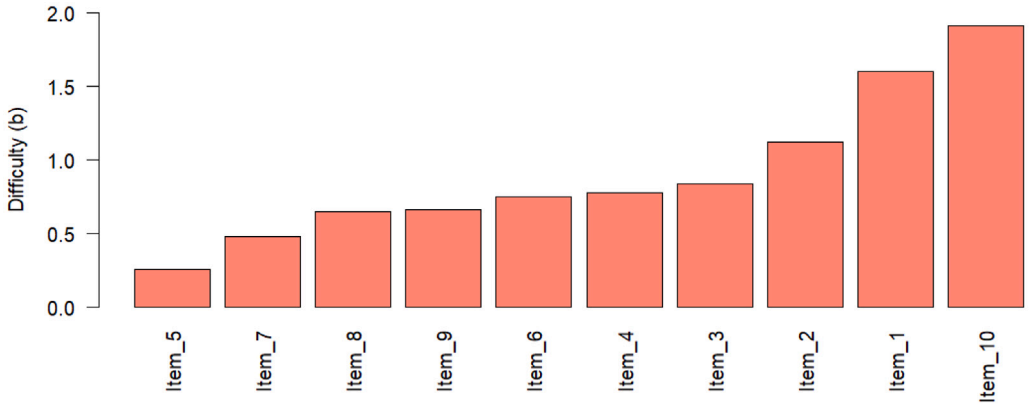


Fig. 4. Difficulty parameter values for the 10 test items.

Table 7
Ability estimates (θ) and standard errors (SE) for selected students.

Student	Ability (θ)	SE
Student_1	0.7065	0.4372
Student_2	−0.0283	0.4983
Student_3	0.4593	0.4499
Student_4	−1.1471	0.6830
Student_5	−0.7787	0.6175
Student_6	−0.6674	0.5979
⋮	⋮	⋮
Student_500	−0.1177	0.5102

Fig. 4 confirms the spread of item difficulty, from Item 5 at the low end to Item 10 at the high end, with the rest distributed in between. This balance shows the test contained both accessible and challenging items, useful for assessing students across ability levels.

Ability estimates of students

The estimated ability parameter (θ) captures each student's algebra proficiency, while the standard error (SE) reflects the precision of these estimates. As shown in Table 7, ability values ranged widely across the cohort: students such as Student 1 ($\theta = 0.71$) demonstrated above-average performance, whereas Student 4 ($\theta = -1.15$) reflected lower ability. Most students clustered around $\theta \approx 0$, indicating average proficiency. The SE values, typically between 0.44 and 0.68, suggest moderate measurement error, with higher uncertainty observed for extreme ability estimates. Together, the results highlight both the variation in student proficiency and the reliability of the ability estimates, which are central to evaluating fairness in the test.

Visualizations of student ability estimates

To further explore the distribution of student ability estimates, a combined visualization was generated that presents three complementary views: a histogram, a density plot, and a boxplot. These together provide a clearer picture of how students performed in the algebra test.

Interpretation: The combined visualization in Fig. 5 shows that most students clustered around average ability levels, with estimates centered near zero. The histogram and density plot reveal a concentration within the range −1 to 1, confirming that the majority of students demonstrated moderate proficiency in algebra. The density curve also suggests a slight skew toward lower ability values, indicating that weaker students outnumbered stronger ones. The boxplot further supports this finding, with the median close to zero and a relatively narrow interquartile range, while also highlighting a few extreme outliers on both ends of the distribution. Taken together, these plots demonstrate that although the cohort contained a spread of abilities, the dominant trend was average performance, with only a small subset of students excelling or performing very poorly.

Visualization of Item Characteristic Curves (ICCs)

Fig. 6 presents the Item Characteristic Curves (ICCs) for all ten items. The curves illustrate variation in item difficulty and discrimination: items such as 9 and 10 showed steep slopes and high discrimination, while items like 7 were flatter and low in differentiating ability levels. In terms of difficulty, items positioned to the right (e.g., 1 and 10) required higher ability, whereas items to the left (e.g., 5) were relatively less difficult.

Fig. 6 The ICCs confirm that the test included a balanced mix of easy, moderate, and difficult items, with varying discrimination power, enabling effective assessment across the full range of student abilities.

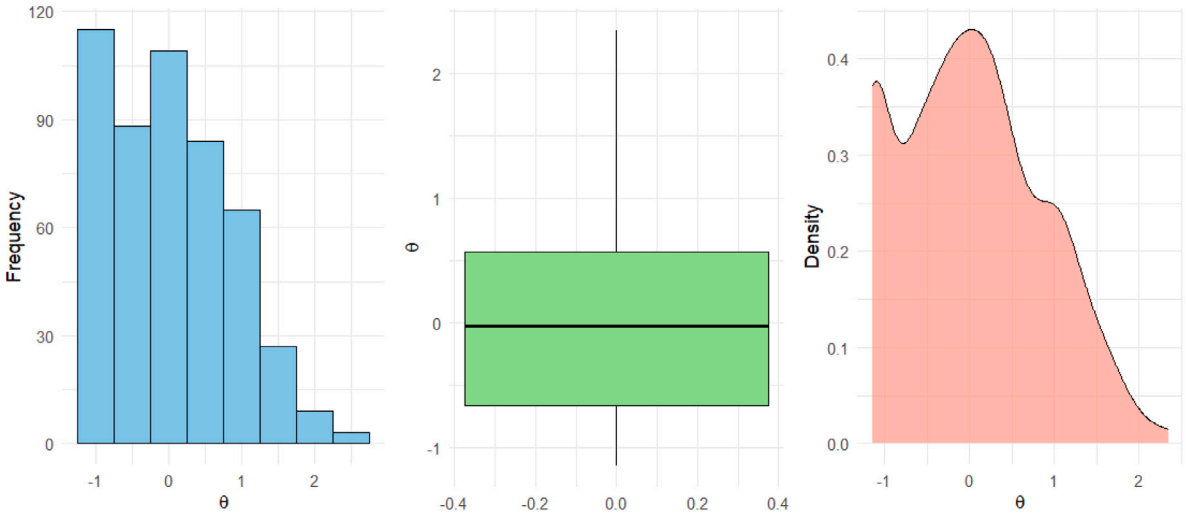


Fig. 5. Combined visualization of student ability estimates: histogram, density plot, and boxplot.

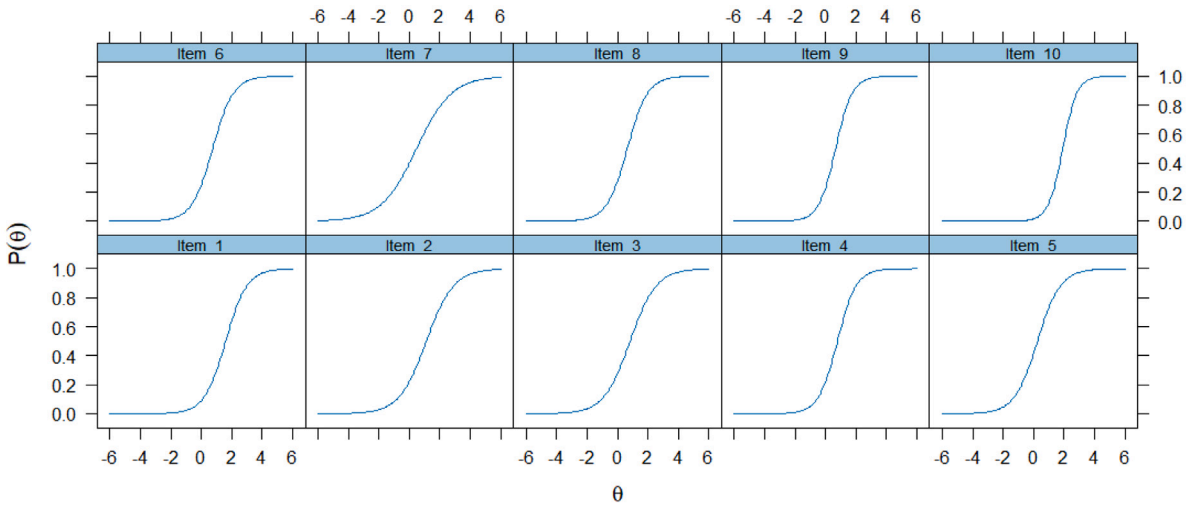


Fig. 6. Item Characteristic Curves for all items.

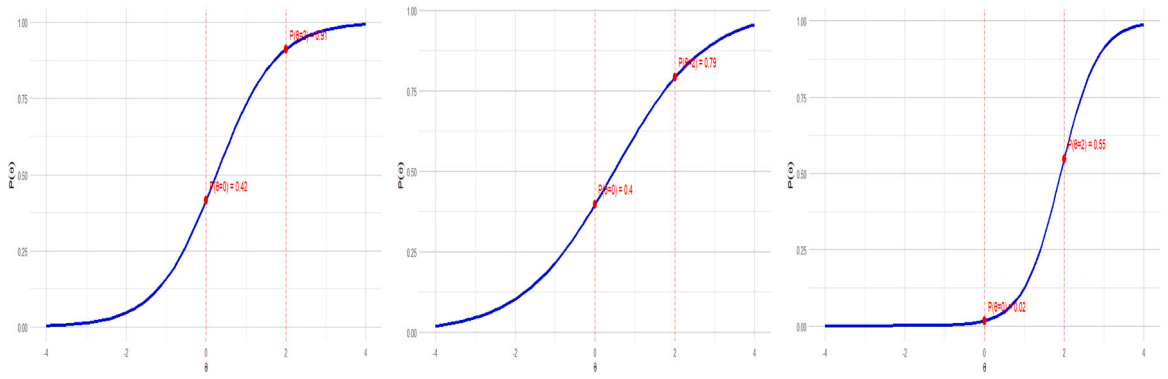
Detailed visualization of selected item characteristic curves

To provide a deeper understanding of how specific items functioned in the test, individual Item Characteristic Curves (ICCs) were examined. In particular, Items 5, 7, and 10 were selected to highlight differences in item difficulty and discrimination. The blue S-shaped curve in each graph shows the probability of a correct response as a function of student ability (θ), while the red lines indicate the probabilities at ability levels $\theta = 0$ and $\theta = 2$.

Interpretation: The selected ICCs in Fig. 7 highlight key contrasts in item functioning. Item 5, positioned to the left, was relatively less difficult: even students of average ability ($\theta = 0$) had about a 42% chance of success, rising sharply to over 90% at $\theta = 2$. Item 7, while similar in baseline probability, had a flatter slope, indicating low discrimination level; its probability only rose to about 79% at $\theta = 2$. Item 10, by contrast, was the most difficult, with virtually no chance of success at $\theta = 0$ (about 2%) and only 55% even at $\theta = 2$, making it demanding even for higher-ability students. Together, these plots show that the test contained a mix of easier, moderately difficult, and highly challenging items, each contributing differently to measuring the full range of student abilities.

Test characteristic curve with observed scores

Fig. 8 compares the model-predicted Test Characteristic Curve (TCC) with the observed average scores. The results show close alignment: as ability (θ) increases, predicted and observed scores rise monotonically, confirming that the 2PL model captures the relationship between ability and performance. Minor deviations occur at extreme ability levels, but overall the fit is strong and supports model validity.



(a) Item 5 (b) Item 7 (c) Item 10

Fig. 7. Item Characteristic Curves for selected items (5, 7, and 10), illustrating differences in difficulty and discrimination.

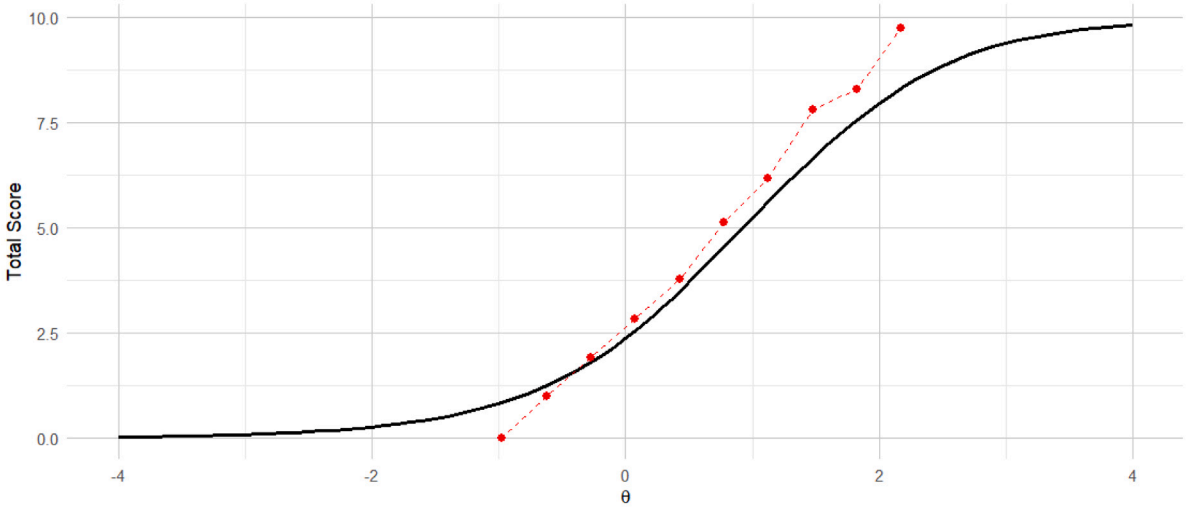


Fig. 8. Test Characteristic Curve with observed student scores.

Test Information Curve (TIC)

The Test Information Curve (Fig. 9) indicates that the test is most precise in the ability range $0 < \theta < 2$, where information peaks around 4.5. This means the test best differentiates students of average to slightly above-average ability. Precision decreases at the extremes, suggesting the need for additional items to better capture very low- or high-ability students.

Standard error of ability estimates

The ability estimates and standard errors (Table 7) show the expected pattern: measurement is most precise near the average ability range and less precise at the extremes. This is confirmed by Fig. 10, where the U-shaped curve shows lowest SE values ($\sim$ 0.42–0.45) for students with moderate ability, and higher SEs for very weak or very strong students. Thus, the test provides its greatest accuracy around the center of the ability distribution.

Item-level fit analysis

To find out how well each question adhered to the Two-Parameter Logistic (2PL) model's presumptions, item-level fit was evaluated. The S-X^2 statistic, the corresponding RMSEA value, and the p-value showing statistical significance of item misfit were the three diagnostics that were investigated. Table 8 describes the output.

A majority of the ten items in the table showed strong fit: nine items had non-significant p-values ($p > 0.05$), indicating no substantial deviation from the model's expected response patterns, and seven items had RMSEA = 0.000. With $S - X^2$ values ranging from 1.182 to 5.304, RMSEA = 0.000, and p-values between 0.505 and 0.978, items 1, 3, 5, 6, 7, and 9 demonstrated excellent match. Two items, however, presented signs of possible mismatch. Item 8 had an acceptable RMSEA (0.041) but a marginal p-value

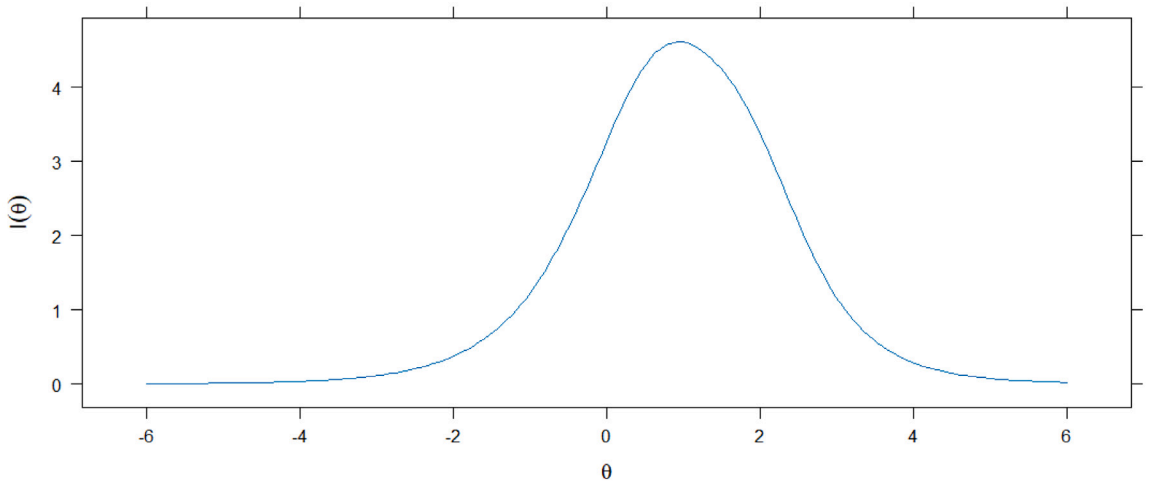


Fig. 9. Test Information Curve (TIC) for the 10-item test.

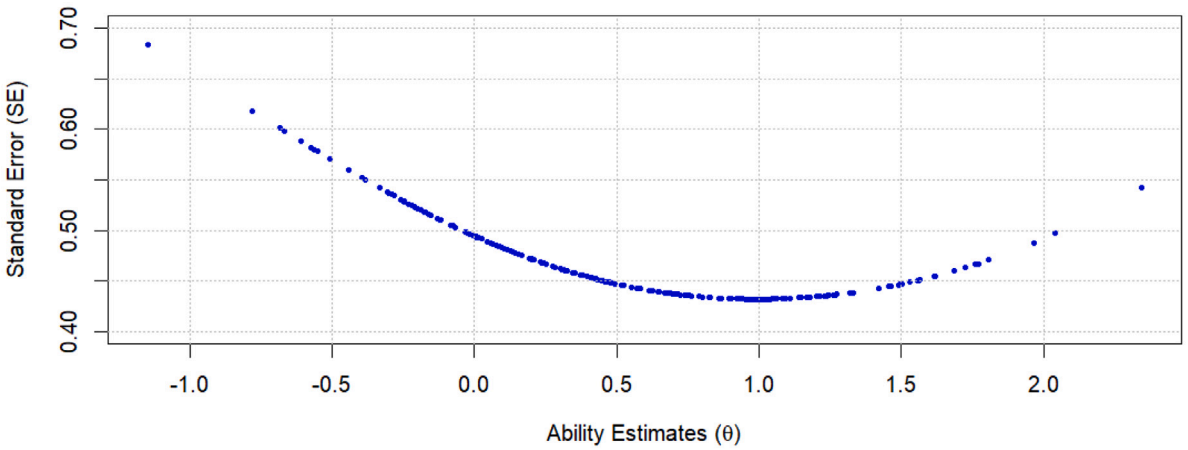


Fig. 10. Standard Error (SE) of Ability Estimates.

Table 8
Item-level fit statistics (S-X^2, RMSEA, *p*-value).

Item	S-X^2	RMSEA	*p*-value
1	1.182	0.000	0.978
2	16.100	0.058	0.013
3	4.173	0.000	0.653
4	6.976	0.018	0.323
5	2.569	0.000	0.861
6	3.599	0.000	0.731
7	2.410	0.000	0.878
8	11.053	0.041	0.087
9	5.304	0.000	0.505
10	12.064	0.045	0.061

(0.087), whereas Item 2 had a borderline RMSEA (0.058) and a significant *p*-value (0.013). These items could need to be reviewed for problems like unclear item construction or poor alignment with the latent feature.

Justification of dichotomization and its implications

In order to enable analysis under the Two-Parameter Logistic (2PL) IRT framework, which assumes binary item responses, the scores were dichotomized even though the original item responses were recorded on a 0–10 scale. The ≥ 5 threshold was selected in

accordance with the course's facilitator's test assessment protocol, which states that mastery of the learning outcome requires a score of at least 5. Dichotomization is frequently used in educational testing when the primary objective is to assess the functioning and fairness of items with a single correct response as opposed to polytomous problems that call for multiple correct response with rating scales or partial-credit scoring, [9,13]. As the algebra test was designed with single-correct-response items, item-fit diagnostics were used to evaluate the effect of the dichotomous transformation on the fundamental assumptions of the Two-Parameter Logistic (2PL) model. Although dichotomization may result in some loss of information compared to polytomous IRT models, the dichotomized data adequately satisfied the model assumptions for the intended analysis, as the majority of items demonstrated satisfactory fit under the 2PL framework. However, this constraint is acknowledged and in order to preserve score granularity, future research may take into account graded response or partial credit models.

Conclusion

This study applied the Two-Parameter Logistic (2PL) Item Response Theory (IRT) model to evaluate the quality of items in a 10 item computer based Algebra test administered to 500 students. The analysis provided detailed insights into item discrimination and difficulty, student ability estimates, and overall test performance.

The Item discrimination parameter ranged from 0.87 to 2.14 with an average of $\bar{a} = 1.46$, indicating that most items moderately to strongly differentiated between high and low ability students. Difficulty parameters spanned 0.26 to 1.91 with a mean of $\bar{b} = 0.91$, showing a balanced mix of easy, moderate, and difficult items. The hardest item was Item 10 (b = 1.91), while Item 5 was the easiest (b = 0.26).

At the test level, the Test Information Function peaked at approximately 4.5 around ability levels $\theta \approx 1$, demonstrating that the test was most precise for students with average to slightly above average ability.

This is confirmed by the ability estimations' wide range, with the majority of participants grouping around $\theta = 0$. The 2PL model's suitability was further supported by the model fit results: Based on S-X^2 and RMSEA values, eight out of ten items had an adequate fit; however, items 2 (p = 0.013) and 8 (p = 0.087) revealed indications of marginal misfit. These numerical outcome collectively indicate that the test provided reasonably fair and reliable measurement, while also highlighting specific items requiring revision.

Overall, the findings highlight that the algebra test was effective in measuring student performance across a broad range of abilities while also identifying specific items that may require revision to enhance fairness and diagnostic precision.

Study limitations

Notwithstanding the study's positives, its drawbacks are acknowledged. First of all, the study only used a 10-item test, which may restrict measurement accuracy along the entire trait measure (θ) continuum. This is especially noticeable at the ability trait's upper and lower extremes, where the Standard Errors were largest. Furthermore, the study's reliance on a single cohort of 500 students from a single academic program may limit its generalizability; therefore, future assessments' robustness and fairness evaluation would be strengthened by broadening the item pool, adding more diverse samples, and connecting parameters across semesters.

CRediT authorship contribution statement

Rhydal Esi Eghan: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing, Supervision. **Edward Osei-Sarpong:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Gaston Edem Awashie:** Methodology, Writing – review & editing, Supervision. **Reindorf Nartey Borkor:** Investigation, Writing – review & editing. **Evans Yaokumah:** Formal analysis, Methodology. **Aditta Abigail N'ganomah:** Formal analysis, Methodology.

Declaration of competing interest

Th authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Fowler, D.H. Smith, C. Emeka, M. West, C. Zilles, Are we fair?: Quantifying score impacts of computer science exams with randomized question pools, in: Proceedings of the 53rd ACM Technical Symposium on Computer Science Education, SIGCSE 2022, ACM, 2022, pp. 647–653, http://dx.doi.org/10.1145/3478431.3499388.

[2] X. Duan, X. Ye, S. Manoharan, An online system for creating personalized assessments to mitigate cheating, in: 2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering, (TALE), IEEE, 2024, pp. 1–8, http://dx.doi.org/10.1109/tale62452.2024.10834362.

[3] I.A. Onyejiekwe, R.O. Okoye, Development of mathematics achievement test using item response theory., Int. J. Res. Publ. Rev. 5 (10) (2024) 1422–1432, http://dx.doi.org/10.55248/gengpi.5.1024.2839.

[4] M. Buhl, L.B. Andreasen, Learning potentials and educational challenges of massive open online courses (moocs) in lifelong learning, Int. Rev. Educ. 64 (2) (2018) 151–160, http://dx.doi.org/10.1007/s11159-018-9716-z.

[5] M. Segado, A. Adair, J. Stewart, Y. Ma, B. Drury, D. Pritchard, A multidimensional bayesian irt method for discovering misconceptions from concept test data, Front. Psychol. 16 (2025) http://dx.doi.org/10.3389/fpsyg.2025.1506320.

[6] S.E. Woo, J. LeBreton, M. Keith, L. Tay, Bias, fairness, and validity in graduate admissions: A psychometric perspective, 2020, http://dx.doi.org/10.31234/osf.io/w5d7r.
[7] N. Abraham, A. ElBassiouny, Educational ability testing (gre/mat/mcat/lsat), 2020, http://dx.doi.org/10.1002/9781119547167.ch155.
[8] Y.-H. Chen, I.Y. Li, C. Cao, Y. Wang, Accuracy of attribute estimation in the crossed random effects linear logistic test model: impact of q-matrix misspecification, Front. Educ. 10 (2025) http://dx.doi.org/10.3389/feduc.2025.1506674.
[9] Y. Chen, X. Li, J. Liu, Z. Ying, Item response theory—a statistical framework for educational and psychological measurement, Statist. Sci. 40 (2) (2025) 167–194.
[10] Y.-H. Chen, I.Y. Li, C. Cao, Y. Wang, Accuracy of attribute estimation in the crossed random effects linear logistic test model: impact of q-matrix misspecification, in: Frontiers in Education, vol. 10, Frontiers Media SA, 2025, p. 1506674.
[11] Y. Liu, E.M. Schulz, L. Yu, Standard error estimation of 3pl irt true score equating with an mcmc method, J. Educ. Behav. Stat. 33 (3) (2008) 257–278.
[12] S.P. Reise, H. Du, E.F. Wong, A.S. Hubbard, M.G. Haviland, Matching irt models to patient-reported outcomes constructs: The graded response and log–logistic models for scaling depression, Psychometrika 86 (3) (2021) 800–824, http://dx.doi.org/10.1007/s11336-021-09802-0.
[13] X. Wu, N. Li, R. Wu, H. Liu, Cognitive analysis and path construction of chinese students' mathematics cognitive process based on cda, Sci. Rep. 15 (1) (2025) 4397.