Item Response Theory For Trait Assessment In Randomized Item Pool For Computer Based Test

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

[a] *Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*

Abstract

The rapid adoption of computer-based examinations (CBE) has raised concerns regarding fairness and reliability in assessing student performance. This study investigates fairness and test takers traits in CBE through the application of Item Response Theory (IRT), focusing on the two-parameter logistic (2PL) model. Data were collected from 500 students who responded to 10 Algebra test items. Raw scores were dichotomized into binary outcomes and analyzed to estimate each item difficulty (b_j), discrimination (a_j), and student ability (θ_i). The Item Characteristic Curves (ICCs) is used to assess the estimates indicated, and results indicates that most items discriminated effectively between low- and high-ability examinees, though some items exhibited weak slopes, suggesting potential fairness concerns. Test-level analyses using the Test Characteristic Curve (TCC) and Test Information Function (TIF) showed that the exam was most reliable for candidates with average ability levels, while measurement precision decreased at the extremes of the ability scale. The Standard Error (SE) confirmed these findings, with higher uncertainty in ability estimates for very high- and low-performing students. These results highlight that fairness in CBE is influenced by the quality of individual items and the overall distribution of test information.

Keywords: Computer-Based Examinations, Randomized item Pools, Item Response Theory, 2PL Model, Item Difficulty, Item Discrimination

1. 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].

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

[*] Corresponding Author
Email address: esi.eghan@knust.edu.gh (Rhydal Esi Eghan)

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.

Among the available frameworks, IRT has gained prominence because of its ability to link item characteristics and student ability on a common scale [6]. The Two-Parameter Logistic (2PL) model, in particular, evaluates both item difficulty—the level of ability required to answer correctly and item discrimination-the degree to which an item distinguishes between high and low ability examinees. By calibrating tests based on these parameters, IRT provides a systematic way of identifying imbalances and ensuring comparability across different exam versions [1].

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 [7]. 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 [8].

Building on this foundation, the present study applies the 2PL model of IRT to assess fairness in randomized item pools. Using a simulated dataset, the analysis focuses on whether item characteristics difficulty and discrimination are distributed equitably and whether ability estimates provide a fair representation of test-taker performance. This work thus contributes a quantitative dimension to ongoing discussions of fairness, situating IRT as a critical tool for evaluating and enhancing equity in CBEs.

The remainder of this paper is organized as follows. Section 2 presents the mathematical framework and outlines the key attributes of the Item Response Theory (IRT) model. Section 3 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 4 reports and discusses the results, highlighting both item-level and test-level analyses. Finally, Section 5 concludes the study and discusses the implications of the findings for fairness in computer-based examinations.

2. 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 [9].

2.1. 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.

2.2. The Two-Parameter Logistic (2PL) Model

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.

2.3. 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 [9]:

- 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.

2.4. 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:

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

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].

3.1. 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.

3.2. 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.

3.3. 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.

3.4. 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.

3.5. 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.

4. Results Discussion

This section presents and interprets the outcomes of the analysis conducted using the Two-Parameter Logistic (2PL) Item Response Theory model [10]. 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.

4.1. Data Preparation

A total of 500 first-year students participated in the Algebra test, which comprised ten items. Raw scores (0–10 scale) provided an initial overview of student performance variation (Table 1). To fit the Item Response Theory (IRT) framework, scores were dichotomized: values ≥ 5 were coded as 1 (correct), and values < 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.

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

4.1.1. Distribution of Raw Scores

The overall performance distribution was heavily skewed toward the lower end of the scale. As shown in Table 3 and Figure 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 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

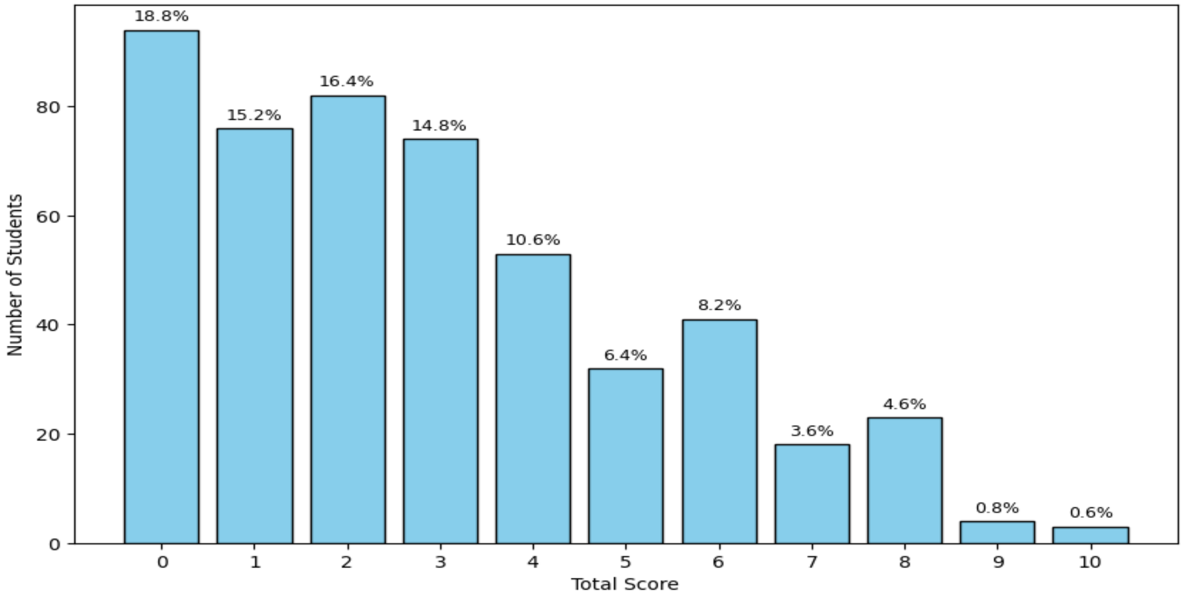


Figure 1: Histogram of raw score distribution across the 10-item Algebra test.

4.1.2. Item-Level Distribution of Correct Responses

Substantial variation in item difficulty was observed across the ten test items. As shown in Table 4 and Figure 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.

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

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

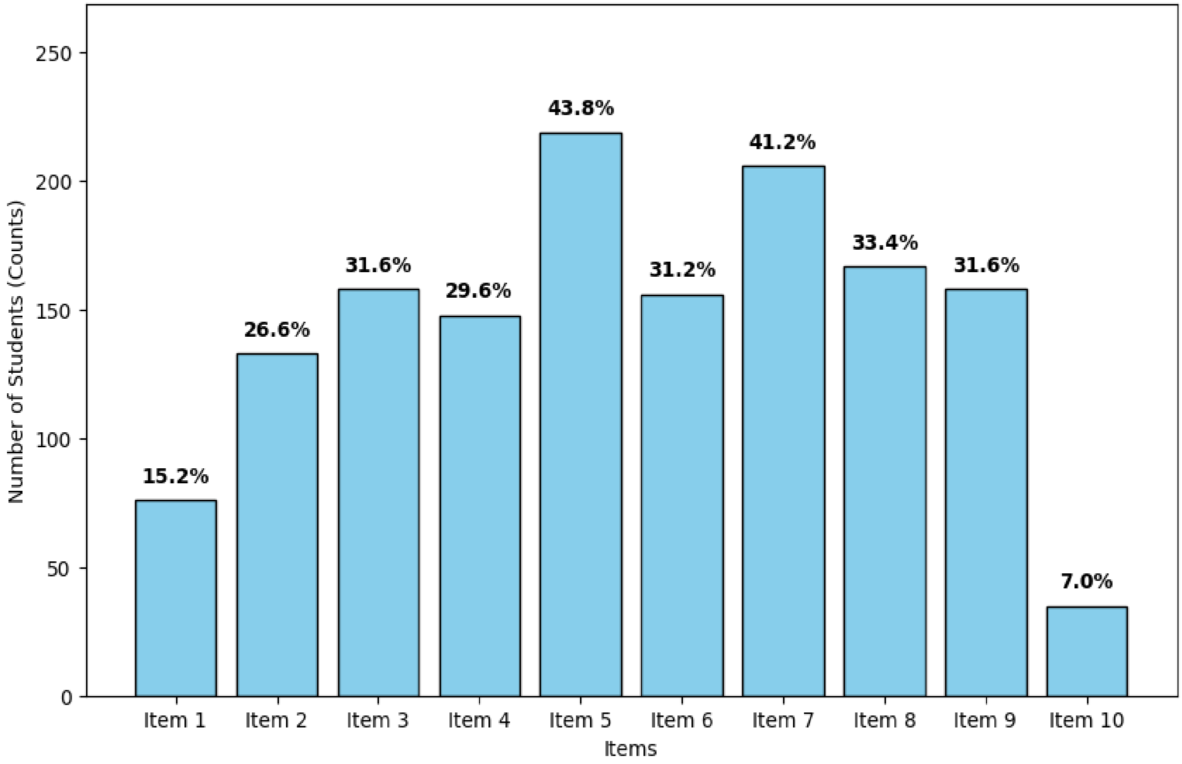


Figure 2: Proportion of correct responses across test items.

4.2. 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.

4.2.1. 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): $-6 \leq \theta \leq 6$

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: 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

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.

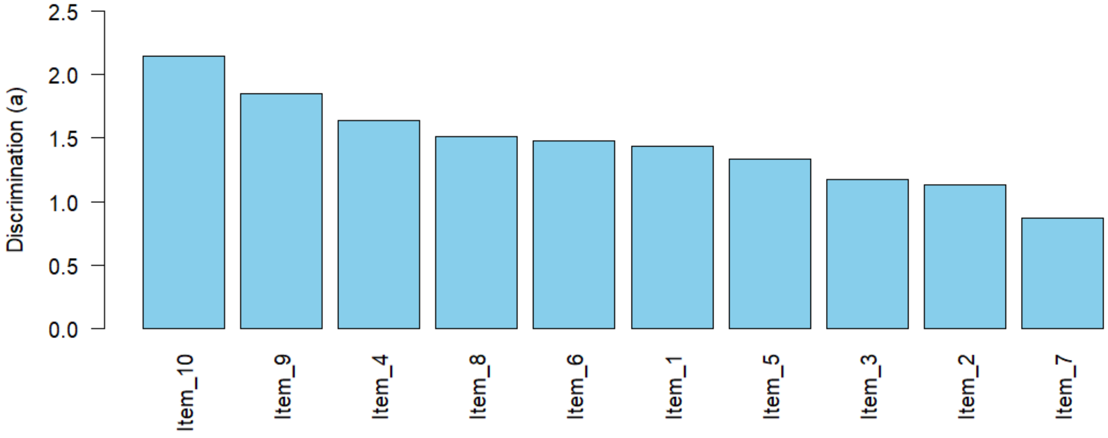


Figure 3: Discrimination parameter values for the 10 test items.

Figure 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: 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

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.

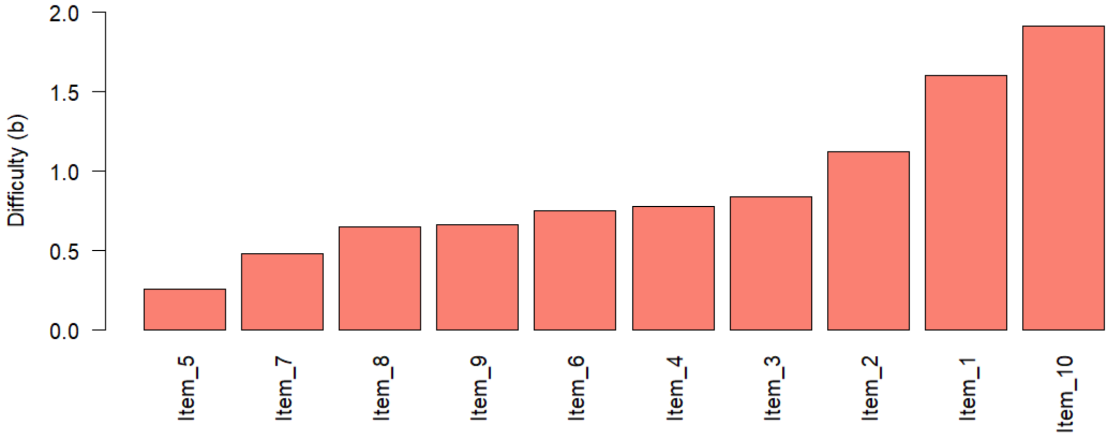


Figure 4: Difficulty parameter values for the 10 test items.

Figure 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.

4.2.2. 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.

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

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.

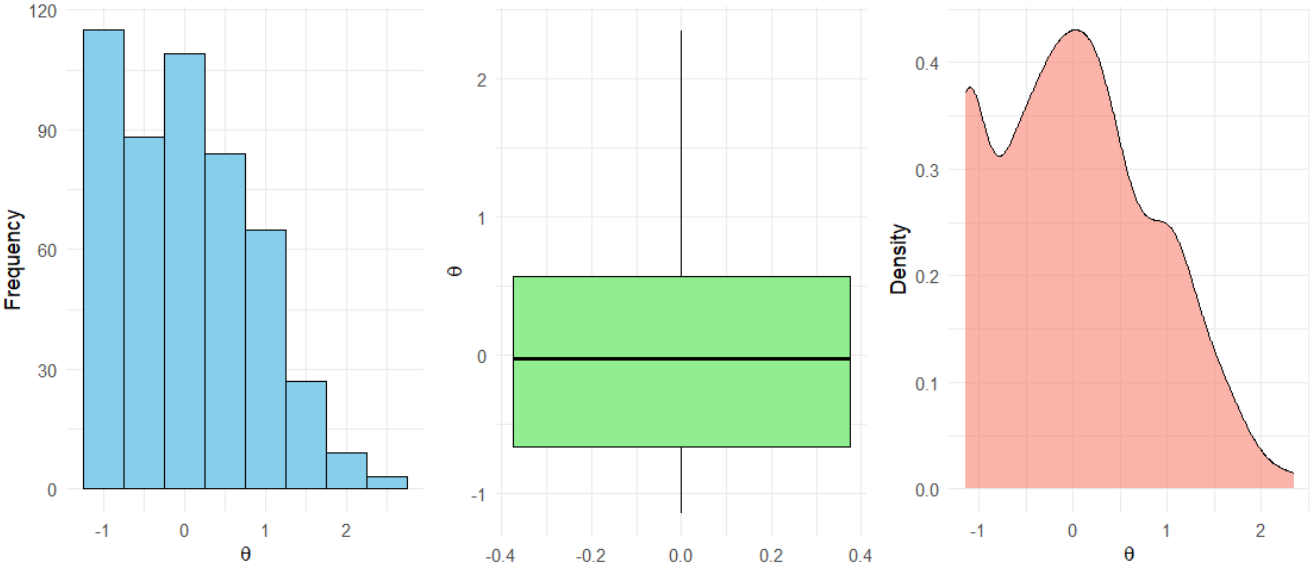


Figure 5: Combined visualization of student ability estimates: histogram, density plot, and boxplot.

Interpretation: The combined visualization in Figure 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.

4.2.3. Visualization of Item Characteristic Curves (ICCs)

Figure 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.

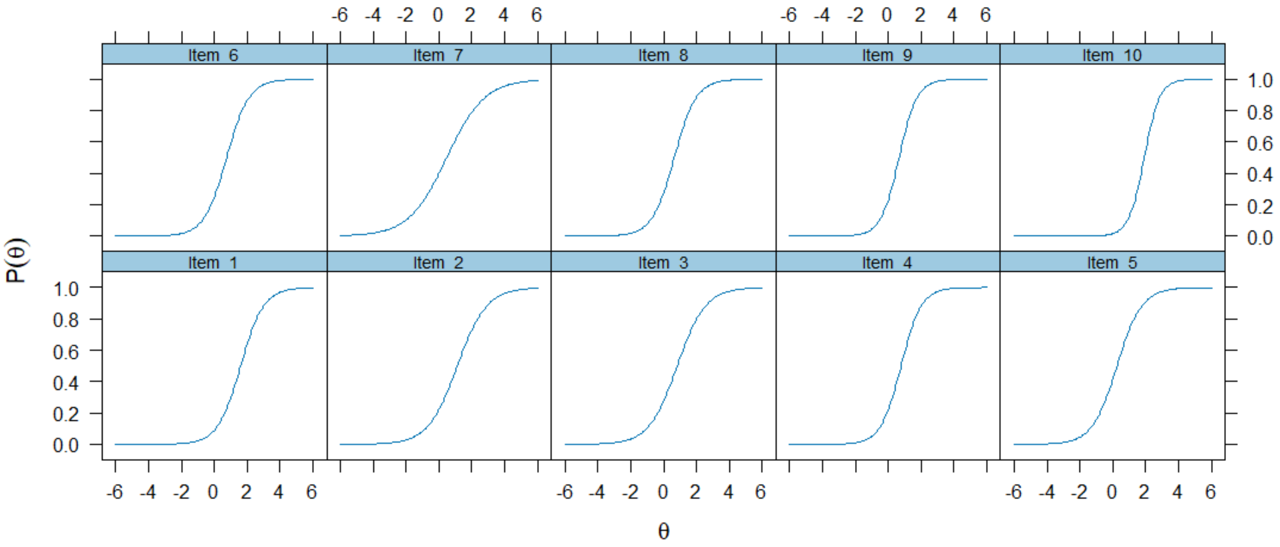
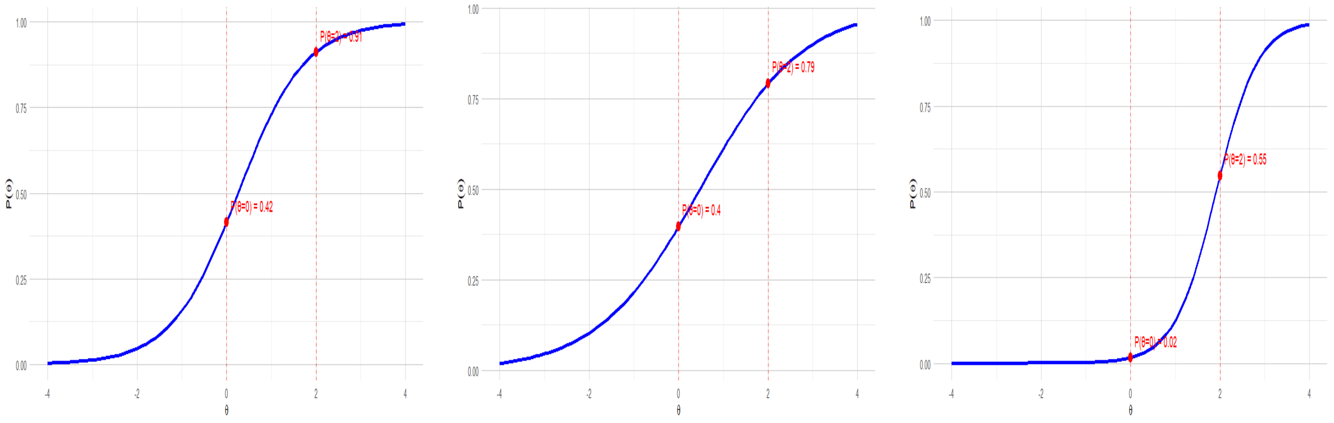


Figure 6: Item Characteristic Curves for all items.

Figure 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.

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 $\theta = 2$.



(a) Item 5

(b) Item 7

(c) Item 10

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

Interpretation: The selected ICCs in Figure 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.

4.2.4. Test Characteristic Curve with Observed Scores

Figure 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.

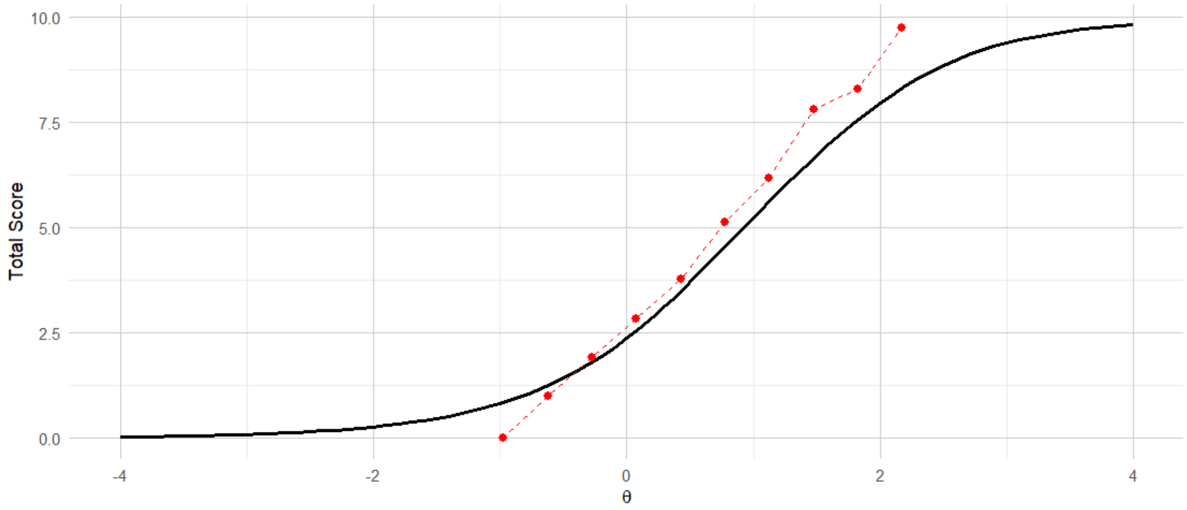


Figure 8: Test Characteristic Curve with observed student scores

4.2.5. Test Information Curve (TIC)

The Test Information Curve (Figure 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.

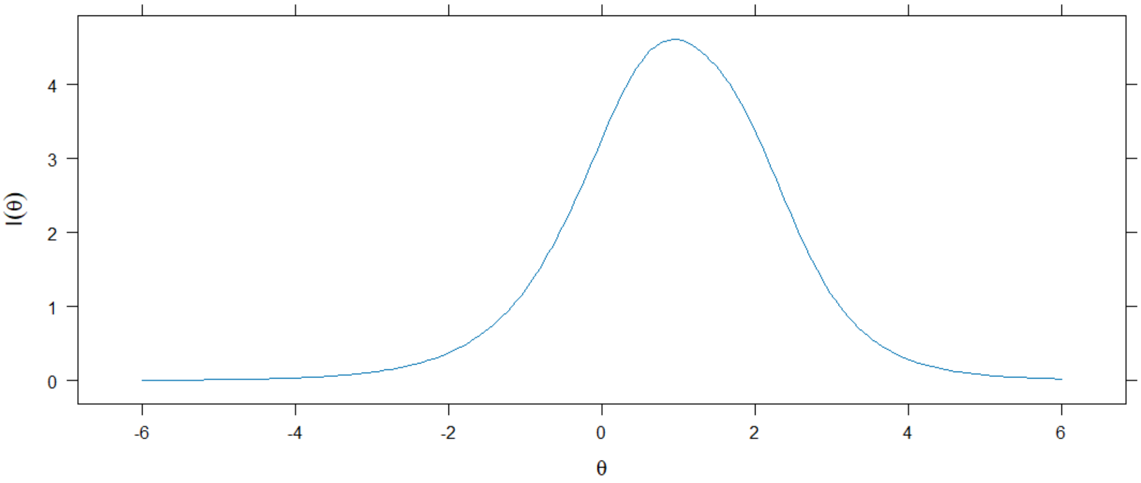


Figure 9: Test Information Curve (TIC) for the 10-item test

4.2.6. 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 Figure 10, where the U-shaped curve shows lowest SE values (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.

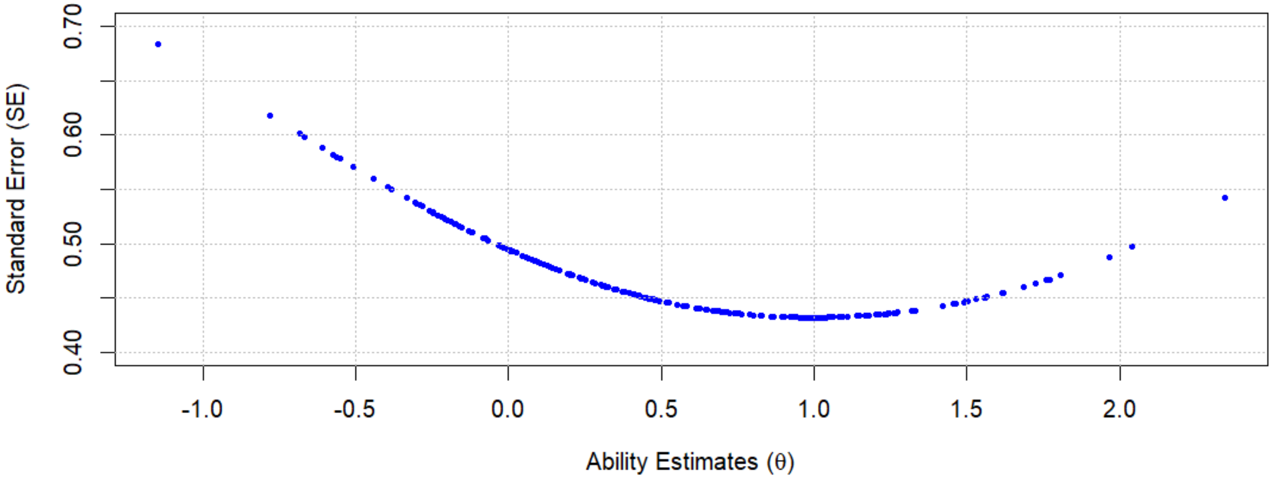


Figure 10: Standard Error (SE) of Ability Estimates

5. Conclusion

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

The results showed that most items demonstrated moderate to strong discrimination, with Items 9 and 10 standing out as highly discriminative, while Item 7 displayed relatively weaker performance. In terms of difficulty, Item 5 emerged as the easiest, whereas Item 10 was the most challenging. The distribution of student abilities revealed a wide variation across the cohort, with most students clustering around average ability but a few at the extremes. The Test Characteristic Curve and Test Information Curve further confirmed that the 2PL model adequately captured student performance and that the test was most informative for students within the average to slightly above-average ability range. Model fit statistics supported the overall adequacy of the 2PL framework, although a small number of items (e.g., Item 2) showed signs of misfit.

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. Future work could extend this analysis by incorporating larger item pools, exploring multidimensional IRT models, or linking item parameters across different cohorts to improve generalizability. Such improvements would strengthen the reliability of assessment and provide educators with more accurate tools for evaluating student proficiency in algebra.

Declaration of Interest Statement

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.

CREDIT author statement

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

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, p. 647–653. doi:10.1145/3478431.3499388.
URL 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, p. 1–8. doi:10.1109/tale62452.2024.10834362.
URL 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., International Journal of Research Publication and Reviews 5 (10) (2024) 1422–1432. doi:10.55248/gengpi.5.1024.2839.
URL 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, International Review of Education 64 (2) (2018) 151–160. doi:10.1007/s11159-018-9716-z.
URL 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, Frontiers in Psychology 16 (Jan. 2025). doi:10.3389/fpsyg.2025.1506320.
URL http://dx.doi.org/10.3389/fpsyg.2025.1506320

[6] X. Wu, N. Li, R. Wu, H. Liu, Cognitive analysis and path construction of chinese students' mathematics cognitive process based on cda, Scientific Reports 15 (1) (Feb. 2025). doi:10.1038/s41598-025-89000-5.
URL http://dx.doi.org/10.1038/s41598-025-89000-5

[7] S. E. Woo, J. LeBreton, M. Keith, L. Tay, Bias, fairness, and validity in graduate admissions: A psychometric perspective (Aug. 2020). doi:10.31234/osf.io/w5d7r.
URL http://dx.doi.org/10.31234/osf.io/w5d7r

[8] N. Abraham, A. ElBassiouny, Educational ability testing (gre/mat/mcat/lsat) (Sep. 2020). doi:10.1002/9781119547167.ch155.
URL http://dx.doi.org/10.1002/9781119547167.ch155

[9] 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, Frontiers in Education 10 (Feb. 2025). doi:10.3389/

feduc.2025.1506674.
URL http://dx.doi.org/10.3389/feduc.2025.1506674

[10] 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. doi:10.1007/s11336-021-09802-0.
URL http://dx.doi.org/10.1007/s11336-021-09802-0