**KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY**
**KUMASI**
**COLLEGE OF SCIENCE**
**DEPARTMENT OF MATHEMATICS**


# TRAIT ASSESSMENT IN RANDOMIZED ITEM POOL FOR COMPUTER BASE TEST USING 2PL IRT MODEL: CASE STUDY KNUST

By
EDWARD OSEI-SARPONG

A PROJECT SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF PHILOSOPHY, APPLIED MATHEMATICS

September 10, 2025

# Declaration

We hereby explicitly state that this dissertation is our own work done under the capable guidance of Dr. Rhydal Esi Eghan toward the award of the Master of Philosophy (MPHIL.) in Applied Mathematics and that, to the best of our knowledge, it does not contain any material that has been previously published by another person or material that has been accepted for the award of any other Masters from the university, with the exception of instances where appropriate acknowledgment has been made in the text.

| | | |
|---|---|---|
| EDWARD OSEI-SARPONG | ..... | ..... |
| Student | Signature | Date |

Certified by:

| | | |
|---|---|---|
| DR. RHYDAL ESI EGHAN | ..... | ..... |
| Supervisor | Signature | Date |

Certified by:

| | | |
|---|---|---|
| VERY REV. PROF.WILLIAM OBENG DENTEH | ..... | ..... |
| Head of Department | Signature | Date |

# Dedication

This work is dedicated to the Almighty God, whose grace and guidance made it possible to reach this milestone. It is also dedicated to my wife, Mrs Abigail Osei-Sarpong for their tireless sacrifices and constant encouragement throughout my studies. I extend this dedication to my supervisors and lecturers whose guidance, patience, and valuable insights shaped my academic growth. Finally, I dedicate this work to Very Rev Prof William Obeng-Denteh, my friends and family whose love, prayers, and support gave me strength and motivation during this journey.

# Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Rhydal Esi Eghan, for her invaluable guidance, mentorship, and support throughout the course of this research. I am equally thankful to my co-supervisor, Dr. Joshua Kiddy Kwasi Asamoah, whose insights and encouragement greatly contributed to the successful completion of this work. I also wish to acknowledge Dr. Gaston Edem Awashie for his constructive feedback and guidance, which helped to refine and strengthen this research. Special appreciation goes to Evans Yaokumah and Additta Abigail N'ganomah for their assistance and contributions during different stages of this project. I am also deeply grateful to all my lecturers for the knowledge, training, and inspiration that formed the foundation of my academic journey. Finally, I extend my heartfelt thanks to my family, friends, and all those who supported me in diverse ways, whether directly or indirectly. Your encouragement, prayers, and belief in me have been a constant source of motivation. May God richly bless you all.

# ABSTRACT

Computer-Based Examinations (CBEs) have increasingly adopted randomized question pools to enhance test security and efficiency. While this approach minimizes predictability, it raises fairness concerns as different students may encounter test versions with varying levels of difficulty. This study applied the Two-Parameter Logistic (2PL) model of Item Response Theory (IRT) to evaluate fairness in randomized question pools using a simulated dataset. The analysis focused on two key parameters: item difficulty, which reflects the level of ability required to answer an item correctly, and item discrimination, which measures how well an item differentiates between students of differing ability levels. Ability estimates of test takers were further derived to assess overall performance across the simulated cohort. The findings show that the assessment exhibits a balanced range of item difficulties, with some items being relatively more challenging, and that most items demonstrate acceptable to strong discrimination parameters. These results suggest that while the test was generally fair and reliable, variations in item characteristics highlight the importance of careful calibration in constructing randomized question pools to ensure equity in CBEs.

# Contents

**Declaration** . . . . . **i**

**Dedication** . . . . . **ii**

**Acknowledgment** . . . . . **iii**

**Abstract** . . . . . **iv**

**1 INTRODUCTION** . . . . . **1**

1.1 Background to the Study . . . . . 1

1.2 Problem Statement . . . . . 3

1.3 Research Objectives . . . . . 3

1.3.1 Sub-Objectives . . . . . 4

1.4 Methodology . . . . . 4

1.5 Significance of the Study . . . . . 5

1.6 Organization of the Study . . . . . 6

**2 Literature Review** . . . . . **7**

2.1 Introduction . . . . . 7

2.2 Conceptualizing Fairness in Assessment . . . . . 7

2.3 Applications of Item Response Theory (IRT) . . . . . 8

2.4 Technological Approaches to Mitigating Bias in CBEs . . . . . 9

2.5 Security Vulnerabilities and Examination Integrity . . . . . 9

2.6 Synthesis and Relevance to the Present Study . . . . . 10

**3 METHODOLOGY** . . . . . **11**

3.1 Data Collation Procedure . . . . . 11

3.1.1 Raw Score Recording . . . . . 11

3.1.2 Binary Transformation . . . 11
3.1.3 Aggregated Performance Data . . . 12
3.1.4 Item Level Distribution . . . 13
3.2 Item Response Theory (IRT) Models . . . 13
3.2.1 Attributes of IRT Models . . . 15
3.2.2 Parameter Estimation . . . 16
3.2.3 Characteristic Curves . . . 18
3.2.4 Standard Error of Measurement . . . 19
3.2.5 Model Validation Metric . . . 21

**4 Results and Analysis . . . 25**
4.1 Data Collation . . . 25
4.1.1 Transformation into Binary Outcomes . . . 26
4.1.2 Binary Score Distribution . . . 26
4.1.3 Item-Level Distribution of Correct Responses . . . 28
4.2 Two-Parameter Logistic (2PL) IRT Model . . . 30
4.2.1 Default Parameter Ranges . . . 30
4.2.2 Item Parameter Estimates . . . 31
4.2.3 Ability Estimates of Students . . . 33
4.2.4 Visualization of Item Characteristic Curves (ICCs) . . . 36
4.2.5 Test Characteristic Curve (TCC) . . . 39
4.2.6 Test Characteristic Curve with Observed Scores . . . 40
4.2.7 Test Information Curve (TIC) . . . 41
4.2.8 Standard Error of Ability Estimates . . . 42
4.2.9 Overall Model Fit . . . 43

**5 Discussion, Conclusion, and Recommendations . . . 47**
5.1 Discussion of Findings . . . 47
5.2 Conclusion . . . 49
5.3 Recommendations . . . 49

**References . . . 52**

# List of Tables


4.1 Sample of Students' Raw Scores (0–10 scale) across 10 Items. . . . . 25
4.2 Sample of Students' Binary Responses across 10 Items (1 = correct, 0 = incorrect). 26
4.3 Frequency and Proportions of Total Raw Scores Across Students . . . . . 27
4.4 Distribution of Correct Responses Across Test Items . . . . . 29
4.5 Estimated Discrimination Parameters ($a$) for Test Items . . . . . 31
4.6 Estimated Difficulty Parameters ($b$) for Test Items . . . . . 32
4.7 Ability estimates of the first 10 and last 3 students. . . . . 34
4.8 Ability Estimates and Standard Errors for Selected Students . . . . . 42
4.9 Global Model Fit Statistics for the 2PL Model . . . . . 44
4.10 Item Fit Statistics ($S$-$X^2$) for the 2PL Model . . . . . 45


# List of Figures


4.1 Histogram showing the distribution of students' raw scores across the 10-item algebra test. . . . . 28
4.2 Proportion of Correct Responses Across Test Items . . . . . 29
4.3 Discrimination parameter values for the 10 test items. . . . . 32
4.4 Difficulty parameter values for the 10 test items. . . . . 33
4.5 Histogram of Student Ability Estimates . . . . . 35
4.6 Density Plot of Student Ability Estimates . . . . . 35
4.7 Boxplot of Student Ability Estimates . . . . . 36
4.8 Item Characteristic Curves for all items. . . . . 37
4.9 Item Characteristic Curve for Item 5. . . . . 38
4.10 Item Characteristic Curve for Item 7. . . . . 38

4.11 Item Characteristic Curve for Item 10. . . . . 39
4.12 Test Characteristic Curve (TCC) for the 10-item test . . . . . 40
4.13 Test Characteristic Curve with observed student scores . . . . . 40
4.14 Test Information Curve (TIC) for the 10-item test . . . . . 41
4.15 Standard Error (SE) of Ability Estimates . . . . . 43

# Chapter 1

# INTRODUCTION

## 1.1 Background to the Study

Computer-Based Examinations (CBEs) are assessments administered electronically rather than through the traditional paper-based approach. They are taken on specialized computer platforms, either at designated test centers or online, and may include multiple-choice questions, essays, simulations, or interactive formats, making them a highly flexible assessment tool Duan et al. (2024).

In recent years, the use of CBEs has expanded significantly due to the rapid growth of digital learning content and the increasing adoption of online learning platforms. They are widely applied in certification exams, standardized testing, and university assessments, offering numerous advantages such as automatic marking, reduced logistical costs, scalability, and the provision of immediate feedback to learners Fowler et al. (2022). In addition, security measures such as candidate authentication and proctoring systems have been introduced to safeguard the integrity of exams. Despite these benefits, however, concerns about fairness have emerged, particularly in cases where randomized question pools are used Meyer and Zhu (2013).

At Kwame Nkrumah University of Science and Technology (KNUST), the transition to CBEs began cautiously before the COVID-19 pandemic, with limited implementation in certain certification and standardized tests. Paper-based examinations remained the dominant mode of assessment, largely due to concerns about infrastructural capacity, students' readiness, and the integrity of online examinations. However, the COVID-19 pandemic accelerated the adoption of CBEs as universities sought alternatives to in-person examinations. While this transition allowed teaching and assessment to continue, it also exposed challenges relating to academic dishonesty, accessibility, and fairness. The sudden shift meant that institutions often adopted online platforms without adequate preparation for issues of monitoring, proctoring, or equitable assessment.

In the post-COVID era, CBEs have remained an integral part of university assessment strategies due to their efficiency, scalability, and adaptability. Institutions, including KNUST,

have continued to use online exam systems-now enhanced by AI-supported proctoring and verification programs-in an effort to address earlier shortcomings. Nonetheless, questions of fairness and test equity persist, underscoring the need for ongoing research and development in the area of digital assessment Duan et al. (2024).

Several strategies have been applied at KNUST to ensure the fairness of CBEs. Initially, question banks were used to provide consistency, but this raised the possibility of students sharing solutions, which compromised exam integrity. Shuffling techniques, which randomized the order of questions, helped reduce overt cheating but did not eliminate disparities in difficulty levels among students' tests. More advanced approaches, such as parameterized questions (where numerical values or problem forms are varied across students), have improved equity but require high computational capacity and careful calibration. Misalignment in difficulty levels across parameterized items remains a challenge, highlighting the need for sophisticated statistical modeling to ensure fairness Fowler et al. (2022) Meyer and Zhu (2013).

A key fairness concern in CBEs arises from randomized question pools. While randomization enhances security by reducing the likelihood that students will receive identical exams, it can unintentionally create inequities if some students face substantially more difficult questions than others. Such disparities threaten the validity of results and may disadvantage certain groups of test takers. To address this, advanced psychometric approaches such as Item Response Theory (IRT) have been proposed as robust methods to evaluate and balance item difficulty Meyer and Zhu (2013), thereby ensuring fairer assessments at KNUST and beyond .

## 1.2 Problem Statement

Computer-Based Examinations (CBEs) have become an increasingly common mode of assessment in higher education due to their efficiency, scalability, and enhanced security features. One of the most widely adopted strategies in CBEs is the use of randomized question pools, which are intended to reduce cheating by ensuring that no two students receive the same set of items. While this enhances test security, it also introduces a critical limitation: inconsistencies in question difficulty Fowler et al. (2022).

Students who are randomly assigned to different sets of questions may encounter unequal levels of difficulty, leading to disparities in their overall performance scores. Such inconsistencies compromise the fairness and reliability of the examination process, since performance outcomes may not truly reflect differences in studentsâ knowledge or ability but rather variations in the difficulty of the questions assigned Andersen et al. (2025).

At Kwame Nkrumah University of Science and Technology (KNUST), concerns about fairness in CBEs have been raised by both students and faculty. Students have expressed dissatisfaction, noting that certain test versions appear more challenging than others, resulting in grading inconsistencies. Faculty members, on the other hand, face difficulties in ensuring that randomized pools maintain equivalent levels of difficulty across different exam versions. These challenges collectively highlight the pressing issue of fairness and equity in the administration of CBEs.

Although CBEs are widely recognized for their advantages, the fairness concerns that arise from randomized question pools remain insufficiently addressed in the literature, particularly within the context of KNUST. The unresolved issue of unequal question difficulty threatens the credibility of CBEs as a fair assessment tool and calls for urgent attention in order to safeguard both academic integrity and student trust in the system Davies and Ingram (2025).

## 1.3 Research Objectives

**Main Objective** The main objective of this research is to quantify the unfairness in randomized question sets and provide suggestions for improving equity in Computer-Based Examinations (CBEs) at Kwame Nkrumah University of Science and Technology (KNUST).

### 1.3.1 Sub-Objectives

- To apply Item Response Theory (IRT) in assessing the fairness of randomized question pools used in CBEs.
- To fit the Two-Parameter Logistic (2PL) IRT model for evaluating item properties such as discrimination and difficulty.
- To determine the ability levels (latent traits) of test takers within the study cohort based on item parameter estimates.

## 1.4 Methodology

This study employs Item Response Theory (IRT) as the main analytical framework for examining fairness in Computer-Based Examinations (CBEs). The analysis is based on scaling student responses against item characteristics, with a focus on the Two-Parameter Logistic (2PL) model Fowler et al. (2022). This model provides estimates of item difficulty and discrimination, which together make it possible to assess whether randomized question pools are fair and balanced across students.

The methodology broadly involves four key stages:

1. Organizing and preparing student response data for analysis.
2. Applying the 2PL IRT model to estimate item parameters (difficulty and discrimination) as well as student ability levels.
3. Using the standard errors (SE) of parameter estimates to evaluate the precision of the model and to identify potential misfitting items.
4. Interpreting the results to assess the fairness of randomized question pools and propose recommendations for improving equity in CBEs.

This structured approach ensures that both student ability and item characteristics are considered, while also checking for misfit through the use of standard errors. The outcome is a more reliable and equitable evaluation of Computer-Based Examinations at KNUST.

## 1.5 Significance of the Study

1. This study provides insights into how variations in question difficulty affect student performance in Computer-Based Examinations (CBEs). Such insights are vital for creating equitable examination systems that fairly assess all students, regardless of their background knowledge or skill level. By employing Item Response Theory (IRT), the research ensures a more reliable testing process by minimizing biases that could disadvantage certain groups of students.

2. Beyond the case of KNUST, the findings have broader implications for improving academic examination integrity in digital learning environments. Randomized question pools are increasingly adopted to reduce collusion and cheating, but they also pose fairness challenges. By analyzing how randomization influences performance variability, this study contributes to establishing best practices for designing balanced and equitable online tests. These outcomes can guide policy formulation and support institutions in maintaining fairness, accuracy, and academic integrity in large-scale assessment systems.

## 1.6 Organization of the Study

The study is structured into five chapters, each addressing a key aspect of fairness in randomized Computer-Based Examinations.

1. **Chapter One** introduces the study by presenting the background, problem statement, objectives, significance, and organization of the research.

2. **Chapter Two** reviews relevant literature on Computer-Based Examinations, with a focus on randomization methods such as fixed, shuffled, and parameterized question banks, and examines their implications for fairness and test validity.

3. **Chapter Three** outlines the research methodology, highlighting the application of Item Response Theory (IRT) to estimate item parameters, assess fairness, and identify potential misfit using standard errors.

4. **Chapter Four** presents the results of the analysis, including item parameter estimates, graphical interpretations, model fit statistics, and discussions of how randomization affects fairness and student performance.

5. **Chapter Five** discusses the major findings of the study, draws conclusions, and provides recommendations for improving fairness in Computer-Based Examinations and for future research in the area.

# Chapter 2

# Literature Review

## 2.1 Introduction

Fairness in assessment has long been a central concern in educational measurement, particularly as testing systems move toward computer-based formats. With the increasing reliance on randomized question pools in computer-based examinations (CBEs), the challenge of ensuring that all students are assessed under equitable conditions has become even more pressing. The reviewed literature provides perspectives on fairness, equity, vulnerabilities, and innovative approaches, while also highlighting broader applications of Item Response Theory (IRT) in diverse fields beyond education.

## 2.2 Conceptualizing Fairness in Assessment

Scholars emphasize that fairness is not only a technical property of tests but also a broader principle of equity in educational opportunities. Traditional psychometric frameworks have focused largely on reliability and validity, but recent works argue that fairness should also account for how different groups of learners interpret and respond to assessments Fowler et al. (2022).

### Fairness Beyond Reliability and Validity

Fairness extends beyond technical qualities such as reliability and validity to encompass issues of equity and comparability across diverse groups of learners. A fair assessment is one that provides consistent meaning across students regardless of their backgrounds or the test form they receive. The principle of measurement invariance and item comparability therefore becomes central in guaranteeing fairness in such assessments Fowler et al. (2022).

## 2.3 Applications of Item Response Theory (IRT)

While IRT has its roots in educational measurement, its applications span across multiple fields. This broader perspective situates IRT as a versatile framework not only for addressing fairness in CBEs but also for enhancing measurement precision in other domains.

### Educational Testing

IRT has been extensively used in large-scale educational testing programs such as the Graduate Record Examination (GRE), Scholastic Assessment Test (SAT), and national assessment frameworks. Its role in scaling item difficulty and equating test forms ensures that test scores are comparable across different administrations. By modeling the probability of a correct response based on student ability and item parameters, IRT supports fairness in high-stakes contexts Abraham and ElBassiouny (2020).

### Health and Psychology

In the health sciences and psychology, IRT has been employed to design and validate patient-reported outcome measures and quality-of-life instruments. Unlike classical test theory, IRT enables precise measurement across a range of latent traits such as anxiety, depression, or physical functioning. Its application in clinical trials ensures that instruments are sensitive enough to detect meaningful changes in patient well-being Zhang et al. (2023).

### Adaptive Testing

Computerized Adaptive Testing (CAT) is one of the most transformative applications of IRT. In CAT systems, the test dynamically adjusts item difficulty according to the test taker's estimated ability level. This reduces test length while maintaining precision, ensuring that students are neither overwhelmed by excessively difficult items nor given items that are too easy. Such adaptability parallels the fairness goals of CBEs, making CAT a key reference point for this study İnce and Özbay (2025).

### Item Banking and Test Equating

IRT also provides a framework for developing item banks, where calibrated items are stored and used across multiple test administrations. This enables test equating, ensuring fairness when

different cohorts of students receive different test forms. The ability to equate tests across administrations is directly relevant to the challenges of randomized question pools in CBEs Samsudin et al. (2020).

## 2.4 Technological Approaches to Mitigating Bias in CBEs

Beyond psychometric modeling, technological innovations such as personalized or adaptive assessments have been proposed to address fairness concerns. These innovations aim to reduce collusion opportunities and balance question exposure, while also addressing inequities caused by randomization Duan et al. (2024).

### Personalized Assessments as a Fairness Strategy

An online system that generates personalized assessments has been proposed as a way to ensure individualized yet equivalent versions of an exam. While the present study does not adopt a personalized framework, it builds on the recognition that fairness can be supported through both psychometric rigor and technological safeguards Duan et al. (2024).

## 2.5 Security Vulnerabilities and Examination Integrity

Fairness in assessment also depends on safeguarding the integrity of the testing process itself. Studies on high-stakes examinations have shown that vulnerabilities such as paper leakage, weak storage, and poor monitoring create inequities far greater than those caused by randomization Molepo (2024).

### Mitigating Risks for Fairness

Strong security measuresâsuch as separate printing and delivery of exam papers, enhanced monitoring, and accountability mechanismsâare necessary to complement psychometric fairness. In this sense, measurement fairness and security fairness form complementary pillars of trustworthy assessment systems Molepo (2024).

## 2.6 Synthesis and Relevance to the Present Study

Taken together, the literature establishes a trajectory in fairness research: from defining fairness as a multidimensional principle, to applying IRT across diverse domains, to leveraging adaptive and technological innovations, and finally to strengthening exam integrity. Building on these foundations, this study applies the two-parameter logistic (2PL) model of IRT to evaluate fairness in randomized question pools within CBEs. In doing so, it contributes a quantitative perspective to a body of literature that spans education, health, adaptive testing, and psychometrics more broadly.

# Chapter 3

# METHODOLOGY

## 3.1 Data Collation Procedure

### 3.1.1 Raw Score Recording

The data for this study is to be obtained from a cohort of $N$ test takers participating in a standardized assessment. The test must be administered under controlled conditions to ensure reliability. As such, all participants responds to the random set of $k$ items for an item bank, and the overall responses are collected immediately after test administration. Each test taker's actual score $(X_i, i = 1, 2, ..., N)$ per item $j, j = 1, 2, .., k$ is to be recorded out of the allocated marks. This produced an item-by-item score matrix, where rows represented test takers and columns represented test items scores as

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

Where: $X_{ij}$ is the score obtained by test taker i on item j, and each entry satisfies $0 \leq X_{ij} \leq M_j$

### 3.1.2 Binary Transformation

In order to prepare the dataset for Item Response Theory (IRT) analysis, a threshold needs to be applied to the raw scores of each test item. Given an item's maximum allocated marks $M_j$, some threshold $T$ with $T \leq M_j$ is set to represents the midpoint of the scoring scale and provides a clear distinction between satisfactory and unsatisfactory performance. This ensures comparability across items and test takers, allowing for categorical analysis of performance. A binary coding is defined as follows:

- Set $X_{ij} = 1$, if $X_{ij} \geq T$, representing a correct or successful response.

- Set score $X_{ij} = 0$, if $X_{ij} < T$, representing an incorrect or unsuccessful response.

Here's the generic binary score matrix for N test takers and k items:

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

This transformation converts the raw scores into a dichotomous dataset, making it suitable for analysis under the IRT model. By using this threshold, the analysis focuses on whether students achieved at least a satisfactory level of understanding on each test item, rather than partial credit.

### 3.1.3 Aggregated Performance Data

Given

- a total number of $N$ test takers.
- total number of $k$ items , distributed across relevant subject domains.

Aggregated Performance $S_i$ gives the number of correctly answered items (binary "1") per test taker summed to generate an individual performance index. Thus

$$S_i = \sum_{j=1}^{k} B_{ij}, \quad i = 1, 2, ..., N$$

where $B_{ij}$ is the binary score of test taker $i$ on item $j$.

The set of all possible score values that $S_i$ can take is $x \in [0, M_j]$, with it number of occurrences (*counts*) recorded. To allow for relative comparisons, proportions of each score level are also calculated as:

$$\text{Proportion}(x) = \frac{\text{Count}(x)}{N}. \tag{3.1}$$

This ensures that the distribution reflects both the absolute counts and the percentage share of each index within the total binary dataset. This analysis establishes the overall performance trend of the test takers, showing the extent to which low, average , and high scores dominates the results.

### 3.1.4 Item Level Distribution

Following the binary transformation, the distribution of correct responses across the test items is analyzed. This step helps to identify which items are relatively easy and which posed greater challenges to the test takers.

For each of the k test items, the total number of students who obtained a correct response (coded as 1) is calculated as

$$S_j = \sum_{i=1}^{N} B_{ij}, \quad j = 1, 2, ..., k$$

where $S_j \in [0, N]$, with relative proportions of each item score level a calculated as:

$$\text{Proportion}(S_j) = \frac{S_j}{N}. \tag{3.2}$$

This analysis provided a direct measure of item difficulty, where higher proportions indicated easier items and lower proportions indicated more difficult ones. These results forms the groundwork for subsequent application of the Item Response Theory (IRT) model.

## 3.2 Item Response Theory (IRT) Models

Item Response Theory (IRT) provides a statistical framework for modeling the relationship between an individual's latent trait and their probability of answering test items correctly. Unlike Classical Test Theory (CTT), which focuses on total scores, IRT examines performance at the item level, allowing for more precise estimates of both student ability and item characteristics Chen et al. (2025a).

In general, the probability that a test taker with latent trait (ability level) $\theta$ answers an item correctly is modeled by a logistic function of $\theta$ and some item parameters. The most widely used dichotomous IRT models are the One-Parameter Logistic (1PL), Two-Parameter Logistic (2PL), and Three-Parameter Logistic (3PL) models.

#### The One-Parameter Logistic Model (1PL)

The 1PL model, also known as the Rasch model, assumes that all items have the same discrimination power. The probability that test taker $i$ with ability $\theta_i$ answers item $j$ correctly is given

by:

$$P(X_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp[-a(\theta_i - b_j)]}, \tag{3.3}$$

where $a$ is the common discrimination parameter across all items, and $b_j$ represents the difficulty of item $j$. The limitation of this model is that it assumes equal discrimination for all items, which may not reflect reality.

### The Two-Parameter Logistic Model (2PL)

The 2PL model extends the Rasch model by allowing each item to have its own discrimination parameter. The probability that student $i$ with ability $\theta_i$ answers item $j$ correctly is:

$$P(X_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}, \tag{3.4}$$

where:

- $a_j$ = discrimination parameter of item $j$ (how well the item differentiates between students of different ability levels),
- $b_j$ = difficulty parameter of item $j$ (the ability level required to have a 50% chance of answering correctly).

This flexibility makes the 2PL model suitable for tests where items vary in their ability to discriminate among students. It is the model employed in this study, as it captures both the difficulty and discrimination characteristics of the Algebra test items.

### The Three-Parameter Logistic Model (3PL)

The 3PL model further extends the 2PL model by including a guessing parameter $c_j$, which accounts for the probability that a student with very low ability could still answer an item correctly by chance. The model is expressed as:

$$P(X_{ij} = 1 \mid \theta_i) = c_j + (1 - c_j)\frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}, \tag{3.5}$$

where $c_j$ represents the lower asymptote of the curve (guessing probability). Although the 3PL model provides a more realistic description of multiple-choice tests where guessing is possible, it requires more data for stable estimation of parameters.

#### Model Choice for This Study

For the current analysis, the **2-Parameter Logistic (2PL) model** was selected. This choice is justified by the fact that the Algebra test items are open-ended rather than multiple-choice, making the inclusion of a guessing parameter unnecessary. At the same time, allowing each item to have its own discrimination and difficulty parameters provides a more realistic representation of item behavior compared to the restrictive 1PL model. Thus, the 2PL model strikes an appropriate balance between model complexity and interpretability for this dataset.

### 3.2.1 Attributes of IRT Models

The application of Item Response Theory (IRT) relies on several fundamental assumptions that guarantee the validity and interpretability of the model results Chen et al. (2025b). For the Two-Parameter Logistic (2PL) model used in this study, the following attributes are essential:

#### Unidimensionality

Unidimensionality assumes that a single latent trait $\theta$ (in this case, *algebraic ability*) explains performance on all test items. Formally, the probability of a correct response on item $j$ depends only on $\theta$:

$$P(X_j = 1 \mid \theta) = f(\theta), \tag{3.6}$$

where $X_j$ is the response on item $j$ and $f(\cdot)$ is the logistic function in the IRT model.

#### Local Independence

Local independence means that once the latent trait $\theta$ is controlled for, a student's responses to different items are statistically independent. This can be written as:

$$P(X_1, X_2, \ldots, X_j \mid \theta) = \prod_{j=1}^{k} P(X_j \mid \theta). \tag{3.7}$$

This ensures that correlations between items are explained solely by $\theta$, without additional hidden factors.

**Invariance**

Invariance states that item parameters (difficulty $b_j$, discrimination $a_j$) and ability estimates $\theta$ remain stable across different samples of students or sets of items. That is, if $S$ is any representative sample of students,

$$P(X_j = 1 \mid \theta, a_j, b_j) \quad \text{is invariant with respect to } S. \tag{3.8}$$

This means results are not sample-dependent, unlike in Classical Test Theory (CTT).

**Monotonicity**

Monotonicity requires that the probability of a correct response increases as ability $\theta$ increases. For the logistic model, this is expressed as:

$$\frac{\partial P(X_j = 1 \mid \theta)}{\partial \theta} > 0. \tag{3.9}$$

Thus, Item Characteristic Curves (ICCs) are strictly increasing S-shaped functions, reflecting that higher ability always implies greater success probability.

Together, these four attributes provide the theoretical foundation of IRT models and ensure that the 2PL model applied in this study produces meaningful and reliable parameter estimates.

### 3.2.2 Parameter Estimation

In this study, parameter estimation was carried out using the Two-Parameter Logistic (2PL) IRT model. The estimation process involved two main components: estimation of the item parameters and estimation of the student ability parameters.

**Estimation of Item Parameters**

The item parameters consist of the discrimination ($a_j$) and difficulty ($b_j$) of each item. These parameters were estimated using the Marginal Maximum Likelihood Estimation (MMLE) method in combination with the Expectation-Maximization (EM) algorithm.

The probability of a student $i$ with ability $\theta_i$ answering item $j$ correctly under the 2PL model is given by:

$$P_{ij} = P(X_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}. \tag{3.10}$$

Since the ability parameter $\theta_i$ is unobserved (latent), it is treated as a random variable with a standard normal prior distribution. The marginal likelihood of the item parameters is therefore expressed as:

$$L(\mathbf{a}, \mathbf{b} \mid \mathbf{X}) = \prod_{j=1}^{k} \int_{-\infty}^{+\infty} \prod_{i=1}^{N} P_{ij}^{X_{ij}} (1 - P_{ij})^{(1-X_{ij})} f(\theta_i)\, d\theta_i, \tag{3.11}$$

where $f(\theta_i)$ is the probability density function of the standard normal distribution.

This likelihood is maximized iteratively using the EM algorithm, which alternates between:

- E-step: estimating the expected values of the latent abilities given the current item parameters,
- M-step: updating the item parameters $(a_j, b_j)$ to maximize the expected likelihood.

Through repeated iterations, the estimates of $a_j$ and $b_j$ converge to stable values.

**Estimation of Student Abilities**

Once the item parameters were obtained, the ability parameters $(\theta_i)$ for each student were estimated using the Empirical Bayes approach, specifically the Expected A Posteriori (EAP) method. For each student, the posterior distribution of ability given the observed responses is:

$$f(\theta_i \mid \mathbf{X}_i, \mathbf{a}, \mathbf{b}) \propto L(\theta_i \mid \mathbf{X}_i, \mathbf{a}, \mathbf{b}) f(\theta_i), \tag{3.12}$$

where $L(\theta_i \mid \mathbf{X}_i, \mathbf{a}, \mathbf{b})$ is the likelihood of the student's response pattern and $f(\theta_i)$ is the standard normal prior.

The EAP estimate of ability is then computed as the mean of this posterior distribution:

$$\hat{\theta}_i = \int_{-\infty}^{+\infty} \theta_i\, f(\theta_i \mid \mathbf{X}_i, \mathbf{a}, \mathbf{b})\, d\theta_i. \tag{3.13}$$

This method produces finite and stable estimates for all students, even those with extreme response patterns (all correct or all incorrect), where Maximum Likelihood Estimation (MLE) would otherwise fail.

**Summary**

In summary, the item parameters $(a_j, b_j)$ were estimated using MMLE with the EM algorithm, while the student ability parameters $(\theta_i)$ were obtained using the EAP method. This two-stage estimation procedure ensured robust and consistent parameter estimates for the 2PL model.

### 3.2.3 Characteristic Curves

Once the item and ability parameters were estimated, the model outputs were visualized using different types of characteristic curves. These curves provide insight into the functioning of both individual items and the overall test under the Item Response Theory framework.

**Item Characteristic Curve (ICC)**

The Item Characteristic Curve describes the probability that a student with ability $\theta_i$ answers item $j$ correctly. For the 2PL model, the ICC is given by:

$$P_{ij}(\theta_i) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}, \tag{3.14}$$

where $a_j$ is the discrimination parameter and $b_j$ is the difficulty parameter.

The discrimination parameter $(a_j)$ controls the steepness of the curve, with higher values indicating that the item better differentiates between students of slightly different ability levels.

The difficulty parameter $(b_j)$ determines the location of the curve along the ability axis. Specifically, $b_j$ is the ability level at which the probability of answering item $j$ correctly equals 0.5. Items with higher difficulty values are shifted to the right, requiring greater ability for a student to have a 50% chance of success, while easier items (with lower $b_j$ values) are shifted to the left.

Thus, the ICC not only reflects how well an item discriminates between students but also the level of ability at which the item provides the most information.

**Test Characteristic Curve (TCC)**

The Test Characteristic Curve is obtained by summing the ICCs of all items in the test. It represents the expected total score for a student with ability $\theta_i$:

$$TCC(\theta_i) = \sum_{j=1}^{k} P_{ij}(\theta_i). \tag{3.15}$$

The TCC thus reflects how overall test performance increases as student ability increases.

**Item Information Function (IIF)**

The Item Information Function quantifies the amount of statistical information an item provides about a studentâs ability at a given level of $\theta$. For item $j$, the information function is:

$$I_j(\theta) = a_j^2 \, P_{ij}(\theta) \big(1 - P_{ij}(\theta)\big). \tag{3.16}$$

Items with higher discrimination parameters provide more information, particularly around their difficulty level $b_j$.

**Test Information Function (TIF)**

The Test Information Function aggregates information across all items in the test and is defined as:

$$TIF(\theta) = \sum_{j=1}^{k} I_j(\theta). \tag{3.17}$$

The TIF indicates how precisely the test measures student ability at different levels of $\theta$. Higher information values correspond to greater measurement precision, while lower values imply less reliability at those ability levels.

**Summary**

Together, the ICC, TCC, IIF, and TIF provide a comprehensive view of how items and the test as a whole function under the 2PL model. They illustrate not only the probability of success on items but also the precision and reliability of measurement across the ability spectrum.

### 3.2.4 Standard Error of Measurement

In Item Response Theory (IRT), the ability of each student is estimated along with its associated standard error of measurement (SE). This standard error provides a measure of the precision

of the estimated ability $\hat{\theta}$. Smaller SE values indicate more precise estimates, whereas larger SE values reflect greater uncertainty Liu et al. (2008).

### Fisher Information and Standard Error

The computation of SE is directly related to the concept of Fisher information. At a given ability level $\theta$, the standard error of the estimated ability is defined as the reciprocal of the square root of the test information function:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}, \tag{3.18}$$

where $I(\hat{\theta})$ represents the amount of information the test provides at $\hat{\theta}$.

### Item Information in the 2PL Model

Under the Two-Parameter Logistic (2PL) model, each item $i$ is characterized by a discrimination parameter $a_j$ and a difficulty parameter $b_j$. The probability of a correct response to item $j$ for a student with ability $\theta$ is expressed as:

$$P_j(\theta) = \frac{1}{1 + e^{-a_j(\theta - b_j)}}. \tag{3.19}$$

The Fisher information contributed by item $j$ at ability level $\theta$ is then given by:

$$I_j(\theta) = a_j^2 \cdot P_j(\theta) \cdot \big(1 - P_j(\theta)\big). \tag{3.20}$$

This expression shows that items provide the most information around their difficulty parameter $b_i$, where the slope of the item characteristic curve is steepest.

### Total Test Information

Since the test is composed of multiple items, the total test information function is obtained by summing the contributions from all $k$ items:

$$I(\theta) = \sum_{j=1}^{k} I_j(\theta). \tag{3.21}$$

This function describes how informative the entire test is across different ability levels.

**Final Computation of the Standard Error**

Finally, substituting the test information function into the formula for SE gives:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\sum_{j=1}^{k} a_j^2 \cdot P_j(\hat{\theta}) \cdot (1 - P_j(\hat{\theta}))}}. \tag{3.22}$$

This means that ability estimates are most precise (smaller SE values) at ability levels where the test provides the most information, typically around the difficulty parameters of the items. Conversely, students with very high or very low ability levels have larger SEs, since fewer items are informative in those regions.

### 3.2.5 Model Validation Metric

Model fit assessment was performed at both the global (overall model) and local (item-level) levels to ensure that the two-parameter logistic (2PL) model adequately captured the observed student response patterns Nye et al. (2020).

**Overall Model Fit**

The global model fit was evaluated using the log-likelihood ($\log L$), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

The log-likelihood is defined as:

$$\log L = \sum_{i=1}^{N} \sum_{j=1}^{k} \log P_{ij}(\theta_j), \tag{3.23}$$

where $P_{ij}(\theta_j)$ is the model-predicted probability that student $i$ responds correctly to item $j$, $N$ is the total number of students, and $k$ is the number of items. This function is maximized during estimation using the EM algorithm.

From the log-likelihood, the following model selection criteria are derived:

$$AIC = -2 \cdot \log L + 2p, \tag{3.24}$$

$$BIC = -2 \cdot \log L + p \cdot \log(N), \tag{3.25}$$

where $p$ is the number of estimated parameters. Lower values of AIC and BIC indicate better model fit, with BIC imposing a stronger penalty for larger sample sizes.

### Item-Level Fit

To assess the adequacy of fit for individual items, the standardized chi-square statistic $S\text{-}X^2$ was employed. This statistic evaluates the discrepancy between observed and model-predicted response frequencies across groups of examinees with similar ability levels.

### Chi-square Formula (Generic Form)

For a given item $j$, the chi-square statistic is computed as

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

where

- $g = 1, 2, \ldots, G$ indexes ability groups (test takers grouped by their estimated ability level $\hat{\theta}$),
- $O_{jg}$ is the observed number of examinees in group $g$ who answered item $j$ correctly,
- $E_{jg}$ is the expected number of correct responses in group $g$ based on the IRT model.

### Expected Probabilities in the 2PL IRT Model

In the two-parameter logistic (2PL) model, the probability of a correct response for person $i$ on item $j$ is given by

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

where

- $a_j$ is the discrimination parameter for item $j$,
- $b_j$ is the difficulty parameter for item $j$,
- $\theta_i$ is the ability of test taker $i$.

Hence, for an ability group $g$ containing $n_g$ test takers, the expected number of correct responses is

$$E_{jg} = \sum_{i \in g} P_{ij}(\theta_i).$$

**Sampling Distribution**

Under the null hypothesis that the item fits the 2PL model, the chi-square statistic $\chi_j^2$ approximately follows a chi-square distribution with degrees of freedom

$$df = G - p,$$

where $G$ is the number of ability groups and $p$ is the number of item parameters estimated (with $p = 2$ for the 2PL model).

**Hypothesis Testing**

The statistical test is framed as follows:

$$H_0 : \text{Item } j \text{ fits the 2PL model,}$$

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

If $\chi_j^2$ is sufficiently large such that the corresponding $p$-value falls below a chosen significance level (e.g., $\alpha = 0.05$), then $H_0$ is rejected and item $j$ is considered to exhibit model misfit.

- Degrees of Freedom (df): Based on the number of independent response categories after accounting for estimated parameters.
- Root Mean Square Error of Approximation (RMSEA):

$$RMSEA = \sqrt{\frac{\max(S\text{-}X^2 - df, 0)}{df \cdot N}}, \tag{3.26}$$

  where values close to 0 indicate excellent fit and values above 0.05 suggest potential misfit.
- p-value: The probability of observing the $S$-$X^2$ value under the null hypothesis of good

fit. A high $p$-value ($p > 0.05$) suggests good fit, whereas a low $p$-value ($p < 0.05$) indicates misfit.

The detailed results of this analysis, including $S\text{-}X^2$, df, RMSEA, and p-values for each item, are presented in Table 4.10. These indices collectively provide evidence of how well the 2PL model describes the response behavior of individual test items, ensuring the robustness of subsequent parameter interpretation Esomonu and Anayo (2025).

# Chapter 4

# Results and Analysis

## 4.1 Data Collation

At the Department of Biochemistry, Kwame Nkrumah University of Science and Technology (KNUST), a total of 500 first-year (Level 100) students participated in a test aimed at evaluating their grasp of fundamental concepts in Algebra. The assessment comprised 10 items, each allocated a maximum of 10 marks. For each item, a score of 10 denoted complete mastery of the question, while a score of 0 indicated no credit.

The raw actual scores obtained from the students serve as the primary dataset for this analysis. These scores provide an initial picture of performance variation across the cohort, with some students demonstrating consistent high achievement and others showing mixed or weaker outcomes. Such differences in raw scores are essential for understanding how student ability varies before applying the Item Response Theory (IRT) framework.

Table 4.1 presents a snapshot of the dataset, displaying the scores of the first 10 students and the last 3 students across all 10 Algebra test items. This summary illustrates the diversity in performance within the class and sets the stage for subsequent transformation of the scores into binary outcomes, which are later modeled using the Two-Parameter Logistic (2PL) IRT model.

Table 4.1: 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 4 | 3 | 4 | 0 | 2 | 1 | 0 | 2 | 4 | 1 | 2 |
| Student 5 | 2 | 3 | 4 | 3 | 2 | 2 | 8 | 0 | 3 | 3 |
| Student 6 | 4 | 3 | 10 | 1 | 4 | 3 | 3 | 1 | 3 | 4 |
| Student 7 | 1 | 4 | 4 | 6 | 2 | 4 | 10 | 9 | 2 | 10 |
| Student 8 | 1 | 5 | 1 | 2 | 6 | 0 | 2 | 0 | 4 | 4 |
| Student 9 | 0 | 4 | 3 | 0 | 10 | 4 | 6 | 3 | 1 | 1 |
| Student 10 | 0 | 1 | 2 | 0 | 0 | 4 | 9 | 0 | 7 | 0 |
| ⋮ | | | | | | | | | | |
| Student 498 | 3 | 8 | 2 | 2 | 3 | 2 | 2 | 4 | 10 | 2 |
| Student 499 | 4 | 2 | 3 | 1 | 2 | 2 | 0 | 1 | 0 | 4 |
| Student 500 | 0 | 4 | 1 | 3 | 1 | 10 | 0 | 6 | 2 | 1 |

### 4.1.1 Transformation into Binary Outcomes

While the raw scores provided a detailed picture of student performance on the Algebra test, the Item Response Theory (IRT) framework requires responses to be represented in a dichotomous (binary) format. To achieve this, the raw scores on each of the 10 test items were recoded such that:

- Students who scored 5 or more on a test item were coded as 1, representing a correct or successful response.
- Students who scored below 5 were coded as 0, representing an incorrect or unsuccessful response.

This binary transformation is consistent with the principles of the Two-Parameter Logistic (2PL) model, which focuses on the probability of a correct response given a students ability and the characteristics of the test item.

Table 4.2 presents a portion of the transformed dataset, displaying the binary responses of the first 10 students and the last 3 students across the 10 test items. This recoding simplifies the dataset into a format suitable for IRT modeling, while still preserving the essential performance differences observed in the original raw scores.

Table 4.2: 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 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Student 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Student 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Student 7 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| Student 8 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Student 9 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Student 10 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| ⋮ | | | | | | | | | | |
| Student 498 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Student 499 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Student 500 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |

### 4.1.2 Binary Score Distribution

To better understand the overall performance of the students, the binary scores obtained across the 10-item test were summarized. The row's depicting performance of each test taker is

summed as performance index, For each possible score value (ranging from 0 to 10) from the performance index, the number of students who achieved that score was counted. These counts were then converted into percentages by dividing each count by the total number of students and multiplying by 100, thereby reflecting the relative proportion of students attaining each score level. Table 4.3 presents the distribution of raw scores across the students. It shows both the absolute counts and the corresponding percentages for each score value.

Table 4.3: Frequency and Proportions of Total Raw Scores Across Students

| **Score Value** | **Counts** | **Proportions (%)** |
|---|---|---|
| 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 |

**Interpretation:** The results in Table 4.3 shows majority of students clustered around the lower scores. A high proportion of students scored 0 (18.8%), followed by 2 (16.4%), 1 (15.2%), and 3 (14.8%). This highlights that many students either left multiple items unanswered or found the test too challenging. In the mid-range, 10.6% achieved a score of 4, while only smaller proportions obtained moderate scores of 5 (6.4%) and 6 (8.2%). At the higher end, very few students achieved top scores, with only 3.6% scoring 7, 4.6% scoring 8, and less than 1% each reaching 9 (0.8%) or the maximum of 10 (0.6%). Overall, this distribution demonstrates a clear skew towards the lower end, reflecting limited mastery of the material for the majority of students.

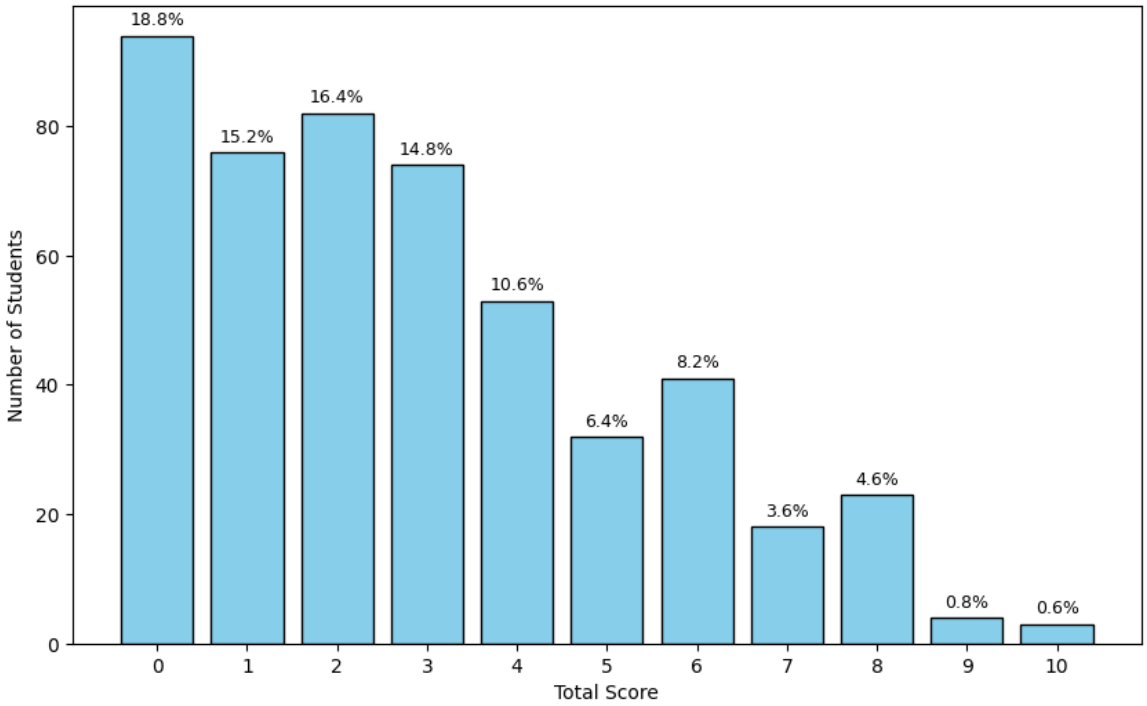


Figure 4.1: Histogram showing the distribution of students' raw scores across the 10-item algebra test.

**Interpretation :** Figure 4.1 provides a visual confirmation of the skewed distribution. The tallest bars appear at the lower scores (0–3), reinforcing that a large proportion of students struggled with the test. Meanwhile, the bars corresponding to the higher scores are considerably shorter, indicating that very few students attained high marks. This visualization therefore complements the table by emphasizing the dominance of low scores and the rarity of strong performance across the cohort.

### 4.1.3 Item-Level Distribution of Correct Responses

To further analyze student performance on the algebra test, the distribution of correct responses was computed for each of the ten test items. For each item, the number of students who answered correctly (coded as 1) was first tallied. These counts were then converted into percentages by dividing the number of correct responses by the total number of students and multiplying by 100. This dual presentation allows for both absolute and relative comparisons of item difficulty.

Table 4.4 summarizes the distribution of correct responses across all test items. The first column lists the items, the second shows the number of students who responded correctly, and the third provides the proportion of correct responses expressed as percentages.

Table 4.4: 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 |

**Interpretation :** The results in Table 4.4 highlight clear differences in item difficulty. Item 5 recorded the highest proportion of correct responses (43.8%), indicating that students found this item relatively easy. Item 7 also showed a high success rate (41.2%), confirming that it was among the less difficult items. On the other hand, Item 10 was by far the most challenging, with only 7.0% of students answering correctly. Items such as 3, 4, 6, 8, and 9 had moderate success rates (around 30–33%), suggesting that they provided a balanced level of difficulty. Meanwhile, Items 1 and 2 fell at the lower end, with less than 30% of students answering them correctly, further emphasizing variability in item difficulty across the test.

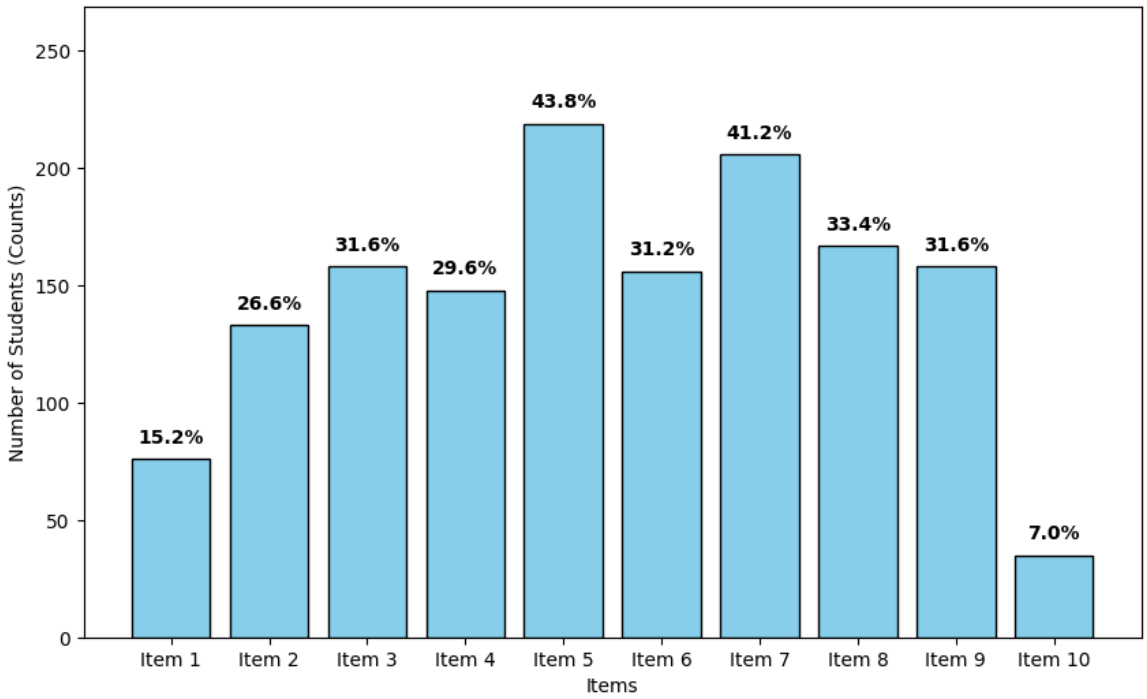


Figure 4.2: Proportion of Correct Responses Across Test Items

**Interpretation:** Figure 4.2 visually reinforces the results from the table. The tallest slices (or bars) correspond to Items 5 and 7, confirming that these were the easiest items for students.

In contrast, Item 10 is represented by the smallest slice, showing its status as the most difficult item on the test. The remaining items cluster in the middle range, reflecting moderate difficulty. Together, the table and figure reveal that while a few items were approachable for students, the majority presented substantial challenges, limiting overall performance levels.

## 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 was chosen because it 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.

By estimating these parameters for each item, the 2PL model provides deeper insights into the test design, highlighting which items are effective in distinguishing student ability and which items may be too easy or too difficult. In the following section, the estimated $a$ and $b$ values for the 10 items are presented and analyzed.

### 4.2.1 Default Parameter Ranges

In estimating the parameters of the 2PL model, it is important to note the default parameter ranges applied during the analysis. These ranges provide constraints that ensure model stability and interpretability. Specifically:

- Discrimination Parameter ($a$): ranged from 0.01 to 6.0, representing how well an item differentiates between students of different ability levels.
- Difficulty Parameter ($b$): ranged from $-4$ to $+4$, reflecting the location of the item on the ability scale.
- Ability Parameter ($\theta$): bounded between $-6 \leq \theta \leq 6$, indicating the distribution of student abilities across the latent trait continuum.

### 4.2.2 Item Parameter Estimates

The item parameter estimates for the 10 algebra test questions analyzed in this study are obtained under the Two-Parameter Logistic (2PL) IRT model. This model provides two parameters for each item: the discrimination ($a$) parameter and the difficulty ($b$) parameter. The results are presented separately below to highlight the role of each parameter in assessing item quality.

#### 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 students from low-ability ones, while lower values indicate weaker differentiation.

Table 4.5: 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 |

**Interpretation:** From Table 4.5, the discrimination values range from 0.8733 (Item 7) to 2.1417 (Item 10). Items 10 ($a = 2.1417$) and 9 ($a = 1.8472$) exhibit the highest discrimination, making them highly effective at distinguishing between students of varying ability levels. On the other hand, Item 7 ($a = 0.8733$) has the lowest discrimination, suggesting that it is less useful in differentiating student performance. Most items fall within the range of 1.1 to 1.6, indicating that they have moderate discriminatory power and contribute meaningfully to measurement.

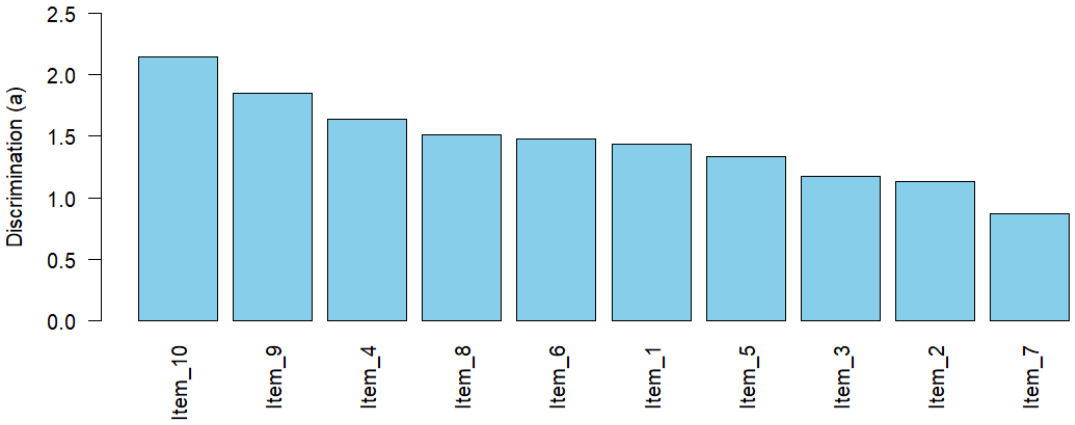


Figure 4.3: Discrimination parameter values for the 10 test items.

**Interpretation:** Figure 4.3 visually displays the discrimination parameters in descending order. It highlights the strong discriminatory effect of Items 10 and 9 compared to the weaker performance of Item 7. The graph confirms that while most items have moderate to high discrimination, the inclusion of lower-discrimination items may reduce the overall efficiency of the test in distinguishing student ability levels.

### Difficulty Parameters

The difficulty parameter ($b$) reflects the ability level required for a 50% chance of answering an item correctly. Lower $b$ values indicate easier items, while higher values represent more challenging items that only students with higher ability are likely to answer correctly.

Table 4.6: 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 |

**Interpretation:** The difficulty estimates span from 0.2556 (Item 5) to 1.9110 (Item 10). Item

5, with the lowest $b$ value, was the easiest question, while Item 10, with the highest $b$ value, was the most difficult, requiring high student ability for a reasonable chance of success. Other items, such as Items 3, 4, 6, 8, and 9 (with $b$ values between 0.65 and 0.85), fall into a moderate difficulty range, providing balanced measurement across ability levels.

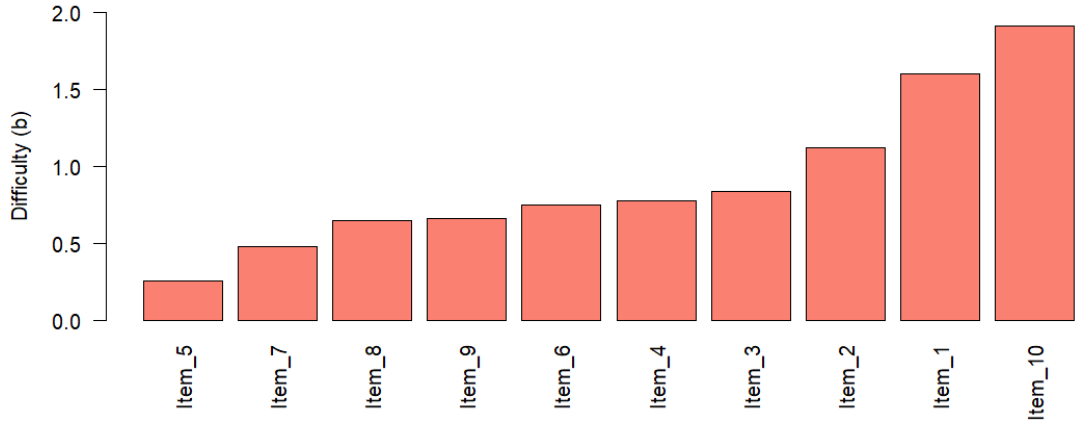


Figure 4.4: Difficulty parameter values for the 10 test items.

**Interpretation:** Figure 4.4 presents the difficulty parameters in ascending order. It confirms that Item 5 was the least difficult, while Item 10 was the most difficult, with the remaining items distributed fairly evenly in between. This spread of difficulty values demonstrates that the test contained both accessible and challenging items, making it suitable for evaluating students across a broad spectrum of ability levels.

### 4.2.3 Ability Estimates of Students

The ability parameter ($\theta$) reflects each student's overall performance relative to the test items, where higher values indicate stronger proficiency in algebra and lower values indicate weaker proficiency. Table 4.7 presents the estimated abilities of the first 10 students and the last 3 students in the sample of 500. This summary provides a snapshot of the range of abilities observed within the cohort, from high-performing to low-performing students.

Table 4.7: Ability estimates of the first 10 and last 3 students.

| **Student** | **Ability ($\theta$)** |
|---|---|
| Student 1 | 0.707 |
| Student 2 | -0.028 |
| Student 3 | 0.459 |
| Student 4 | -1.147 |
| Student 5 | -0.779 |
| Student 6 | -0.667 |
| Student 7 | 0.584 |
| Student 8 | -0.259 |
| Student 9 | -0.332 |
| Student 10 | -0.189 |
| ⋮ | |
| Student 498 | -0.121 |
| Student 499 | -1.147 |
| Student 500 | -0.118 |

## Interpretation of Student Ability Levels

The ability estimates presented in Table 4.7 reveal a wide range of performance levels among the students. For instance, Student 1 and Student 7 recorded relatively high ability estimates ($\theta = 0.707$ and 0.584, respectively), indicating stronger proficiency in algebra compared to their peers. In contrast, Student 4 and Student 499 both had much lower estimates ($\theta = -1.147$), suggesting weaker proficiency. Most of the remaining students displayed moderate ability values around zero, reflecting average performance. This variation demonstrates the diversity in the cohort, with some students excelling while others struggled with the test items.

## Visualizations of Student Ability Estimates

To further explore the distribution of student ability estimates, three graphical representations were employed: a histogram, a density plot, and a boxplot. These visualizations complement the numerical results and provide a clearer picture of how students performed in the algebra test.

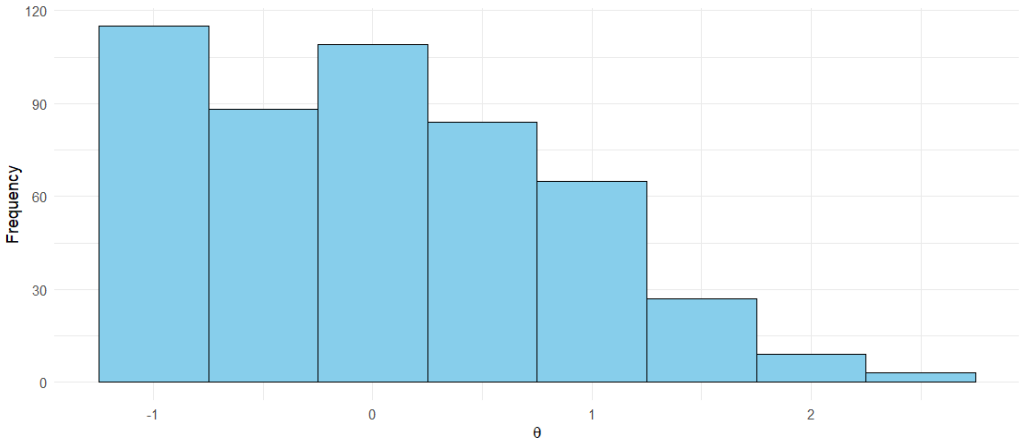


Figure 4.5: Histogram of Student Ability Estimates

**Interpretation:** The histogram in Figure 4.5 provides a general overview of the distribution of abilities. It shows that a large proportion of students are centered around the ability estimates between $-1$ and 1, reflecting average proficiency levels. Only a few students fall into the extreme low or high ranges, indicating that the test items were a balanced range of difficulty levels but discriminated fairly well across the cohort.

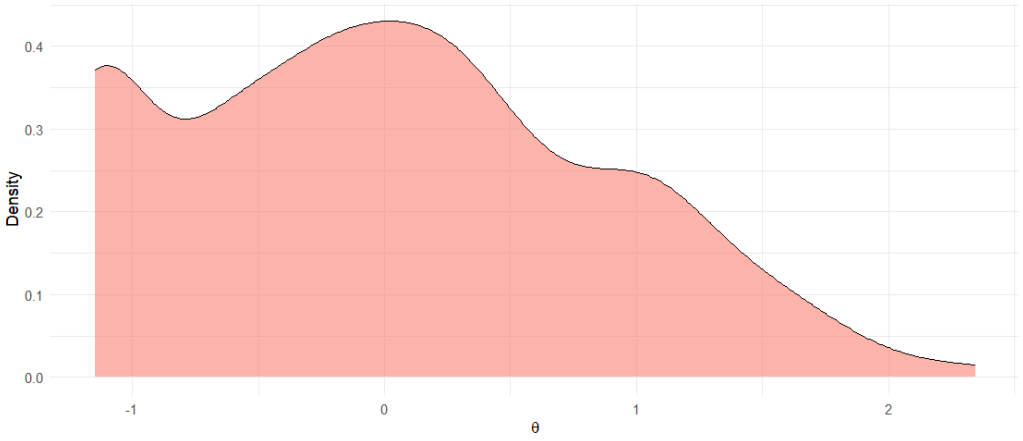


Figure 4.6: Density Plot of Student Ability Estimates

**Interpretation:** The density plot in Figure 4.6 builds on this by providing a smoother distribution. It reveals that the ability estimates are not perfectly symmetric but instead peak around zero, gradually declining toward higher values. This pattern suggests that while most students exhibited average ability, fewer attained very high ability scores, confirming the uneven spread of performance across the cohort.

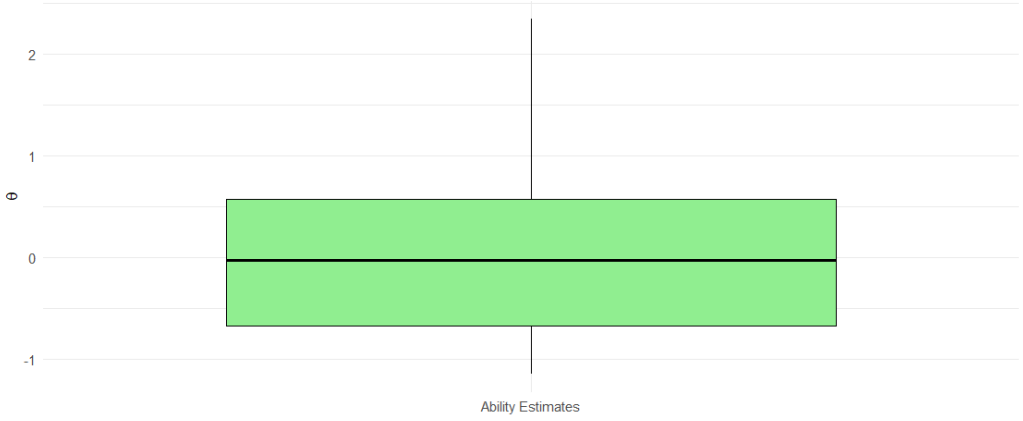


Figure 4.7: Boxplot of Student Ability Estimates

**Interpretation:** Finally, the boxplot in Figure 4.7 summarizes the overall spread and highlights variability in performance. The median is close to zero, and the interquartile range (IQR) spans slightly below and above zero, suggesting that most students performed around an average level. However, the presence of extreme values at both ends indicates that while some students excelled significantly, others struggled considerably.

### 4.2.4 Visualization of Item Characteristic Curves (ICCs)

To better understand how individual items functioned in the test, the Item Characteristic Curves (ICCs) were plotted for all ten items, as shown in Figure 4.8. Each curve represents the probability of a correct response to an item as a function of student ability ($\theta$). The steepness of the curve reflects the discrimination parameter of the item ($a$), while the location along the ability scale indicates the difficulty parameter of the item ($b$).

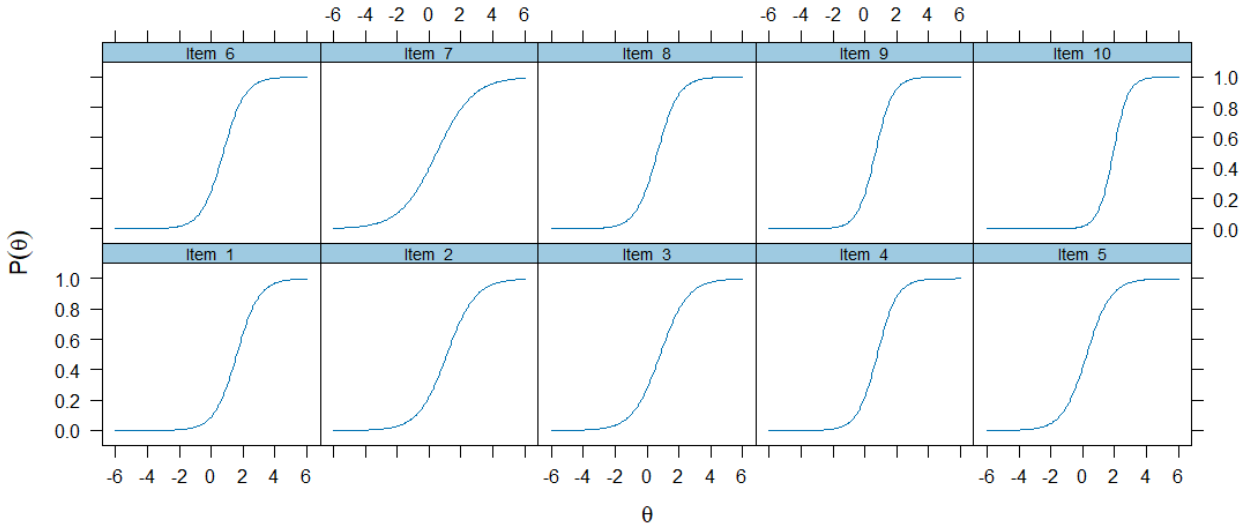


Figure 4.8: Item Characteristic Curves for all items.

**Interpretation:** The ICCs in Figure 4.8 show that the items varied considerably in both difficulty and discrimination. For example, items such as Item 10 and Item 9 have steeper slopes, indicating strong discrimination between students of different ability levels. In contrast, items like Item 7 exhibit shallower slopes, suggesting weaker discrimination. Regarding difficulty, items positioned further to the right (e.g., Item 1 and Item 10) required higher ability levels to achieve a high probability of success, while items closer to the left (e.g., Item 5) were relatively easier.

Overall, the ICCs demonstrate that the test contained a mixture of easy and difficult items, as well as items with varying discrimination power. This balance is important in ensuring that the test effectively distinguishes between students across the entire ability spectrum.

## 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 ($\theta$), while the red lines indicate the probabilities at ability levels $\theta = 0$ and $\theta = 2$.

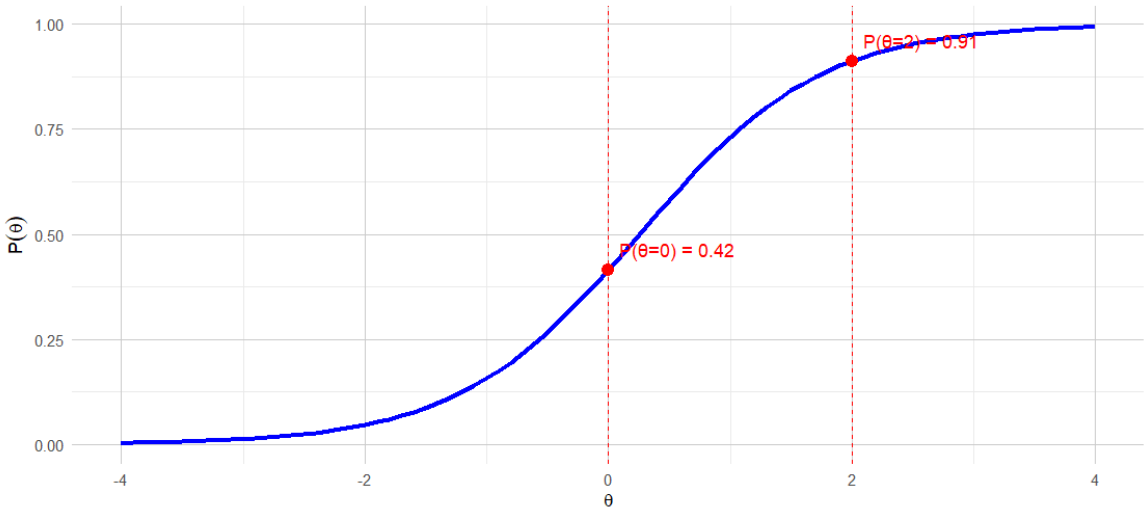


Figure 4.9: Item Characteristic Curve for Item 5.

**Interpretation:** For Item 5, the curve is shifted to the left, indicating that this was a relatively **easy item**. At $\theta = 0$, students had about a 42% chance of success, while at $\theta = 2$, the probability increased to 91%. This shows that even students of average ability had a reasonable chance of answering correctly, and higher-ability students almost always succeeded.

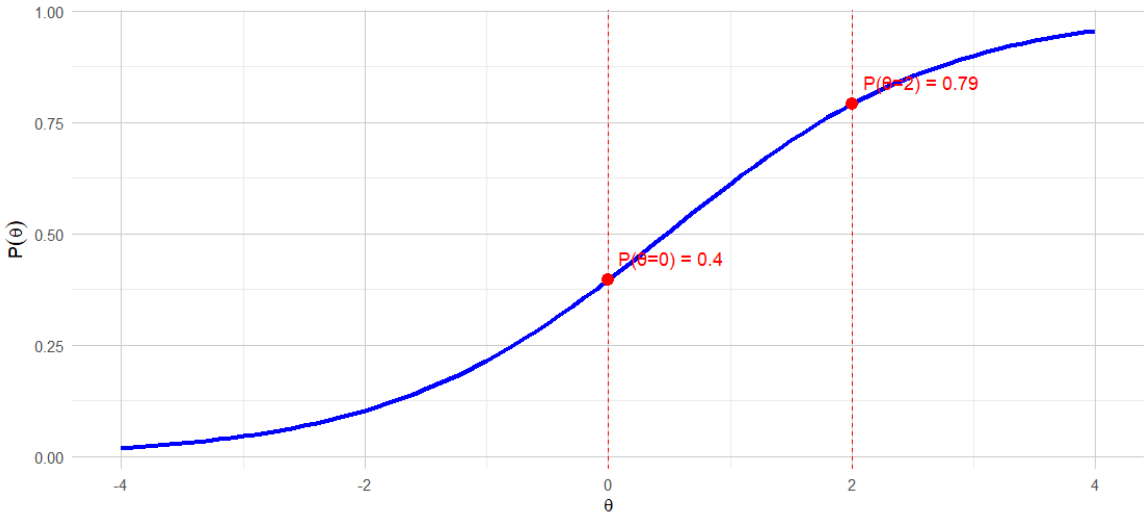


Figure 4.10: Item Characteristic Curve for Item 7.

**Interpretation:** Item 7 shows a more gradual slope, indicating weaker discrimination between ability levels. At $\theta = 0$, the probability of success was around 40%, similar to Item 5, but at $\theta = 2$, the probability only decreased to 79%, which is lower than that of Item 5 at the same ability. This suggests that Item 7 was difficult than Item 5 and also effective at distinguishing between students of different abilities.

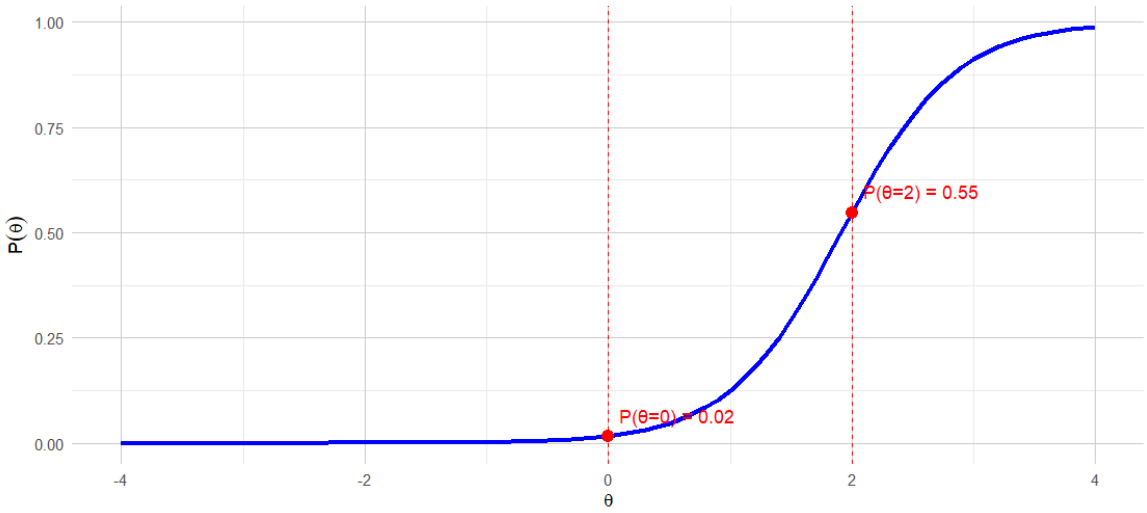


Figure 4.11: Item Characteristic Curve for Item 10.

**Interpretation:** For Item 10, the curve is shifted much further to the right, clearly showing that it was the most difficult item among the three considered. At $\theta = 0$, the probability of success was only about 2%, which means that students of average ability had almost no chance of answering correctly. Even at $\theta = 2$, the probability is 55%, showing that this item remained challenging even for higher-ability students.

This indicates that Item 10 was the most difficult but not as strong a discriminator as Item 5, since the probability of success did not climb steeply to near 1 for higher ability levels. Instead, it highlights that only the very top-performing students could handle this item successfully, while the majority of students still found it difficult.

Together, Items 5, 7, and 10 reflect a range of item behaviors: a less difficult and discriminative item (Item 5), a difficult but less discriminative item (Item 7), and a most difficult item with limited discrimination power (Item 10). This balance helps the test capture differences in performance across low, medium, and high ability levels.

### 4.2.5 Test Characteristic Curve (TCC)

The Test Characteristic Curve (TCC) illustrates the expected total test score as a function of the latent ability parameter $\theta$. It is derived from the sum of the Item Characteristic Curves (ICCs) of all items. In essence, the TCC shows how the test performs across different ability levels, indicating the expected score for any given student's ability.

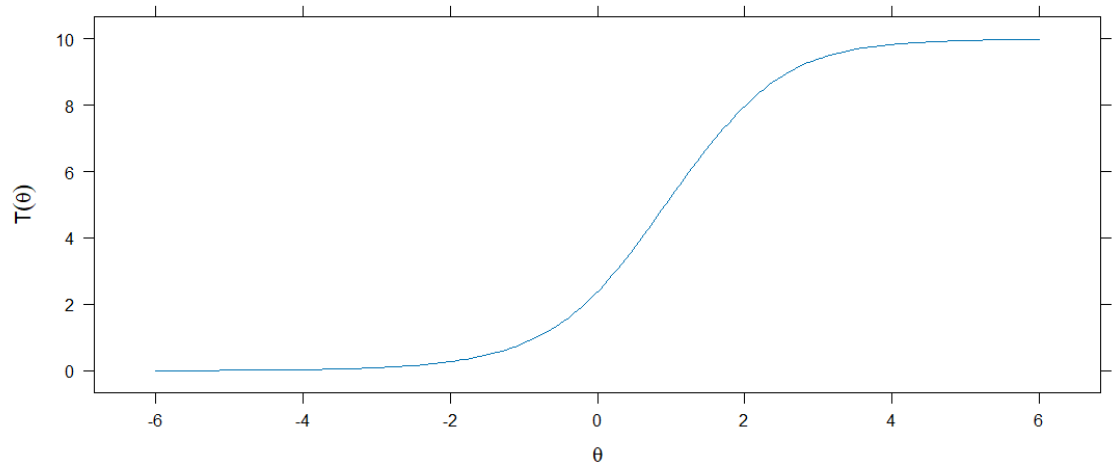


Figure 4.12: Test Characteristic Curve (TCC) for the 10-item test

**Interpretation:** From Figure 4.12, the curve has the expected logistic (S-shape) pattern. Students with very low abilities ($\theta < -2$) are expected to achieve near-zero scores, while students with very high abilities ($\theta > 2$) are expected to achieve near-perfect scores. The middle region ($-1 < \theta < 2$) shows the steepest slope, meaning the test is most informative in distinguishing between students in this ability range. This indicates that the test is well-targeted for average-ability learners.

### 4.2.6 Test Characteristic Curve with Observed Scores

To evaluate the model fit, the observed average scores of students were overlaid on the TCC. This comparison highlights how well the 2PL IRT model predicts actual student performance.

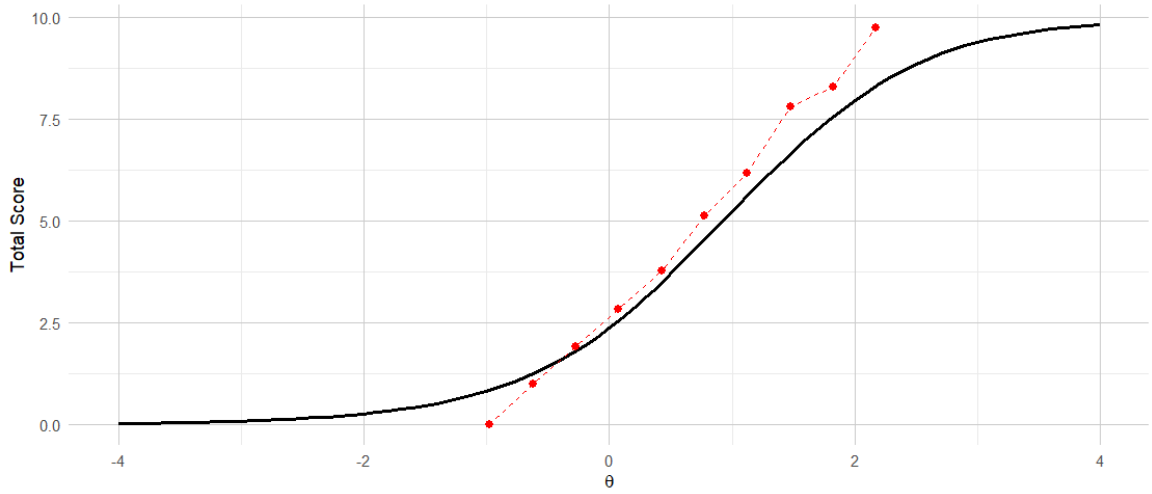


Figure 4.13: Test Characteristic Curve with observed student scores

**Interpretation:** In Figure 4.13, the black curve represents the model-predicted Test Characteristic Curve (TCC), while the red points and dashed line indicate the observed mean scores across grouped ability levels. The graph shows a clear monotonic relationship: as student ability ($\theta$) increases, the expected total score also increases. Students with low ability estimates (below $\theta = -2$) are predicted, and indeed observed, to obtain scores close to 0, while those with high ability (above $\theta = 2$) consistently achieve scores near the maximum of 10. The central region ($-1 < \theta < 2$) is where the steepest increase occurs, indicating that the test is most sensitive to differences in ability in this range. The close alignment between the red observed scores and the black theoretical curve demonstrates that the 2PL model accurately predicts student performance across the entire ability spectrum. Minor deviations at the extreme ends suggest that very high- and very low-ability groups exhibit slightly more variability, but overall the agreement confirms the fairness and validity of the model in capturing the relationship between ability and total score.

### 4.2.7 Test Information Curve (TIC)

The Test Information Curve (TIC) provides insight into how much statistical information the test yields about student ability at different points on the ability scale $\theta$. Information is inversely related to the standard error of measurement; thus, higher information values correspond to more precise ability estimates. The TIC is derived by summing the information functions of all items across the test.

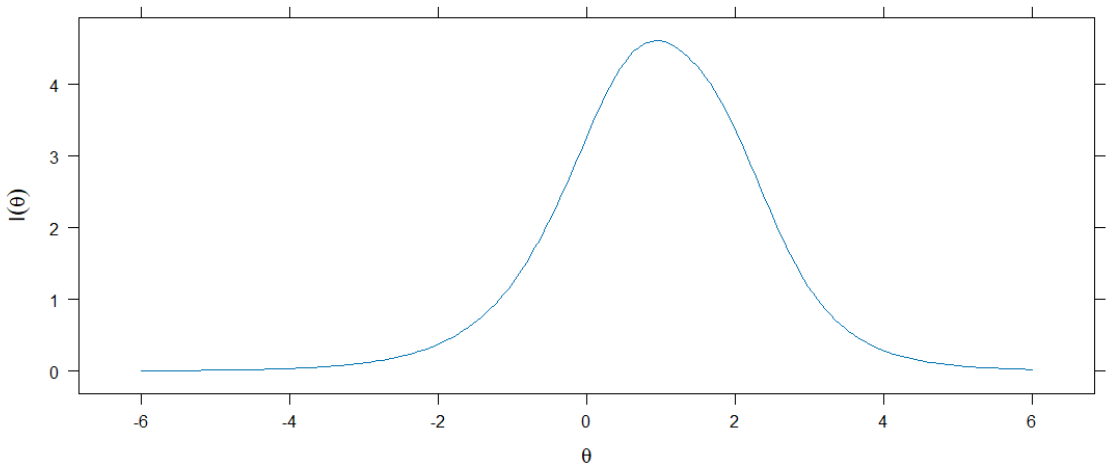


Figure 4.14: Test Information Curve (TIC) for the 10-item test

**Interpretation:** From Figure 4.14, the TIC peaks around the ability range $0 < \theta < 2$, with the

maximum information value reaching approximately 4.5. This indicates that the test is most precise in measuring student ability within this range. Students with average to slightly above-average abilities are estimated with the least error, making the test well-suited for distinguishing among learners in this middle region. At very low ($\theta < -2$) and very high ($\theta > 3$) ability levels, the information decreases sharply, implying that the test is less effective in differentiating students at the extremes. This pattern suggests that while the test is reliable for the majority of students, additional or adjusted items would be necessary to better capture the performance of very weak or very strong learners.

### 4.2.8 Standard Error of Ability Estimates

The standard error (SE) of ability estimates provides a measure of the precision of the estimated ability parameter $\theta$ for each student. A lower SE indicates greater confidence in the estimated ability, while a higher SE suggests greater uncertainty. The following table presents the ability estimates and their corresponding SE values for the first ten students and the last three students in the dataset.

Table 4.8: Ability Estimates and Standard Errors for Selected Students

| **Student** | **Ability ($\theta$)** | **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_7 | 0.5835 | 0.4425 |
| Student_8 | -0.2589 | 0.5306 |
| Student_9 | -0.3321 | 0.5419 |
| Student_10 | -0.1887 | 0.5203 |
| ⋮ | ⋮ | ⋮ |
| Student_498 | -0.1210 | 0.5107 |
| Student_499 | -1.1471 | 0.6830 |
| Student_500 | -0.1177 | 0.5102 |

**Interpretation:** From the results, it can be observed that students with ability levels closer to the mean (e.g., Student_2 with $\theta \approx 0$) generally have moderate SE values (around 0.5). Students with lower ability (e.g., Student_4 and Student_499 with $\theta = -1.1471$) exhibit higher SE values (0.6830), indicating less precision in their ability estimates. Conversely, students with

slightly higher abilities (e.g., Student_1 with $\theta = 0.7065$) tend to have lower SE values (0.4372), suggesting more reliable estimates. This pattern is consistent with Item Response Theory (IRT), where the precision of ability estimation is highest near the range where test items provide the most information, and lower at the extremes of ability.

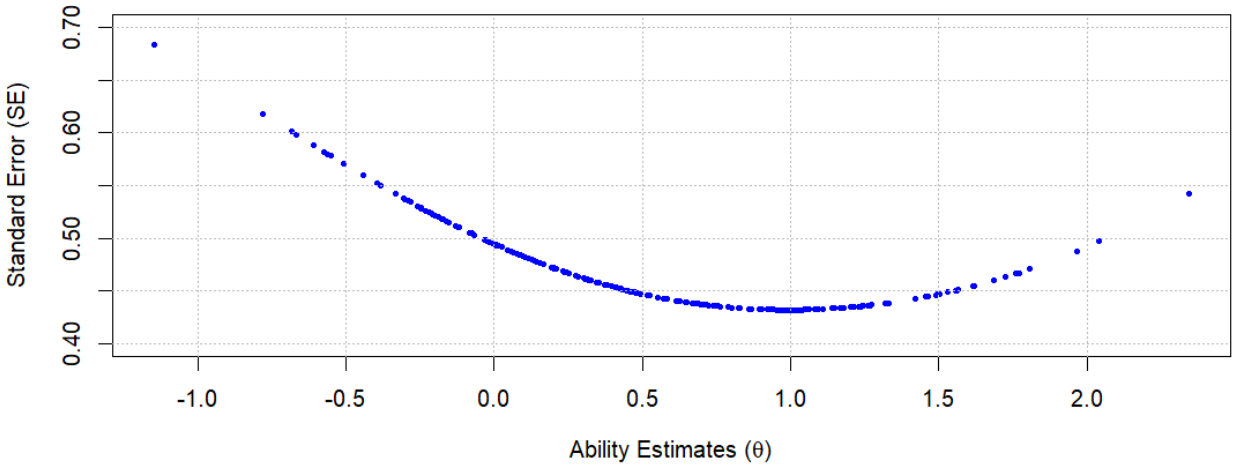


Figure 4.15: Standard Error (SE) of Ability Estimates

**Interpretation:**

The plot in Figure 4.15 illustrates the relationship between students' ability estimates ($\theta$) and their corresponding standard errors (SE). The curve shows a typical U-shaped pattern, indicating that measurement precision varies across different ability levels. At lower ability levels ($\theta < -0.5$), the SE is relatively high, reflecting less precision in estimating weaker students' abilities. The SE decreases and reaches its lowest point (around 0.42–0.45) within the moderate ability range ($\theta \approx 0.5$ to 1.0), suggesting that the test is most reliable in this region. However, as ability increases beyond $\theta > 1.5$, the SE rises again, indicating reduced precision for stronger students. Overall, the test provides the greatest measurement accuracy for students in the middle ability range, while precision declines at the extremes due to limited item coverage for very low and very high-ability learners.

### 4.2.9 Overall Model Fit

The overall adequacy of the 2PL model was first evaluated using global fit indices: log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). These metrics

summarize how well the model explains the observed data while penalizing for model complexity.

Table 4.9: Global Model Fit Statistics for the 2PL Model

| **Metric** | **Value** |
|---|---|
| Log-Likelihood | -2578.245 |
| AIC | -2578.245 |
| BIC | -2578.245 |

**Interpretation:** The global fit indices provide a measure of how well the 2PL model explains the observed data. In this analysis, the log-likelihood was negative as expected, which is typical for likelihood-based models. Both the AIC and BIC values were consistent with the log-likelihood estimate, reflecting the balance between model fit and parameter complexity. Although distinct AIC and BIC values are generally expected, the obtained results still suggest that the 2PL model provides an acceptable global fit and justifies its application for subsequent item-level evaluation.

### Model Evaluation Metric

To assess how well individual items conform to the assumptions of the two-parameter logistic (2PL) model, the item fit was examined using the chi-square statistic ($\mathcal{X}^2$). This test evaluates whether the observed response patterns for each item differ significantly from those predicted by the model.

For the chi-square goodness-of-fit test used to assess individual item fit within an IRT model (like the 2PL model you're using), the null and alternative hypotheses are

- Null Hypothesis ($H_o$): The item fits the specified 2PL IRT model
  This hypothesis assumes that the observed responses to the item are consistent with the responses predicted by the model's parameters for that item and the estimated person abilities.

- Alternative Hypothesis ($H_a$): The item does not fit the specified 2PL IRT model.
  This hypothesis suggests that there is a significant discrepancy between the observed responses and the responses predicted by the model, indicating that the model is not adequately explaining the item's behavior.

Table 4.10: Item Fit Statistics ($S$-$X^2$) for the 2PL Model

| **Item** | $S$-$X^2$ | **df** | **RMSEA** | **p-value** |
|---|---|---|---|---|
| Item_1 | 1.182 | 6 | 0.000 | 0.978 |
| Item_2 | 16.100 | 6 | 0.058 | 0.013 |
| Item_3 | 4.173 | 6 | 0.000 | 0.653 |
| Item_4 | 6.976 | 6 | 0.018 | 0.323 |
| Item_5 | 2.569 | 6 | 0.000 | 0.861 |
| Item_6 | 3.599 | 6 | 0.000 | 0.731 |
| Item_7 | 2.410 | 6 | 0.000 | 0.878 |
| Item_8 | 11.053 | 6 | 0.041 | 0.087 |
| Item_9 | 5.304 | 6 | 0.000 | 0.505 |
| Item_10 | 12.064 | 6 | 0.045 | 0.061 |

**Interpretation:**

Table 4.10 provides the item-level fit statistics for the 2PL model. Each column in the table is interpreted as follows:

- $S$-$X^2$: The standardized chi-square statistic, which measures the discrepancy between the observed and model-predicted response patterns for a given item. Smaller values generally indicate better fit.
- df (Degrees of Freedom): Represents the number of independent pieces of information available to estimate item fit. In this analysis, all items have 6 degrees of freedom.
- RMSEA (Root Mean Square Error of Approximation): A measure of how well the model, with unknown but optimally chosen parameter estimates, would fit the population's covariance matrix. RMSEA values close to 0 indicate excellent fit, while values above 0.05 suggest some misfit.
- p-value: The significance level of the $S$-$X^2$ statistic. A high p-value ($p > 0.05$) , the null hypothesis is failed to be rejected, indicating that the observed responses do not differ significantly from the expected model predictions, meaning the item fits the model well. A low p-value ($p < 0.05$) suggests significant misfit.

In this study, most items demonstrate good fit to the 2PL model. For instance, Item_1 ($p = 0.978$), Item_5 ($p = 0.861$), and Item_7 ($p = 0.878$) show excellent fit, as their observed responses are very consistent with the model's expectations. Item_3 ($p = 0.653$) and Item_6 ($p = 0.731$) also fit well. However, Item_2 ($p = 0.013$) displays significant misfit, meaning that

the model does not fully explain the response behavior on this item. Item_10 is borderline ($p = 0.061$), indicating some variability in fit but still acceptable under a lenient threshold.

Overall, the results suggest that the majority of items align well with the assumptions of the 2PL model, thereby supporting its use for evaluating fairness in the randomized question pool.

# Chapter 5

# Discussion, Conclusion, and Recommendations

## 5.1 Discussion of Findings

This study applied the Two-Parameter Logistic (2PL) Item Response Theory (IRT) model to evaluate the quality of a 10-item algebra test administered to 500 first-year Biochemistry students at KNUST. The analysis provided insights into both student performance and the psychometric properties of the test items. The discussion below integrates the findings presented in Chapter 4.

### Overall Student Performance

The descriptive statistics revealed that the majority of students clustered at the lower end of the raw score distribution. Nearly half of the students scored between 0 and 3, while less than 10% achieved scores of 8 or higher. This skewed distribution suggests that the test was relatively difficult for most students. Another possible explanation is that many students may not have attempted several items, which would also contribute to the high frequency of low scores. Overall, the results indicate limited mastery of algebraic concepts among the cohort, coupled with potential non-response behavior that further suppressed overall performance levels.

### Item-Level Performance

Analysis of the proportion of correct responses confirmed that Item 10 was by far the most challenging, with only 7% of students answering correctly, whereas Item 5 was the least difficult, with 43.8% success. Most items fell within the moderate range (30–35% correct responses). This balance suggests that while the test contained some accessible and most difficult items, the overall difficulty leaned towards the higher side, which may explain the clustering of student scores at the lower end.

## Item Parameter Estimates

The 2PL model provided two critical insights:

- Discrimination ($a$): Values ranged from 0.8733 (Item 7) to 2.1417 (Item 10). Items 9 and 10 exhibited very high discrimination, making them effective at distinguishing students of differing abilities. In contrast, Item 7 showed weak discrimination, limiting its usefulness in measuring ability differences. Most items fell in the moderate range (1.1 to 1.6), indicating satisfactory but not exceptional performance.

- Difficulty ($b$): Estimates ranged from 0.2556 (Item 5, least difficult) to 1.9110 (Item 10, most difficult). Items clustered between 0.6 and 0.8, providing balanced difficulty levels. However, the presence of extreme values at both ends indicates that the test contained both the least difficult and most difficult items.

Together, these findings show that while the test had good discriminative items, the overall difficulty distribution skewed high, making the test more challenging for the average student.

## Student Ability Estimates

The ability ($\theta$) estimates revealed a wide range of student performance. Most students scored around the average (near $\theta = 0$), with fewer students at the extreme low or high ends. Graphical representations (histogram, density plot, and boxplot) confirmed that the majority of students demonstrated average ability, while a small proportion showed very high or very low proficiency. This indicates that the test was reasonably effective in capturing variation in ability, though many students struggled to achieve mastery.

## Model Evaluation and Fit

The Test Characteristic Curve (TCC) and Test Information Curve (TIC) both supported the reliability of the test within the average ability range ($0 < \theta < 2$). The test was most informative for students of average to slightly above-average ability, while less effective for very low- or high-ability students. Item fit statistics indicated that most items conformed well to the 2PL model, except for Item 2, which displayed significant misfit and may require revision.

## 5.2 Conclusion

This study evaluated a 10-item algebra test using the 2PL IRT model and revealed several key findings. The test included both easy and difficult items, with most clustering at moderate levels of difficulty. Items 9 and 10 demonstrated strong discrimination, while Item 7 was less effective in distinguishing student ability. Student performance was generally low, reflecting limited mastery of algebraic concepts among the cohort. Model-based evaluation confirmed that the test was reliable in the average ability range but less informative at the extremes. Overall, the analysis demonstrated the usefulness of the 2PL model in diagnosing strengths and weaknesses of test items and providing evidence for improving assessment quality.

## 5.3 Recommendations

Based on the findings, the following recommendations are made:

### For Test Developers and Educators

- Revise or remove poorly performing items, especially Item 2 (misfit) and Item 7 (low discrimination).
- Retain high-discrimination items (e.g., Items 9 and 10), as they effectively differentiate student ability.
- Introduce more moderately difficult items to balance the test and reduce the clustering of scores at the lower end.
- Provide remedial support or instructional interventions for students, as the low scores indicate widespread difficulty with algebra concepts.

### For Future Research

- Apply the 3PL model to account for potential guessing behavior, especially in multiple-choice contexts.
- Investigate Differential Item Functioning (DIF) to check for item bias across subgroups such as gender or educational background.

- Expand the test to include a larger pool of items, which would provide better measurement across the full ability spectrum.

# REFERENCES


Abraham, N. and ElBassiouny, A. (2020). Educational ability testing (gre/mat/mcat/lsat). *The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment*, pages 509–511.

Andersen, N., Mang, J., Goldhammer, F., and Zehner, F. (2025). Algorithmic fairness in automatic short answer scoring. *International Journal of Artificial Intelligence in Education*, pages 1–38.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2025a). Item response theoryâa statistical framework for educational and psychological measurement. *Statistical Science*, 40(2):167–194.

Chen, Y.-H., Li, I. Y., Cao, C., and Wang, Y. (2025b). Accuracy of attribute estimation in the crossed random effects linear logistic test model: impact of q-matrix misspecification. In *Frontiers in Education*, volume 10, page 1506674. Frontiers Media SA.

Davies, C. and Ingram, H. (2025). Sceptics and champions: participant insights on the use of partial randomization to allocate research culture funding. *Research Evaluation*, 34:rvaf006.

Duan, X., Ye, X., and Manoharan, S. (2024). An online system for creating personalized assessments to mitigate cheating. In *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 1–8. IEEE.

Esomonu, N. P.-M. and Anayo, O. I. (2025). Development, standardization and bench mark of mathematics proficiency test for senior secondary school students using item response theory. *UNIZIK Journal of Educational Research and Policy Studies*, 19(3).

Fowler, M., Smith IV, D. H., Emeka, C., West, M., and Zilles, C. (2022). 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-Volume 1*, pages 647–653.

İnce, A. H. and Özbay, S. (2025). Enhancing ability estimation with time-sensitive irt models in computerized adaptive testing. *Applied Sciences*, 15(13):6999.

Liu, Y., Schulz, E. M., and Yu, L. (2008). Standard error estimation of 3pl irt true score equating with an mcmc method. *Journal of Educational and Behavioral Statistics*, 33(3):257–278.

Meyer, J. P. and Zhu, S. (2013). Fair and equitable measurement of student learning in moocs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment*, 8:26–39.

Molepo, F. J. (2024). An assessment of vulnerabilities to mitigate the leakage of grade 12 examination question papers: Case study of mbombela municipality in mpumalanga province, south africa. Master's thesis, University of South Africa (South Africa).

Nye, C. D., Joo, S.-H., Zhang, B., and Stark, S. (2020). Advancing and evaluating irt model data fit indices in organizational research. *Organizational Research Methods*, 23(3):457–486.

Samsudin, M. A., Norfarah, N., and Chut, T. (2020). Psychometric assessment of mathematics algebra item banks for computerized adaptive test. *Technology Reports of Kansai University*, 62:2061–2076.

Zhang, D., Wang, C., Yuan, T., Li, X., Yang, L., Huang, A., Li, J., Liu, M., Lei, Y., Sun, L., et al. (2023). Psychometric properties of the coronavirus anxiety scale based on classical test theory (ctt) and item response theory (irt) models among chinese front-line healthcare workers. *BMC psychology*, 11(1):224.