Optimizing Pharmacotherapy Exam Items and Assessing Student Proficiency: A Comparative Analysis between Item Response Theory and Classical Test Theory

by

Ping Nie (5473217)

To obtain the degree of Master of Science
in Applied Mathematics: Stochastics specialisation
at the Delft University of Technology,
to be defended publicly on November 24, 2023

Thesis Committee:

Prof. Dr. N.V. Budko, TU Delft, Chair
Prof. Dr. N. Parolya, TU Delft
Dr. T. Preijers, Erasmus MC

PREFACE

I would like to express my gratitude to those who supported me through the completion of this thesis. Their support, guidance, and encouragement have been invaluable throughout this journey.

The idea for this thesis is the motivation to utilize a better method to measure the outcome of the exams. The thesis is organized into four chapters. In Chapter 1, I provide an introduction to the background and context of the research. In Chapter 2, I introduce the mathematical framework. Chapter 3 exhibits the results of implementation. Chapter 4 concludes the thesis by summarizing key findings and suggesting avenues for future research.

Ping Nie
November 2023

ABSTRACT

This paper primarily employs Item Response Theory (IRT) to estimate item characteristics and the proficiency levels of students as reflected in the exam results. The process includes the application of algorithms for item characteristic parameter estimation and the utilization of statistical techniques for item selection from an item pool. Additionally, Classical Test Theory (CTT) is also used to gain insights into the item characteristics. The mathematical frameworks behind every algorithm and model will be introduced in detail. Our ultimate objective is to create an item bank that unifies all items from various exam versions onto a common scale. The technique to put items from different versions on the same scale is the test equating technique, and it will also be described in detail.

After a comprehensive analysis to determine the appropriate model within the framework of Item Response Theory (IRT), the Rasch model has been selected for all versions of the exams. One item from version 9 has been removed from the item pool due to its unsatisfactory item fit. Based on the Item-Map plot, the conclusion can be made that all three exam versions appear to be relatively easy for students to answer, as evidenced by the high bar of the exams. Ultimately, the item bank has been successfully established through the application of two test-equating methods, with results indicating its reliability.

TABLE OF CONTENTS

PREFACE ii

ABSTRACT iii

CHAPTER

1 Introduction **1**

2 Methods **3**

2.1 Item Response Theory 3
2.1.1 Linear Mixed Model 3
2.1.2 IRT Models 4
2.1.3 Local Independence 6
2.1.4 Parameter Estimation 7
2.1.5 Ability Estimation 11
2.1.6 Item Fit 12
2.1.7 Test Equating 13
2.2 Classical Test Theory 16
2.2.1 Reliability Coefficient 17
2.2.2 Item Analysis 17

3 Results **19**

3.1 Data 19
3.2 Classical Test Theory 21
3.3 Item Response Theory 26
3.3.1 Parameter Estimation 26
3.3.2 Local Dependency Detection 29
3.3.3 Item Fit 34
3.3.4 Person Fit 41
3.3.5 Item-Person Map 42
3.3.6 Item Characteristic Curve 42
3.4 Test Equating 43
3.4.1 Separate Calibration 43
3.4.2 Concurrent Calibration 50
3.4.3 Comparison 53

4 Discussion & Conclusion **55**

APPENDICES 57

BIBLIOGRAPHY 61

CHAPTER 1

Introduction

An examination or test is a form of educational evaluation designed to measure a student's understanding and proficiency in a particular subject or discipline. There are various forms of questions in an exam that students can take, such as multiple-choice questions, single-choice questions, and subjective questions. The primary goal of using examinations is to provide a standardized way to evaluate students' academic progress and determine their level of proficiency.

The student's proficiency is revealed by their scores on an exam in the most common sense. It is commonly believed that a high score reflects a high level of proficiency, while a low score suggests a lower level of proficiency. This perspective is in line with the principles of Classical Test Theory (CTT). CTT utilizes measures of item characteristics, such as item difficulty and item discrimination, to assess the items. Their values are easy to compute and are dependent upon the distribution of student proficiency within a sample [6]. Modern research focuses on the Item Response Theory (IRT) model, also known as the latent trait model, which has a strong mathematical basis and considers the proficiency/ability of students as the latent trait and measures the relationship between an individual's proficiency/ability and the probability of correctly responding to test items involving the item characteristics. This relationship can be illustrated by a plot called Item Characteristic Curve [11].

In this work, we have datasets from Erasmus Medical Center (Erasmus MC). The data is the scores that students obtained from each item/question during the final pharmacotherapy exam. The questions are all single-choice questions and students will get 1 point for correctly answering the question and 0 points for failing to answer the question. We have three versions of test papers that were administered to students on several dates. And between each two exam papers, there are some common items/questions which means there are some items appearing more than once in those exam papers.

Our main objective is to use the Item Response Theory (IRT) model to estimate item parameters and the ability of students for three versions of exams. Subsequently, we will

utilize statistical analyses to evaluate the model's goodness of fit and detect the violation of assumptions that may require adjustment to enhance the model's accuracy.

Following this, we employ test equating methods to establish a common scale for all three versions of items, utilizing common items as a bridge [1]. This enables us to create an item bank and conduct a comparative analysis of item parameters, helping us distinguish between difficult and easy items among all the items.

Another objective is to conduct an analysis using Classical Test Theory (CTT), the traditional approach, and subsequently, perform a comparative assessment of the results obtained from the Item Response Theory (IRT) analysis. By employing both methodologies, we aim to gain a comprehensive understanding of the assessment's performance, including the measurement of test reliability, item difficulty, and item discrimination. This comparative analysis will provide valuable insights into the advantages and limitations of each approach, helping us make decisions for future research.

However, the Item Response Theory model is more strict with the sample size while the Classical Test Theory is not [3]. Usually for Classical Test Theory to get stable parameter estimates, it can be achieved by a sample size of 100 to 200 [4]. Since our data is only collected from several exams from Erasmus Medical Center. The small sample size may not provide very stable results under the IRT. However, in this thesis, we want to provide the methodology and some suggestions to Erasmus Medical Center for future exam designs and data collection.

The structure of this thesis is as follows. Chapter 2 provides the mathematical framework of the models, the parameter estimation algorithm, and test equating techniques. In Chapter 3, we thoroughly analyze the results derived from employing these methods. The Conclusion and Discussion Chapter offers a comparative analysis of IRT and CTT and provides valuable insights into the future possibilities for the optimization of the exam items within the medical center.

CHAPTER 2

Methods

In this chapter, we are going to introduce the mathematical framework of both the IRT and CTT models, their parameter estimation algorithms, the item fit statistics, and some other techniques that will be used during the implementation process.

2.1 Item Response Theory

Item Response Theory Models are special cases of what is called *generalized linear* or *nonlinear mixed models*. Generalized linear mixed models (GLMMs) or nonlinear mixed models (NLMMs) is an appropriate way to model repeated binary data. Models with mixing of fixed effects (that do not vary over persons) and random effects (that do vary over persons) are called mixed models [28]. From the perspective of IRT, the latent traits (abilities) can be considered as random effects and the item difficulties can be considered as fixed effects.

2.1.1 Linear Mixed Model

In this model, two different symbols are used for the predictors: X for predictors with a fixed effect, and Z for predictors with a random effect.

$$Y_{pi} = \sum_{k=0}^{K} \beta_k X_{ik} + \sum_{j=0}^{J} \theta_{pj} Z_{ij} + \varepsilon_{pi} \tag{2.1}$$

where Y_{pi} is the observed response variable;
k $(k = 0, ..., K)$ is an index for predictors with a fixed effect;
j $(j = 0, ..., J)$ is an index for predictors with a random effect;
X_{ik} is the value of predictor k for item i;
Z_{ij} is the value of predictor j for item i;

β_k is the fixed regression weight of predictor k, an overall intercept for $k = 0$, and predictor-specific effects for $k = 1, \ldots, K$;
θ_{pj} is the random regression weight of predictor j for person p, a person-specific intercept for $j = 0$, and person-specific slopes for $j > 0$;
ε_{pi} is the error term for person p and item i. It is assumed that ε_{pi} has an independent normal distribution with mean 0 , and variance σ_ε^2, the same for all persons and items.

It's obvious that the linear mixed model has a continuous error term that requires continuous outcomes. However, the data in item response models is categorical data, particularly binary data in our case.

A special case of the linear mixed model *random-intercepts model* will be introduced, which can be applied to produce item response theory models

$$U_{pi} = \sum_{k=0}^{K} \beta_k X_{ik} + \theta_{p0} Z_{i0} + \varepsilon_{pi} \tag{2.2}$$

where $(\beta_1, \ldots, \beta_k, \ldots, \beta_K)$ as multiple fixed slopes, β_0 as an overall intercept, and θ_{p0} as the random deviation from the overall intercept.

2.1.2 IRT Models

Since the responses to the items are scored as correct or incorrect, the dichotomous response is denoted by $U_{pi} = \{0, 1\}$ by test takers $p = 1, \ldots, P$ on items $i = 1, \ldots, I$. The distributions of $U_{pi} = 1$ are Bernoulli distributions with the parameter $\pi_{pi} \in [0, 1]$ as the success parameters. Thus, the expected value of this binary response variable is π_{pi}.

There is a continuous variable V_{pi} with its expected value η_{pi}. This V_{pi} follows the linear mixed model formula. Then we use the link function $g(.)$ to link the expected value of the binary response π_{pi} to the expected value of the underlying continuous variable η_{pi}. The most common link function used in Item Response Theory model is the logit function.

$$L(\pi_{pi}) = \eta_{pi} = \log \frac{\pi_{pi}}{1 - \pi_{pi}} \tag{2.3}$$

Rasch Model

First of all, the random intercept $\theta_{p0}\mathrm{Z}_{i0}$ from (2.2) corresponds to the person parameter in the IRT. It is often denoted by θ_p, and called 'ability'.

Then, when the item predictors are dummy variables that are used to identify the items ($X_{ik} = 1$ if $i = k$, and $X_{ik} = 0$ if $i \neq k$), then the fixed effects term $\Sigma_k^K \beta_k X_{ik}$ from (2.2) corresponds to the item parameter. It is denoted by β_i $\left(\Sigma_k^K \beta_k X_{ik} = -\beta_i\right)$, and called 'item

difficulty'.

$$\eta_{pi} = \theta_p - \beta_i \tag{2.4}$$

Using (2.3), we obtain

$$\begin{aligned} \log \frac{\pi_{pi}}{1-\pi_{pi}} &= \theta_p - \beta_i \\ \pi_{pi} &= \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \end{aligned} \tag{2.5}$$

Since π_{pi} is the probability of a test taker answering the item correctly, it can be replaced by $p(Y_{pi} = 1 \mid \beta_i, \theta_p)$.

Therefore, the Rasch Model is

$$p(U_{pi} = 1 \mid \beta_i, \theta_p) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \tag{2.6}$$

From the descriptive point of view, the log odds version of the Rasch model (the first formula in (2.5)) shows that the natural logarithm of the odds ratio of the probability of correct response to the probability of incorrect response is modeled by the difference between person parameter θ_p and the item difficulty β_i.

Two Parameter Logistic Model

The extension for the Rasch model is to include the latent item predictors which is the item discrimination α_i.

$$P(U_{pi} = 1 \mid \alpha_i, \beta_i, \theta_p) = \frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]} \tag{2.7}$$

The item discrimination α_i here is the multiplier of the difference between ability and item difficulty [10]. From the descriptive point of view, it says that the difference between the ability and item difficulty depends on the discriminating power of the item.

Specifically, the impact of the difference between the ability and item discrimination is stronger on the probability when the item discrimination is higher.

The parameters

The α_i is the item discrimination parameter. This parameter reflects how well an item can distinguish individuals with different levels of the latent trait. Therefore, when α_i increases, the item will be more sensitive to individual differences in the latent trait. From the perspective of the Item Characteristics Curve (ICC) which is the plot of the formulas (2.6) or (2.7) to illustrate the relationship between the latent trait and the probability of

answering items correctly, the increased item discrimination will make the ICC steeper [28].

The β_i is the item difficulty parameter. This parameter represents how challenging the item is. Therefore, when β_i increases, the item discrimination parameter tends to be lower since it is less useful for differentiating individuals at lower trait levels. In addition, the value of the item difficulty means that the examinee with the same value of ability will have a 50% possibility of giving the correct response for that item.

GLMM or NLLM

If the link function is the identity function, then the IRT model is still a linear mixed model. Since the logit function is used to generate the IRT model, the IRT model is what we call *Generalised Linear Mixed Model* or *Nonlinear Mixed Model*.

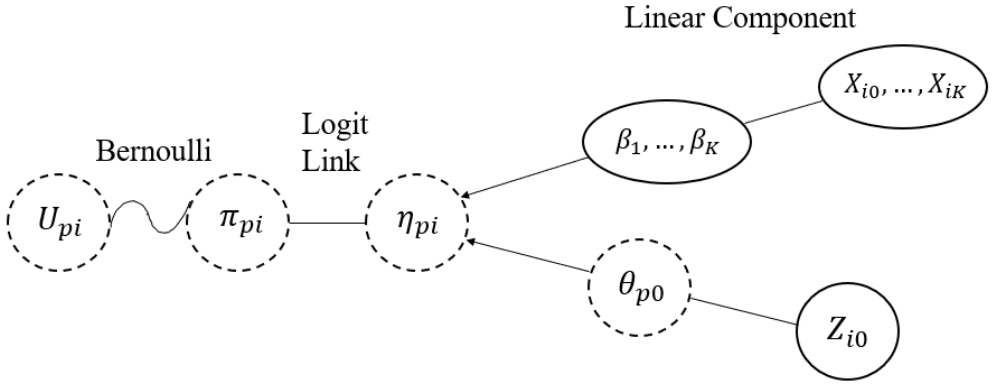

Figure 2.1: Graphical Representation

2.1.3 Local Independence

Local Independence is an important assumption and concept of the IRT model. Local Independence says that item responses only depend on latent traits and not on the responses to other exam items. Thus, exam designers prefer to design the exam paper with items that provide unique information regarding the examinee's skills or knowledge.

Detecting Local Item Dependency

The Q3 statistic is a residual-based statistic. It measures the correlations of the residuals between any two of the items in an exam [22].

For any item response logistic model, if the local independence is met, then

$$P_{ij}(\theta_p) = P_i(\theta_p) P_j(\theta_p) \tag{2.8}$$

The procedure for calculating Q_3 is to first remove the nonlinear effects of $\hat{\theta}$ from the item scores. Define

$$d_{ip} = u_{ip} - \hat{p}_i\left(\hat{\theta}_p\right) \tag{2.9}$$

where u_{ip} is the score of the pth examinee on the ith item; $\hat{p}_i\left(\hat{\theta}_p\right)$ is the observed probability of examinees that correctly answer the items.

and

$$Q_{3,ij} = r_{d_i d_j} \tag{2.10}$$

which is the correlation (taken over all the test takers) of these scores. In our case, the Q_3 statistic for the Rasch model, any pair of items with residuals over 0.20 may violate LID.

2.1.4 Parameter Estimation

In typical IRT applications, both item parameters and latent trait levels are unknown and must be estimated from the same dataset. The marginal maximum likelihood (MML) method is the most popular method that is widely used nowadays.

Marginal Maximum Likelihood

Under this method, the latent traits are handled by expressing the response pattern probabilities. First, the likelihood of a response pattern of person p to I items $\left(X_{1_p}, ..., X_{I_p}\right)$ conditional on the latent trait θ_p and the vector of item difficulty of every item $\underline{\beta}$, is the product of their response probabilities (local independence).

$$P\left(X_{1_p}, \ldots, X_{i_p} \mid \theta_p, \underline{\beta}\right) = \Pi_i {P_{i_p}}^{X_{i_p}} {Q_{i_p}}^{1-X_{i_p}} \tag{2.11}$$

In the following, the $P\left(X_{1_p}, \ldots, X_{i_p} \mid \theta_p, \underline{\beta}\right)$ will be denoted as $P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right)$ due to notational convenience.

For each response pattern $\underline{X}_p$, there is a unique probability $P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right)$. The persons who produce the pattern are regarded as replicates. The number of persons with this pattern is denoted by n_p. The latent trait levels θ_p are unknown for the persons who are observed, but we can still specify the probability of this latent trait by $P(\theta_p)$. The probability may be known in advance if the assumption of the distribution of latent traits is made.

Assume that the probability can be generated from a population distribution, and the distribution of latent traits is a standard normal distribution. Then the marginal probability of the response pattern is:

$$P\left(\underline{X}_p \mid \underline{\beta}\right) = \Sigma_p^n P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right) P\left(\theta_p\right) \tag{2.12}$$

The $P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right)$ is computed from the IRT model, as in (2.11).

For a continuous variable, the expected value for a response pattern is to integrate across the range of latent traits. (2.12) becomes

$$P\left(\underline{X}_p\right) = \int P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right) g(\theta) d\theta \tag{2.13}$$

A posterior distribution of θ_p given $\underline{X}_p$ can also be computed. This posterior distribution combines the information from the (assumed) distribution of ability and the likelihood function based on the observed responses. It can be computed using Bayes theorem [15]:

$$P\left(\theta_p \mid \underline{X}_p, \underline{\beta}\right) = \frac{P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right) g(\theta)}{\int P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right) g(\theta) d\theta} \tag{2.14}$$

The marginal likelihood function and its logarithm are

$$\begin{aligned} \mathcal{L} &= \Pi_p^n P\left(\underline{X}_p\right) \\ \log \mathcal{L} &= \sum_p^n \log P\left(\underline{X}_p\right) \end{aligned} \tag{2.15}$$

For a 2PL model, the vector of item parameters $\underline{\beta} = (\alpha_i, \beta_i)$. And the marginal likelihood equation for α_i,

$$\frac{\partial}{\partial \alpha_i}(\log \mathcal{L}) = 0 \tag{2.16}$$

Then

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_i} \log \mathcal{L} &= \sum_p^n \frac{\partial}{\partial \alpha_i} \left(\log P\left(\underline{X}_p\right)\right) \\
&= \sum_p^n \left[P\left(\underline{X}_p\right)\right]^{-1} \int \frac{\partial}{\partial \alpha_i} \left[P\left(\underline{X}_p \mid \theta_p, \underline{\beta}\right)\right] g(\theta) d\theta \\
&= \sum_p^n \left[P\left(\underline{X}_p\right)\right]^{-1} \int \frac{\partial}{\partial \alpha_i} \left[\log P\left(\underline{X}_p \mid \theta, \underline{\beta}\right)\right] P\left(\underline{X}_p \mid \theta, \underline{\beta}\right) g(\theta) d\theta \\
&= \sum_p^n \int \frac{\partial}{\partial \alpha_i} \left[\log P\left(\underline{X}_p \mid \theta, \underline{\beta}\right)\right] \left[\frac{P\left(\underline{X}_p \mid \theta, \underline{\beta}\right) g(\theta)}{P\left(\underline{X}_p\right)}\right] d\theta \\
&\overset{\text{(i)}}{=} \sum_p^n \int \frac{\partial}{\partial \alpha_i} \left[\log P\left(\underline{X}_p \mid \theta, \underline{\beta}\right)\right] \left[P\left(\theta \mid \underline{X}_p, \underline{\beta}\right)\right] d\theta \\
&\overset{\text{(ii)}}{=} \sum_p^n \int \frac{\partial}{\partial \alpha_i} \left[\log \Pi_i {P_{i_p}}^{X_{i_p}} {Q_{i_p}}^{1-X_{i_p}}\right] \left[P\left(\theta_p \mid \underline{X}_p, \underline{\beta}\right)\right] d\theta
\end{aligned}
\tag{2.17}
$$

where (i) follows the posterior distribution equation (2.14); (ii) follows the equation (2.11).

Then, we substitute the probability P_{i_p} and Q_{i_p} with (2.7). After some deduction, the equation is as follows:

$$
\frac{\partial}{\partial \alpha_i} \log \mathcal{L} = \sum_p^n \int \left[X_{ip} - P_i\left(\theta_p\right)\right] \left(\theta_p - \beta_i\right) \left[P\left(\theta_p \mid \underline{X}_p, \underline{\beta}\right)\right] d\theta = 0 \tag{2.18}
$$

Notice that the integral is difficult to evaluate. Thus, a method called the quadrature approximate approach can be employed here to approximate such integrals. As the distribution of latent traits is assumed to be the standard normal distribution, the Gaussian quadrature method is used here.

Gaussian quadrature is analogous to dividing a normal distribution into segments, with a representative value and a probability of occurrence.

The midpoint of each rectangle on the ability scale, $Y_k (k = 1, 2, \ldots, q)$, is called a "node". Each node has an associated weight $A\left(Y_k\right)$ which takes into account the height of the density function $g(\theta)$ in the neighborhood of Y_k and the width of the rectangles.

Returning to equation (2.18), this equation can be rewritten in the form of Gaussian quadrature:

$$
\alpha_i : \sum_k^q \sum_p^n \left[X_{ip} - P_i\left(Y_k\right)\right] \left(Y_k - \beta_i\right) \left[P\left(Y_k \mid \underline{X}_p, \underline{\beta}\right)\right] = 0. \tag{2.19}
$$

$$\beta_i : (-\alpha_i) \sum_k^q \sum_p^n \left[X_{ip} - P_i \left(Y_k \right) \right] \left[P \left(Y_k \mid \underline{X}_p, \underline{\beta} \right) \right] = 0 \tag{2.20}$$

Next, we begin defining the following two quantities:

$$\bar{n}_{ik} = \sum_p^n P \left(Y_k \mid \underline{X}_p, \underline{\beta} \right) = \sum_p^n \left[\frac{\prod_i^I P_i \left(Y_k \right)^{X_{ip}} Q_i \left(Y_k \right)^{1-X_{ip}} A \left(Y_k \right)}{\sum_k^q \prod_i^I P_i \left(Y_k \right)^{X_{ip}} Q_i \left(Y_k \right)^{1-X_{ip}} A \left(Y_k \right)} \right] \tag{2.21}$$

$$\bar{r}_{ik} = \sum_p^n X_{ip} P \left(Y_k \mid \underline{X}_p, \underline{\beta} \right) = \sum_p^n \left[\frac{\prod_i^I X_{ip} P_i \left(Y_k \right)^{X_{ip}} Q_i \left(Y_k \right)^{1-X_{ip}} A \left(Y_k \right)}{\sum_k^q \prod_i^I P_i \left(Y_k \right)^{X_{ip}} Q_i \left(Y_k \right)^{1-X_{ip}} A \left(Y_k \right)} \right] \tag{2.22}$$

where $\bar{n}_{ik}$ is the expected number of examinees at ability level Y_k and $\bar{r}_{ik}$ is the expected number of correct responses to item i at ability level Y_k.

Since,

$$L \left(X_k \right) = \prod_i^I P_i \left(Y_k \right)^{X_{ip}} Q_i \left(Y_k \right)^{1-X_{ip}} \tag{2.23}$$

Thus, the equation (2.14) under the quadrature form is

$$P \left(Y_k \mid \underline{X}_p, \underline{\beta} \right) = \frac{L \left(Y_k \right) A \left(Y_k \right)}{\sum_k^q L \left(Y_k \right) A \left(Y_k \right)} \tag{2.24}$$

Then, the equations (2.21) and (2.22) become

$$\bar{n}_{ik} = \sum_p^n P \left(Y_k \mid \underline{X}_p, \underline{\beta} \right) = \sum_p^n \left[\frac{L \left(Y_k \right) A \left(Y_k \right)}{\sum_k^q L \left(Y_k \right) A \left(Y_k \right)} \right] \tag{2.25}$$

$$\bar{r}_{ik} = \sum_p^n X_{ip} P \left(Y_k \mid \underline{X}_p, \underline{\beta} \right) = \sum_p^n \left[\frac{X_{ip} L \left(Y_k \right) A \left(Y_k \right)}{\sum_k^q L \left(Y_k \right) A \left(Y_k \right)} \right] \tag{2.26}$$

Finally, (2.19) and (2.20) can be rewritten using $\bar{n}_{ik}$ and $\bar{r}_{ik}$ as follows

$$\alpha_i : \sum_k^q \left(Y_k - \beta_i \right) \left[\bar{r}_{ik} - \bar{n}_{ik} P_i \left(Y_k \right) \right] = 0 \tag{2.27}$$

$$\beta_i : \quad (-a_i) \sum_k^q \left[\bar{r}_{ik} - \bar{n}_{ik} P_i \left(Y_k \right) \right] = 0 \tag{2.28}$$

EM Algorithm

The EM algorithm is an iterative procedure for finding maximum likelihood estimates in the presence of unobserved random variables in probability models [29]. E represents the expectation step and M represents the maximization step. Under the IRT setting, we want to find the maximum likelihood estimates with the latent trait. Thus, there are two steps under the EM algorithm [15]. When it is employed for our case:
E-step: Use provisional item parameters to generate $\bar{n}_{ik}$ and $\bar{r}_{ik}$, the expected number of examinees and the expected number of correct responses.
M-step: Use $\bar{n}_{ik}$ and $\bar{r}_{ik}$ to calculate the new item parameters based on the maximum likelihood functions (2.27) and (2.28).

The EM cycles are continued until the estimates become stable to the required number of places.

2.1.5 Ability Estimation

For this ability estimation process, the maximum likelihood procedures are used. Similar to parameter estimation, the procedure is also an iterative process. It begins with an a priori value for the ability of the examinee of which the distribution was assumed in the process of parameter estimation and the known values of the item parameters [11]. We can use the 2PL formula (2.7) to compute the probability of correct response to each item for that examinee.

The ability estimation equation for the 2PL model is shown as follows,

$$\hat{\theta}_{s+1} = \hat{\theta}_s - \frac{\sum_{i=1}^{I} a_i \left[u_i - P_i\left(\hat{\theta}_s\right) \right]}{-\sum_{i=1}^{J} a_i^2 P_i\left(\hat{\theta}_s\right) Q_i\left(\hat{\theta}_s\right)} \tag{2.29}$$

where $\hat{\theta}_s$ is the provisional or estimated ability of the examinee within iteration s, a_i is the item discrimination parameter of item $i, i = 1, 2, \ldots, I$, and $u_i = \{0, 1\}$ is the response made by the examinee to item i; $P_j\left(\hat{\theta}_s\right)$ is the probability of correct response to item j, $Q_j\left(\hat{\theta}_s\right) = 1 - P_j\left(\hat{\theta}_s\right)$ is the probability of incorrect response to item j.

From this equation, we can consider the second term on the right side is the adjustment term. It measures the difference between the examinee's item response to item i and the probability of correct response at $\hat{\theta}_s$. Consequently, the ultimate goal is to find the ability estimation that can let the values of $P_j\left(\hat{\theta}_s\right)$ for all items simultaneously that minimizes the sum. After many times of iteration, the adjustment term become small enough so that the $\hat{\theta}_{s+1}$ won't change, then the $\hat{\theta}_{s+1}$ is the estimated ability.

2.1.6 Item Fit

Item fit statistics are calculated to assess whether the individual item fits the Rasch model [20]. Residual-based fit statistics are widely used to assess Rasch model fit.

Infit Outfit Mean Squares Statistics

First, the standard residuals between the model and responses are

$$Z_{ip} = \frac{X_{ip} - E\left(X_{ip}\right)}{\sqrt{\mathrm{Var}\left(X_{ip}\right)}} \tag{2.30}$$

with the responses $X_{ip} = \{0, 1\}$, $E\left(X_{ip}\right) = P\left(X_{ip} = 1\right) = p_{ip}$ and $\mathrm{Var}\left(X_{ip}\right) = p_{ip}\left(1 - p_{ip}\right)$.

Then, the outfit (unweighted mean square) and infit (weighted mean square) are calculated as means of the squared residuals,

$$U_i = \sum_p Z_{ip}^2/n, \quad V_i = \frac{\sum_p Z_{ip}^2 \cdot w_{ip}}{\sum_p w_{ip}} \tag{2.31}$$

The weights w_{ip} used to calculate the infit statistics are equal to $\mathrm{Var}\left(X_{ip}\right)$.

The outfit is highly influenced by a few outliers (unexpected responses) and the infit is more sensitive to the overall pattern of responses. The mean square fit statistics have a chi-square distribution and an expected value of 1 [25].

To detect misfitting items the fit statistics are either compared to rule-of-thumb critical values or transformed to test statistics which can be compared with the values of the purported distribution. Linacre [17] suggested that both mean squares statistics values between 0.5 and 1.5 are acceptable. Then, the Wilson-Hilferty cube root transformation [21] can be used to improve the approximation of a chi-square variable to the standard normal distribution. The outcome called standardized fit statistics [2] is shown as follows:

$$z\left(U_i\right) = \left(U_i^{\frac{1}{3}} - 1\right)\left(\frac{3}{\sigma_i}\right) + \left(\frac{\sigma_i}{3}\right), \quad z'\left(V_i\right) = \left(V_i^{\frac{1}{3}} - 1\right)\left(\frac{3}{\sigma_i'}\right) + \left(\frac{\sigma_i'}{3}\right), \tag{2.32}$$

where σ_i and σ_i' stand respectively for the standard deviation of U_i and V_i. The formulas of them are explicitly given in [27].

The p-values of the standardized fit statistics are very close to those of a standard normal variable. Thus, it can be interpreted as a classical t-statistic, where a value of 1.96 corresponds to a two-sided significance of 5%.

Please note that this item fit method can also be applied to responses from each person,

then this method can be called person fit.

2.1.7 Test Equating

There are some common items between each two of our exam papers. The examinees who took the Pharmacotherapy exams were from different groups of class or from different grades. In this case, the test equating method that should be applied is the Non-Equivalent Common Item Equating [1].

Non-equivalent Common Item Equating

The parameter estimation that results from non-equivalent groups of examinees is always on different scales. When conducting equating with nonequivalent groups, the parameters from different scales should be put on the same scale.

if an IRT model fits a set of data, then any linear transformation of the θ-scale also fits the set of data, provided that the item parameters also are transformed [1].

Define two scales: scale I and scale J. They are the 2PL logistic IRT scales determined by two non-equivalent groups. When an IRT model is used to fit two different exam versions, the linear transformation $\varphi : \Theta_I \mapsto \Theta_J$ is assumed. A linear equation is used to convert the IRT score:

$$\theta_{Ji} = A\theta_{Ii} + B \tag{2.33}$$

The relations between item parameters on the two exam versions are

$$a_{Jj} = a_{Ij}/A \tag{2.34}$$

$$b_{Jj} = Ab_{Ij} + B \tag{2.35}$$

where A and B are equating coefficients that will be estimated after the separate calibration of two exam versions, θ_{Ji} and θ_{Ii} are values of θ for examinee i on Scale J and Scale I, and where a_{Jj} and b_{Jj} are the item parameters for item j on Scale J and a_{Ij} and b_{Ij} are the item parameters for item j on Scale I.

Appropriateness of Scale Transformations

The 2PL model (2.7) under scale J should be converted to scale I by the linear transformation (2.33)-(2.35). The 2PL model formula is rewritten with scale J

$$\frac{\exp\left[a_{Jj}\left(\theta_{Ji} - b_{Jj}\right)\right]}{1 + \exp\left[a_{Jj}\left(\theta_{Ji} - b_{Jj}\right)\right]} \tag{2.36}$$

Now substitute all the terms with the formula (2.33)-(2.35) as follows:

$$\frac{\exp\left\{\frac{a_{Ij}}{A}\left[A\theta_{Ii}+B-\left(Ab_{Ij}+B\right)\right]\right\}}{1+\exp\left\{\frac{a_{Ij}}{A}\left[A\theta_{Ii}+B-\left(Ab_{Ij}+B\right)\right]\right\}}=\frac{\exp\left[a_{Ij}\left(\theta_{Ii}-b_{Ij}\right)\right]}{1+\exp\left[a_{Ij}\left(\theta_{Ii}-b_{Ij}\right)\right]}. \tag{2.37}$$

Apparently, the resulting expression is the 2PL model formula with the scale I, which indicates the A and B in formula (2.33)-(2.35) provides the scale transformation.

A and B constants

For any two items j and j^*, the A and B constants can be expressed as follows:

$$A=\frac{b_{Jj}-b_{Jj^*}}{b_{Ij}-b_{Ij^*}}=\frac{a_{Ij}}{a_{Jj}} \tag{2.38}$$

and

$$B=b_{Jj}-Ab_{Ij}=\theta_{Ji}-A\theta_{Ii}. \tag{2.39}$$

After the separate calibration from both versions, the groups of item parameters are calibrated and applied to calculate the A and B constants.

$$\begin{aligned} A&=\frac{\sigma\left(b_J\right)}{\sigma\left(b_I\right)},\\ &=\frac{\mu\left(a_I\right)}{\mu(a_J)},\\ &=\frac{\sigma\left(\theta_J\right)}{\sigma\left(\theta_I\right)}, \end{aligned} \tag{2.40}$$

and

$$\begin{aligned} B&=\mu\left(b_J\right)-A\mu\left(b_I\right),\\ &=\mu\left(\theta_J\right)-A\mu\left(\theta_I\right). \end{aligned} \tag{2.41}$$

where $\mu\left(b_J\right),\mu\left(b_I\right),\mu\left(a_I\right)$, and $\mu\left(a_J\right)$ are the means that are defined over common items with parameters on both scale I and J and $\sigma\left(b_J\right)$ and $\sigma\left(b_I\right)$ are the standard deviations that are defined over common items with parameters on both scale I and J.

Moments Methods to Estimate Equating Coefficients

These methods use the mean and the standard deviation of the common items to obtain the equating coefficients A and B. The mean/mean method and mean/sigma method are introduced by Kolen and Brennan [1] with early descriptions from Marco [12] and Loyd and Hoover [5].

The first two equations in (2.40) can be used to calculate the constant A respectively and the first equation in (2.41) can be used to calculate the constant B. If the first equation in (2.40) is used, the method is called the mean/sigma method. If the second equation in (2.41) is used, the method is called the mean/mean method.

Common-Item Equating to a Calibrated Pool

When several exam versions need to be equated together, the common item equating to a calibrated pool method can be employed. A calibrated pool is a set of items coming from different versions whose parameters are expressed on the same scale. When a new form is constructed, some items from the calibrated item pool are included. The parameters that result from estimating this new version are transformed to the scale that was established for the pool and is then included in the pool.

Suppose that a pool is formed by 3 versions (labeled 1, 2, and 3) and that they enter the pool in order 1, 2, and 3 [14]. We take the Rasch model as an example and apply the mean/mean method here to do the conversion. Then the conversion of the item difficulty parameters of version 2 to the scale of version 1 is b_2+B_{21}. Let B_{31} be the equating coefficient for converting from the scale of version 3 to the scale of version 1, B_{32} the equating coefficient for converting from the scale of version 3 to the scale of version 2, n_{13} the number of common items between version 1 and 3 and n_{23} the number of common items between version 2 and 3.

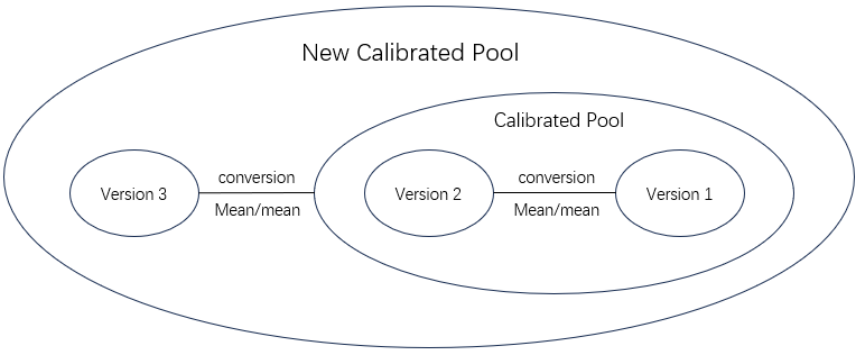

Figure 2.2: Common-Item Equating to a Calibrated Pool

Thus, parameters of version 3 can be converted on the scale of the pool (composed of versions 1 and 2) using the equation $b_3 + B^*_{31}$, where

$$B^*_{31} = \frac{n_{13}B_{31} + n_{23}\left(B_{21} + B_{32}\right)}{n_{13} + n_{23}}. \tag{2.42}$$

The above methods can be called the separate calibration, since we estimate the parameters for the common item separately and then use the transformation to put the items on the same scale. There is another method called concurrent calibration, which is a method that doesn't estimate the parameters separately but simutaneously.

2.1.7.1 Concurrent Calibration

Another method to equate different versions of exams is to put two datasets into one dataset and estimate the parameters simultaneously. In this case, we should rearrange the order of the items to make them between the unique items. The illustration [8] is in the Figure 2.3.

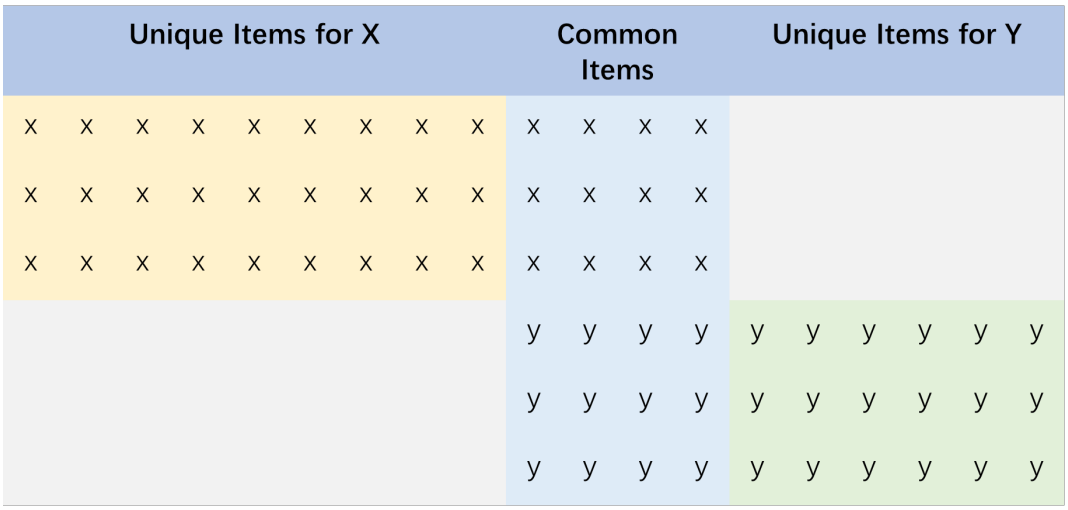

Figure 2.3: Concurrent Calibration

2.2 Classical Test Theory

Classical Test Theory is also called true score theory. It assumes that each person has a true score T that would be obtained if there were no errors in measurement. A person's true score is the expected score over an infinite number of independent scale administrations [3]. However, true scores can't be observed, only the observed scores can. The relationship between observed scores and true scores is that the observed score (X) = true score (T) + some error (E). This model assumes that the expected value of the random errors is zero and that the random errors are uncorrelated with the true score [23].

The true score reflects the concept of the trait or ability of interest. There should be a monotonically increasing relationship between true scores and observed scores so that higher

responses reflect higher values of the concept.

2.2.1 Reliability Coefficient

Reliability, in the context of educational testing, is intimately tied to the notion of consistency. It is a term employed to evaluate the consistency of individuals' scores on educational assessments [13].

Before Cronbach's alpha was developed, the split-half coefficients were employed in most cases. In this method, the test is randomly split into two halves, and the sum scores of the two halves are compared as if they were two separate administrations of the same test score [7]. The correlation between the sum scores of the two halves is an estimate of the reliability of the half test. The high split-half coefficient indicates high reliability. High reliability indicates that an examinee would perform equally well on both halves of the exam.

Cronbach's alpha is the average of all possible split-half estimates. The limitation of split-alpha coefficients is that the estimate depends on the way the split is made since there are multiple ways to divide the items into two halves. Therefore, Cronbach's alpha removes in a way the arbitrariness of how to split an exam.

$$\begin{aligned} \alpha^C &= \frac{p}{p-1} \frac{\operatorname{Var}\left(\sum_{j=1}^{p} X_j\right) - \sum_{j=1}^{p} \operatorname{Var}\left(X_j\right)}{\operatorname{Var}\left(\sum_{j=1}^{p} X_j\right)} \\ &= \frac{p}{p-1} \frac{\sum \sum_{j \neq k} \sigma_{jk}}{V_t}, \end{aligned} \tag{2.43}$$

where X_j is the observed scores of item j for $j = 1, \ldots, p$; V_t is the variance of the total score; σ_{jk} is the covariance of the pair (X_j, X_k).

2.2.2 Item Analysis

Item Difficulty

Item difficulty is the proportion of the number of examinees who give a correct response to the items among all the examinees. It means item difficulty is a measure of the proportion of examinees that answered the item correctly. The item difficulty parameter is called p-value and it can be simply calculated by dividing the number of examinees who gave the correct response by the total number of examinees who responded to the item [9].

$$p_j = \frac{\sum_{i=1}^{n} X_{ij}}{n} \tag{2.44}$$

X_{ij} is the observed score for examinee i and item j; n is the total number of items.

Item Discrimination

Item discrimination refers to the degree to which an item differentiates correctly among examinees in the behavior that the exam is designed to measure. It is an index that assesses an item's capability to differentiate between examinees who are good at acquiring knowledge and those who are not. The discrimination parameter can be described as the correlation between the performance of an item and the performance of the total test. The formula used here is the point biserial correlation, which is a special case of Pearson correlation. The formula is as follows:

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} \tag{2.45}$$

where M_1 is the mean of the total test scores for those whose dichotomous response was 1 and M_0 is the mean of the total test scores for those whose dichotomous response was 0; s_n is the standard deviation of all scores on the total exam.

CHAPTER 3

Results

3.1 Data

The datasets are collected from Erasmus Medical Center (EMC). The data is the binary score that examinees obtain on every item during the pharmacotherapy tests. Simply Speaking, the examinees will obtain 1 point for giving the correct response and 0 for giving the wrong response.

Items	**A1.1.817it**	**A1.4.1017j**	**A4.1.817iu**	**A1.1.817fm**	**A2.1.198**
1	1	1	1	0	1
2	0	0	1	1	1
3	0	1	1	1	1
4	1	1	1	1	1
5	1	1	1	1	1
6	0	1	1	0	1
7	0	1	0	0	0

Table 3.1: Example

For this specific pharmacotherapy exam, EMC has several versions of exam papers, and each of them has 60 items in total. Between each two versions, there are some common items, which means some items appear in more than one version of the exam. Different versions of exams were administered at different times to different groups of medical students.

	Version1	**Version5**	**Version9**
Number	174	95	261

Table 3.2: Number of Samples in each version

The total score of the exam is 60. Examinees who can give correct responses to at least 85% of the items can pass the exam. In our case, examinees who obtain 51 out of 60 can

pass the exam. The numbers of examinees who passed or failed the exam of three versions are shown in Table3.3. The frequency of the total scores is shown in Figure3.1.

	Version1	**Version5**	**Version9**
Pass	112	40	162
Fail	62	55	99
In total	174	95	261

Table 3.3: The number of examinees who passed/failed the exam

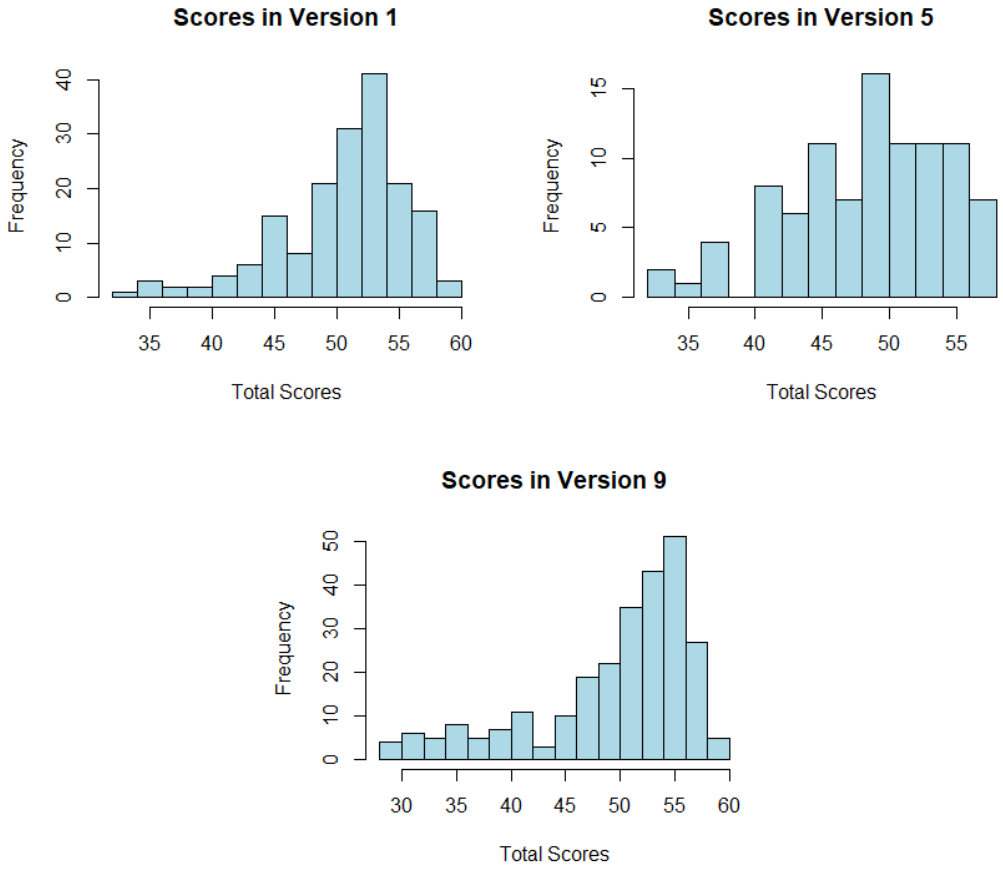

Figure 3.1: Histograms of the examinees' score for exam papers of three versions

3.2 Classical Test Theory

Internal Consistency

Under the classical test theory, the first thing is to test the reliability, which is to test the internal consistency across items within an exam. The method is Cronbach's alpha and the result is shown in

Cronbach's Alpha					
version1	value	version5	value	version9	value
All Items	0.7666	All Items	0.796	All Items	0.8754
A1.1.817it	0.7619	A1.4.817dx	0.7858	A2.817dg	0.8734
A1.4.1017j	0.7632	A2.1.817dw	0.7973	A3.4.817fl	0.8756
A4.1.817iu	0.7627	A4.1.817s	0.7944	A4.1.817s	0.874
A3.4.817fl	0.7681	A4.4.2.817fk	0.7974	A4.3.2.817jr	0.8741
A6.4.1.817ck	0.7547	A5.4.2.817fz	0.7899	A5.4.817r	0.8728
A1.1.817fm	0.7623	A6.2.2.817cp	0.7956	A6.2.2.817cj	0.8724
A2.1.198	0.7663	A6.4.1.817ev	0.796	A6.4.1.817ck	0.8724
A5.2.4.817p	0.7629	A7.2.1.817dy	0.7966	A6.4.1.817ew	0.8779
A4.4.2.817ix	0.767	A7.4.2.817dv	0.7962	A7.2.1.817dy	0.875
B1.1.052	0.7661	B1.1.052	0.7992	B1.1217ar	0.8703
B.1017m	0.7639	B.1017m	0.7919	B2.2.051	0.8749
B.1017n	0.7601	B2.3.817fp	0.795	B2.2.1.817f	0.8715
B2.1217au	0.7621	B3.3.817iy	0.7936	B3.4.032	0.8735
B2.2.3.817a	0.7597	B3.4.072	0.7941	B3.7.fu	0.8721
B3.7.fu	0.7617	B3.7.fu	0.7936	B5.6.817ez	0.8739
B.5.5.3	0.7663	B5.3.817ey	0.784	B6.1.817jb	0.8693
B5.6.036	0.7597	B5.6.817ez	0.7947	B8.3.077	0.8716
B6.1.079	0.7601	B6.3.817js	0.7946	B8.7.817fv	0.871
C1.1.1217bf	0.7624	C1.4.1217bj	0.7907	C1.2.1.817hk	0.8747
C1.1.2.069	0.7623	C1.6.068	0.7951	C1.2.1217az	0.8764
C1.2.1217bg	0.7611	C2.1.3.817ek	0.7931	C1.2.817fa	0.8731
C1.2.4.158	0.764	C2.2.4_1217ab	0.7894	C1.7.817fg	0.8751
C1.3.2.817jj	0.76	C2.6.817jk	0.7924	C2.6.817fj	0.875
C1.4.817ci	0.762	C4.1.1.817hu	0.7956	C4.1.3.029	0.8693
C2.6.817fj	0.7581	C4.2.028	0.7937	C4.1.3.031	0.8798

C4.1.3.031	0.7663	C4.4.817hp	0.7898	C5f.817hs	0.8753
C4.4.171	0.7661	C6.6.817dp	0.7899	C8.8b.817w	0.8698
D.1017y	0.7629	D1.1.817bu	0.788	D3.3.1.817ee	0.8736
D2.2.165	0.7637	D3.4.817by	0.7856	D4.2.101	0.8713
D4.3.108	0.7662	D5.1217ae	0.8	D6.2.817hz	0.8682
E.2.1	0.7616	E1.1217bk	0.7969	E1.1217bk	0.872
E1.1217bb	0.7585	E1.1217bn	0.7853	E1.817gh	0.872
E8.817gk	0.7689	E1.817gi	0.795	E4.1.055	0.8739
E2.2	0.764	E4.1.056	0.7884	E4.4.2.817ie	0.8724
E4.1.055	0.7658	E5.2.2.045	0.7935	E6.1.1.126	0.8738
E5.2.2.124	0.7627	E6.1.1.1217w	0.7874	E8.817gk	0.8723
F1.1217bw	0.7556	F2.1.1.817j	0.7888	F3.1	0.8733
F.1017q	0.7666	F3.1217by	0.7959	F5.1217ca	0.8722
F4.4.3	0.7649	F4.4.2	0.7955	F6.1.1.817da	0.8738
G.1017a	0.7688	G1.1.1.817am	0.793	G.1017d	0.8702
G.1017c	0.7675	G3.4.817ca	0.7906	G1.1.817ai	0.8738
G6.5.026	0.7668	G4.4.1.043	0.7944	G3.3.040	0.8716
G3.5.817cb	0.7559	G4.4.2.064	0.796	G3.817cc	0.8755
G4.2.214	0.7591	G4.5.216	0.7889	G4.2.213	0.8762
G5.6.025	0.7664	G5.5.167	0.789	G4.2.214	0.8719
H.1	0.7672	H.2.084	0.7938	H2.1	0.8793
H1.817ba	0.7673	H3.817ah	0.7938	H2.817bb	0.8735
H3.1	0.7612	H4	0.7908	H3.817je	0.8705
I1.2.817dj	0.7598	I2.1.090	0.7961	I1.2.817dj	0.8731
I1.817dl	0.7619	I2.2.006	0.7916	I2_1217an	0.8719
I5.1.010a	0.7667	I5.817di	0.7868	I5.1.008	0.8759
J.2.2	0.7659	J1.1.2.817io	0.7946	J1.1.1.817im	0.8743
J1.1.817x	0.766	J1.2.817cv	0.7956	J1.2.208	0.8733
J1.1.1.817il	0.7691	J2.2.1217c	0.8016	J2.817eb	0.8764
K.1017u	0.7682	K3.3.817bi	0.7965	K.318d	0.8742
K2.3.817en	0.7629	k3.817aa	0.7971	K3.817ab	0.8756
K4.5.817eq	0.7674	K4.2.817eo	0.8006	K4.5.817ac	0.8744
L1_1217k	0.7653	L2.1217o	0.7943	L2.318a	0.8732
L2.318a	0.7628	L3.2.1_1217g	0.7932	L3.2.1_1217p	0.8749
L3_1217s	0.7641	L3_1217q	0.7922	L3_1217r	0.8767

Table 3.4: Cronbach's alpha for three versions of exams

We calculate Cronbach's alpha for three versions and simultaneously calculate them excluding one item at a time to see if Cronbach's alpha increases and decreases. If Cronbach's alpha has distinct changes after excluding one item, this item can be considered as the item that affects internal consistency.

For all three versions, Cronbach's alpha is beyond the acceptable level of 0.70. For some items in Version 5 and all the items in Version 9, Cronbach's alpha is beyond 0.8, which is considered excellent for internal consistency.

Item Analysis

The item difficulty in CTT can be obtained by calculating the proportion of correct answers for each item, which is called the p-value. The item discrimination in CTT can be obtained by the point biserial correlation between the item response and the total score.

After looking into the datasets, there is one item in each version of the exam to which all the examinees give the correct responses. In this case, this item should be removed since the all-correct responses won't provide any information about the item discrimination. Therefore, we will use 59 items to do the item analysis.

Item Parameters for CTT								
V1	Diff	Disc	V5	Diff	Disc	V9	Diff	Disc
A1.1.817it	0.897	0.310	A1.4.817dx	0.695	0.499	A2.817dg	0.866	0.354
A1.4.1017j	0.908	0.268	A2.1.817dw	0.979	-0.028	A4.1.817s	0.973	0.342
A4.1.817iu	0.937	0.289	A4.1.817s	0.989	0.279	A4.3.2.817jr	0.973	0.322
A3.4.817fl	0.983	-0.034	A4.4.2.817fk	0.979	-0.041	A5.4.817r	0.843	0.397
A6.4.1.817ck	0.902	0.538	A5.4.2.817fz	0.937	0.451	A6.2.2.817cj	0.866	0.425
A1.1.817fm	0.977	0.378	A6.2.2.817cp	0.726	0.231	A6.4.1.817ck	0.931	0.455
A2.1.198	0.983	0.105	A6.4.1.817ev	0.516	0.243	A6.4.1.817ew	0.402	0.164
A5.2.4.817p	0.925	0.281	A7.2.1.817dy	0.874	0.142	A7.2.1.817dy	0.870	0.235
A4.4.2.817ix	0.983	0.053	B1.1.052	0.895	0.015	B1.1217ar	0.851	0.560
B1.1.052	0.759	0.230	B.1017m	0.832	0.325	B2.2.051	0.651	0.309
B.1017m	0.897	0.247	B2.3.817fp	0.832	0.220	B2.2.1.817f	0.770	0.473
B.1017n	0.477	0.377	B3.3.817iy	0.979	0.309	B3.4.032	0.954	0.373

B2.1217au	0.948	0.323	B3.4.072	0.758	0.267	B3.7.fu	0.889	0.455
B2.2.3.817a	0.713	0.375	B3.7.fu	0.895	0.255	B5.6.817ez	0.950	0.322
B3.7.fu	0.920	0.321	B5.3.817ey	0.768	0.557	B6.1.817jb	0.843	0.619
B.5.5.3	0.977	0.114	B5.6.817ez	0.968	0.201	B8.3.077	0.755	0.470
B5.6.036	0.511	0.385	B6.3.817js	0.421	0.278	B8.7.817fv	0.904	0.547
B6.1.079	0.925	0.384	C1.4.1217bj	0.768	0.366	C1.2.1.817hk	0.893	0.247
C1.1.1217bf	0.443	0.334	C1.6.068	0.463	0.266	C1.2.1217az	0.858	0.141
C1.1.2.069	0.937	0.308	C2.1.3.817ek	0.347	0.311	C1.2.817fa	0.897	0.378
C1.2.1217bg	0.649	0.352	C2.2.4_1217ab	0.768	0.404	C1.7.817fg	0.563	0.309
C1.2.4.158	0.741	0.282	C2.6.817jk	0.842	0.310	C2.6.817fj	0.751	0.282
C1.3.2.817jj	0.856	0.363	C4.1.1.817hu	0.884	0.177	C4.1.3.029	0.793	0.602
C1.4.817ci	0.948	0.328	C4.2.028	0.758	0.280	C4.1.3.031	0.571	0.069
C2.6.817fj	0.724	0.408	C4.4.817hp	0.842	0.398	C5f.817hs	0.996	0.146
C4.1.3.031	0.448	0.258	C6.6.817dp	0.926	0.433	C8.8b.817w	0.797	0.572
C4.4.171	0.736	0.236	D1.1.817bu	0.863	0.466	D3.3.1.817ee	0.805	0.352
D.1017y	0.971	0.319	D3.4.817by	0.789	0.516	D4.2.101	0.847	0.498
D2.2.165	0.943	0.249	D5.1217ae	0.421	0.136	D6.2.817hz	0.870	0.717
D4.3.108	0.989	0.117	E1.1217bk	0.779	0.178	E1.1217bk	0.801	0.447
E.2.1	0.920	0.325	E1.1217bn	0.884	0.580	E1.817gh	0.920	0.482
E1.1217bb	0.661	0.402	E1.817gi	0.979	0.184	E4.1.055	0.759	0.341
E8.817gk	0.856	0.115	E4.1.056	0.905	0.476	E4.4.2.817ie	0.847	0.425
E2.2	0.954	0.238	E5.2.2.045	0.726	0.291	E6.1.1.126	0.927	0.319
E4.1.055	0.718	0.247	E6.1.1.1217w	0.811	0.469	E8.817gk	0.866	0.432
E5.2.2.124	0.701	0.314	F2.1.1.817j	0.705	0.422	F3.1	0.958	0.402
F1.1217bw	0.931	0.554	F3.1217by	0.989	0.086	F5.1217ca	0.843	0.435
F.1017q	0.805	0.202	F4.4.2	0.968	0.139	F6.1.1.817da	0.774	0.343
F4.4.3	0.977	0.204	G1.1.1.817am	0.884	0.283	G.1017d	0.766	0.545
G.1017a	0.897	0.087	G3.4.817ca	0.768	0.370	G1.1.817ai	0.943	0.324
G.1017c	0.943	0.089	G4.4.1.043	0.926	0.213	G3.3.040	0.862	0.482
G3.5.817cb	0.816	0.460	G4.4.2.064	0.979	0.084	G3.817cc	0.996	0.069
G4.2.214	0.885	0.395	G4.5.216	0.737	0.418	G4.2.213	0.858	0.160
G5.6.025	0.931	0.144	G5.5.167	0.884	0.440	G4.2.214	0.881	0.468
H.1	0.977	0.054	H.2.084	0.916	0.245	H2.1	0.421	0.092
H1.817ba	0.805	0.185	H3.817ah	0.811	0.268	H2.817bb	0.579	0.387
H3.1	0.655	0.351	H4	0.442	0.375	H3.817je	0.870	0.562

I1.2.817dj	0.862	0.369	I2.1.090	0.989	0.068	I1.2.817dj	0.870	0.372
I1.817dl	0.931	0.318	I2.2.006	0.916	0.348	I2_1217an	0.854	0.458
I5.1.010a	0.908	0.150	I5.817di	0.874	0.515	I5.1.008	0.985	0.017
J.2.2	0.966	0.143	J1.1.2.817io	0.779	0.247	J1.1.1.817im	0.931	0.275
J1.1.817x	0.994	0.150	J1.2.817cv	0.968	0.129	J1.2.208	0.958	0.407
J1.1.1.817il	0.885	0.087	J2.2.1217c	0.737	0.047	J2.817eb	0.900	0.106
K.1017u	0.989	-0.074	K3.3.817bi	0.789	0.186	K.318d	0.575	0.349
K2.3.817en	0.609	0.322	k3.817aa	0.863	0.132	K3.817ab	0.977	0.096
K4.5.817eq	0.908	0.123	K4.2.817eo	0.895	-0.043	K4.5.817ac	0.969	0.280
L1_1217k	0.977	0.182	L2.1217o	0.958	0.218	L2.318a	0.762	0.378
L2.318a	0.718	0.310	L3.2.1_1217g	0.863	0.278	L3.2.1_1217p	0.858	0.246
L3_1217s	0.845	0.255	L3_1217q	0.421	0.340	L3_1217r	0.552	0.229

Table 3.5: The Item Difficulty and Item Discrimination for Classical Test Theory Analysis

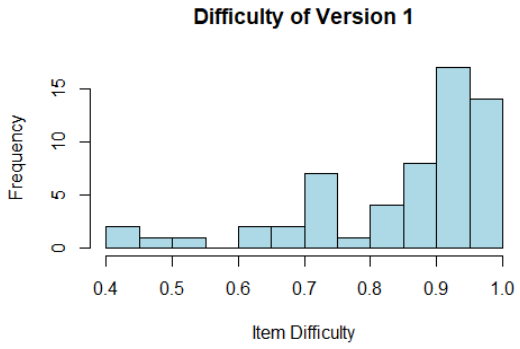

Figure 3.2: CTT Item Difficulty for V1

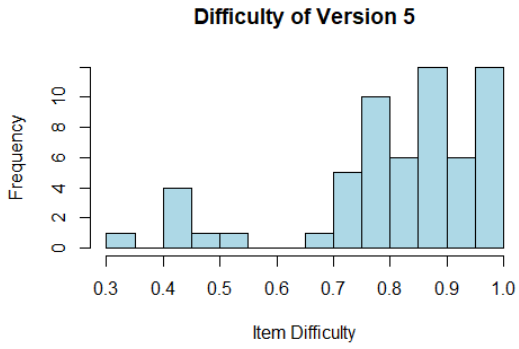

Figure 3.3: CTT Item Difficulty for V5

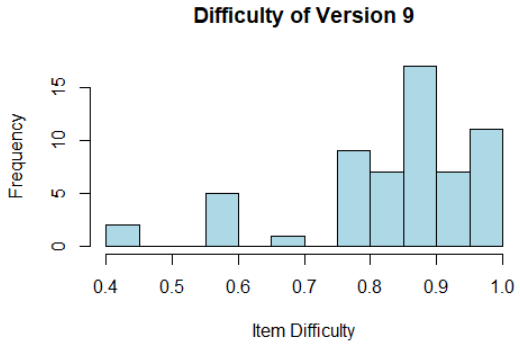

Figure 3.4: CTT Item Difficulty for V9

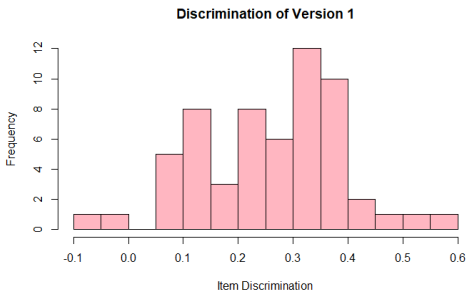

Figure 3.5: CTT Item Discrimination for V1

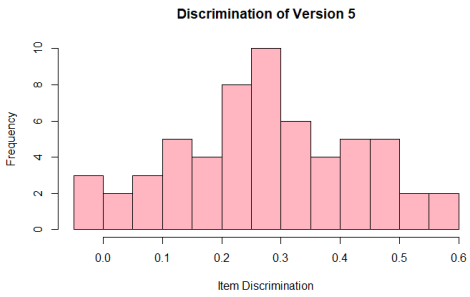

Figure 3.6: CTT Item Discrimination for V5

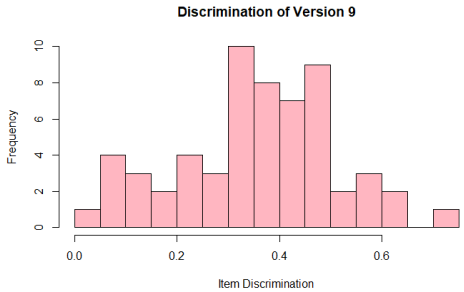

Figure 3.7: CTT Item Discrimination for V9

From the above tables and graphs. It is evident that the item difficulty and item discrimination in CTT depend on the characteristics of the students and are irregularly patterned. An item that is easy for one group of students may be difficult for another group. Therefore their values are relative values. And also they are easy to compute since there are no complicated models.

3.3 Item Response Theory

In this section, we will employ Item Response Theory to analyze the response data. We have two models to select: the Rasch model and the 2PL model

3.3.1 Parameter Estimation

First, we estimate the parameters using the Marginal Maximum Likelihood (MML) for both the Rasch model and 2PL model, the result is shown in

	Rasch	2PL			Rasch	2PL	
Version1	Diff	Disc	Diff	Version9	Diff	Disc	Diff
A1.1.817it	-2.357	0.948	-2.623	A2.817dg	-2.186	0.997	-2.202
A1.4.1017j	-2.494	0.884	-2.926	A4.1.817s	-4.042	4.267	-1.995
A4.1.817iu	-2.918	0.997	-3.108	A4.3.2.817jr	-4.042	2.714	-2.247
A3.4.817fl	-4.296	-0.581	7.244	A5.4.817r	-1.975	1.247	-1.72
A6.4.1.817ck	-2.424	2.577	-1.569	A6.2.2.817cj	-2.186	1.311	-1.838
A1.1.817fm	-3.999	3.117	-2.291	A6.4.1.817ck	-2.999	2.391	-1.854
A2.1.198	-4.296	0.786	-5.511	A6.4.1.817ew	0.489	0.221	1.811
A5.2.4.817p	-2.731	1.237	-2.502	A7.2.1.817dy	-2.223	0.631	-3.232
A4.4.2.817ix	-4.296	0.45	-9.203	B1.1217ar	-2.042	1.985	-1.41
B1.1.052	-1.266	0.48	-2.508	B2.2.051	-0.734	0.65	-1.055
B.1017m	-2.357	0.723	-3.257	B2.2.1.817f	-1.428	1.278	-1.242
B.1017n	0.106	0.779	0.131	B3.4.032	-3.458	2.278	-2.121
B2.1217au	-3.138	1.435	-2.593	B3.7.fu	-2.426	1.605	-1.814
B2.2.3.817a	-1.005	0.768	-1.332	B5.6.817ez	-3.369	1.61	-2.445
B3.7.fu	-2.647	1.146	-2.557	B6.1.817jb	-1.975	2.536	-1.266
B.5.5.3	-3.999	0.852	-4.792	B8.3.077	-1.329	1.262	-1.166
B5.6.036	-0.048	0.958	-0.062	B8.7.817fv	-2.609	2.446	-1.637
B6.1.079	-2.731	1.552	-2.179	C1.2.1.817hk	-2.47	0.794	-2.96
C1.1.1217bf	0.261	0.54	0.455	C1.2.1217az	-2.113	0.276	-6.628
C1.1.2.069	-2.918	1.228	-2.677	C1.2.817fa	-2.515	1.38	-2.037
C1.2.1217bg	-0.683	0.872	-0.824	C1.7.817fg	-0.288	0.58	-0.475
C1.2.4.158	-1.165	0.6	-1.891	C2.6.817fj	-1.304	0.645	-1.866
C1.3.2.817jj	-1.959	0.957	-2.172	C4.1.3.029	-1.586	2.133	-1.086
C1.4.817ci	-3.138	1.497	-2.528	C4.1.3.031	-0.326	0.001	-222.104
C2.6.817fj	-1.068	0.985	-1.173	C5f.817hs	-6.042	6.809	-2.268
C4.1.3.031	0.235	0.454	0.478	C8.8b.817w	-1.614	1.78	-1.181
C4.4.171	-1.132	0.364	-2.899	D3.3.1.817ee	-1.67	0.801	-1.995
D.1017y	-3.766	2.038	-2.531	D4.2.101	-2.008	1.775	-1.453
D2.2.165	-3.023	1.005	-3.196	D6.2.817hz	-2.223	4.599	-1.286
D4.3.108	-4.712	1.29	-4.043	E1.1217bk	-1.642	1.287	-1.413
E.2.1	-2.647	1.215	-2.458	E1.817gh	-2.819	2.458	-1.745
E1.1217bb	-0.739	0.775	-0.976	E4.1.055	-1.353	0.875	-1.516
E8.817gk	-1.959	0.181	-9.913	E4.4.2.817ie	-2.008	1.111	-1.882
E2.2	-3.266	1.154	-3.109	E6.1.1.126	-2.937	1.265	-2.49

E4.1.055	-1.037	0.456	-2.15	E8.817gk	-2.186	1.348	-1.808
E5.2.2.124	-0.944	0.681	-1.38	F3.1	-3.554	2.822	-2.014
F1.1217bw	-2.821	2.995	-1.724	F5.1217ca	-1.975	1.148	-1.812
F.1017q	-1.56	0.306	-4.716	F6.1.1.817da	-1.454	0.73	-1.873
F4.4.3	-3.999	1.093	-3.919	G.1017d	-1.403	1.719	-1.05
G.1017a	-2.357	0.16	-13.568	G1.1.817ai	-3.208	1.714	-2.262
G.1017c	-3.023	0.114	-24.69	G3.3.040	-2.149	1.685	-1.583
G3.5.817cb	-1.641	1.253	-1.521	G3.817cc	-6.042	1.316	-4.808
G4.2.214	-2.231	1.357	-1.947	G4.2.213	-2.113	0.471	-3.989
G5.6.025	-2.821	0.553	-4.94	G4.2.214	-2.342	1.665	-1.724
H.1	-3.999	0.226	-16.683	H2.1	0.395	0.132	2.411
H1.817ba	-1.56	0.302	-4.777	H2.817bb	-0.363	0.855	-0.434
H3.1	-0.711	0.688	-1.032	H3.817je	-2.223	2.353	-1.439
I1.2.817dj	-2.009	1.099	-2.017	I1.2.817dj	-2.223	1.152	-2.027
I1.817dl	-2.821	1.173	-2.673	I2_1217an	-2.077	1.379	-1.701
I5.1.010a	-2.494	0.384	-6.122	I5.1.008	-4.628	0.331	-12.748
J.2.2	-3.573	0.725	-4.923	J1.1.1.817im	-2.999	0.918	-3.203
J1.1.817x	-5.415	2.062	-3.4	J1.2.208	-3.554	2.68	-2.047
J1.1.1.817il	-2.231	-0.047	43.661	J2.817eb	-2.561	0.413	-5.491
K.1017u	-4.712	-1.643	3.494	K.318d	-0.345	0.66	-0.505
K2.3.817en	-0.491	0.586	-0.819	K3.817ab	-4.205	0.547	-7.105
K4.5.817eq	-2.494	0.324	-7.193	K4.5.817ac	-3.899	1.831	-2.594
L1_1217k	-3.999	1.258	-3.536	L2.318a	-1.378	0.996	-1.406
L2.318a	-1.037	0.636	-1.602	L3.2.1_1217p	-2.113	0.604	-3.19
L3_1217s	-1.862	0.701	-2.644	L3_1217r	-0.232	0.506	-0.437

Table 3.6: Item Parameter Estimation for Version 1 and 9 under IRT

Version 5							
Item	Diff	Item	Diff	Item	Diff	Item	Diff
A1.4.817dx	-0.927	B5.6.817ez	-3.683	E1.1217bn	-2.242	H3.817ah	-1.621
A2.1.817dw	-4.105	B6.3.817js	0.358	E1.817gi	-4.105	H4	0.261
A4.1.817s	-4.816	C1.4.1217bj	-1.344	E4.1.056	-2.479	I2.1.090	-4.816
A4.4.2.817fk	-4.105	C1.6.068	0.165	E5.2.2.045	-1.098	I2.2.006	-2.614
A5.4.2.817fz	-2.937	C2.1.3.817ek	0.71	E6.1.1.1217w	-1.621	I5.817di	-2.137

A6.2.2.817cp	-1.098	C2.2.4_1217ab	-1.344	F2.1.1.817j	-0.983	J1.1.2.817io	-1.41
A6.4.1.817ev	-0.073	C2.6.817jk	-1.859	F3.1217by	-4.816	J1.2.817cv	-3.683
A7.2.1.817dy	-2.137	C4.1.1.817hu	-2.242	F4.4.2	-3.683	J2.2.1217c	-1.157
B1.1.052	-2.355	C4.2.028	-1.28	G1.1.1.817am	-2.242	K3.3.817bi	-1.478
B.1017m	-1.776	C4.4.817hp	-1.859	G3.4.817ca	-1.344	k3.817aa	-2.039
B2.3.817fp	-1.776	C6.6.817dp	-2.765	G4.4.1.043	-2.765	K4.2.817eo	-2.355
B3.3.817iy	-4.105	D1.1.817bu	-2.039	G4.4.2.064	-4.105	L2.1217o	-3.377
B3.4.072	-1.28	D3.4.817by	-1.478	G4.5.216	-1.157	L3.2.1_1217g	-2.039
B3.7.fu	-2.355	D5.1217ae	0.358	G5.5.167	-2.242	L3_1217q	0.358
B5.3.817ey	-1.344	E1.1217bk	-1.41	H.2.084	-2.614		

Table 3.7: Item Parameter Estimation for Version 5 under IRT

Notice that we only estimate the parameters in Version 5 under the Rasch model. This is due to the small number of samples (95) we have. Since we have 60 items, we need more data to implement 2PL models.

Throughout the obtained results of 2PL in Table 3.6, it is evident that certain items exhibit significantly high difficulty parameters, for example, G.1017a, G.1017c, H.1, J1.1.1.817il, and C4.1.3.031, which means these items are challenging, to the extent that even proficient students may struggle to provide correct responses. This is due to the discrimination values being close to 0.

It is crucial to emphasize that, during the estimation process, the number of Expectation-Maximization (EM) iterations exceeded 3000. This suggests that the specified model is operating with very limited empirical data. Empirical analysis further underscores the limitation of the sample size, particularly in the context of examining 60 items. To be specific, there are only 174 and 261 samples available for assessing these 60 items.

Since the 2PL model requires more information compared to the Rasch model, the decision has been made to use the Rasch model for the subsequent analysis in light of these limitations.

3.3.2 Local Dependency Detection

In this section, we employ the traditional Q_3 method to compute residuals for pairs of items. This analysis helps us assess whether there is local item dependency and guides our decision on whether to retain or remove specific items.

Q3	A1.4.1017j	A6.4.1.817ck	A1.1.817fm	A2.1.198	A5.2.4.817p	B2.1217au	B3.7.fu	B.5.5.3	C1.3.2.817jj	C1.4.817ci	D.1017y	D4.3.108	E.2.1	E2.2	F1.1217bw	F4.4.3	G.1017c	J.2.2	L1_1217k
A1.1.817fm		0.24																	
B2.1217au	0.22		0.24																
B6.1.079		0.32	0.32		0.27	0.26													
D.1017y				0.22															
D2.2.165	0.22																		
D4.3.108								0.34											
E.2.1					0.26														
E2.2								0.32				0.22							
F1.1217bw		0.22							0.21		0.29								
F4.4.3				0.26															
G3.5.817cb			0.23																
G4.2.214													0.21						
I1.2.817dj										0.20	0.37								
I1.817dl															0.38				
I5.1.010a			0.28																
J.2.2											0.31								
J1.1.817x											0.43				0.25			0.39	
K.1017u																	0.22		
L1_1217k				0.26			0.20			0.28				0.31		0.21			
L2.318a																			0.20
L3_1217s																			0.22

Table 3.8: Q3 residual values exceeding 0.2 for version 1

Q3	A5.4.2.817fz	A7.2.1.817dy	B3.7.fu	B6.3.817js	C1.4.1217bj	C2.1.3.817ek	C4.2.028	C6.6.817dp	D1.1.817bu	E1.1217bn	E4.1.056	E5.2.2.045	E6.1.1.1217w	F2.1.1.817j	F3.1217by	F4.4.2
A1.4.817dx			0.206727					0.213351								
A4.1.817s	0.3459922		0.241161							0.213673						
A5.4.2.817fz		0.209879						0.328525	0.288111		0.246762					
A6.2.2.817cp						0.234387	0.326917									
A6.4.1.817ev																
B.1017m														0.230048		
B1.1.052												0.334759				0.325204
B2.3.817fp										0.260929				0.256856		0.22189
B3.3.817iy											0.399672					
B3.7.fu				0.248789												
B5.3.817ey					0.20564					0.258182				0.267191		
B5.6.817ez																0.292735
C1.6.068																
C2.2.4_1217ab																
C2.6.817jk																
C4.2.028													0.258696			
C6.6.817dp									0.240599		0.528763		0.384908			
D1.1.817bu											0.399476		0.219324		0.244551	
D3.4.817by										0.204581						
D5.1217ae																
E1.1217bk																
E1.1217bn																
E1.817gi																
E4.1.056															0.309576	
E5.2.2.045																0.270578
F2.1.1.817j																0.245712

Table 3.9: Q3 residual values exceeding 0.2 for version 5, PART 1

Q3	G1.1.1.817am	G3.4.817ca	G4.4.1.043	G4.4.2.064	G5.5.167	H.2.084	H3.817ah	I2.1.090	I2.2.006	I5.817di	J1.1.2.817io	J1.2.817cv	K3.3.817bi	k3.817aa	K4.2.817eo	L2.1217o	L3.2.1_1217g
A1.4.817dx																	
A4.1.817s					0.217113	0.285281											
A5.4.2.817fz													0.30901				
A6.2.2.817cp																	
A6.4.1.817ev			0.250854														
B.1017m								0.220425									
B1.1.052																	
B2.3.817fp								0.224058									
B3.3.817iy												0.361405					
B3.7.fu																	
B5.3.817ey					0.201407				0.265813								
B5.6.817ez											0.30098					0.232544	
C1.6.068		0.291118															
C2.2.4_1217ab										0.266119							
C2.6.817jk								0.22991									
C4.2.028																	0.219268562
C6.6.817dp				0.211441													
D1.1.817bu					0.206476					0.270159							
D3.4.817by																	
D5.1217ae											0.328684						
E1.1217bk									0.226448	0.231339							
E1.1217bn							0.202204			0.441004		0.26111					
E1.817gi							0.268417		0.462553	0.352199							
E4.1.056	0.243281																
E5.2.2.045																	
F2.1.1.817j																	
F3.1217by										0.257437					0.295116		0.249262422
F4.4.2																0.244548	
G1.1.1.817am								0.280279				0.285962					
G3.4.817ca																0.204046	
G4.4.1.043				0.230178													
G5.5.167									0.410146								
H.2.084											0.249097						
H3.817ah										0.281437							
I2.1.090															0.298576	0.490511	
I2.2.006										0.236484							
J1.1.2.817io														0.208486			

Table 3.10: Q3 residual values exceeding 0.2 for version 5, PART 2

Q3	A4.1.817s	A4.3.2.817jr	A6.4.1.817ck	B1.1217ar	B3.4.032	B3.7.fu	B5.6.817ez	B6.1.817jb	B8.3.077	B8.7.817fv	C4.1.3.029	C8.8b.817w	D6.2.817hz	E1.817gh	F3.1	J1.2.208
A4.3.2.817jr	0.3572531															
B5.6.817ez		0.34016														
B8.7.817fv	0.2385423															
C1.2.817fa		0.250578														
C4.1.3.029									0.205508	0.242794						
C5f.817hs	0.3533781															
D6.2.817hz	0.2280548							0.271552		0.218385						
E6.1.1.126			0.289424													
E8.817gk												0.249366				
F3.1												0.223499		0.239045		
F5.1217ca							0.238328									
F6.1.1.817da				0.205513												
G1.1.817ai									0.251177						0.276242	
G3.817cc					0.27724											
G4.2.214					0.257641											
H3.817je						0.237812		0.330222			0.27198		0.308759			
J1.1.1.817im							0.215939									
J1.2.208	0.3671693	0.239393														
K4.5.817ac																0.215179

Table 3.11: Q3 residual values exceeding 0.2 for version 9

Based on empirical analysis, the critical value for Q_3 is determined to be 0.2. The four tables provided above list item pairs that do not adhere to the principle of local independence, based on their Q_3 values. The highest Q_3 value observed is 0.39 for version 1, 0.53 for version 5, and 0.37 for version 9. It's important to note that each version contains a total of 1711 pairs of items. However, version 1 has 35 pairs with Q_3 values exceeding 0.2, version 5 has 77 pairs with Q_3 values exceeding 0.2, and version 9 has 28 pairs with Q_3 values above 0.2.

The assumptions of local independence can be violated through response dependency. It occurs when items are linked in some way, such that the response on one item governs the response on another because of similarities in, for example, item content or response content [16].

Therefore, we delve into the content of the items to check for any overlapping content or inner relationships between pairs of items. It turns out that all the pairs of items are incomparable and independent of each other. Thus, it is advisable to keep all the items.

3.3.3 Item Fit

In this section, the infit, outfit, standardization, and p-value can be used to test item fit simultaneously.

The expected value of the Infit and Outfit mean squares is 1. Values larger than 1 mean underfit and indicate unexpected responses, for example, lucky guesses and careless mistakes. Values smaller than 1 mean overfit and indicate there are overly predictable outliers. Underfit is easier to detect than overfit, and Outfit is easier to explain than Infit since Infit is sensitive to the pattern of inlying observations [18]. As indicated in the Methods section, Lincare suggested that both mean squares statistics values between 0.5 and 1.5 are acceptable, the interpretation is as follows:

Interpretation of parameter-level mean-square fit statistics:	
>2.0	Distorts or degrades the measurement system.
1.5 - 2.0	Unproductive for construction of measurement, but not degrading.
0.5 - 1.5	Productive for measurement.
<0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations.

The expected value of the Z-standardized Infit and Ourfit is 0. Less than 0 indicates models are too predictable. More than 0 indicates there is a lack of predictability. However, if mean-squares are acceptable, Zstd can be ignored.

The general principle is:
(1) Investigate outfit before infit.
(2) Investigate Mean-squares before z-standardized.

(3) Investigate high values before low or negative values.

The Outfit and Infit, their z-standardization, and their p-values for both version 1 and version 9 are shown in table (3.12) - (3.14).

item - V1	outfit	z.outfit	outfit.p	infit	z.infit	infit.p
A1.1.817it	0.909	-0.296	0.767	0.984	-0.017	0.986
A1.4.1017j	0.896	-0.313	0.754	1.006	0.099	0.921
A4.1.817iu	0.789	-0.558	0.577	0.995	0.071	0.943
A3.4.817fl	**1.732**	1.095	0.274	1.08	0.326	0.744
A6.4.1.817ck	0.619	-1.638	0.101	0.873	-0.584	0.559
A1.1.817fm	**0.456**	-0.951	0.342	0.967	0.095	0.924
A2.1.198	1.035	0.282	0.778	1.054	0.284	0.776
A5.2.4.817p	0.798	-0.609	0.543	1.004	0.099	0.921
A4.4.2.817ix	1.083	0.349	0.727	1.067	0.306	0.760
B1.1.052	1.003	0.062	0.951	1.037	0.408	0.683
B.1017m	0.923	-0.238	0.812	1.018	0.159	0.874
B.1017n	0.919	-1.637	0.102	0.93	-1.745	0.081
B2.1217au	0.713	-0.709	0.478	0.982	0.043	0.966
B2.2.3.817a	0.904	-0.969	0.333	0.952	-0.578	0.563
B3.7.fu	0.758	-0.811	0.417	0.988	0.027	0.978
B.5.5.3	0.795	-0.168	0.867	1.06	0.284	0.776
B5.6.036	0.913	-1.809	0.070	0.912	-2.179	**0.029**
B6.1.079	0.703	-0.983	0.326	0.953	-0.115	0.908
C1.1.1217bf	0.943	-1.051	0.293	0.961	-0.904	0.366
C1.1.2.069	0.751	-0.693	0.488	0.989	0.05	0.960
C1.2.1217bg	0.919	-1.094	0.274	0.96	-0.629	0.529
C1.2.4.158	0.977	-0.17	0.865	1.006	0.096	0.924
C1.3.2.817jj	0.852	-0.734	0.463	0.962	-0.199	0.842
C1.4.817ci	0.664	-0.874	0.382	0.987	0.06	0.952
C2.6.817fj	0.899	-0.978	0.328	0.93	-0.821	0.412
C4.1.3.031	1.005	0.104	0.917	1.007	0.169	0.866
C4.4.171	1.013	0.159	0.874	1.033	0.394	0.694
D.1017y	**0.476**	-1.058	0.290	1	0.144	0.886
D2.2.165	0.815	-0.431	0.666	1.018	0.159	0.874
D4.3.108	0.652	-0.18	0.857	1.053	0.31	0.757

E.2.1	0.788	-0.686	0.493	0.98	-0.01	0.992
E1.1217bb	0.921	-1.001	0.317	0.932	-1.042	0.297
E8.817gk	1.195	1.01	0.312	1.087	0.602	0.547
E2.2	0.704	-0.673	0.501	1.028	0.194	0.846
E4.1.055	1.015	0.184	0.854	1.028	0.363	0.717
E5.2.2.124	0.956	-0.443	0.658	0.989	-0.121	0.904
F1.1217bw	**0.477**	-1.933	0.053	0.879	-0.407	0.684
F.1017q	1.039	0.306	0.760	1.051	0.465	0.642
F4.4.3	0.774	-0.209	0.834	1.023	0.209	0.834
G.1017a	1.226	0.933	0.351	1.088	0.503	0.615
G.1017c	1.055	0.27	0.787	1.082	0.372	0.710
G3.5.817cb	0.829	-1.097	0.273	0.904	-0.753	0.451
G4.2.214	0.771	-1.009	0.313	0.948	-0.233	0.816
G5.6.025	0.962	-0.013	0.990	1.065	0.333	0.739
H.1	1.082	0.333	0.739	1.071	0.305	0.760
H1.817ba	1.158	1.055	0.291	1.053	0.479	0.632
H3.1	0.936	-0.834	0.404	0.962	-0.582	0.561
I1.2.817dj	0.852	-0.711	0.477	0.957	-0.222	0.824
I1.817dl	0.8	-0.56	0.575	0.979	0.002	0.998
I5.1.010a	1.173	0.7	0.484	1.054	0.322	0.747
J.2.2	0.999	0.159	0.874	1.047	0.248	0.804
J1.1.817x	**0.424**	-0.335	0.738	1.004	0.328	0.743
J1.1.1.817il	1.281	1.198	0.231	1.086	0.522	0.602
K.1017u	**2.008**	1.203	0.229	1.078	0.344	0.731
K2.3.817en	0.972	-0.43	0.667	0.975	-0.449	0.653
K4.5.817eq	1.054	0.289	0.773	1.08	0.436	0.663
L1_1217k	0.676	-0.41	0.682	1.041	0.246	0.806
L2.318a	0.965	-0.312	0.755	0.99	-0.089	0.929
L3_1217s	0.961	-0.155	0.877	1.02	0.188	0.851

Table 3.12: Outfit, Infit, and their standardization for Version 1

item - V5	outfit	z.outfit	outfit.p	infit	z.infit	infit.p
A1.4.817dx	0.819	-1.432	0.152	0.875	-1.271	0.204
A2.1.817dw	**2.215**	1.33	0.184	1.072	0.333	0.739

A4.1.817s	**0.288**	-0.431	0.666	1.001	0.333	0.739
A4.4.2.817fk	**1.787**	1.022	0.307	1.078	0.341	0.733
A5.4.2.817fz	0.556	-0.964	0.335	0.913	-0.125	0.901
A6.2.2.817cp	1.008	0.101	0.920	1.046	0.44	0.660
A6.4.1.817ev	1.038	0.546	0.585	1.032	0.519	0.604
A7.2.1.817dy	1.172	0.642	0.521	1.062	0.343	0.732
B1.1.052	1.362	1.063	0.288	1.121	0.535	0.593
B.1017m	0.931	-0.213	0.831	0.977	-0.067	0.947
B2.3.817fp	0.944	-0.157	0.875	1.045	0.308	0.758
B3.3.817iy	**0.384**	-0.641	0.522	0.994	0.222	0.824
B3.4.072	0.921	-0.412	0.680	1.025	0.235	0.814
B3.7.fu	0.894	-0.205	0.838	1.009	0.122	0.903
B5.3.817ey	0.721	-1.658	0.097	0.842	-1.201	0.230
B5.6.817ez	0.78	-0.103	0.918	1.019	0.219	0.827
B6.3.817js	1.016	0.214	0.831	0.988	-0.156	0.876
C1.4.1217bj	1.017	0.153	0.878	0.952	-0.319	0.750
C1.6.068	1.029	0.408	0.683	0.995	-0.054	0.957
C2.1.3.817ek	0.998	0.016	0.987	0.97	-0.332	0.740
C2.2.4_1217ab	0.849	-0.823	0.411	0.936	-0.442	0.658
C2.6.817jk	0.933	-0.184	0.854	0.985	-0.019	0.985
C4.1.1.817hu	1.011	0.142	0.887	1.057	0.311	0.756
C4.2.028	0.964	-0.156	0.876	1.013	0.143	0.886
C4.4.817hp	0.86	-0.492	0.623	0.936	-0.282	0.778
C6.6.817dp	0.599	-0.946	0.344	0.924	-0.125	0.901
D1.1.817bu	0.715	-1.022	0.307	0.899	-0.423	0.672
D3.4.817by	0.769	-1.206	0.228	0.865	-0.913	0.361
D5.1217ae	1.132	1.545	0.122	1.109	1.561	0.119
E1.1217bk	1.218	1.143	0.253	1.063	0.487	0.626
E1.1217bn	0.531	-1.668	0.095	0.843	-0.608	0.543
E1.817gi	0.582	-0.271	0.786	1.034	0.28	0.779
E4.1.056	0.55	-1.347	0.178	0.908	-0.248	0.804
E5.2.2.045	0.944	-0.334	0.738	1.009	0.114	0.909
E6.1.1.1217w	0.756	-1.153	0.249	0.897	-0.606	0.545
F2.1.1.817j	0.872	-0.929	0.353	0.925	-0.704	0.481
F3.1217by	0.732	0.162	0.871	1.047	0.38	0.704

F4.4.2	0.81	-0.053	0.958	1.047	0.268	0.789
G1.1.1.817am	0.858	-0.359	0.720	1.001	0.086	0.931
G3.4.817ca	0.865	-0.727	0.467	0.959	-0.267	0.789
G4.4.1.043	0.834	-0.264	0.792	1.031	0.202	0.840
G4.4.2.064	0.962	0.249	0.803	1.055	0.309	0.757
G4.5.216	0.835	-1.065	0.287	0.929	-0.576	0.565
G5.5.167	0.696	-0.956	0.339	0.913	-0.293	0.770
H.2.084	0.859	-0.241	0.810	1.007	0.124	0.901
H3.817ah	0.985	0.006	0.995	1.013	0.135	0.893
H4	0.951	-0.611	0.541	0.936	-0.968	0.333
I2.1.090	0.807	0.235	0.814	1.05	0.382	0.702
I2.2.006	0.681	-0.773	0.440	0.968	-0.006	0.995
I5.817di	0.628	-1.326	0.185	0.876	-0.5	0.617
J1.1.2.817io	0.989	0.002	0.998	1.032	0.271	0.786
J1.2.817cv	1.359	0.693	0.488	1.037	0.249	0.803
J2.2.1217c	1.237	1.46	0.144	1.158	1.316	0.188
K3.3.817bi	1.133	0.711	0.477	1.06	0.448	0.654
k3.817aa	1.174	0.676	0.499	1.07	0.395	0.693
K4.2.817eo	**1.609**	1.615	0.106	1.14	0.603	0.547
L2.1217o	0.673	-0.415	0.678	1.026	0.211	0.833
L3.2.1_1217g	0.846	-0.475	0.635	1.009	0.114	0.909
L3_1217q	0.967	-0.373	0.709	0.959	-0.586	0.558

Table 3.13: Outfit, Infit, and their standardization for Version 5

item - V9	outfit	z.outfit	outfit.p	infit	z.infit	infit.p
A2.817dg	1.006	0.096	0.924	1	0.041	0.967
A4.1.817s	**0.33**	-1.54	0.124	0.981	0.066	0.947
A4.3.2.817jr	**0.357**	-1.439	0.150	0.992	0.097	0.923
A5.4.817r	0.874	-0.683	0.495	0.985	-0.096	0.924
A6.2.2.817cj	0.811	-0.944	0.345	0.954	-0.319	0.750
A6.4.1.817ck	**0.497**	-1.88	0.060	0.915	-0.37	0.711
A6.4.1.817ew	1.249	3.23	**0.001**	1.144	3.056	**0.002**
A7.2.1.817dy	1.138	0.712	0.476	1.096	0.755	0.450
B1.1217ar	0.634	-2.234	**0.025**	0.848	-1.3	0.194

B2.2.051	1.038	0.51	0.610	1.073	1.157	0.247
B2.2.1.817f	0.883	-0.926	0.354	0.928	-0.814	0.416
B3.4.032	**0.489**	-1.457	0.145	0.963	-0.059	0.953
B3.7.fu	0.724	-1.257	0.209	0.925	-0.476	0.634
B5.6.817ez	0.569	-1.218	0.223	1.002	0.092	0.927
B6.1.817jb	0.584	-2.728	**0.006**	0.8	-1.828	0.068
B8.3.077	0.861	-1.205	0.228	0.932	-0.802	0.423
B8.7.817fv	0.53	-2.163	**0.031**	0.851	-0.913	0.361
C1.2.1.817hk	1.021	0.168	0.867	1.075	0.54	0.589
C1.2.1217az	1.446	2.11	**0.035**	1.169	1.335	0.182
C1.2.817fa	0.726	-1.176	0.240	0.985	-0.045	0.964
C1.7.817fg	1.064	1.028	0.304	1.06	1.19	0.234
C2.6.817fj	1.109	0.949	0.343	1.1	1.201	0.230
C4.1.3.029	0.655	-2.787	**0.005**	0.816	-2.045	**0.041**
C4.1.3.031	1.33	4.798	**0.000**	1.286	5.145	**0.000**
C5f.817hs	**0.267**	-1.098	0.272	0.804	-0.009	0.993
C8.8b.817w	0.693	-2.388	**0.017**	0.843	-1.699	0.089
D3.3.1.817ee	1.013	0.136	0.892	1.028	0.313	0.754
D4.2.101	0.648	-2.178	**0.029**	0.906	-0.787	0.431
D6.2.817hz	**0.401**	-3.784	**0.000**	0.721	-2.312	**0.021**
E1.1217bk	0.859	-0.979	0.328	0.949	-0.499	0.618
E1.817gh	0.586	-1.624	0.104	0.902	-0.496	0.620
E4.1.055	0.987	-0.07	0.944	1.046	0.568	0.570
E4.4.2.817ie	0.877	-0.649	0.516	0.955	-0.353	0.724
E6.1.1.126	0.8	-0.605	0.545	1.001	0.069	0.945
E8.817gk	0.801	-1.003	0.316	0.945	-0.391	0.696
F3.1	**0.373**	-1.856	0.063	0.952	-0.088	0.930
F5.1217ca	0.856	-0.798	0.425	0.949	-0.415	0.678
F6.1.1.817da	1.036	0.322	0.747	1.043	0.51	0.610
G.1017d	0.707	-2.608	**0.009**	0.867	-1.58	0.114
G1.1.817ai	0.614	-1.163	0.245	1.006	0.103	0.918
G3.3.040	0.695	-1.686	0.092	0.912	-0.678	0.498
G3.817cc	0.578	-0.376	0.707	0.816	0.01	0.992
G4.2.213	1.177	0.935	0.350	1.17	1.341	0.180
G4.2.214	0.688	-1.533	0.125	0.915	-0.578	0.563

H2.1	1.377	4.978	**0.000**	1.182	3.881	**0.000**
H2.817bb	0.994	-0.076	0.939	0.989	-0.19	0.849
H3.817je	0.566	-2.47	**0.014**	0.849	-1.169	0.242
I1.2.817dj	0.795	-1.011	0.312	1.001	0.051	0.959
I2_1217an	0.793	-1.125	0.261	0.926	-0.582	0.561
I5.1.008	1.461	0.81	0.418	1.085	0.333	0.739
J1.1.1.817im	1.022	0.174	0.862	1.019	0.163	0.871
J1.2.208	**0.403**	-1.723	0.085	0.943	-0.12	0.904
J2.817eb	1.275	1.105	0.269	1.172	1.089	0.276
K.318d	1.052	0.828	0.408	1.041	0.809	0.419
K3.817ab	1.146	0.434	0.664	1.073	0.313	0.754
K4.5.817ac	0.566	-0.865	0.387	1.002	0.117	0.907
L2.318a	0.978	-0.142	0.887	1.016	0.209	0.834
L3.2.1_1217p	1.107	0.604	0.546	1.095	0.784	0.433
L3_1217r	1.125	2.004	**0.045**	1.11	2.169	**0.030**

Table 3.14: Outfit, Infit, and their standardization for Version 9

According to the general principles, all the Infit values for Version 1, 5, and 9 are between 0.5 and 1.5. Thus, we don't have to check the infit values anymore.

In version 1, the initial focus is on examining items A3.4.817fl and K.1017u. These two items exhibit very low difficulty parameters, indicating that they are relatively easy items. When analyzing the data, it becomes evident that some examinees with high scores and strong abilities have answered these questions incorrectly. This unexpected pattern of responses contributes to the high Outfit values for these items. For items with lower Outfit values, the data shows there are some good matches between the examinees' abilities and their responses, accounting for the lower Outfit values in such cases.

Same as version 1, in version 5, the items A4.4.2.817fk and K4.2.817eo should be checked first. The high outfit values are also due to the unexpected pattern of responses. The data which shows there are some good matches between the examinees' abilities and their responses is causing low Outfit values.

In version 9, we observed no instances of underfit, but there were multiple instances of overfit. Similar patterns were noted in Version 1, where the data revealed some strong connection between examinees' abilities and their responses, particularly for item D6.2.817hz,

[0]The numbers highlighted in red indicate that they fall outside the required range specified by the critical value.

resulting in lower Outfit values. Furthermore, our attention turned towards examining Zstd. Interestingly, only item D6.2.817hz exhibited significant p-values for Zstd. Simultaneously, through an analysis of both theoretical Item Characteristic Curves (ICC) and observed ICC plots, we made the decision to remove item D6.2.817hz from the assessment.

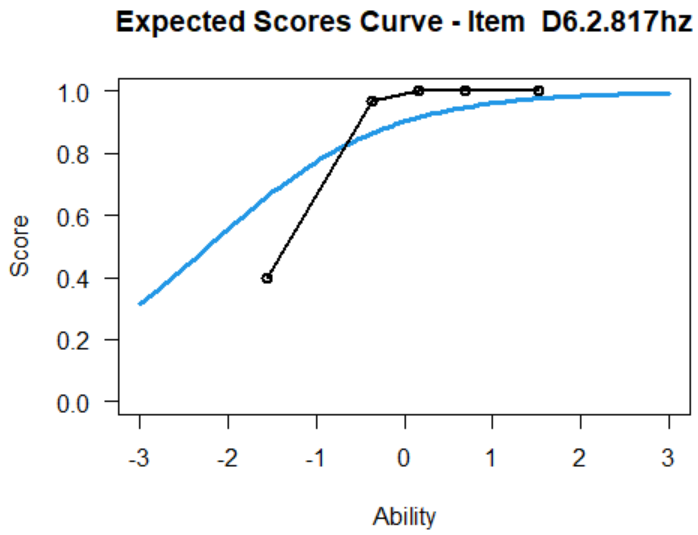

Figure 3.8: Expected and Observed ICC plot

For the other items, making a decision to delete them is not easy. This is because
(1) Sometimes, the underfit or overfit is due to an inefficient number of samples since an insufficient number of samples directly corresponds to a limited variety of response patterns.
(2) Removing the misfitting items only improved the results in the case of severe multidimensionality and a large proportion of misfitting items, and deteriorated them otherwise [24].

3.3.4 Person Fit

If an individual with a high ability fails to correctly respond to a simple item, their performance deviates from the model. Conversely, if a person with low ability successfully answers a highly challenging question, this also represents a deviation from the model. In practical terms, it is expected that a limited number of individuals may not conform closely to the model.

However, as long as the percentage of non-conforming examinees remains below 5%, we consider it acceptable. If less than 5% of respondents exhibit z standardized infit and outfit values exceeding 1.96 and -1.96, respectively, the model is considered satisfactory.

	Version 1		Version 5		Version 9	
	infit.outside	outfit.outside	infit.outside	outfit.outside	infit.outside	outfit.outside
Zstd>1.96	0.98850575	0.96551724	0.97894737	0.97894737	0.95785441	0.97318008
Zstd<-1.96	0.01149425	0.03448276	0.02105263	0.02105263	0.04214559	0.02681992

Table 3.15: the proportion of Zstd exceeding 1.96 and -1.96

As shown in Table 3.15 and Figures 3.9, 3.10, and 3.11 (the green bars indicates the person that fall outside the range between -1.96 and 1.96), less than 5% of respondents exhibit Zstd values exceeding 1.96 and -1.96, suggesting that the model is satisfactory.

3.3.5 Item-Person Map

The item-person map is also called the Wright map. This visualization initially displays the distribution of the latent ability within the analyzed samples. Then, it plots the difficulty of each item on the same theta scale. The alignment of these two plots enables us to assess the extent to which the items cover the latent ability.

Figures (3.12) - (3.14) show that the item difficulties are generally lower than the latent traits in Version 1, 5, and 9. Specifically, it is observed that approximately 50% of the items possess item difficulties that fall below the minimum latent trait value. This observation leads to the conclusion that the items are relatively easy for examinees to answer in general. This conclusion aligns with the fact that the bar for the exams is very high (85%). In order to maintain a proper pass rate, it follows that the exam items are designed to be not too challenging for examinees to answer.

3.3.6 Item Characteristic Curve

Each item has its own item characteristic curve (ICC). Since we are using the Rasch model, there is no item discrimination parameter. The only factor that determines how the ICC looks like is the item difficulty. The difficulty parameter in the ICC represents the level of ability at which there is a 50% chance of an examinee answering the item correctly. The discrimination parameter in the ICC determines how steep or shallow the curve is.

In our case, the point on the ability continuum at which there is a 50% probability of success on the item is different for different items. However, the steepness of the curve remains the same for different items.

Figure 3.15 shows an example of ICCs of three items that have different difficulty parameters. All the ICCs of items in Version 1, 5, and 9 are in the AppendixA.

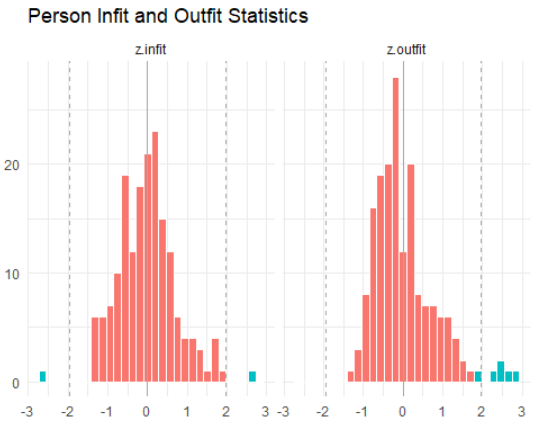

Figure 3.9: Person Fit of V1

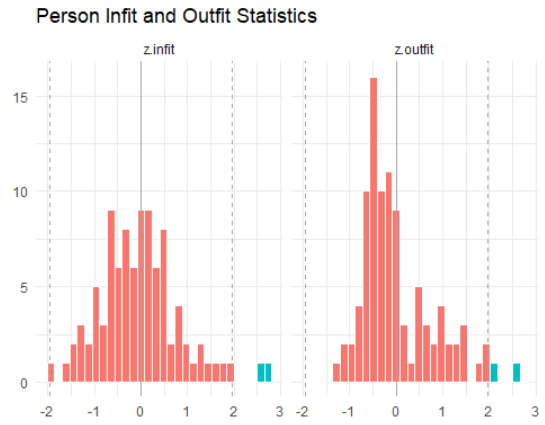

Figure 3.10: Person Fit of V5

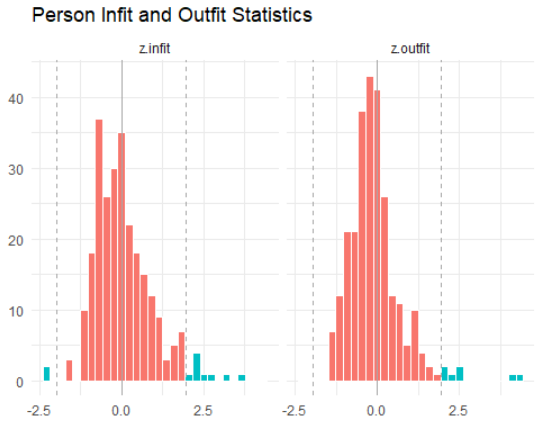

Figure 3.11: Person Fit of V9

3.4 Test Equating

There are two methods to do the test equating which have been introduced in Chapter 2. However, One benefit of separate calibration is that it facilitates examining item parameter estimates for the common items, while concurrent calibration can not. Concurrent calibration could be used as an adjunct to the separate calibration method [1].

After the selection of items from the previous steps, the items in each version and the number of their common items are shown in Table 3.16.

3.4.1 Separate Calibration

First, we construct a scatterplot of the IRT parameter estimates by plotting the parameter estimates for the common items between any two versions.

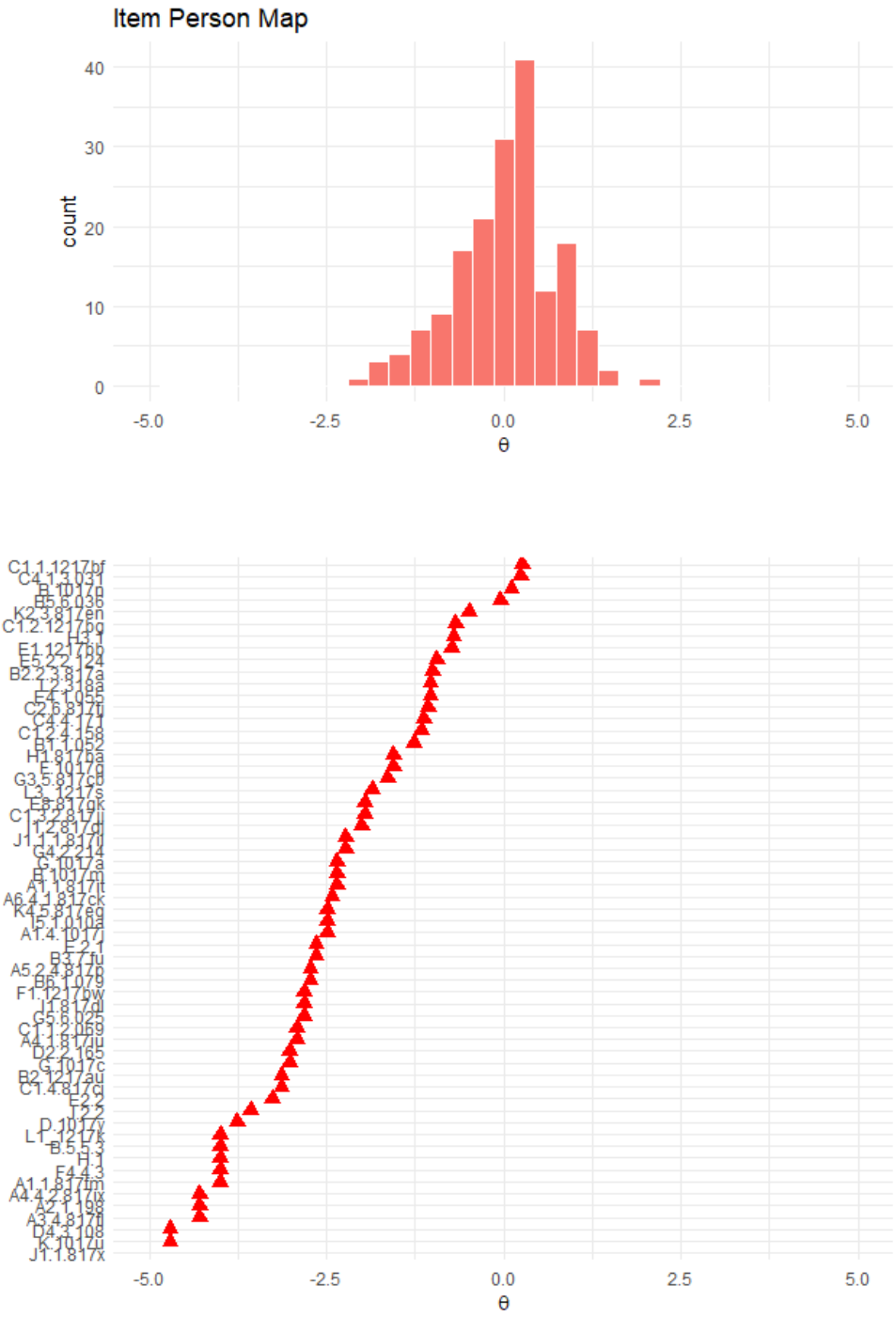

Figure 3.12: Item and Person Fit for Version 1

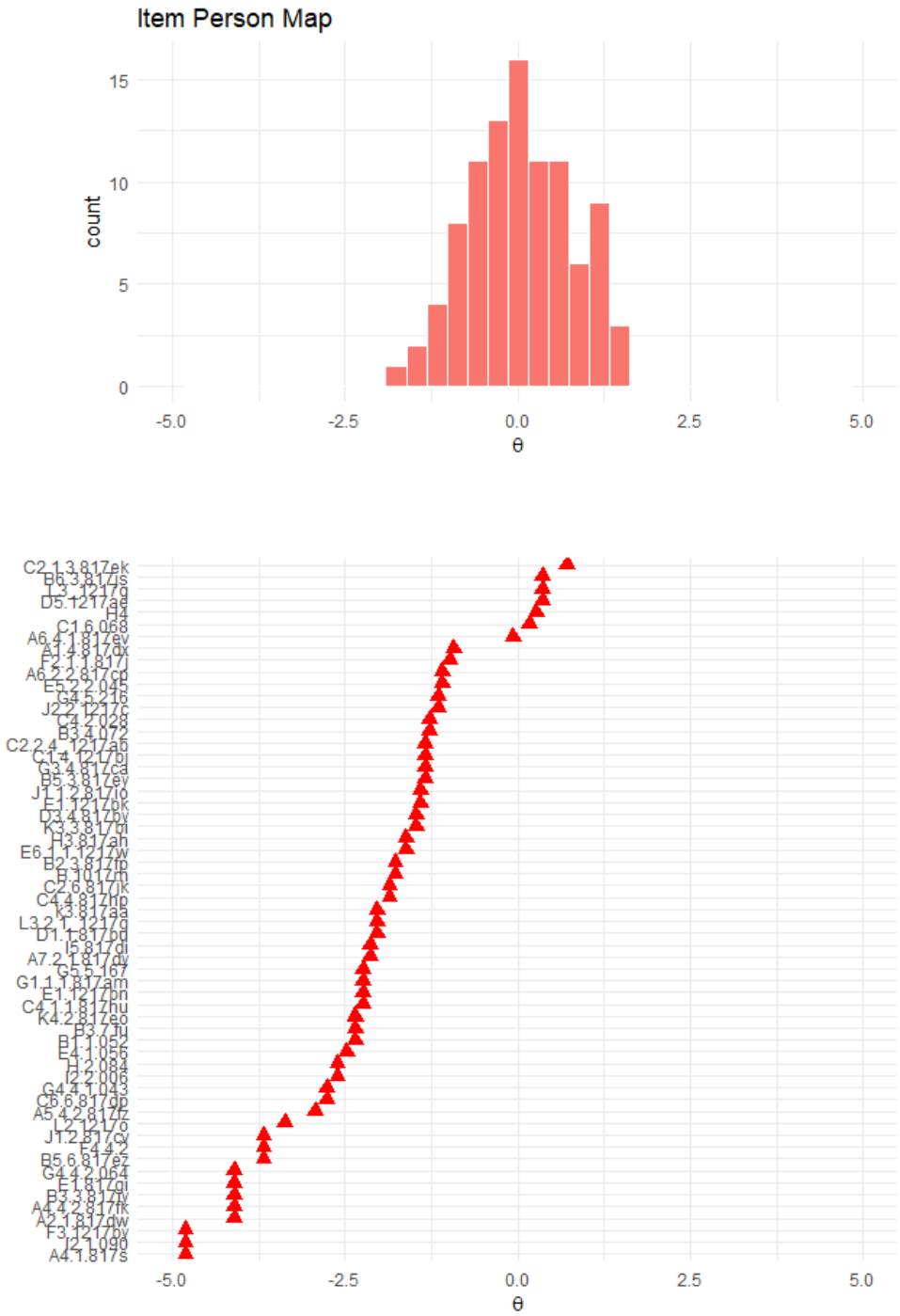

Figure 3.13: Item and Person Fit for Version 5

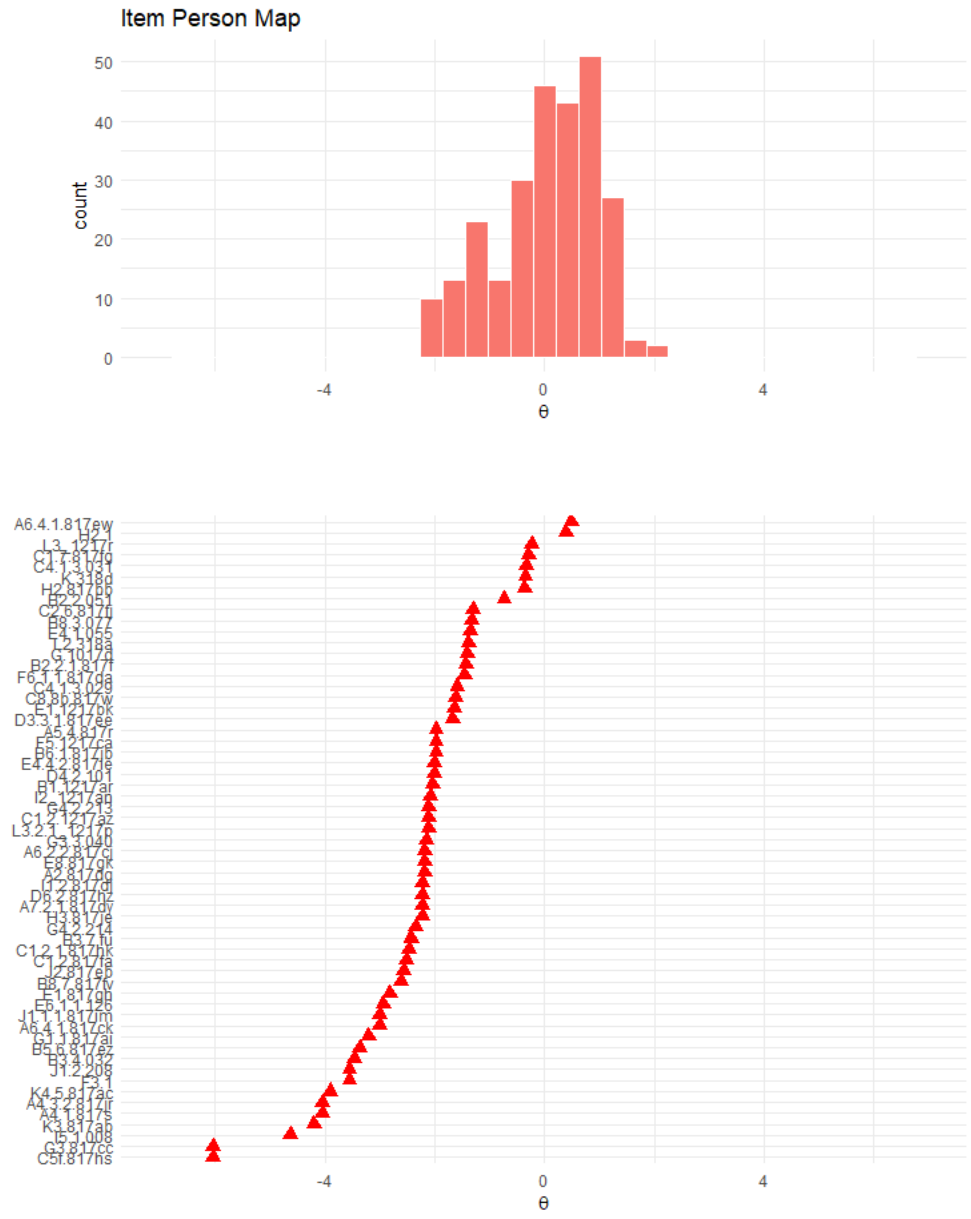

Figure 3.14: Item and Person Fit for Version 9

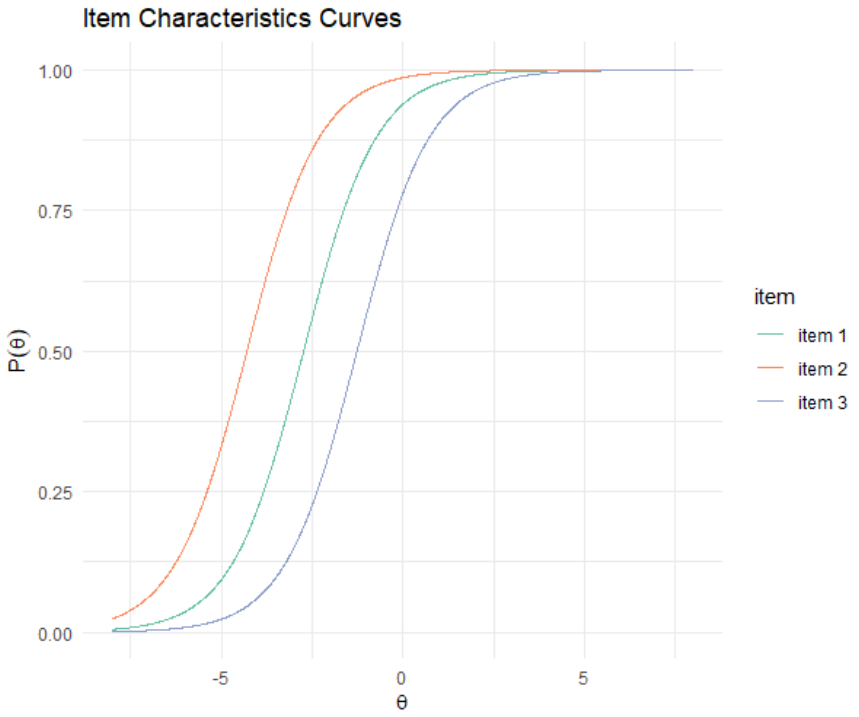

Figure 3.15: ICC Example

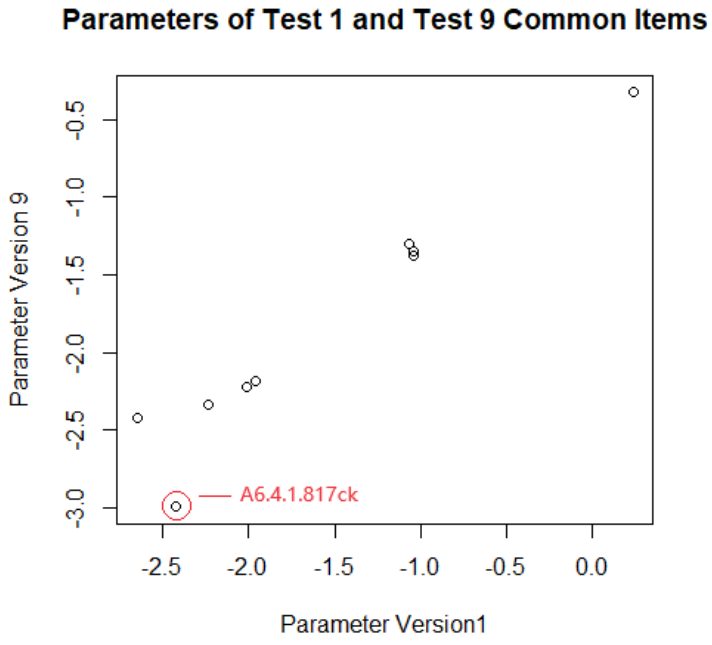

Figure 3.16: Scatterplot of the IRT parameters for common items between version 1 and 9

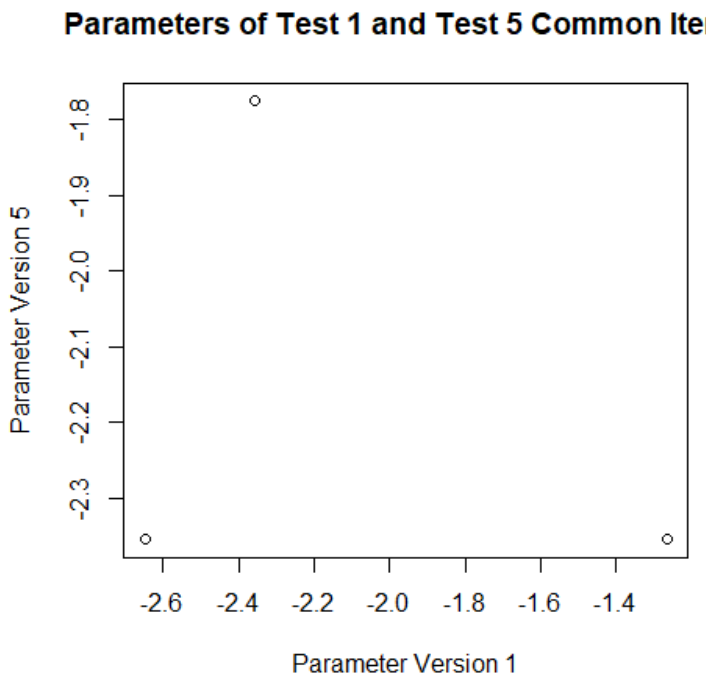

Figure 3.17: Scatterplot of the IRT parameters for common items between version 1 and 5

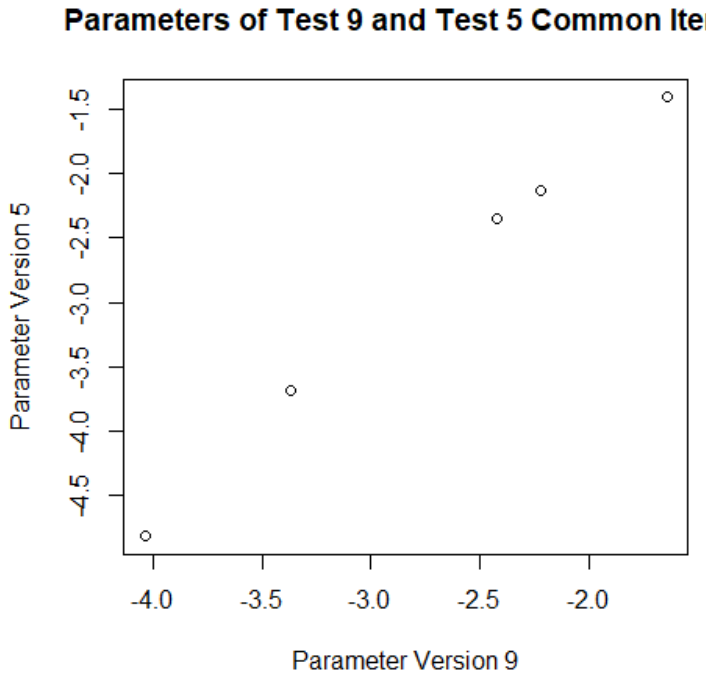

Figure 3.18: Scatterplot of the IRT parameters for common items between version 9 and 5

Common	v1	v5	v9
v1	59	3	9
v5	3	59	5
v9	9	5	58

Table 3.16: Common Items between any two versions

The parameter estimates for the common items are represented in Figure (3.16) - (3.18) to look for outliers - items with estimates that do not appear to lie in a straight line. In those illustrations, it is obvious that item A6.4.1.817ck in Figure 3.16 stands out as an outlier. Specifically, item A6.4.1.817ck has parameter estimates of -2.424 on Version 1 and -2.999 on Version 9. This item needs to be removed from the common item set. In Figure 3.18, all the spots are lying in a straight line which means there is no outlier. However, in Figure 3.17, due to the small number of common items, it's impossible to determine which one is the outlier. Due to this uncertainty, we can not involve version 5 to do the equating. However, whether removing an item or not is a judgemental process.

Mean/Mean Method

The mean/mean method uses equations (2.40) and (2.41) to estimate the A and B constants (scaling constants). Since we are using the Rasch model, the A constant is always 1 and the B constants are calculated.

Common Items	V1	V9	V1 to V9	Average
B3.7.fu	-2.647	-2.426	-2.87013	-2.64806
C2.6.817fj	-1.068	-1.304	-1.29113	-1.29756
C4.1.3.031	0.235	-0.326	0.011875	-0.15706
E8.817gk	-1.959	-2.186	-2.18213	-2.18406
E4.1.055	-1.037	-1.353	-1.26013	-1.30656
G4.2.214	-2.231	-2.342	-2.45413	-2.39806
I1.2.817dj	-2.009	-2.223	-2.23213	-2.22756
L2.318a	-1.037	-1.378	-1.26013	-1.31906
Mean	-1.46913	-1.69225		
A	1			
B	-0.22313			

Table 3.17: Commom-item estimates rescaled using mean/mean method

Since item A6.4.1.817ck has been removed, the number of common items is 8 now. The means after removing this item shown in Table 3.17 are the A- and B-constants. The scale of version 1 can be converted to the scale of version 9 using Equation (2.35). The result is

shown in the column V1 to V9. The item parameter estimates from the original scale and the transformed item parameter estimates from the target scale for the common items are averaged to obtain the final estimates [26]. The final estimates are shown in the column Average in Table 3.17.

3.4.2 Concurrent Calibration

In the concurrent calibration method [19], the parameters of items that belong to both exam versions are estimated together in a single calibration. In the context of non-equivalent common item equating, it's important to acknowledge that the abilities of the two populations are not equivalent. However, in accordance with the principle of Marginal Maximum Likelihood, it is necessary to make an initial assumption about the distribution of abilities within each population. As a consequence of this assumption, the combined population is treated as if it follows a normal distribution, even though it is a simplification that doesn't account for the actual differences in ability distributions between the two groups. However, the simulation study in [8] showed that the amount of bias was small.

Therefore, we ignore the bias and perform the concurrent calibration.

Separate Calibration		Concurrent Calibration	
Item	Difficulty	Item	Difficulty
C5f.817hs	-6.022	C5f.817hs	-5.947
G3.817cc	-6.022	G3.817cc	-5.947
J1.1.817x	-5.63	J1.1.817x	-5.53
D4.3.108	-4.927	D4.3.108	-4.826
K.1017u	-4.927	K.1017u	-4.826
I5.1.008	-4.609	I5.1.008	-4.535
A2.1.198	-4.511	A2.1.198	-4.41
A3.4.817fl	-4.511	A3.4.817fl	-4.41
A4.4.2.817ix	-4.511	A4.4.2.817ix	-4.41
A1.1.817fm	-4.214	A1.1.817fm	-4.112
B.5.5.3	-4.214	B.5.5.3	-4.112
F4.4.3	-4.214	F4.4.3	-4.112
H.1	-4.214	H.1	-4.112
L1_1217k	-4.214	K3.817ab	-4.112
K3.817ab	-4.186	L1_1217k	-4.112
A4.1.817s	-4.024	A4.1.817s	-3.95
A4.3.2.817jr	-4.024	A4.3.2.817jr	-3.95

D.1017y	-3.981	D.1017y	-3.878
K4.5.817ac	-3.881	K4.5.817ac	-3.808
J.2.2	-3.788	J.2.2	-3.685
F3.1	-3.537	F3.1	-3.464
J1.2.208	-3.537	J1.2.208	-3.464
E2.2	-3.481	E2.2	-3.377
B3.4.032	-3.441	B3.4.032	-3.369
B5.6.817ez	-3.353	B5.6.817ez	-3.28
B2.1217au	-3.353	B2.1217au	-3.249
C1.4.817ci	-3.353	C1.4.817ci	-3.249
D2.2.165	-3.238	D2.2.165	-3.133
G.1017c	-3.238	G.1017c	-3.133
G1.1.817ai	-3.192	G1.1.817ai	-3.12
A4.1.817iu	-3.133	A4.1.817iu	-3.027
C1.1.2.069	-3.133	C1.1.2.069	-3.027
F1.1217bw	-3.036	F1.1217bw	-2.93
G5.6.025	-3.036	G5.6.025	-2.93
I1.817dl	-3.036	I1.817dl	-2.93
J1.1.1.817im	-2.984	J1.1.1.817im	-2.913
A5.2.4.817p	-2.946	E6.1.1.126	-2.851
B6.1.079	-2.946	A5.2.4.817p	-2.839
E6.1.1.126	-2.921	B6.1.079	-2.839
E.2.1	-2.862	E.2.1	-2.755
E1.817gh	-2.804	E1.817gh	-2.734
A1.4.1017j	-2.709	A1.4.1017j	-2.6
I5.1.010a	-2.709	I5.1.010a	-2.6
K4.5.817eq	-2.709	K4.5.817eq	-2.6
B3.7.fu	-2.6375	B8.7.817fv	-2.526
B8.7.817fv	-2.596	B3.7.fu	-2.499
A1.1.817it	-2.572	J2.817eb	-2.479
B.1017m	-2.572	A1.1.817it	-2.462
G.1017a	-2.572	B.1017m	-2.462
J2.817eb	-2.548	G.1017a	-2.462
C1.2.817fa	-2.502	C1.2.817fa	-2.433
C1.2.1.817hk	-2.457	C1.2.1.817hk	-2.388

J1.1.1.817il	-2.446	J1.1.1.817il	-2.336
G4.2.214	-2.388	G4.2.214	-2.292
I1.2.817dj	-2.218	A7.2.1.817dy	-2.144
A7.2.1.817dy	-2.212	H3.817je	-2.144
H3.817je	-2.212	I1.2.817dj	-2.131
A2.817dg	-2.174	A2.817dg	-2.107
A6.2.2.817cj	-2.174	A6.2.2.817cj	-2.107
C1.3.2.817jj	-2.174	E8.817gk	-2.087
E8.817gk	-2.174	G3.3.040	-2.071
G3.3.040	-2.138	C1.3.2.817jj	-2.06
C1.2.1217az	-2.102	C1.2.1217az	-2.035
G4.2.213	-2.102	G4.2.213	-2.035
L3.2.1_1217p	-2.102	L3.2.1_1217p	-2.035
L3_1217s	-2.077	I2_1217an	-2
I2_1217an	-2.067	B1.1217ar	-1.966
B1.1217ar	-2.032	L3_1217s	-1.962
D4.2.101	-1.998	D4.2.101	-1.932
E4.4.2.817ie	-1.998	E4.4.2.817ie	-1.932
A5.4.817r	-1.965	A5.4.817r	-1.899
B6.1.817jb	-1.965	B6.1.817jb	-1.899
F5.1217ca	-1.965	F5.1217ca	-1.899
G3.5.817cb	-1.856	G3.5.817cb	-1.739
F.1017q	-1.775	F.1017q	-1.658
H1.817ba	-1.775	H1.817ba	-1.658
D3.3.1.817ee	-1.662	D3.3.1.817ee	-1.598
E1.1217bk	-1.634	E1.1217bk	-1.57
C8.8b.817w	-1.606	C8.8b.817w	-1.543
C4.1.3.029	-1.579	C4.1.3.029	-1.516
B1.1.052	-1.481	F6.1.1.817da	-1.385
F6.1.1.817da	-1.447	B2.2.1.817f	-1.36
B2.2.1.817f	-1.422	B1.1.052	-1.359
G.1017d	-1.397	G.1017d	-1.335
C1.2.4.158	-1.38	B8.3.077	-1.262
C4.4.171	-1.347	C1.2.4.158	-1.257
B8.3.077	-1.323	L2.318a	-1.232

L2.318a	-1.312	C4.4.171	-1.224
E4.1.055	-1.2995	E4.1.055	-1.218
C2.6.817fj	-1.291	C2.6.817fj	-1.205
B2.2.3.817a	-1.22	B2.2.3.817a	-1.096
E5.2.2.124	-1.159	E5.2.2.124	-1.034
E1.1217bb	-0.954	E1.1217bb	-0.826
H3.1	-0.926	H3.1	-0.797
C1.2.1217bg	-0.898	C1.2.1217bg	-0.768
B2.2.051	-0.732	B2.2.051	-0.676
K2.3.817en	-0.706	K2.3.817en	-0.574
H2.817bb	-0.364	H2.817bb	-0.311
K.318d	-0.345	K.318d	-0.293
C1.7.817fg	-0.289	C1.7.817fg	-0.237
B5.6.036	-0.263	L3_1217r	-0.182
L3_1217r	-0.233	B5.6.036	-0.123
C4.1.3.031	-0.153	C4.1.3.031	-0.093
B.1017n	-0.109	B.1017n	0.033
C1.1.1217bf	0.046	C1.1.1217bf	0.191
H2.1	0.391	H2.1	0.436
A6.4.1.817ew	0.484	A6.4.1.817ew	0.528
		A6.4.1.817ck	-2.742

Table 3.18: Rescaled item parameter estimates

3.4.3 Comparison

The items have been arranged in order of their difficulties from the smallest to the largest. The items that are marked red are the ones that are not in the same position as the other ones for the other category. In general, most of the items are in the same position. Even for the unpaired items, their positions don't change a lot from the other ones, which indicates that the equating result is acceptable.

It's evident that there is an item A6.4.1.817ck appearing in the concurrent calibration section while it has already been removed from the separate calibration process. This is the

[0]The item names highlighted in red in one column indicate that they are not in the same position as the other

feature of concurrent calibration in that there is no selection step but putting every item together and doing a single run.

Therefore, we think the separate calibration provides a reliable result, and all the items can become an item bank.

CHAPTER 4

Discussion & Conclusion

The ultimate goal of this thesis is to establish an item bank for the medical center, which will facilitate future exam item allocation and lead to the creation of more effective and comprehensive exam papers for assessing student proficiency. However, the limitations imposed by the sample size pose challenges. For instance, although the medical center has ten different versions of exam papers, most of them lack a sufficiently large sample size, particularly for conducting IRT analysis.

Out of the ten versions of exam papers, versions 1, 5, and 9 initially appeared to have reasonably adequate sample sizes for analysis. However, as suggested in some literature, for example [3], said that the Rasch model needs fewer samples to estimate the parameters than 2PL, while at least 200 samples are needed. In general, it is recommended that a sufficient sample size for IRT analysis should ideally amount to at least five times the number of items or a minimum of 300. it is obvious that during our implementation version 1, 5, and 9 did not provide a sufficient sample size for both Rasch and 2PL as suggested. For version 9, the presented issue is that the response patterns did not converge for 2PL parameter estimation, considering it doesn't provide enough empirical data. In the case of both version 1 and 5, they are able to produce the estimation results. However, to guarantee stable estimations and ensure alignment with version 9, it is better to apply the Rasch model to these versions.

Furthermore, it's important to note that the data collection process was not carried out in an ideal way. Typically, when the intention is to perform IRT analysis and create a solid item bank, it is advisable to administer different exam papers to specific populations with known characteristics. This includes ensuring that these populations are either in the same academic grade or possess equivalent educational backgrounds in the subject that is tested by the exam papers. Additionally, it is recommended to have a sufficient number of common items among the exam papers, ideally at least one-fifth of the total number of items in each paper. The medical centers can intentionally administer the exam papers to targeted students and get enough data to have more stable and accurate results.

Upon conducting our item fit assessment and local dependency detection, most of the items exhibit good fit within the Rasch model. Nevertheless, there are instances where certain item pairs fail to meet the Q_3 critical value threshold, and the item fit statistics fall outside the acceptable range. Throughout our analysis, no evident violations were observed at either the item content or response pattern levels. Consequently, only one item was deleted. Additionally, the deleted item is not considered a bad item since we can only consider that this specific group answered this item has some unexpected response or the amount of sample size is not sufficient to provide stable information.

As the scatterplot of common item parameter estimates was shown, the common item parameter estimates were lying in a straight line. This pattern typically indicates that the items exhibit measurement invariance across these different exam forms, meaning that their relationship with the latent trait under assessment remains stable and parallel. Thus, this is the other way to illustrate that there is a stable measurement during the estimation process.

The item bank under development is undergoing two equating methods: the mean/mean method and concurrent calibration. Our findings indicate that, following the scale conversion and concurrent run, no distinct differences are observed between the two methods. The item bank that is made is reliable and can be used in the future. Since we don't adopt version 5 to do the equating due to the small amount of common items between version 1 and 5. However, the common item parameter estimates scatterplot between versions 5 and 9 has a good outcome, in which all of the spots are in a straight line.

When comparing Item Response Theory (IRT) and Classical Test Theory (CTT) in terms of constructing an item bank, a notable distinction arises. In CTT, there is no method to establish a relationship for converting item difficulty parameters and putting them on a common scale, especially when dealing with non-equivalent populations. CTT equating methods primarily focus on equating students' scores.

In contrast, IRT offers the advantage of equating item characteristics in a lot of situations. This capability allows us to employ IRT effectively in the development of an item bank, making it work for applications like Computer Adaptive Testing (CAT). In CAT, the next item a student will answer is determined by their performance on the last item, exemplifying a dynamic approach to assessment. In the future, after we have a larger item bank with difficulties covering a larger range, we can build a CAT system for the pharmacotherapy exam.

APPENDIX A

Item Characteristic Curves

The Item Characteristic Curves for each item in each version of exams are shown in this appendix.

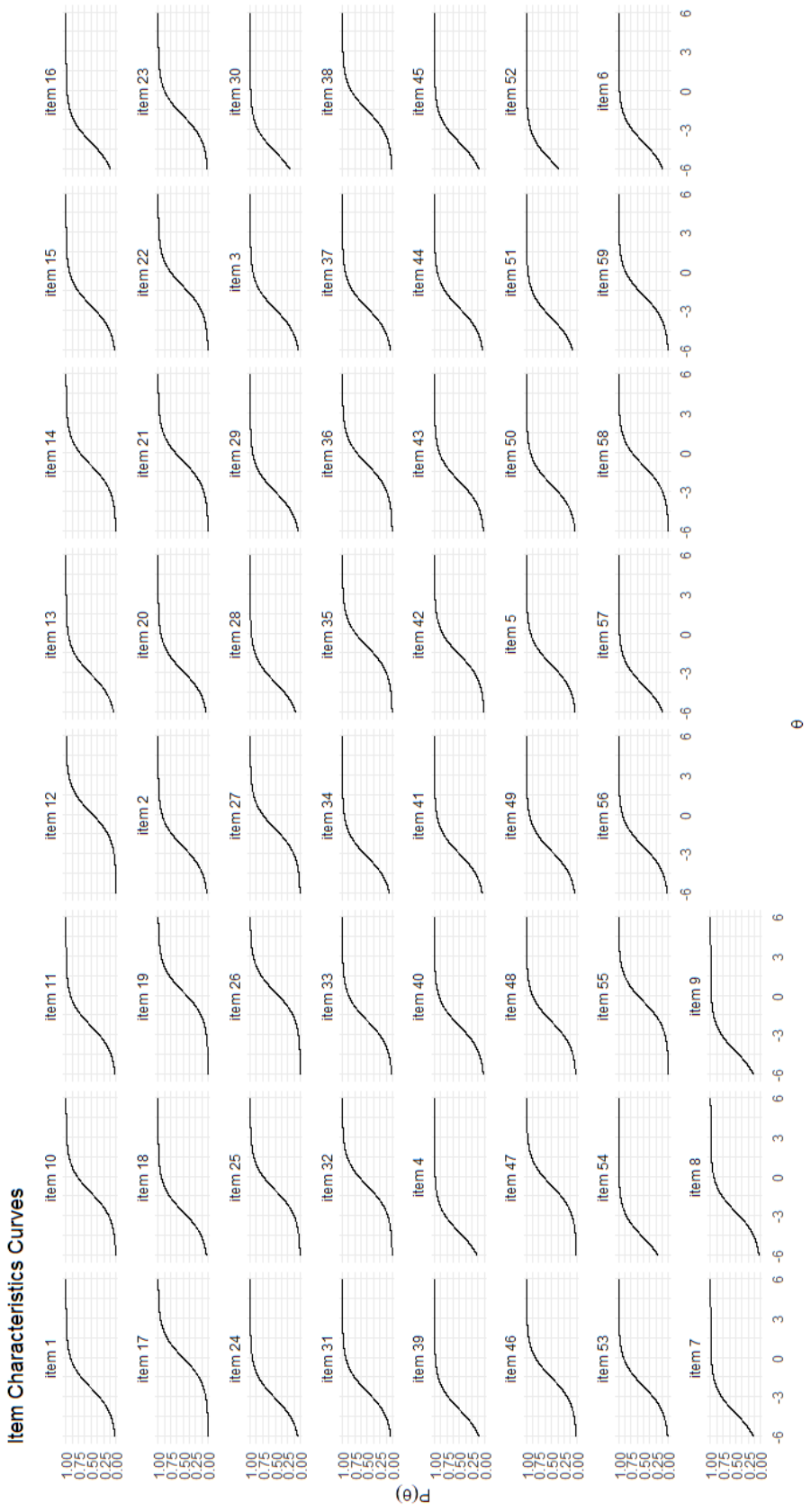

Figure A.1: Item Characteristic Curve for version 1

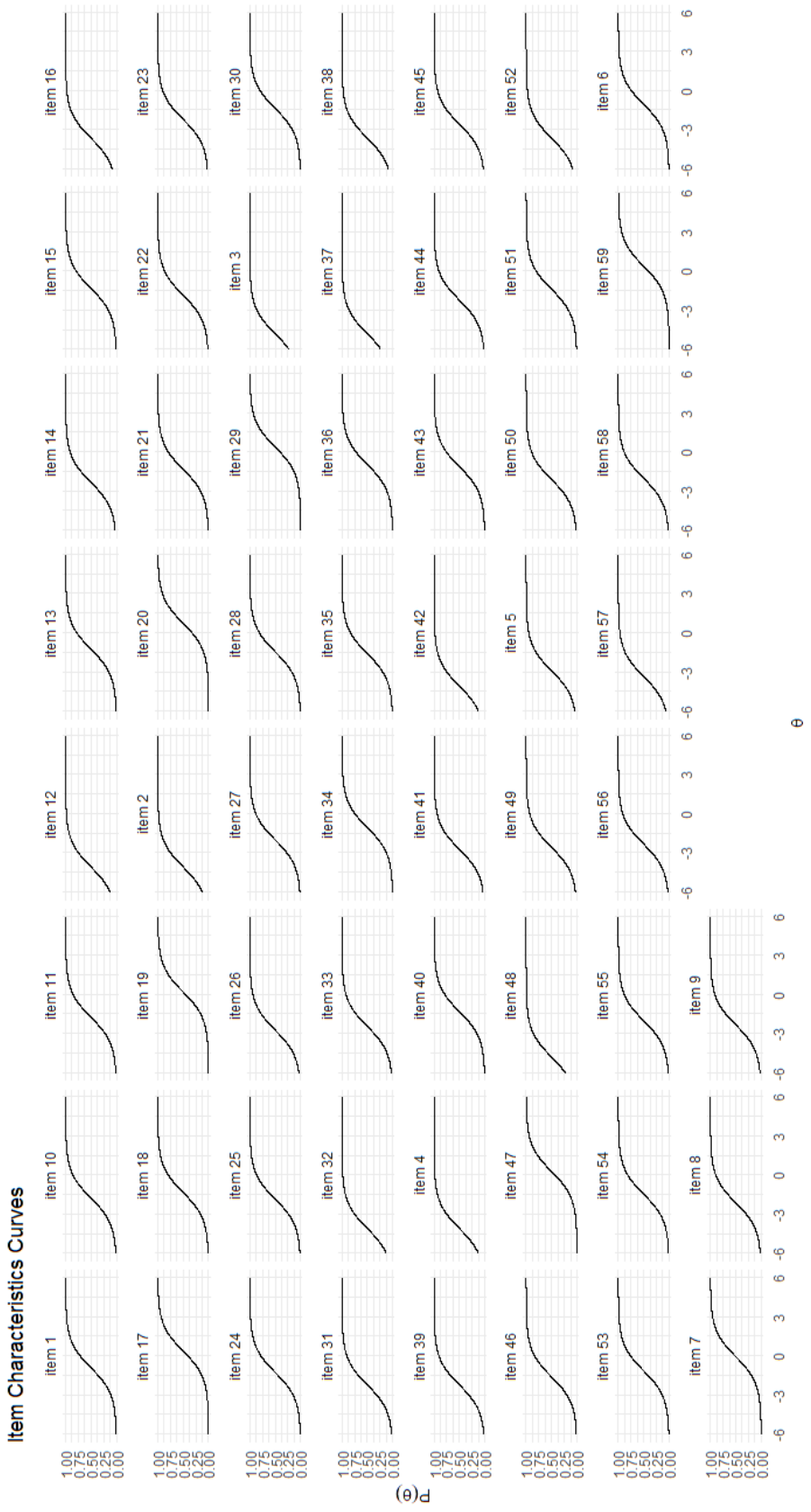

Figure A.2: Item Characteristic Curve for version 5

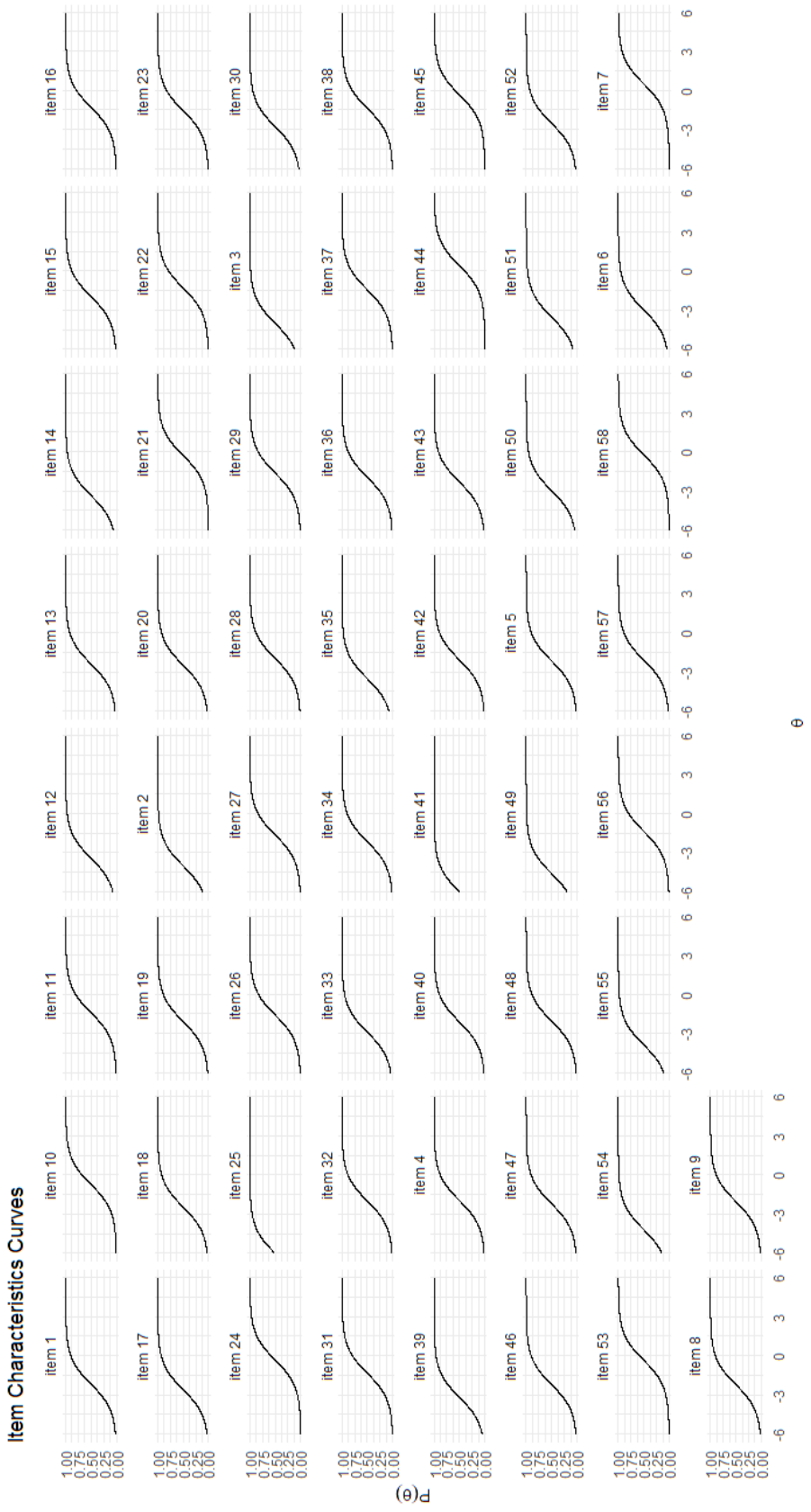

Figure A.3: Item Characteristic Curve for version 9

BIBLIOGRAPHY

[1] Michael J. Kolen, Robert L. Brennan. *Test Equating, Scaling, and Linking: Methods and Practices*. Springer New York, 2014.

[2] Gaborieau JB., Pronello C. Validation of a unidimensional and probabilistic measurement scale for pro-environmental behaviour by travellers. *Transportation 48, 555–593*, 2021.

[3] Cappelleri J. C., Jason Lundy J., Hays R. D. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical therapeutics, 36(5), 648–662*, 2014.

[4] Chang S., Hanson B., Harris D. A standardization approach to adjusting pretest item statistics. In *the Annual Meeting of the National Council on Measurement in Education*, 2000.

[5] Loyd B. H., Hoover H. D. Vertical equating using the rasch model. *Journal of Educational Measurement, 17, 179–193.*, 1980.

[6] Awopeju O. A., Afolabi E. R. I. Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, ESJ, 12(28), 263*, 2016.

[7] Warrens M. J. On cronbach's alpha as the mean of all split-half reliabilities. In *Quantitative psychology research, 293-300*, 2015.

[8] Marie Wiberg Jorge González. *Applying Test Equating Methods: Using R*. Springer Cham, 2017.

[9] Ashraf Z. A., Jaseem K. Classical and modern methods in item analysis of test tools. *International Journal of Research and Review, 7(5), 397-403*, 2020.

[10] Baylor C., Hula W., Donovan N. J., Doyle P. J., Kendall D., Yorkston K. An introduction to item response theory and rasch models for speech-language pathologists. *American journal of speech-language pathology, 20(3), 243–259*, 2011.

[11] Frank B. Baker, Seock-Ho Kim. *The Basics of Item Response Theory Using R*. Springer, 2017.

[12] Marco G. L. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14, 139–160*, 1977.

[13] Traub, Ross E., Rowley, Glenn L. Traub, r.e. and rowley, g.l. *Educational Measurement: Issues and Practice, 10, 37-45*, 1991.

[14] Battauz M. Irt test equating in complex linkage plans. *Psychometrika, 78(3), 464–480*, 2013.

[15] Bock R.D., Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika 46, 443–459*, 1981.

[16] Christensen K. B., Makransky G., Horton M. Critical values for yen's q3: Identification of local dependence in the rasch model using residual correlations. *Applied psychological measurement, 41(3), 178–194*, 2017.

[17] Linacre M. Teaching rasch measurement. *Trans. Rasch Meas. 31, 1630–1631*, 2017.

[18] Linacre J. M. What do infit and outfit, mean-square and standardized mean? In *Rasch Measurement Transactions, 16*. 2002.

[19] Lord F., Wingersky M. Comparison of irt true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8(4), 453–461*, 1984.

[20] Müller M. Item fit statistics for rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications, 7(5)*, 2020.

[21] Wilson E. B., Hilferty M. M. The distribution of chi-square. *Proceedings of the National Academy of Sciences, 17(12), 684-688.*, 1931.

[22] Yen W. M. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8(2), 125-145*, 1984.

[23] Miller M.D. Classical test theory reliability. In *International Encyclopedia of Education, 27-30*. 2010.

[24] Crişan D. R., Tendeiro J. N., Meijer R. R. Investigating the practical consequences of model misfit in unidimensional irt models. *Applied Psychological Measurement, 41(6), 439-455*, 2017.

[25] Smith A.B., Rush R. Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodol 8, 33*, 2008.

[26] Kim S. H., Cohen A. S. A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22(2), 131-143*, 1998.

[27] Wang W. C., Chen C. T. Item parameter recovery, standard error estimates, and fit statistics of the winsteps program for the family of rasch models. *Educational and Psychological Measurement, 65(3), 376-404*, 2005.

[28] Paul Boeck, Mark Wilson. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer, 2004.

[29] Harwell, Michael R., Frank B. Baker, Michael Zwarts. Item parameter estimation via marginal maximum likelihood and an em algorithm: A didactic. *Journal of Educational Statistics, 13(3), 243–271*, 1988.