An Analysis of Item Response Theory

by

Stuart Jones

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 3, 2019

Keywords: Item Response Theory, IRT, statistical learning, machine learning, testing theory, measurement theory

Approved by

Mark Carpenter, Professor in College of Mathematics & Statistics
Nedret Billor, Professor in College of Mathematics & Statistics
Asher Abebe, Professor in College of Mathematis & Statstics

Abstract

This paper examines a field of psychological and educational measurement testing theory called Item Response Theory. The paper delves into the basics of the theory, its theoretical and statistical background, and discusses the usefulness. The primary method of parameter estimation, Birnbaum's Paradigm using Newton-Raphson method and maximum likelihood estimation, is discussed with a brief overview of the mathematics involved. The bulk of the paper focuses on a data set of Medical School Admission Test in Biology (MSATB) data administered to hopeful medical students in the Czech Republic. This data of 1,407 individuals and a subset of 20 questions selected from the overall test make up the set analyzed here. The main tools of Item Response Theory are used upon the data to test different model fits and produce graphics and charts to visualize the results. After the best-fitting model is selected for the data set, it is treated as a item bank to create a five-question subtest designed for pass/fail results; the analysis on this subtest shows it working fairly well in discriminating between ability levels. Finally, statistical machine learning methods are utilized to classify students based on ability level and to identify possible clusterings of students present in the data. Success was found both in classifying students based on performance and in identifying clusters in the data.

Acknowledgments

Special thanks to my adviser, Dr. Mark Carpenter, for his continuous support and assistance in this endeavor and to Dr. Nedret Billor and Dr. Ash Abebe for their time and support in writing this.

Table of Contents

List of Figures

Chapter 1

Introduction

In the fields of educational and psychological testing and measurement, we often encounter situations where there is an underlying variable of interest; in educational settings, this is typically thought of as "achievement" or "mastery level" or "intelligence", and this notion has become central to these areas. We sometimes refer to a person as intelligent or having a high aptitude, and when we do so, others instinctively can understand what is meant by such. Measuring this idea of scholastic achievement, or more generally, intelligence, has proven to be an ongoing desire in academic areas. Yet, academic settings often find themselves plagued with measuring this academic ability objectively and in a way that is statistically sound. The field of item response theory (IRT) has arisen to fill this void by developing a framework that can place modern academic testing in a proper statistics setting to arrive at the most accurate estimate of this ability trait as possible.

Now, if one is going to measure how much of this trait "ability" a person has, we have to create a metric to place the measurement upon. Since this ability trait is, by definition, an abstract construct, assigning it an objective metric with a true origin and giving meaning to the numbers on the scale is a difficult task. In Chapter 4, we will address this problem more rigorously with the idea of test calibration. For now, though, we must arbitrarily define an ability scale measuring some trait θ. This scale will have a midpoint of 0 and a unit of measurement of 1. Notice this is not a ratio scale of measurement, since our 0 is arbitrarily defined - a "2" on this ability scale will not necessarily mean twice as much ability as a "1" on the scale. It is said to be on an interval scale, then, ranging from $-\infty$ to ∞, but generally ability scores will range from -4 to 4. The reason for this is that the curve (defined later) that places ability on this scale will be a cumulative distribution function (cdf) that is typically clustered near 0, as stated by Baker & Kim [1].

The idea to measure this ability and place it on the scale we have created is to administer a test made of items. Each item on this test measures (in theory) this ability that we are interested in and provides information that we can statistically quantify to make an assertion about the true ability level of the person. In practice, these tests are almost all multiple choice due to constraints on scoring (free response items are very difficult to objectively, reliably, and efficiently score), and each of these items is marked either correct or incorrect. This means that each item is dichotomously scored. (There is IRT theory developed for polytomously scored items, but that will not be explored here.)

Item Response Theory differs from traditional "Classical Test Theory" (CTT) in that CTT focuses almost solely on test-wide results and statistics, whereas IRT focuses mostly on the item-level [1]. There are advantages to IRT that CTT is not afforded, mainly that the test, items, and results of CTT are all dependent on each other, meaning that test results cannot be interpreted outside of the individual test administration and the specific students that took it. Under IRT, test items and test-taker ability levels are independent of the individual test administration and of each other [1]. This has given IRT a huge advantage in the last few decades, and it is now the most widely used test analysis method for large-scale testing.

In the modern age, many of the standardized tests now utilize IRT methodology in assessing and scoring their products, such as the GRE and the GMAT. This paper will dive into the inner workings of IRT and explore the statistical underpinnings of the theory. In addition, this paper will use statistical learning theory to compare results to traditional IRT output as possible route of improving upon existing schemes. First, this paper will discuss the ideas that form the foundation of the theory; then, the process of item parameter estimation, ability parameter estimation, and Joint Maximum Likelihood Estimation will be visited. We will see how information theory plays a role in IRT and can be used to develop the best tests, and we will also analyze a data set of test scores using Item Response Theory procedures and compare the results with statistical learning theory techniques. The central question that will be posed here is if modern statistical (and machine) learning techniques can produce equally good or perhaps even better statistical results than traditional Item Response Theory.

Chapter 2

Models of Item Response Theory

As noted in Chapter 1, we will note the ability level or score of an examinee by θ. It stands to reason, then, that at each ability level θ, there is a probability associated with it that can be interpreted as the probability of responding correctly to the particular item that is being tested. We will denote this by $P(\theta)$. If this function is plotted over the entire range of values ($-\infty$ to ∞) that θ can take, the result is a roughly S-shaped curve. [1] See Figure 2.1.

Near an ability level of negative infinity, the probability of a correct response approaches 0. Alternatively, as ability level approaches positive infinity, the probability of a correct response approaches 1. This curve establishes the relationship between the probability of a correct response and the ability scale being used. This curve is called the item characteristic curve (ICC), and every item on the test has its own curve. It is this ICC that much of Item Response Theory builds upon, and the theory of the ICC relates to all other areas of IRT. There are two (or sometimes three) parameters that we are interested in when specifying the ICC. Together, these parameters will totally specify the curve. The first is a location parameter, β, that describes where along the ability scale the median of the curve is located; an examinee has a 0.5 probability of answering correctly (and a 0.5 probability of answering incorrectly) if their ability level coincides with β on that particular item. Since β determines the median of the ICC, it is often referred to as the "item difficulty" parameter [1], as a measure of the item's hardness.

The second parameter of import when discussing the ICC is the scale parameter, $alpha$. It describes the slope of the ICC at β. It is a measure of how quickly the curve changes in steepness as it moves from left to right. The parameter α is often called the "item discrimination" parameter because of its role in differentiating or discriminating between different ability levels. A higher α means a small change in θ near β results in a great change in probability;

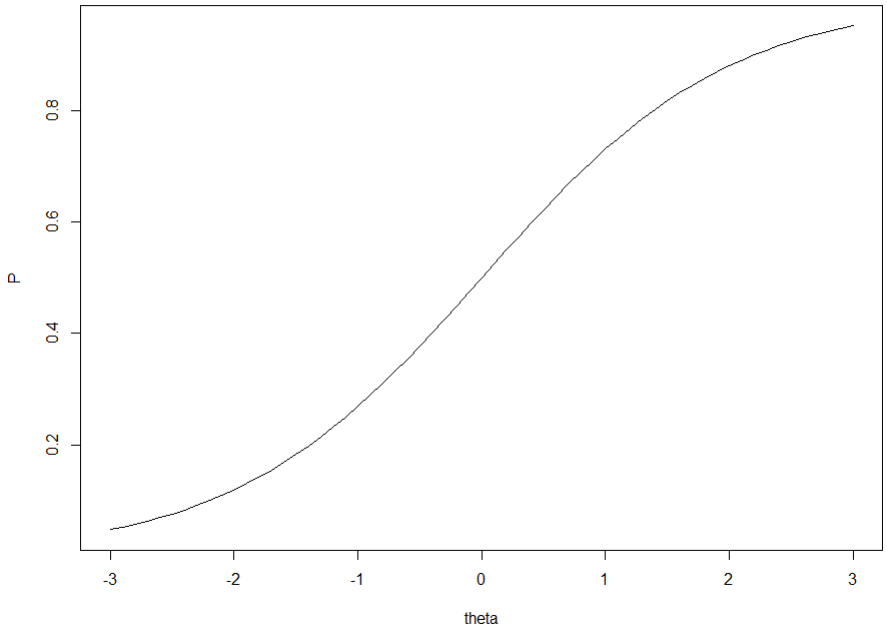


Figure 2.1: A basic Item Characteristic Curve with $\alpha = 1$ and $\beta = 0$

likewise, a lower $alpha$ level means the curve is "flatter" near β and differentiates less between different θ levels. A high α level is, then, useful in situations where a cutoff score is important, as it will highly differentiate examinees on one side or the other. While both α and β theoretically range from $-\infty$ to ∞, they usually occur over only a small range. Typical values of β range from -1 to 1, while typical values of α range from 0.5 to 2 [1].

In Item Response Theory, the ICC is defined formally as $P_i(\theta) = P(\beta_i, \alpha_i, \theta)$ for item i on a test. A wide variety of functions (infinite, in fact) could be used to model the ICC. However, the S shape and upper asymptote of 1 and lower asymptote of 0 lends to the idea of using a cumulative distribution function. In practice, two are used - the normal CDF and the logistic CDF. The normal CDF is used because of its popularity in statistical theory, but the prevailing model used in IRT literature is the logistic CDF. Here, the focus will be on the logistic CDF for the reason that the two can be made virtually equivalent with a simple transformation -It can be shown using gradient descent on the maximum deviations between the distributions that if a scaling constant of 1.702 is multiplied by α for the logistic CDF, the normal and logistic CDF differ by less than 0.01 over $-\infty$ to ∞.

The major models used in Item Response Theory for dichotomous items will be discussed here. The most common model used is the two-parameter logistic (2PL) function [2]:

$$\frac{e^{L_i}}{1+e^{L_i}} = \frac{1}{1+e^{-L_i}}, \tag{2.0.1}$$

where $L_i = \alpha_i(\theta - \beta_i)$ is the logit. Alternatively, if equating to the normal CDF, $L_i = \alpha_i D(\theta - \beta_i)$, where $D = 1.702$. Both forms are used in various settings. Notice the 2PL involves a different item discrimination and difficulty for each item i. This allows great flexibility when finding item parameters and allows most items to fit the model well.

The other commonly-used model is the Rasch model [2], using equation 2.0.1, where $L_i = \theta - \beta_i$. Notice, the item discrimination parameter α is set to 1. This model is used frequently due to its ease of computation. It assumes that all items have a fixed item discrimination of 1, however, which may or may not be a realistic assumption.

Another model used sometimes in place of the Rasch model is the one-parameter logistic (1PL) [2], using equation 2.0.1, where $L_i = \alpha(\theta - \beta_i)$. Here, the α is fixed as well, but at a value other than 1. This is a useful model when all items are roughly the same level of discrimination.

Finally, we have the three-parameter logistic function (3PL) that introduces a third parameter [2]:

$$c_i + (1 - c_i)\frac{1}{1+e^{-L_i}}, \tag{2.0.2}$$

where $L_i = \alpha_i(\theta - \beta_i)$. This new parameter c is often called the pseudo-guessing parameter: It is the probability that an examinee get an item correct merely by guessing, and it represents a lower asymptote other than 0 of the ICC. This model is sometimes used, but it has several problems with use, the most grievous of which is that it is not technically a logistic CDF anymore, so it loses many of the mathematical properties. This also presents problems with accurate item parameter estimation, which will be discussed in Chapter 3.

Chapter 3

Classical Parameter Estimation Techniques

The methods in this chapter are taken from Baker and Kim (2004) [2] and were first established by Birnbaum in the 1960s [4]. The primary goal of placing examinees on an ability scale lends itself to one primary objective: estimating the item and ability parameters for the item characteristic curves of all items on the test and all ability parameters of all examinees who took the test. For the present, it is convenient just to focus on one item's ICC and a group of examinees of one particular, known, ability level.

Let the cumulative logistic distribution function be given by

$$P_{i,j} = P(\alpha, \beta, \theta_j) = \frac{1}{1 + e^{-\alpha+\beta\theta_j}} \tag{3.0.1}$$

for an item i and a group of examinees j with ability level θ_j, where $Z_j = \alpha + \beta\theta_j$ is the logit. (We will omit the index i on the following derivations simply because the calculations are for a fixed item i.) We will be proceeding by maximum likelihood estimation, so some derivatives we will need are:

$$\frac{\partial P_j}{\partial \alpha} = P_j Q_j, \tag{3.0.2}$$

where $Q_j = 1 - P_j$, and

$$\frac{\partial P_j}{\partial \beta} = P_j Q_j \theta_j \tag{3.0.3}$$

Also,

$$\frac{\partial Q_j}{\partial \alpha} = -P_j Q_j \tag{3.0.4}$$

and

$$\frac{\partial Q_j}{\partial \beta} = -P_j Q_j \theta_j \tag{3.0.5}$$

. Suppose the k groups of f_j subjects with known ability scores θ_j are drawn at random from a population of persons and $j = 1, \ldots, k$. The subscript i will be left out here just for clarity, but it is understood that each subject has responded to a single item i. From the f_j people having ability score θ_j, r_j gave the correct response and therefore $f_j - r_j$ gave the wrong response. the observed proportion of correct responses is $\frac{r_j}{f_j}$ and $\frac{f_j - r_j}{f_j}$ is the proportion of incorrect responses. Now, let $R = (r_1, r_2, ...r_k)$ be the vector of observed number of correct responses, where k is the number of groups of examinees grouped by ability. It is assumed under IRT that the observed r_j at each ability level θ_j are distributed binomally with parameters f_j and P_j (shown above), which is the true probability of a correct response .We also assume, as a chief assumption of IRT that items and groups are independent of one another. Then, for the vector of observed correct responses, R, we have the joint probability of R as the likelihood function

$$P(R) = \Pi_{j=1}^{k} \frac{f_j!}{r_j!(f_j - r_j)!} P_j^{r_j} Q_j^{f_j - r_j}$$

Then, the log-likelihood is given by:

$$\log P(R) = constant + \sum_{j=1}^{k} r_j \log P_j + \sum_{j=1}^{k} (f_j - r_j) \log Q_j$$

The first derivatives of $L = \log P(R)$ are:

$$\frac{\partial L}{\partial \alpha} = L_1 = \sum_{j=1}^{k} \frac{r_j}{P_j} P_j Q_j + \sum_{j=1}^{k} \frac{f_j - r_j}{Q_j} (-P_j Q_j)$$

$$= \sum_{j=1}^{k} [r_j Q_j - (f_j - r_j) P_j] = \sum_{j=1}^{k} [r_j Q_j - f_j P_j + r_j P_j]$$

$$= \sum_{j=1}^{k} (r_j - f_j P_j) = \sum_{j=1}^{k} f_j (p_j - P_j),$$

where $p_j = \frac{r_j}{f_j}$. And

$$\frac{\partial L}{\partial \beta} = L_2 = \sum_{j=1}^{k} \frac{r_j}{P_j} P_j Q_j \theta_j + \sum_{j=1}^{k} \frac{f_j - r_j}{Q_j} (-P_j Q_j \theta_j)$$

$$= \sum_{j=1}^{k} (r_j - f_j P_j) \theta_j$$

$$= \sum_{j=1}^{k} f_j (p_j - P_j) \theta_j$$

When we set both derivatives equal to 0, the resulting simultaneous equations can, in theory, be solved for the maximum likelihood estimates for α and β:

$$L_1 = \sum_{j=1}^{k} f_j (p_j - P_j) = 0$$

and

$$L_2 = \sum_{j=1}^{k} f_j (p_j - P_j) \theta_j = 0$$

Solving the equations is particularly difficult, and so a numerical procedure is classically done to solve them, an iterative procedure based on a Taylor series called the Newton-Raphson method. For this method, an initial estimate of α and β are needed, let's say $\hat{\alpha_1}$ and $\hat{\beta}_1$. Then, the Newton-Raphson equation is:

$$\begin{bmatrix} \hat{\alpha_{t+1}} \\ \hat{\beta_{t+1}} \end{bmatrix} = \begin{bmatrix} \hat{\alpha_t} \\ \hat{\beta_t} \end{bmatrix} - \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}_t^{-1} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}_t ,$$

where the inverted matrix is the Hessian, the matrix of second partial derivatives. The equation above is iterated until successive iterations result in a change of parameter estimates less than a specified threshold. When that occurs, the current estimates are considered the parameter estimates for α and β.

The above item estimation procedure assumes that ability is known. To estimate ability, we assume that item parameters are known. In a similar process to the above, we proceed. A given examinee responds to n items on a test and the responses are dichotomously scored u_{ij}, where i designates the ith item and j designates examinee j, yielding a vector of item

responses of length n denoted by $U_{ij} = (u_{1j}, u_{2j}, ..., u_{nj}|\theta_j)$. In IRT, it is assumed that the u_{ij} are statistically independent, so the joint probability of the item responses for examinee j is given by the likelihood function:

$$P(U_j|\theta_j) = \Pi_{i=1}^{n} P_i^{u_{ij}}(\theta_j) Q_i^{1-u_{ij}}(\theta_j)$$

We will somewhat ignore the θ_j notation and assume that P and Q are functions are θ_j. Taking the natural log of the likelihood function, we get:

$$L = \log P(U_j|\theta_j) = \sum_{i=1}^{n} [u_{ij} \log P_{ij} + (1 - u_{ij}) \log Q_{ij}]$$

This time, we only need to take one partial derivative, with respect to θ_j:

$$\frac{\partial L}{\partial \theta_j} = \sum_{i=1}^{n} u_{ij} \frac{1}{P_{ij}} \frac{\partial P_{ij}}{\partial \theta_j} + \sum_{i=1}^{n} (1 - u_{ij}) \frac{1}{Q_{ij}} \frac{\partial Q_{ij}}{\partial \theta_j}$$

Again, note that P_{ij} and Q_{ij} are the above-defined logistic cumulative distribution function probabilities. The Newton-Raphson equation here is:

$$\left[\hat{\theta_j}\right]_{t+1} = \left[\hat{\theta_j}\right]_t - \left[\frac{\partial^2 L}{\partial \theta_j^2}\right]_t^{-1} \left[\frac{\partial L}{\partial \theta_j}\right]_t$$

As with the item estimation procedure, the equation is iterated until successive iterations result in a change of parameter estimates less than a specified threshold, and the resulting ability estimate is finalized.

It may appear obvious the problem - all three parameters are not something we can ever fully know, and assuming that any of them are known is a flaw. Therefore, a modification of the above procedure must be obtained. This is addressed with Birnbaum's Joint Maximum Likelihood Estimation (JMLE) paradigm. In this technique, all n items on a test and all N examinee ability levels are estimated jointly (not simultaneously). For each examinee, their responses can be encoded as a vector u_j. Since each examinee gets a vector, we can assemble all examinees together into a $n \times N$ matrix of item responses whose probability is given by the

likelihood function

$$P(U|\theta) = \Pi_{j=1}^{N} \Pi_{i=1}^{n} P_i^{u_{ij}} Q^{1-u_{ij}},$$

where P and Q are functions of the item characteristic curve dependent on θ. Then, the log likelihood is

$$L = \log P(U|\theta) = \sum_{j=1}^{N} \sum_{i=1}^{n} [u_{ij} \log P_{ij} + (1 - uij \log Q_{ij})]$$

Now, we must take derivatives with respect to each α and β and θ, resulting in $2n + N$ equations to get the maximum likelihood estimates. ($2n$ for both of the item estimates for the n items and N for N examinee ability levels.) The Newton-Raphson method of solving these equations is presented as the following matrix equation:

$$A_{t+1} = A_t - B_t^{-1} F_t$$

where A is the column vector of item and ability parameter estimates (a $2n+N$ vector), B is the matrix of second partial derivatives (a $2n+N \times 2n+N$ matrix) and F is the column vector of the derivatives of the likelihood function (a $2n + N$ vector). To begin the iteration, initial estimates of the item parameters are made, assuming the examinee's abilities are known. Often, the raw scores of each examinee are used. In the second stage, the obtained estimates for item parameters are used to estimate the examinee's ability scores. In each stage, the Newton-Raphson method is iterated until a convergence criterion is reached. This back-and-forth estimation procedure continues until subsequent iterations of each estimate produces results that fall within a convergence criterion. These item parameter estimates and ability parameter estimates are taken as the official parameter estimates.

Chapter 4

Test Creation and Item Selection

One of the most important topics that IRT is used for in applications is its role in test creation and item selection. To begin the discussion, we must introduce the idea of information, also called Fisher's Information. For dichotomously scored items under IRT, the Fisher's Information function I for data assumed to be binomially distributed can be written as [1]

$$I = \frac{n^2}{\sigma^2}, \tag{4.0.1}$$

where n is the number of examinees (trials in a binomial), and σ^2 is the variance. Therefore, as noted by Desjardins [6] , the information function is inversely related to the variance. A high variance at a particular level of θ will mean a low amount of information, and a low variance will mean a high amount of information. Then, Information is a measure (that ranges from 0 to ∞) of with what accuracy one can estimate the parameter at that point. Information in this context becomes very useful because it allows one to gather information about an item and even an entire test in advance. This means a test can be made to have certain properties in advance.

When information is plotted against θ, the result is the Item Information Curve (IIC), which is typically a peaked curve resembling a symmetric distribution function. The maximum of the IIC gives the location of θ where that particular item gives the most information about $\hat{theta}$, and thus, gives the location where that item functions best along the ability scale. As we move away from the maximum, the amount of information decreases and approaches 0 as the graph nears the left and right extremes of the ability scale.

From Desjardins [6], if we sum the information of each item on a test, we can achieve an overall measure of information that the entire test provides:

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta),$$

where $I(\theta)$ is the amount of total test information at an ability level θ, $I_j(\theta)$ is the amount of information for item j at ability level θ, and J is the number of items on the test. It is easy to see, since I is non-negative, as the number of items on a test increase, the test information function increases. Therefore, the more items on the test, the more precise one can measure the unknown ability parameters, and the more test information. If we plot the test information function against θ, we arrive at the Test Information Curve (TIC). This TIC is of primary interest because it gives important characteristics of the test and how it will function and perform at various ability levels.

To begin creating a test using IRT, first we must develop a pool of calibrated test questions. To do this, questions are created and administered to a group of examinees; this is considered "piloting" the test. In practice, the way this is often done is delivering a small set of pilot items mixed into an actual test that examinees take, and examinees do not know which items are "real" and which are pilot items. The pilot items are not scored as part of the test administration, but instead used for future test development. These pilot items are scored, calibrated to the ability metric, and item parameter estimates are made for each item. Over several such "pilot" administrations, a pool of calibrated items is developed.

Next, the IIC are plotted for each item. The purpose of the test comes into play here. [1] If this is meant to be a benchmark test, such as a pass-fail test or a test centered around a certain threshold, then the test should be constructed by picking items for the test that have IIC peaked as much as possible and centered over the threshold for θ. This will maximize the parameter estimate information on each item, and give the best delineation between passing/meeting the threshold or failing/not meeting the threshold. If this is meant to be a general purpose knowledge test to place examinees somewhere on an ability scale (but where is not of particular interest), then items should be selected which have IIC maximums at varied and mixed levels

of θ. This will give a test whose information can, as accurately as possible, place examinees anywhere along the ability metric. In the threshold-based test, the TIF will ideally be a sharp peak at the threshold with a high amount of information, and a low near-0 amount elsewhere. In the general-purpose test, the TIF will ideally be elevated as much as possible above 0 but very flat over the range of θ.

Chapter 5

An Analysis of MSATB Scores

The data set used for this paper comes from Drabinova & Martinkova (2017) [8]. The data are scores from 20 selected questions from a Medical School Admissions Test in Biology (MSATB) in the Czech Republic. The data set consists of responses of 1,407 subjects (484 males, 923 females). Each question has 4 answer choices, and any combination of these answer choices could be correct. For example, a correct answer may be A or BC or BCD. Since the data is still marked all right or all wrong, it is still treated as dichotomous data. A correct answer is coded as a "1" and an incorrect answer is coded as a "0". In addition, male is coded as "0" and female is coded as "1". The study [8] found previously that Item 49 functioned better for females than for males. This item asks about a childhood disease caused by deficiency of vitamin D in childhood with possible answers A. rickets (correct), B. scurvy, C. dwarfism, and D. intellectual disability. It was theorized that women in the Czech Republic tend to be more experienced in looking after children and know about their diseases.

To begin an analysis of the data, some overall descriptive statistics were ran on the data set. The overall mean score of examinees taking the test was 0.55, ie, the proportion that was marked "correct". The minimum score is 0.1 and the max score is 1. The data is symmetric, with the median also 0.55 (after rounding), and a histogram of the data (see Figure 5.1) confirms the data is mostly symmetric. In addition, the overall standard deviation of the item scores was 0.19.

The item-level descriptive statistics were computed next. Of the 20 items on the test, the average difficulty (as found before) was 0.55. The hardest item has a correct-rate of 0.18, while

Figure 5.1: A histogram of the MSATB raw scores

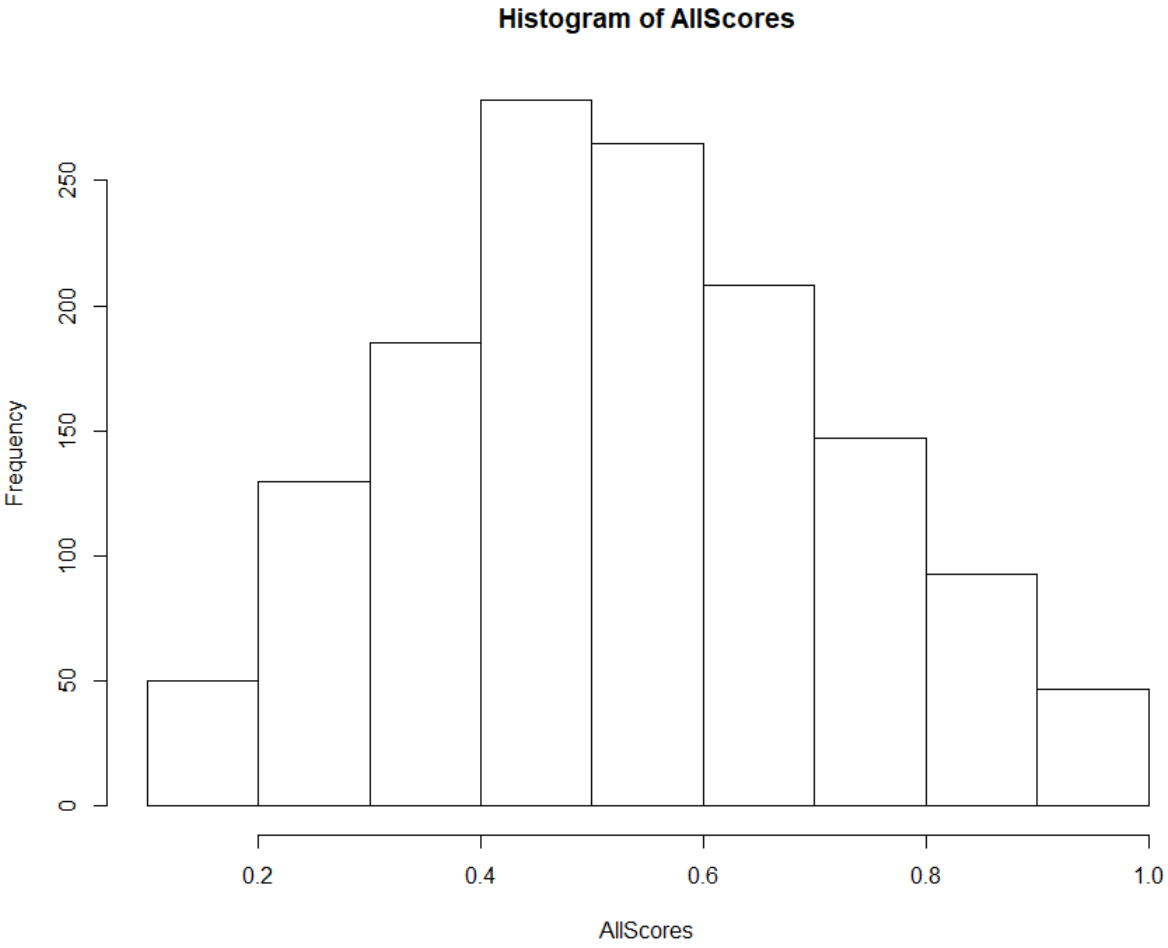
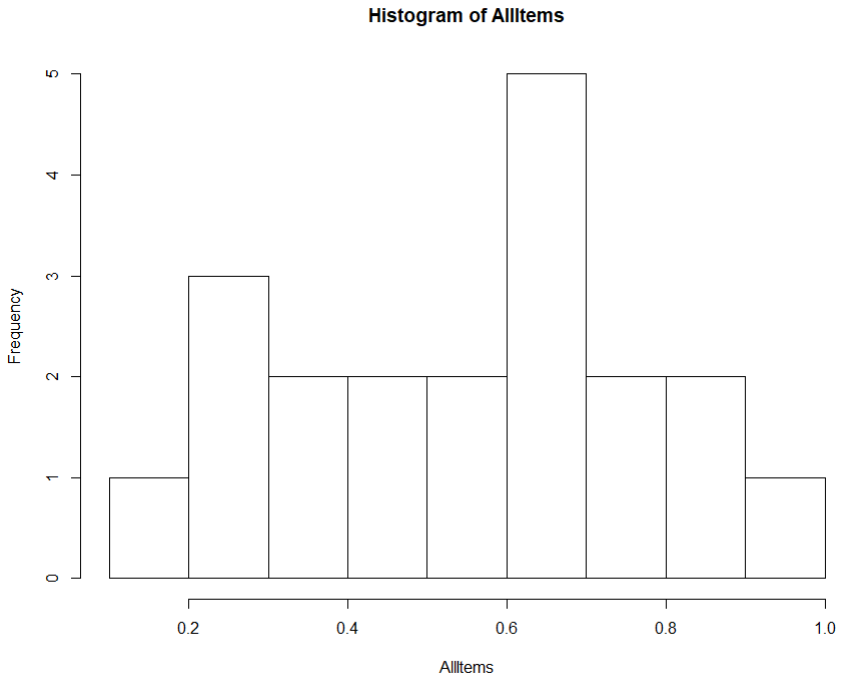


Figure 5.2: A histogram of the MSATB item scores



the easiest item had a correct-rate of 0.92. The median lf 0.58 shows that the distribution of item scores isn't perfectly symmetric, and the histogram (see Figure 5.2) shows this.

It is also of interest to look at differences between male and female. The data were split into two groups, one for male and one for female. When separated into groups, males scored an average percent correct of 56%, while females scores 55%. The standard deviations of 0.19 were almost identical. Histograms for male (Figure 5.3) and female (Figure 5.4) show minor differences.

Next, a t-test was run on the two groups (male and female) to see if the group means significantly differed from one another. The t-test found a 95% confidence interval of -0.01 - 0.03 for the difference of means, so the test found that there was no significant difference between the two means. The obtained p-value was 0.31, and the obtained t test statistic was 1.0152.

With descriptive statistics done, attention turned to estimating item and ability parameters and characteristics of the items and test. To do this, each of four models were fit to the data - Rasch, two-parameter logistic, three-parameter logistic, and four-parameter logistic.

Figure 5.3: A histogram of the MSATB raw scores for Males

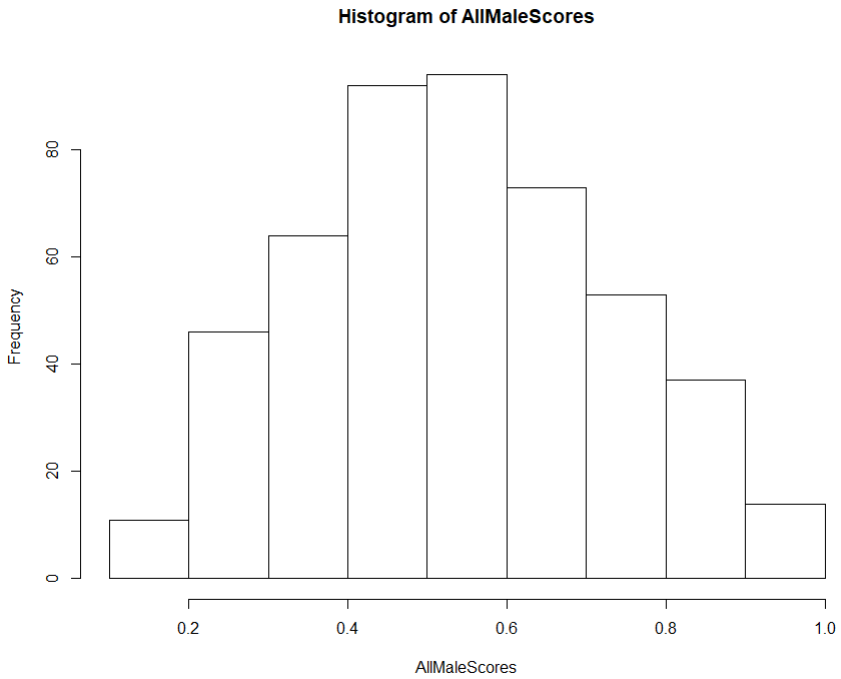
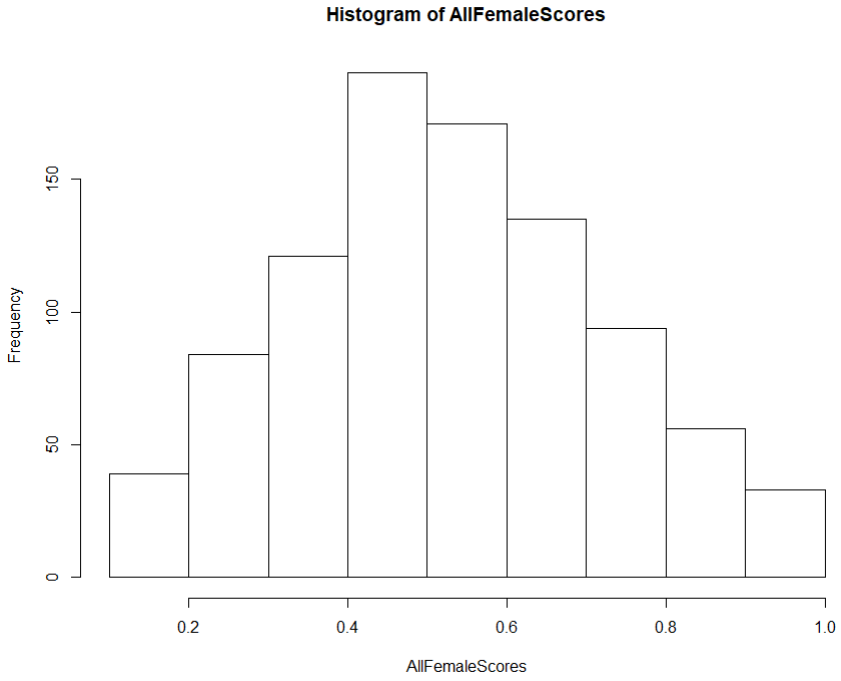


Figure 5.4: A histogram of the MSATB raw scores for Females



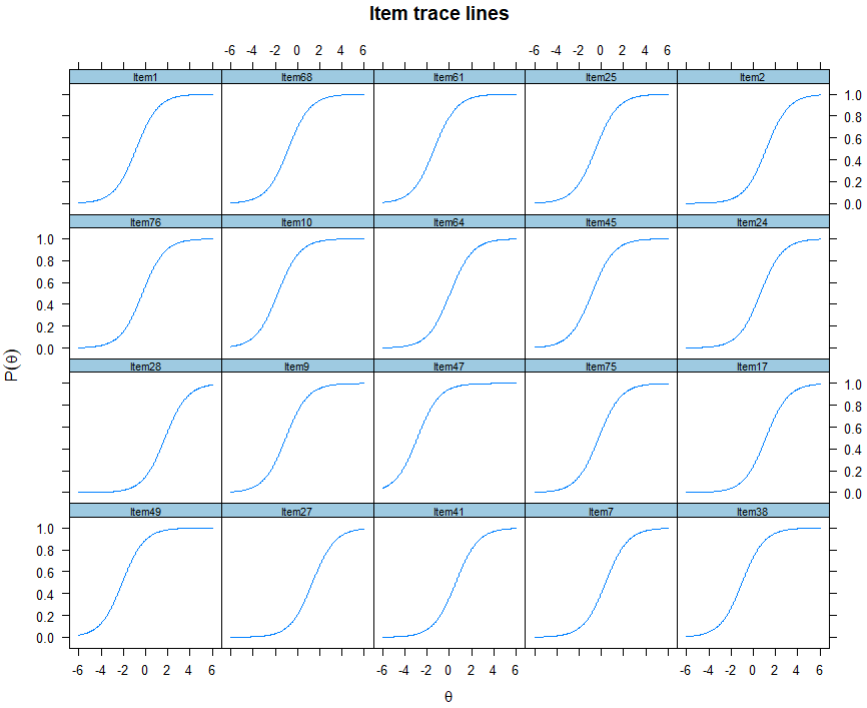


Figure 5.5: Rasch model Item Characteristic Curves

The Rasch model was fit first. The ICC, IIC, TIC and cSEM plot, and expected total score plots are shown in the following figures. Notice that each curve has identical item discrimination slope parameters. This makes this model the simplest of all, but it also produces the worst fit. After the item parameters are estimated, we estimate the ability parameters of the examinees. Once the ability parameters are estimated, it is found that the mean ability parameter is 0.05260.

Diagnostics are run on the Rasch model next. To begin with, three chi-square-based statistics are computed to find the goodness of fit of the data. The chi-square test (specifically Zh) indicates that 13 of the 20 items fit a Rasch model, while the remaining do not.The Zh statistic quantifies how far the observed values deviate from the theoretical values. Finally, the Zh statistic for person-fit is shown, and a score of -2 or higher means the model fits the person data. As we can see, most of the examinees fit the data well.

We next turn our attention to the two-parameter logistic model. The same plots as above are shown below for the 2PL model as well. Notice that, since the discrimination (slope) parameter can now vary, we get very different shapes for some of the ICC, where the slope is steeper or shallower than before. The fact that the parameter is now allowed to vary for

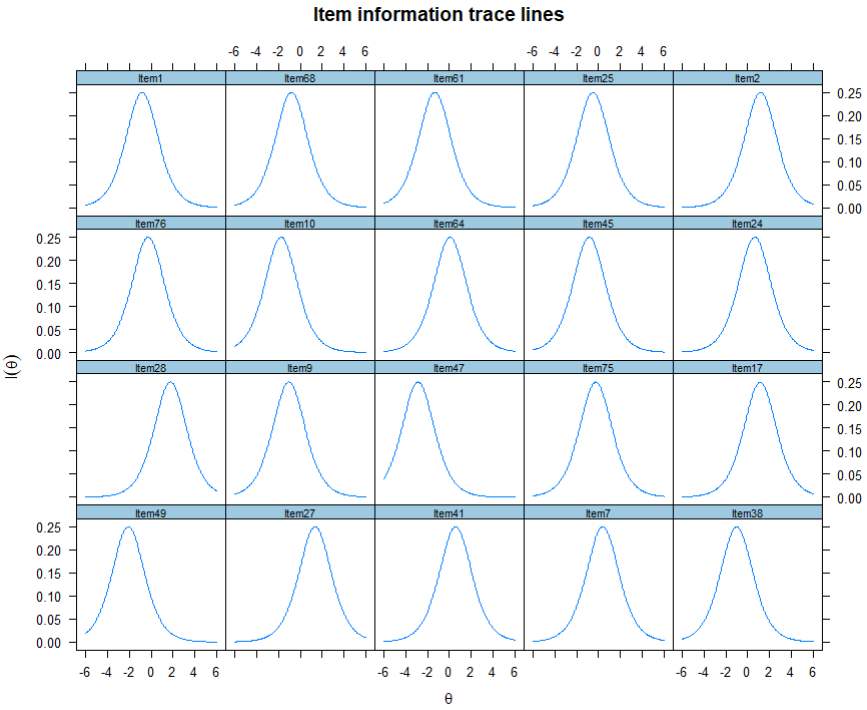


Figure 5.6: Rasch model Item Information Curves

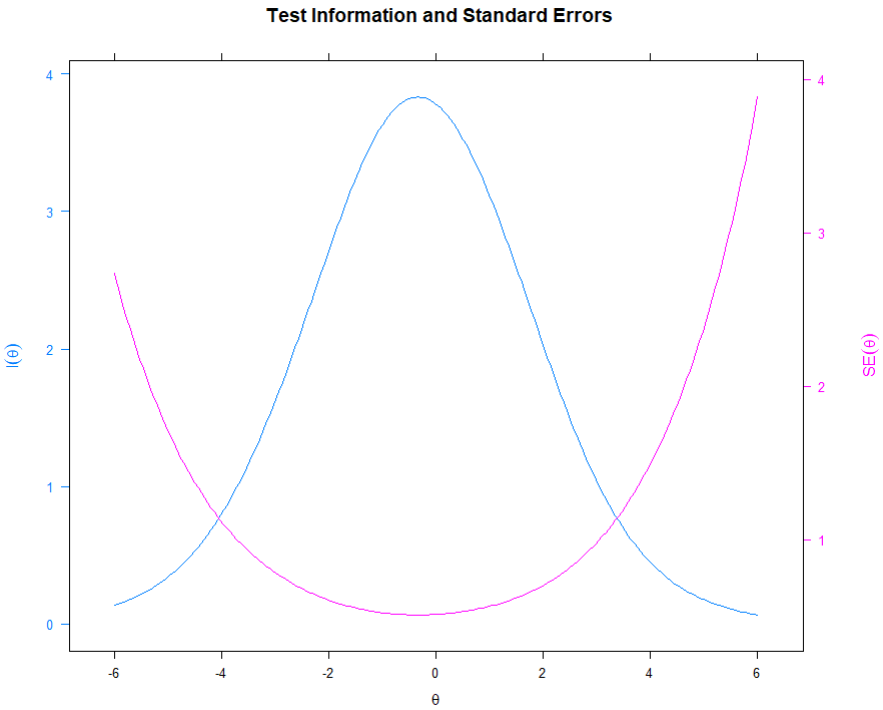


Figure 5.7: Rasch model Test Information Curve

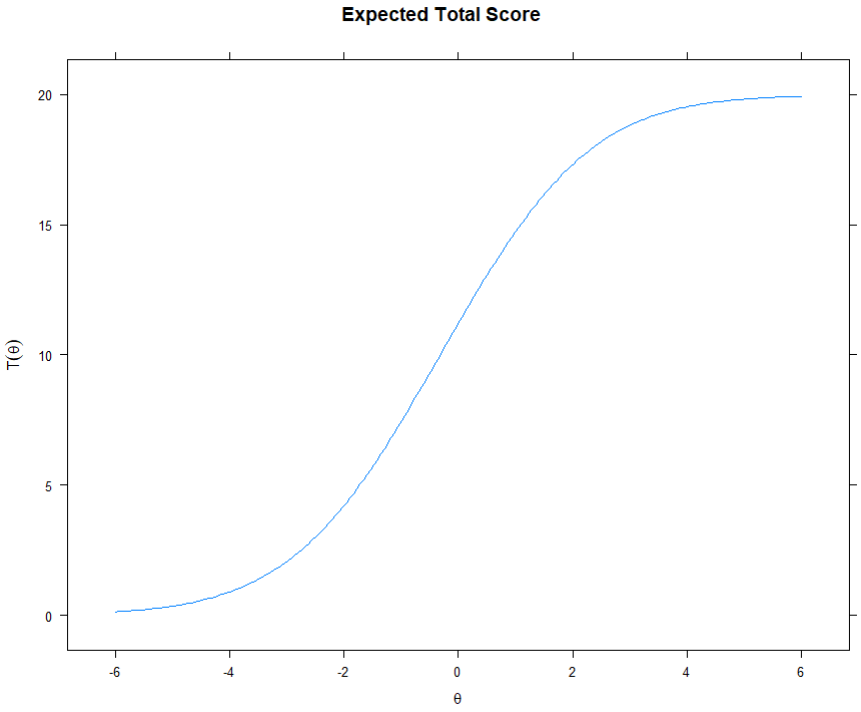


Figure 5.8: Rasch model Test Characteristic Curve

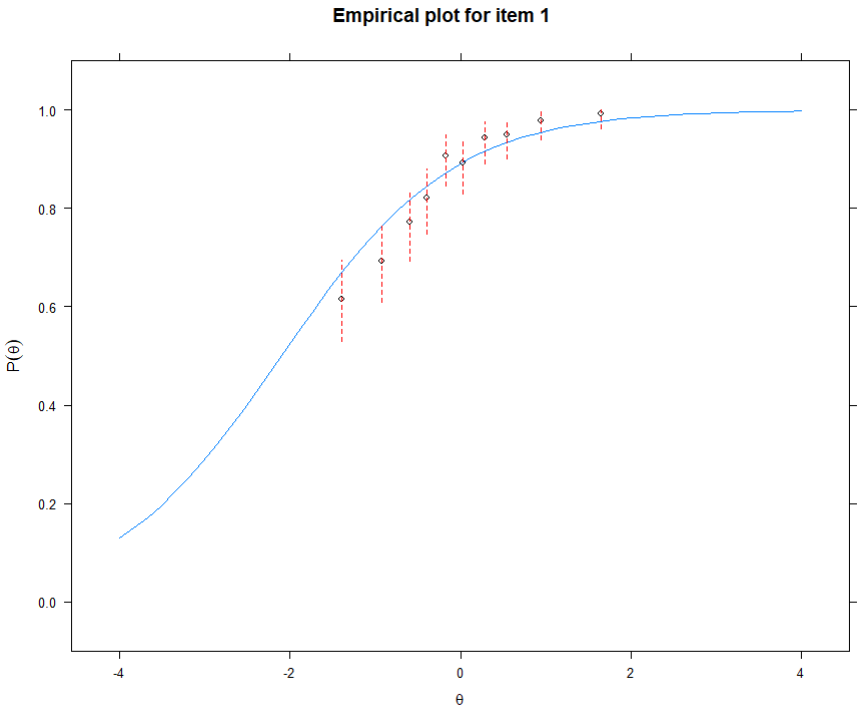


Figure 5.9: Empirical Plot Rasch Model for Item 1

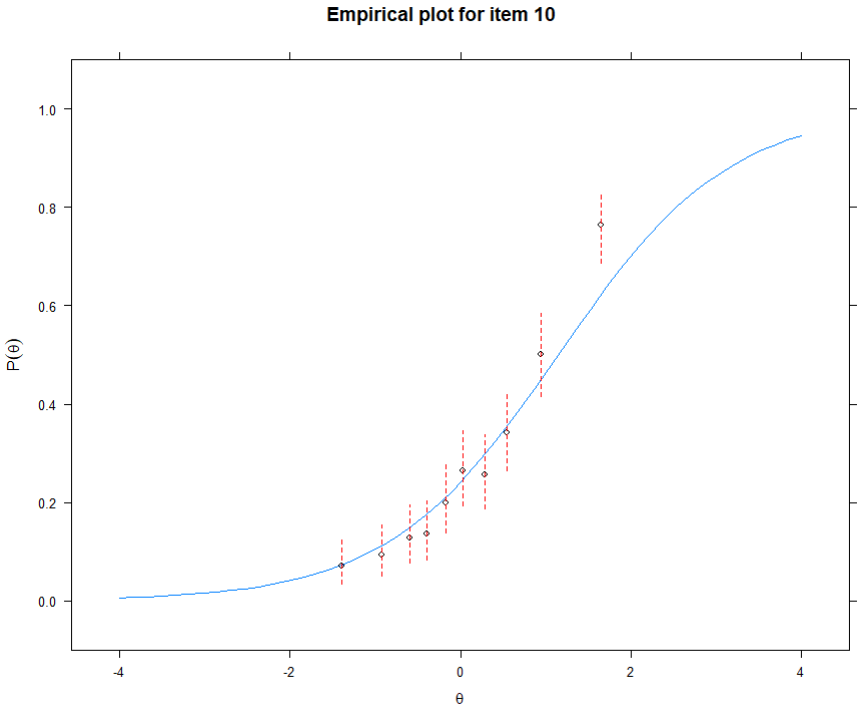


Figure 5.10: Empirical Plot Rasch Model for Item 10

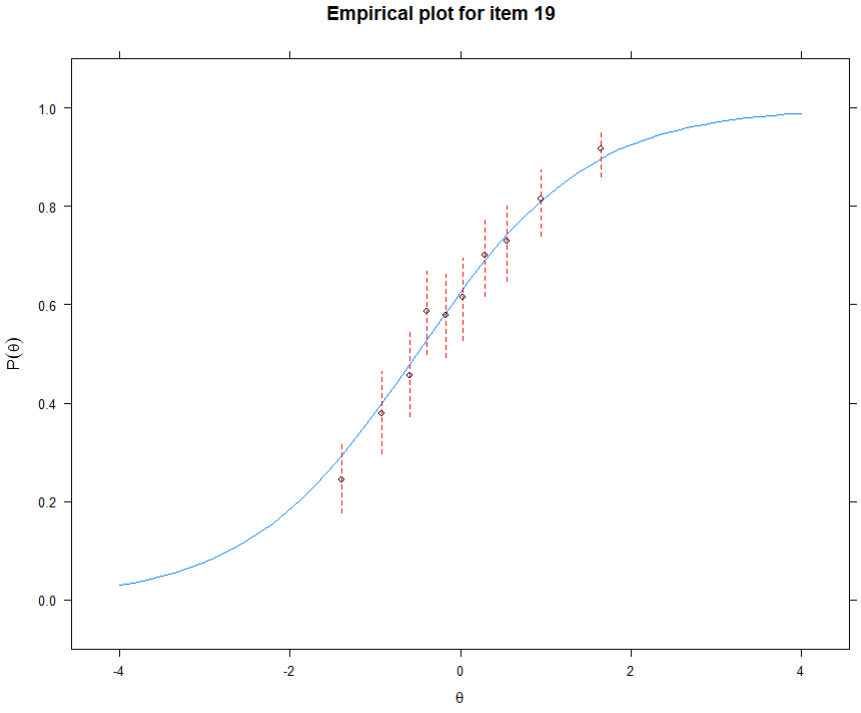


Figure 5.11: Empirical Plot Rasch Model for Item 19

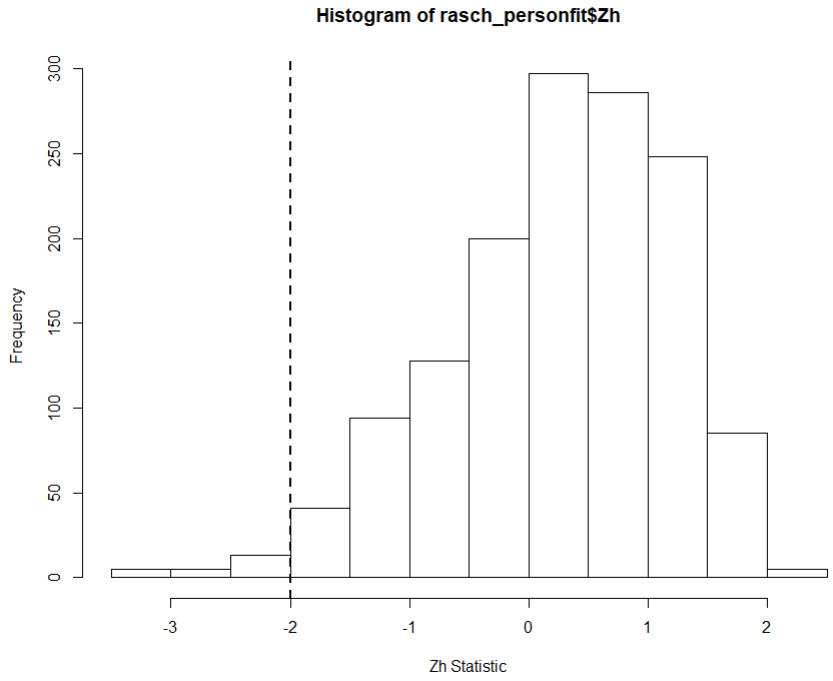


Figure 5.12: Histogram for the Zh Person-Fit Statistic

discrimination also means that it can carry more information, so the ICC are generally more responsive in this case than the Rasch. When ability parameters are estimated, the mean of 0.06 is found, similar to the Rasch model.

Diagnostics are run on the two-parameter model next. The chi-square statistics (specifically the Zh statistic) finds 15 of the 20 items fit the model (at the 90% significance level). (90% is chosen because all of the commonly-used statistics tend to be imperfect and sometimes are overly restrictive, so a 10% alpha level gives room for this error.) Empirical plots of the same items as the Rasch model are shown below, showing how well the model fits the data. In addition, the Zh person-fit statistic histogram is shown, indicating most examinees can be fit under the model.

Third, a three-parameter logistic (3PL) model is fit against the data. In this model, there is a pseudo-guessing parameter c that serves as the lower asymptote of the curve. When this is taken into account, notice how different the Item Characteristic Curves look. The addition of the third parameter really gives the model freedom to take into account many different item features. When ability parameters are estimated, the mean is approximately 0.07, close to the

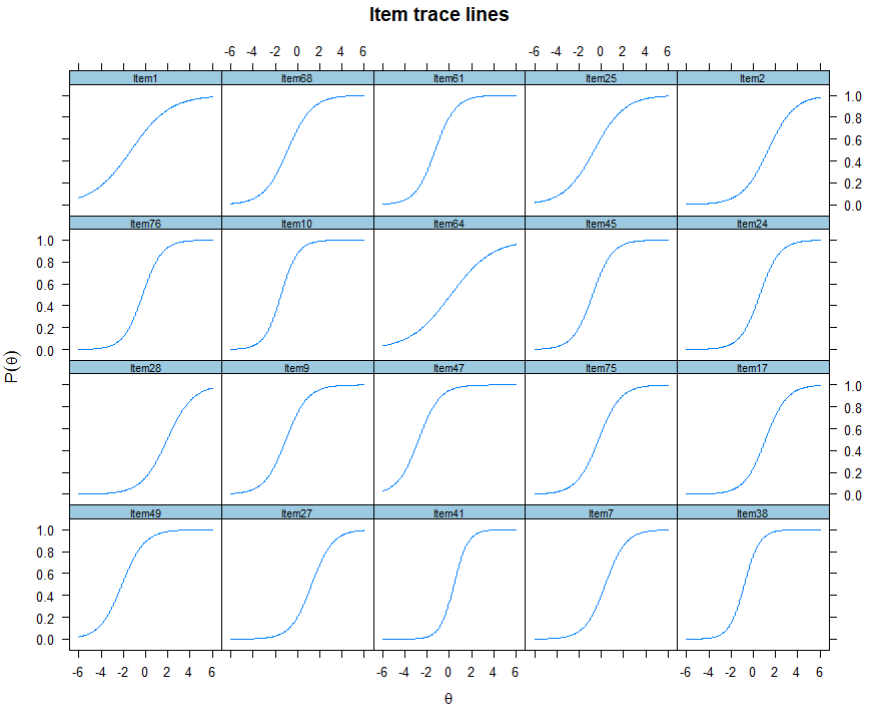


Figure 5.13: Two-Parameter Logistic Item Characteristic Curves

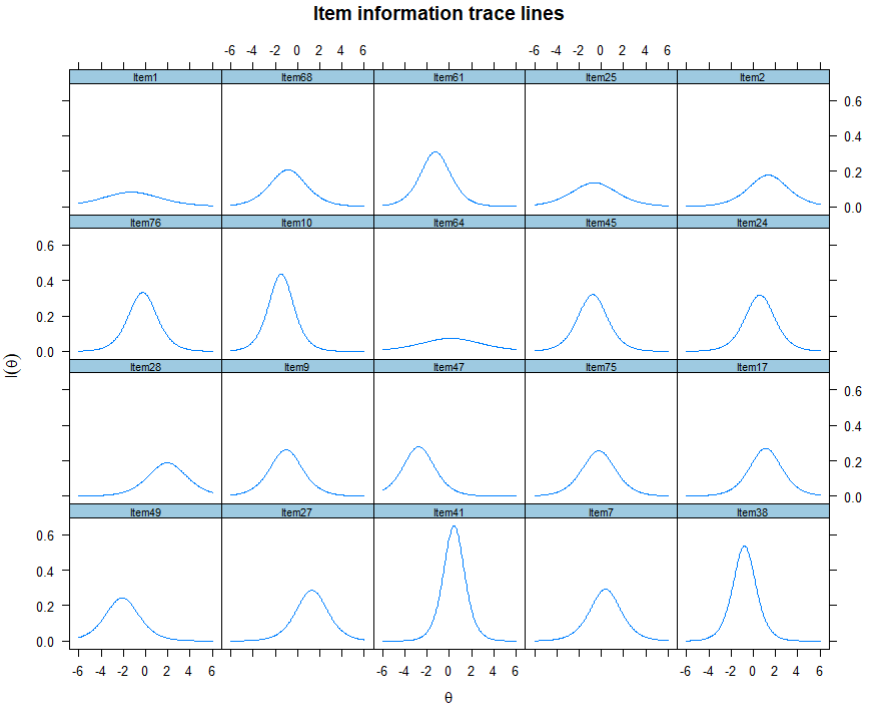


Figure 5.14: Two-Parameter Logistic Item Information Curves

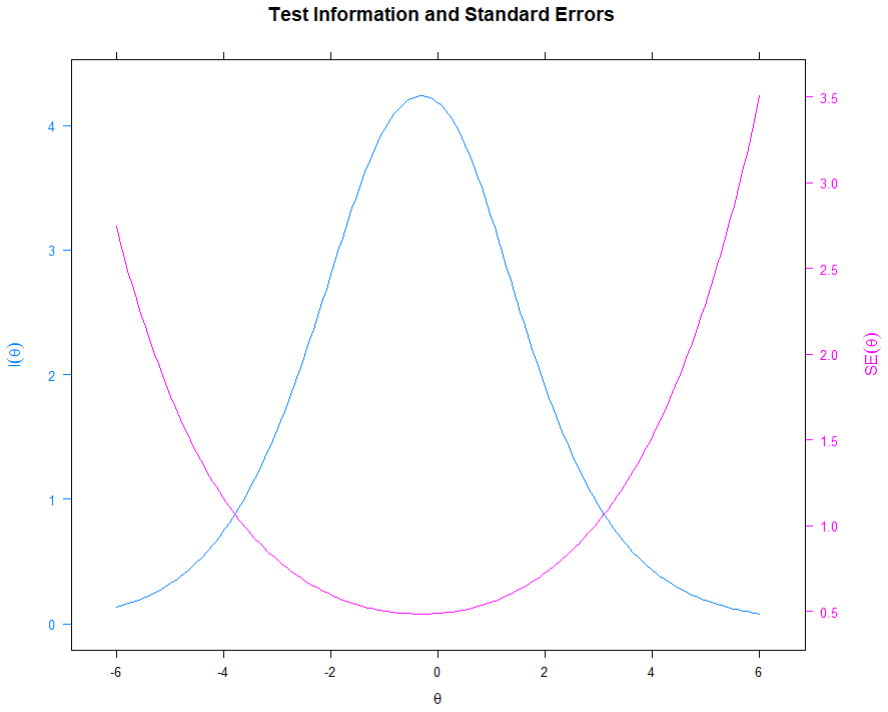


Figure 5.15: Two-Parameter Logistic Test Information Curve

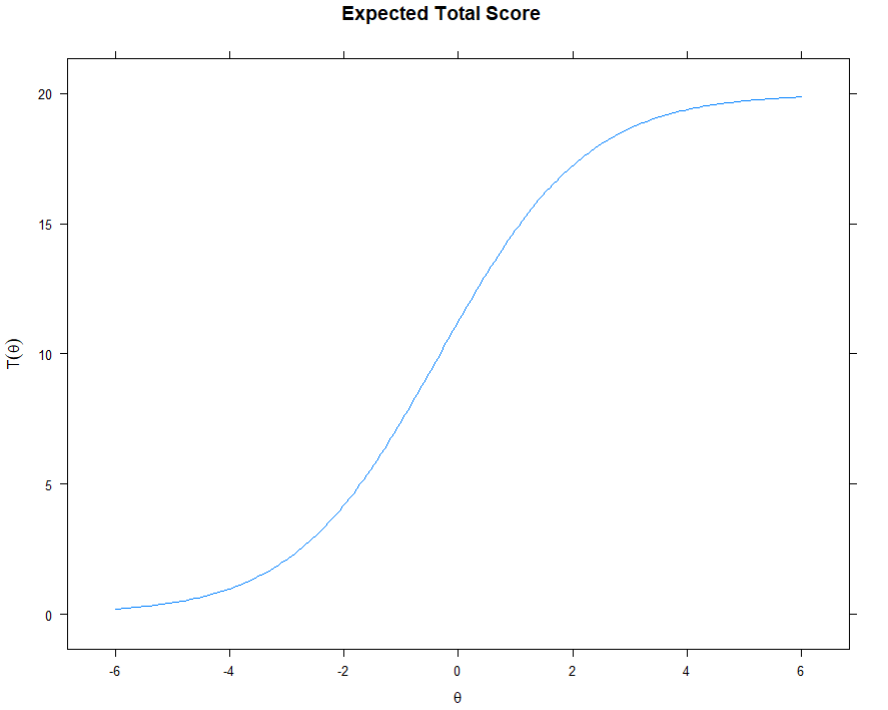


Figure 5.16: Two-Parameter Logistic Test Characteristic Curve

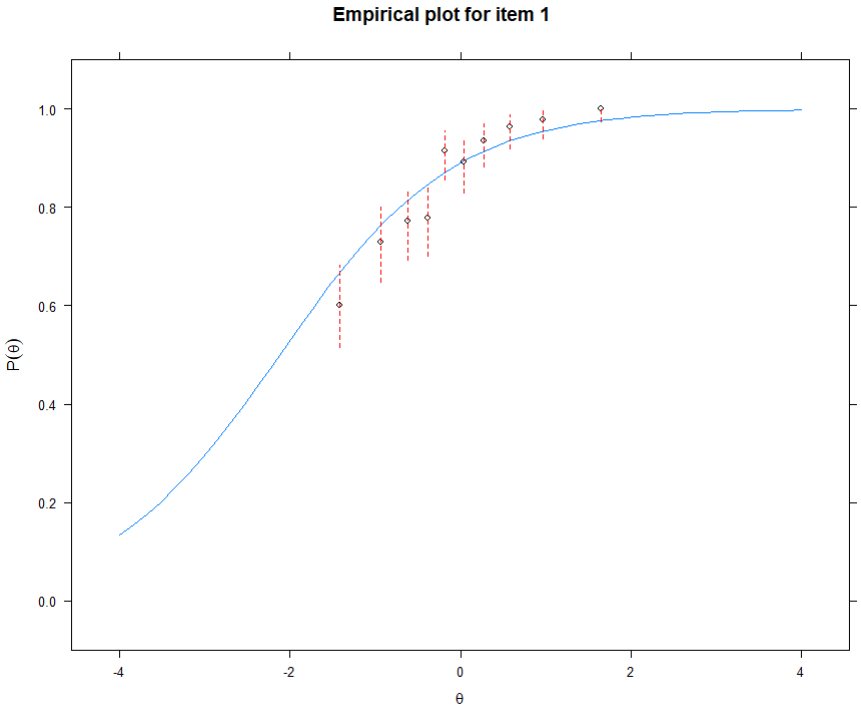


Figure 5.17: Two-Parameter Logistic Empirical Plot for Item 1

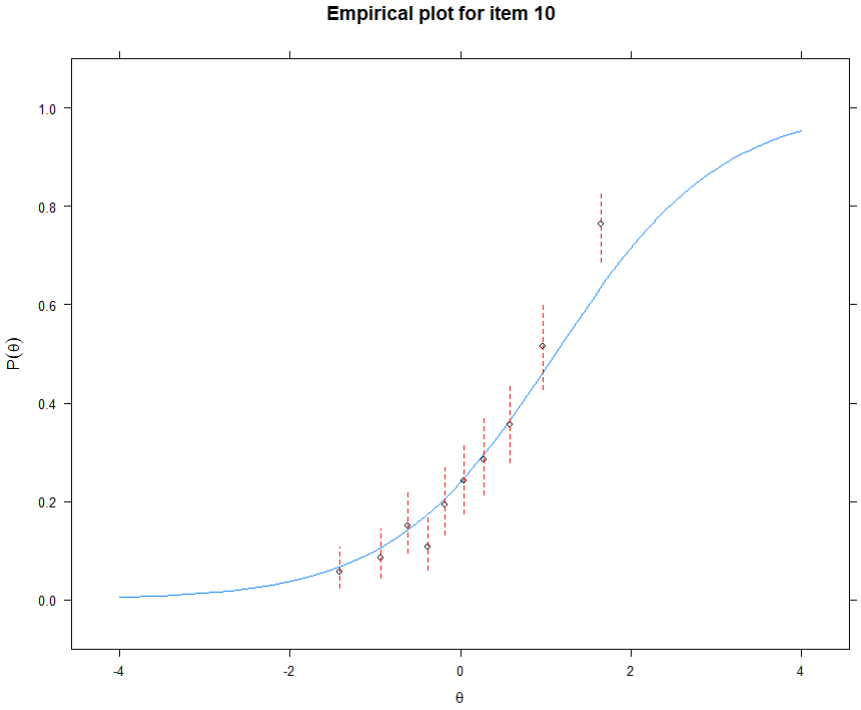


Figure 5.18: Two-Parameter Logistic Empirical Plot for Item 10

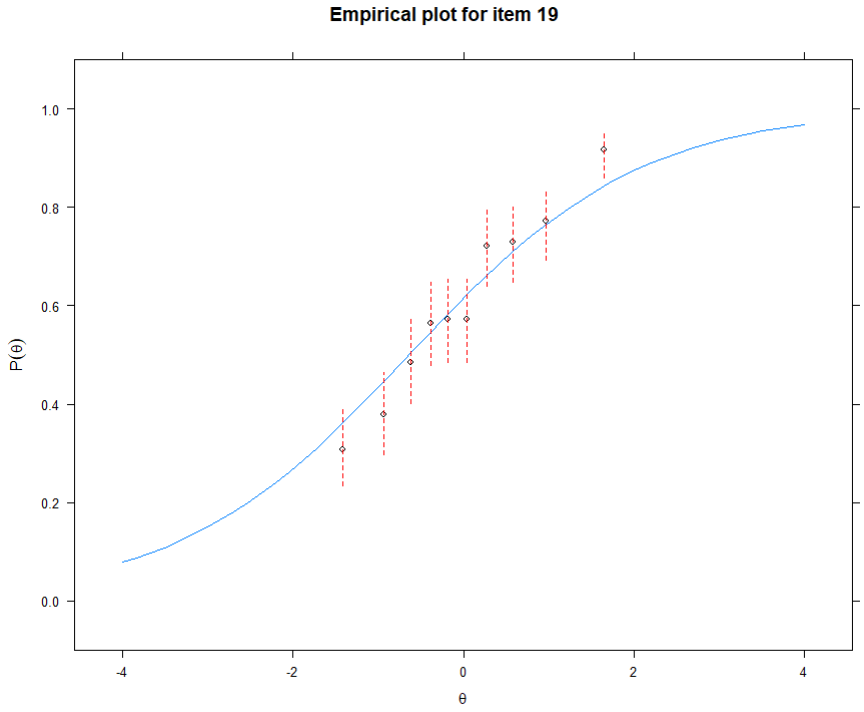


Figure 5.19: Two-Parameter Logistic Empirical Plot for Item 19

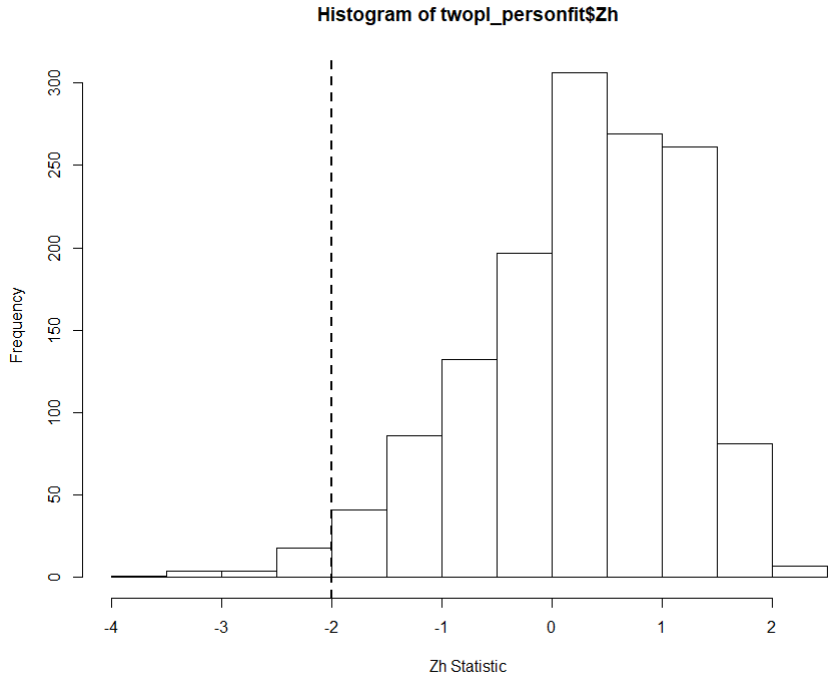


Figure 5.20: Two-Parameter Logistic Zh Person-Fit Statistic

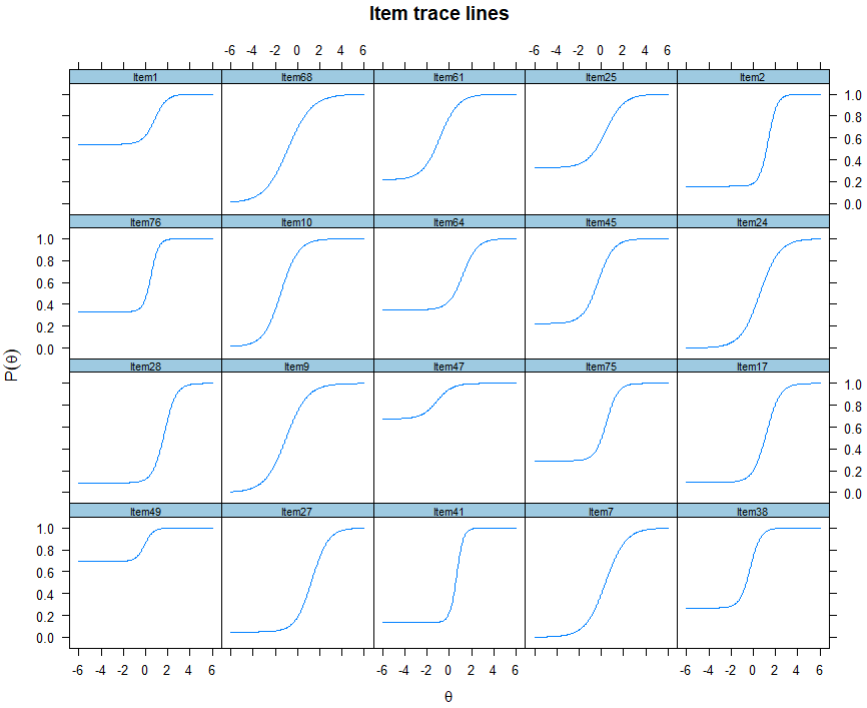


Figure 5.21: Three-Parameter Logistic Item Characteristic Curve

other two models. The plots shown for the previous two models are shown here as well for comparative purposes.

Diagnostics are run next on the 3PL model. The chi-square statistic indicates that 15 of the 20 items fit well in the 3PL model at the 90% level. The plots, as above, are shown below displaying the fit against the model of selected items, as well as the Zh statistic for person-fit. Again, a person-fit Zh statistic of -2 or more is desired to show model fit, and this is largely met by the vast majority of examinees.

Finally, we turn our attention to the four-parameter logistic model (4PL). In this model, there is additionally an upper asymptote that may differ from 1. It can be seen by observing the ICCs below that the presence of both lower and upper asymptotes give rise to a wide variety of S-shaped curves in the ICCs. This could be a benefit, allowing the most flexibility of any of the models to fit the data. However, the addition of a new asymptote may or may not be needed in the model, and this will be explored more below. The mean of ability scores is measured as 0.12. Now, we give the standard plots, as in the other three models:

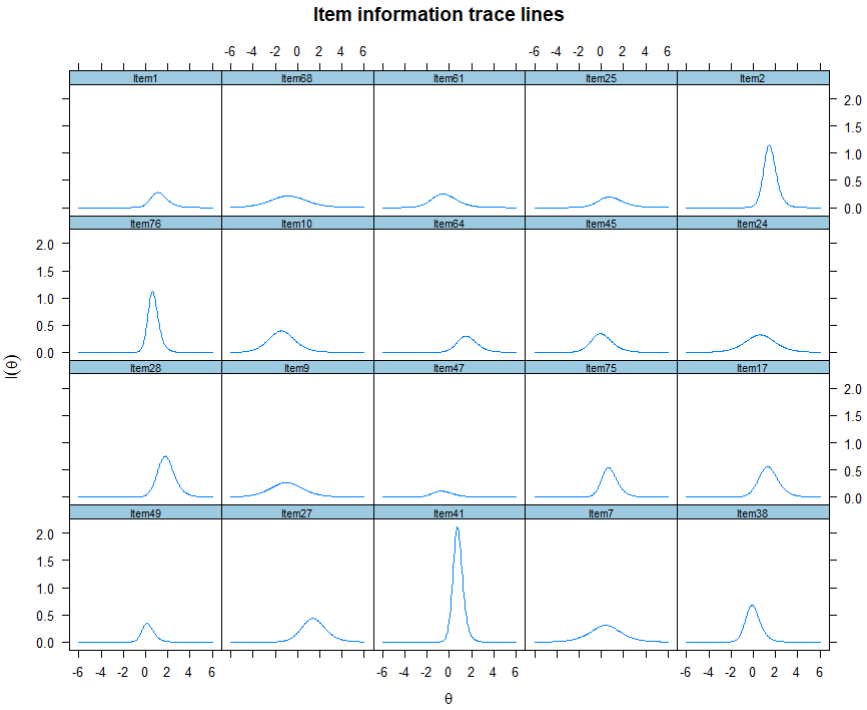


Figure 5.22: Three-Parameter Logistic Item Information Curves

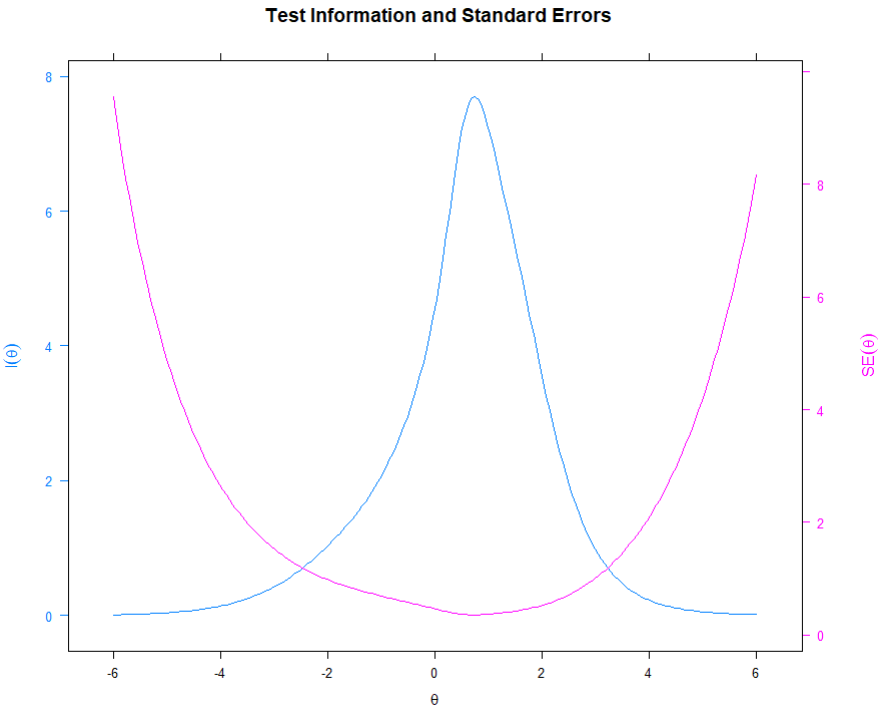


Figure 5.23: Three-Parameter Logistic Test Information Curve

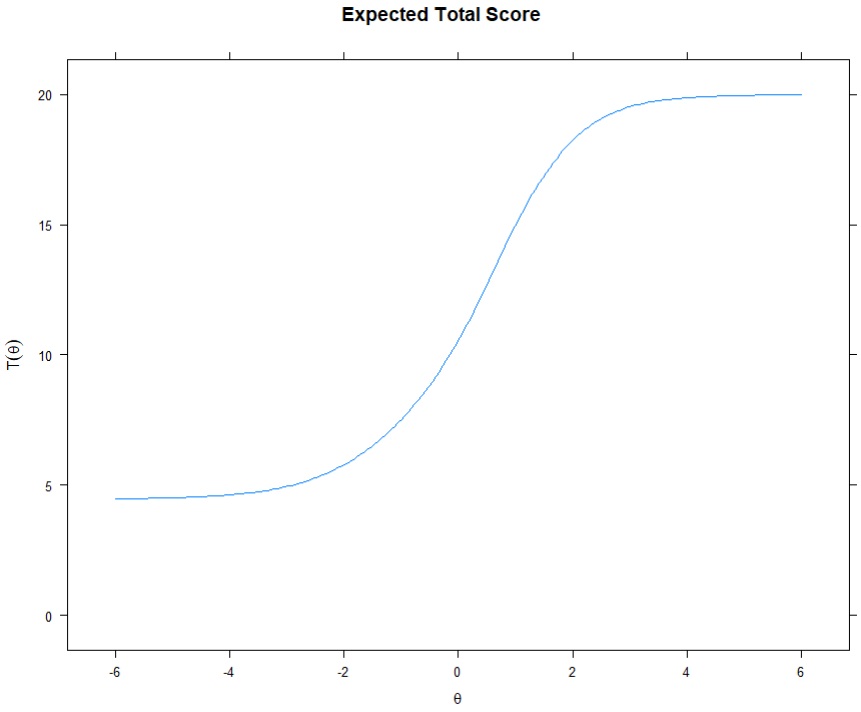


Figure 5.24: Three-Parameter Logistic Test Characteristic Curve

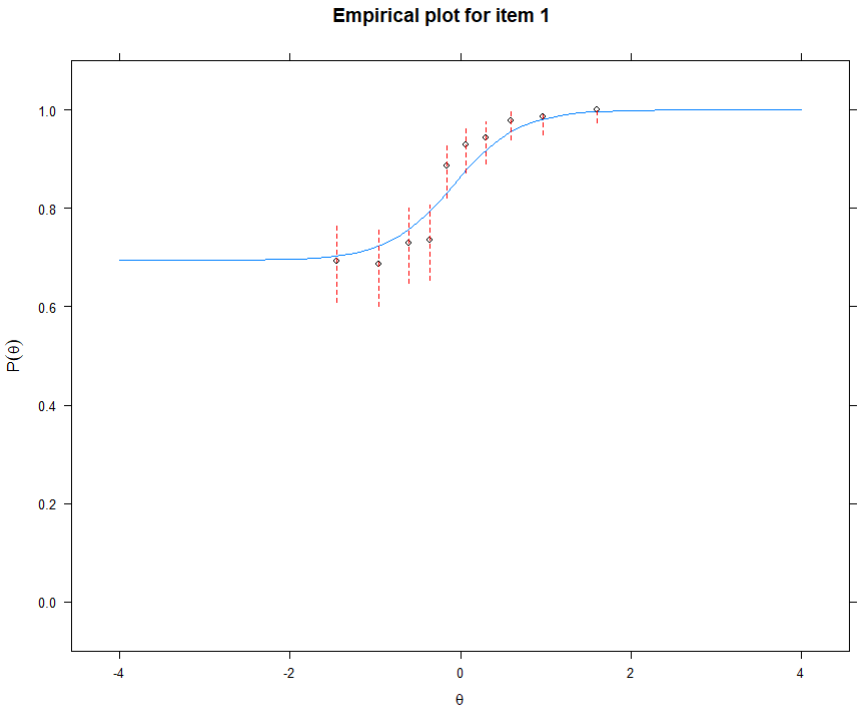


Figure 5.25: Three-Parameter Logistic Empirical Plot for Item 1

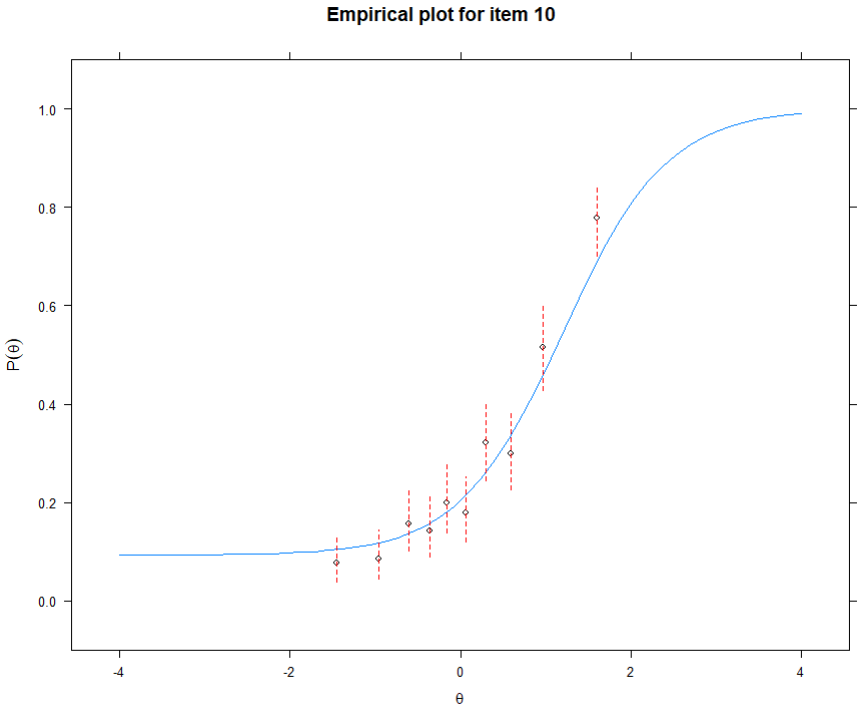


Figure 5.26: Three-Parameter Logistic Empirical Plot for Item 10

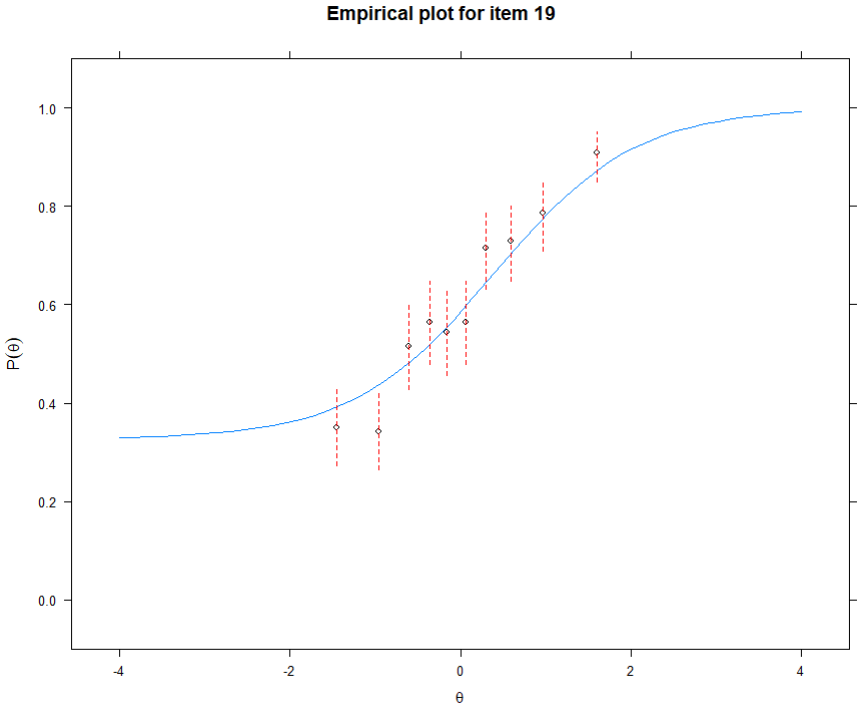


Figure 5.27: Three-Parameter Logistic Empirical Plot for Item 19

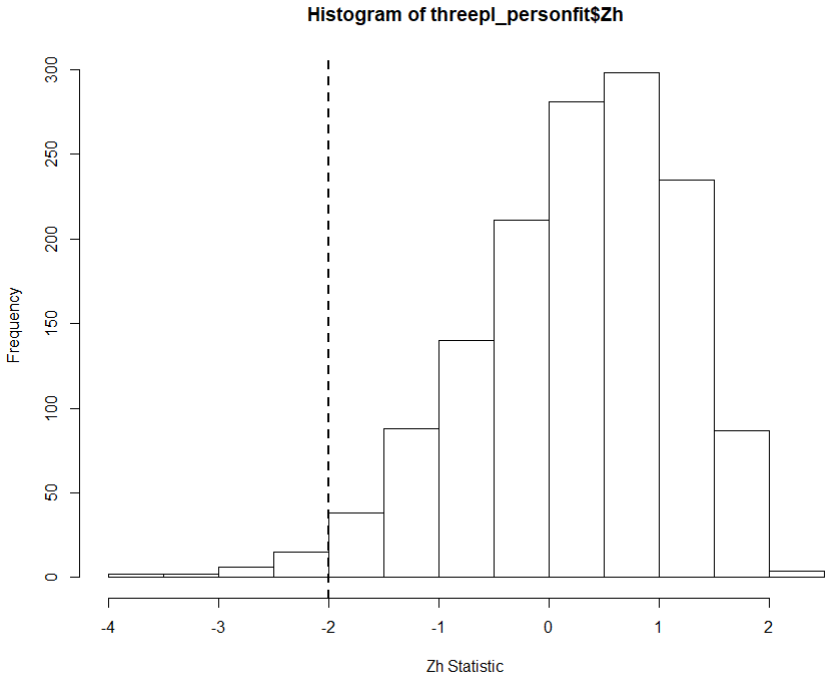


Figure 5.28: Three-Parameter Logistic Zh Person-Fit Statistic

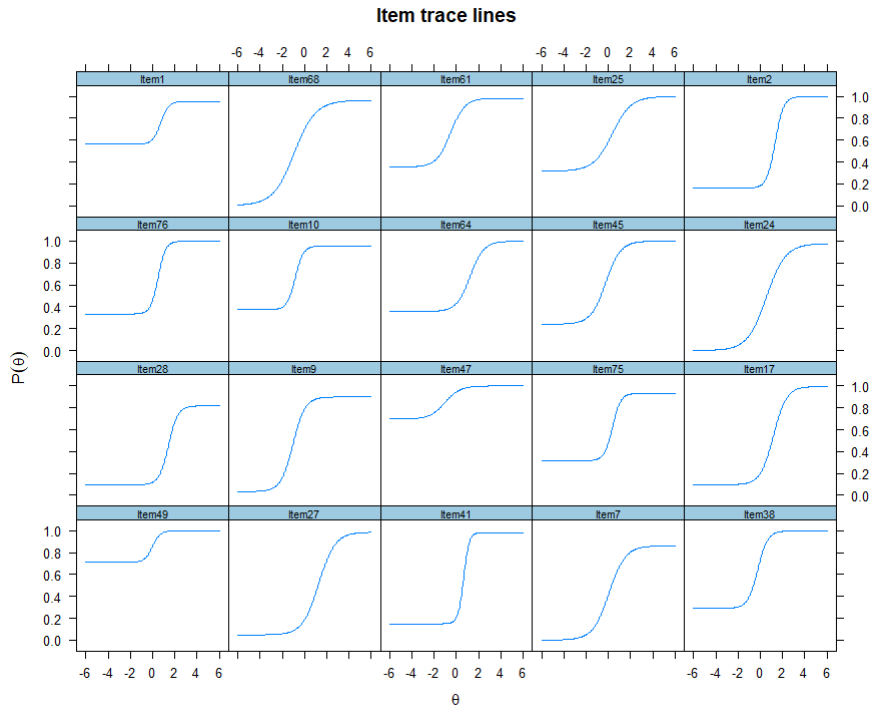


Figure 5.29: Four-Parameter Logistic Item Characteristic Curves

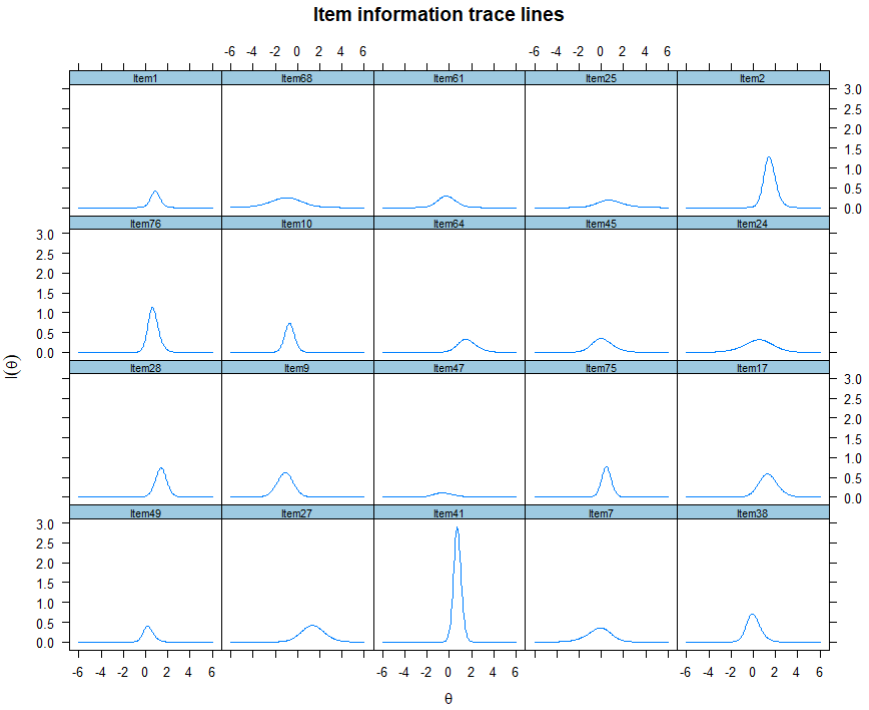


Figure 5.30: Four-Parameter Logistic Item Information Curves

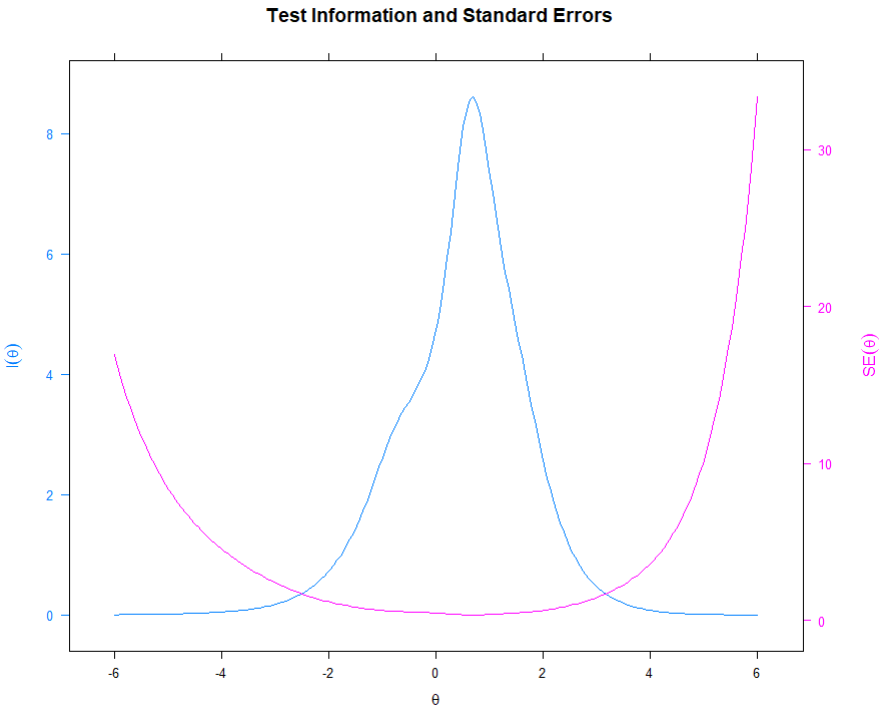


Figure 5.31: Four-Parameter Logistic Test Information Curve

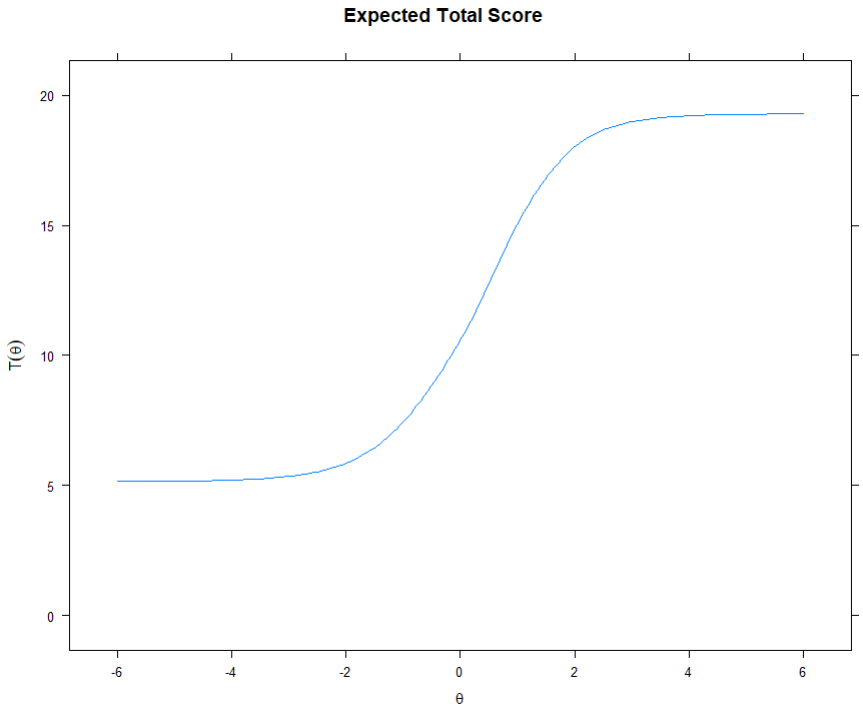


Figure 5.32: Four-Parameter Test Characteristic Curve

The last step is to run the diagnostics on this last model. The Zh statistic at 90% level shows 13 of the 20 questions are adequately fit by the model. The other chi-square-based statistics are more strict but still have about half of the questions fitting the model. The person-fit statistic histogram shows that most of the examinees are adequately represented by the model. The empirical plots are shown, as in the other cases, as well.

So, all four models have been specified. The big question now is "Which model is best?" This is partly answered with the statistician's own knowledge and partly answered with the use of statistical tests. As with the rest of this analysis, the computing will come from the mirt package in R [5]. The comparison will be made pairwise with the "winner" pairing off with the next-higher model.

The first model is the Rasch model versus the 2PL model. The AIC, AICc, and BIC are all lower for the 2PL model. Additionally, the chi-square statistic (based on likelihood ratio test) is 158.769,p $<$.005, so this implies that the 2PL model fits better than the Rasch model.

We repeat this procedure with the 2PL and the 3PL. In this case, the AIC and AICc is lower for the 3PL, but the BIC is (slightly) lower for the 2PL model. The chi-square statistic here is 141.232, p $<$.005, so this would mostly warrant the 3PL being the choice over the 2PL.

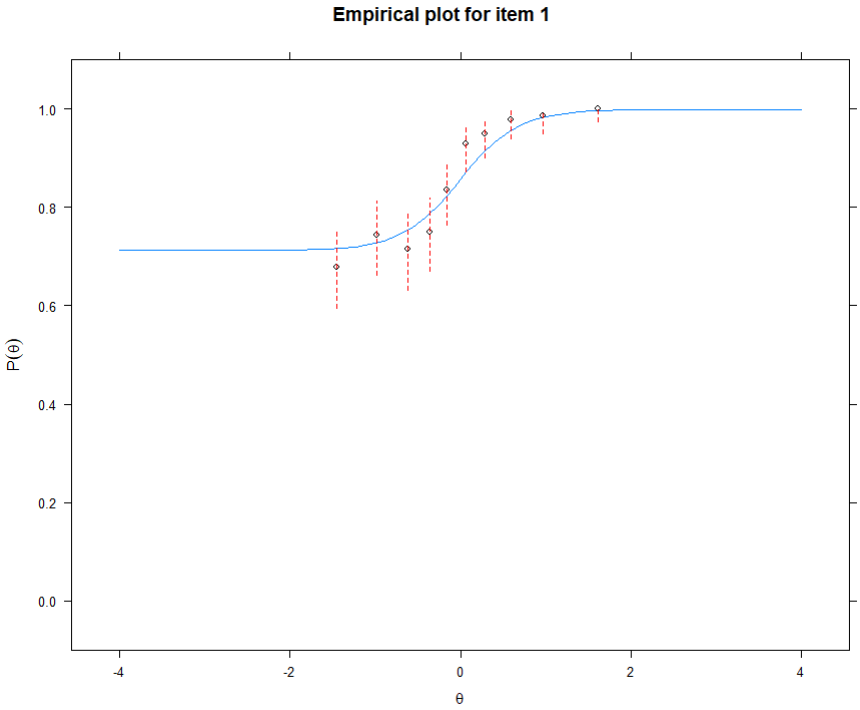


Figure 5.33: Four-Parameter Empirical Plot for Item 1

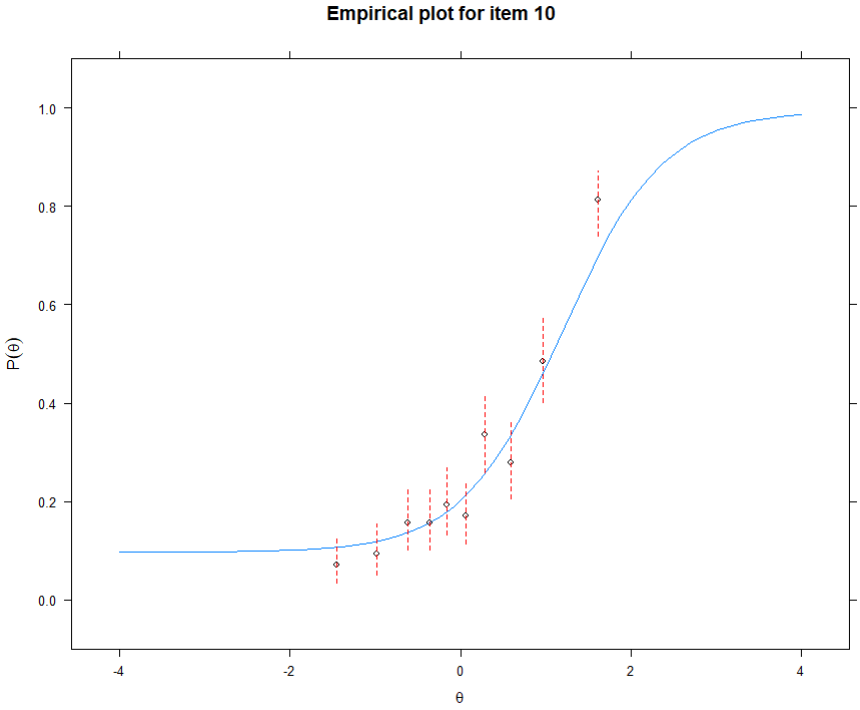


Figure 5.34: Four-Parameter Empirical Plot for Item 10

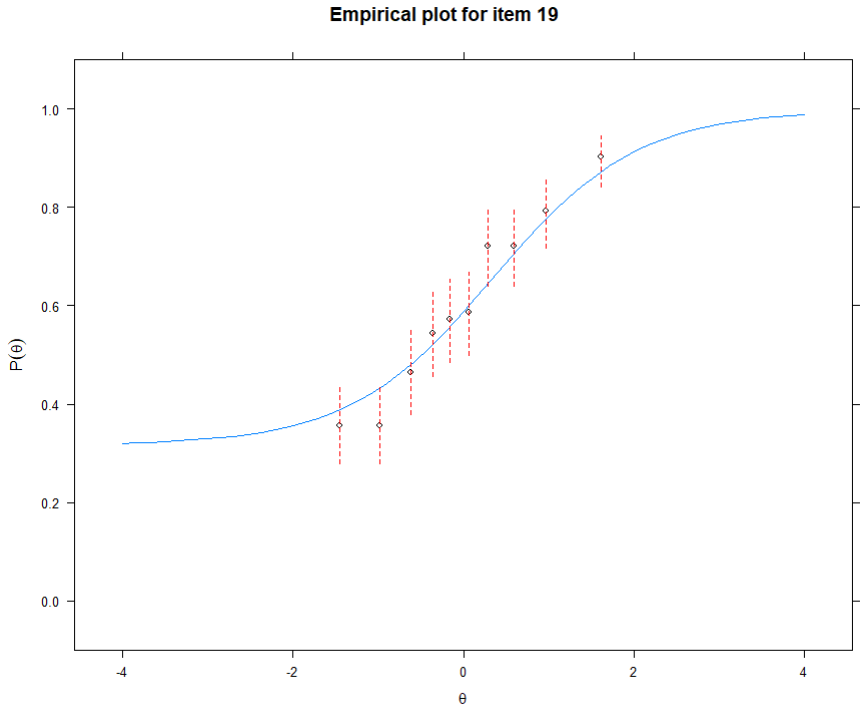


Figure 5.35: Four-Parameter Empirical Plot for Item 19

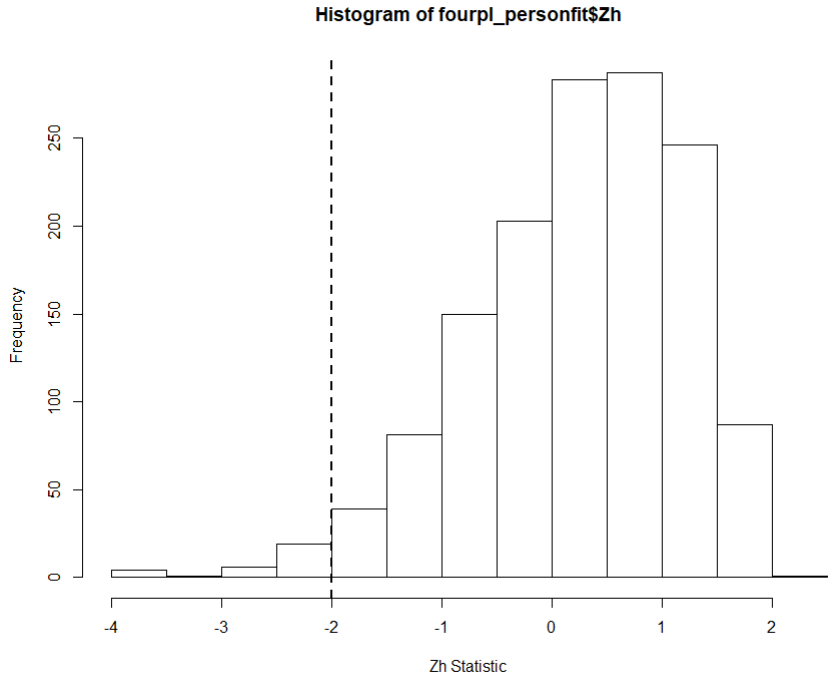


Figure 5.36: Four-Parameter Zh Person-Fit Statistic

This is also likely, as if we look back and examine the Item Characteristic Curves between the 2PL and 3PL, we will note quite a few of the items have distinct lower asymptotes when allowed to do so in the 3PL. The fact that the ICCs respond to the addition of a third parameter lends credence to the idea of the 3PL being superior in this case to the 2PL. So, we select the 3PL to move on.

In our last comparison, we examine the 3PL and 4PL. The 3PL in this case has lower AIC, AICc, and BIC. Additionally, The chi-square is not significant with 12.838 being the test statistic and p = 0.884. This means perhaps the 3PL is effective over the 4PL. Looking at the Item Characteristic Curves for both models, there does appear to be the presence of an upper asymptote in a couple of the items, but in almost every case, the upper asymptote is near to 1. In fact, out of the 20 items, only 3 items in the 4PL item parameter estimates have upper asymptotes below 0.90 (the lowest is 0.81), and only 4 have upper asymptotes below 0.95. The vast majority hover very close to 1. This implies that a fourth parameter is not needed in our model. Therefore, the model that we will select as our final model is the 3PL model.

We also want to briefly look into the idea of test construction and creation. We will treat the existing test questions as our item pool and create a small test from those questions. With an item pool of only 20 questions, we are limited in scope at what can be accomplished, but we will create a 5-question threshold test that attempts to classify examinees as "pass" or "fail" based on the cutoff score of +2 on the ability scale. To do this, we look at the data and carefully select 5 questions that have high information values around +2 and also high discrimination values to help discriminate between slight adjustments in ability level. With a larger pool of questions, an algorithmic approach would need to be programmed to find such items. With such a small pool, however, it is easy to pick items by inspection only. Items selected are Item 27, Item 17, Item 64, Item 1, and Item 2.

The usual plots are shown with this short 5-question subtest. Some important points are of note here. First, notice the all of the ICCs (particularly 2 of them) are very steep, approaching jagged corners instead of smooth curves. This shows the high discrimination that is useful in a threshold type of test. Secondly, the Item Information Curves are peaked near +2, two of them particularly high peaks, yielding much information near that ability level. This is crucial

in creating such a test. This is exemplified in the Test Information Curve (that also shows the Standard Error of Measurement), and notice the extremely sharp peak near +2. This means that this test will provide accurate estimates of ability near +2, which is important because that is the cutoff level for this test's "pass" or "fail". The SEM is also low around +2, which indicates that we can be confident in our estimates near +2. The Expected Total Score/Test Characteristic Curve elevates steeply between +1.5 and +2, showing the expected score jumping from 2 to 4 in a very brief space. All of these indicators show a good test for this purpose. Do note, though, that away from +2, this test's performance rapidly depletes, so examinees scoring very differently than +2 are not likely to be accurately estimated by this test.

Also, it was attempted to create a general-purpose subtest as well (as can be seen in the R code). However, this was much less successful, with the Test Information Curve not creating the needed rounded flat peak necessary for that type of test. The main issue was a lack of questions in the lower ability levels from which to choose. After contemplating about the data set, however, this makes perfect sense; the data is from a medical school admissions test, which means the questions have likely already been chosen to perform well at a (positive) certain ability level for the purposes of discriminating between candidates that should be admitted and those that should not.

After these diagnostics were ran on the subtest, attention turned to prediction. We wanted to see if a classification could be made based on gender.

The labels for gender are present in the original dataset, so the data were split into a test and training set and a logistic regression model ran on the training set. The result here is that the model is able to correctly predict gender 62% of the time, leading to a 38% error rate. An interesting note is that most of the misclassification was on males; of the 151 males in the test set, only 5 were correctly classified as male, and only 8 females were incorrectly classified as male.

To see if a better classifier for gender was possible, a random forest model was run on the data. This produced very similar results of 59% accuracy and about 41% error rate. Principal component analysis is run on the training data set to see if a dimension reduction technique might prove beneficial in the classification. From the original 20 dimensions, it would have

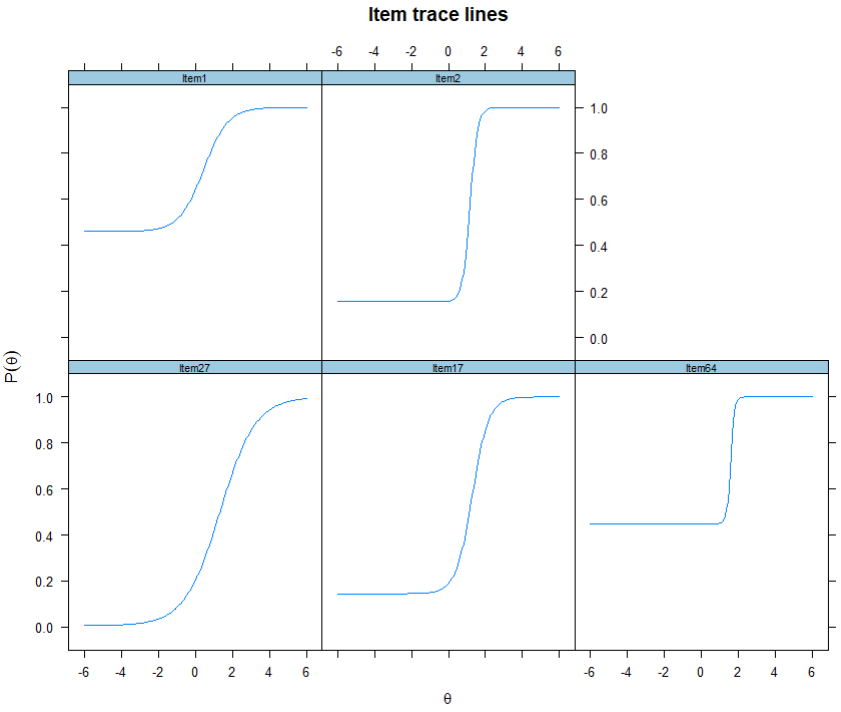


Figure 5.37: Threshold Test Item Characteristic Curves

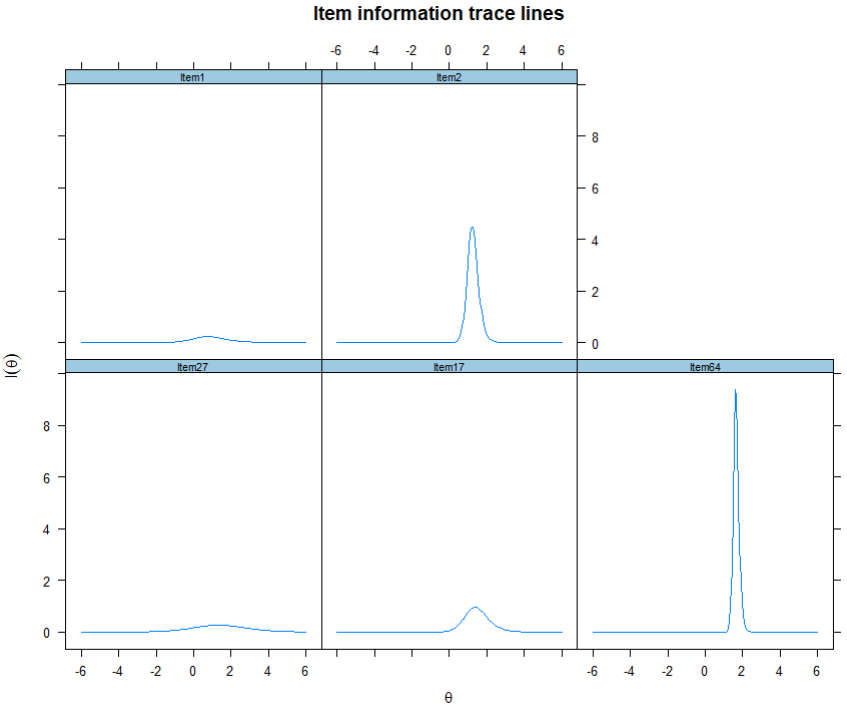


Figure 5.38: Threshold Test Item Information Curves

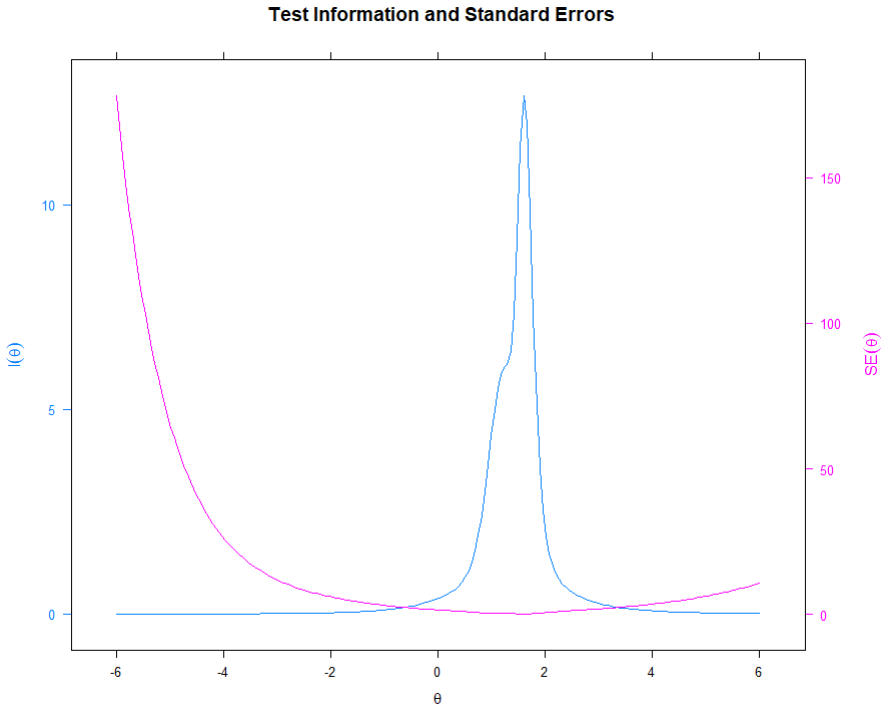


Figure 5.39: Threshold Test Information Curve

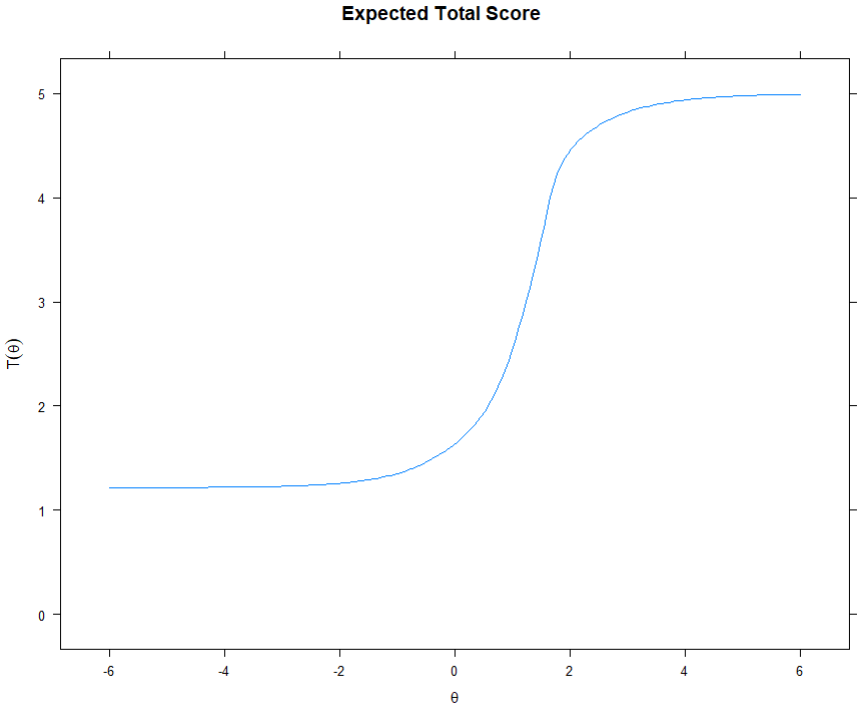


Figure 5.40: Threshold Test Characteristic Curve

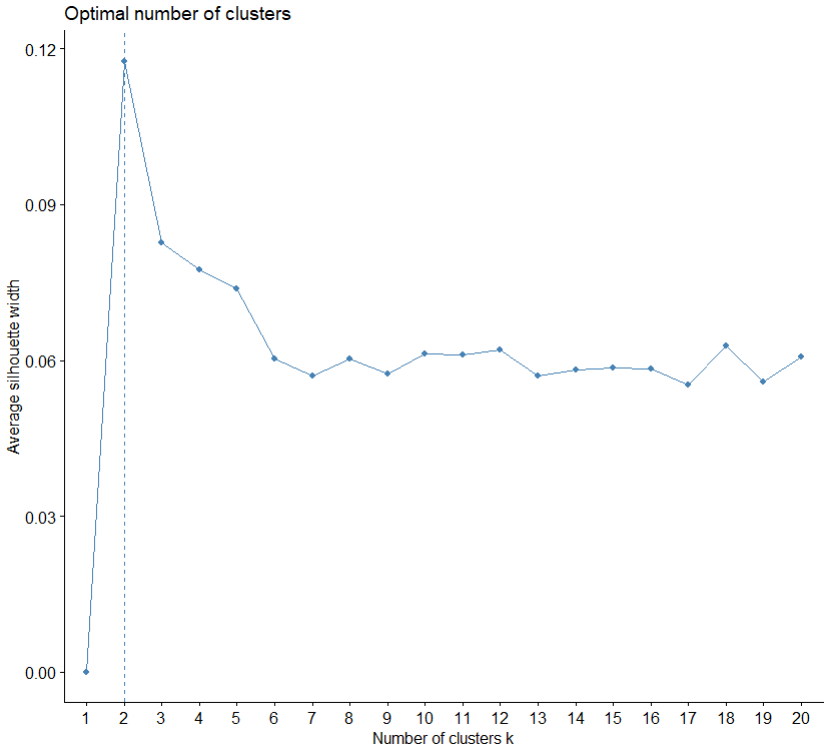


Figure 5.41: Optimal Number of Clusters

taken 17 principal components to reach 90% of the total variance, so principal component reduction looked less desirable. A k-nearest-neighbor model was tested against the data next, and this achieved similar results of 59% accuracy/ 41% error rate. The final approach attempted on classifying gender was a neural network. This achieved the best accuracy rate out of all of the statistical learning approaches, with an accuracy of 69% and error rate of 31%.

So, since there was some success (albeit far from perfect success) in classifying examinees based on gender, it supported the idea that overall, based on all of the data, there was likely some limited bias to gender. The less-than-ideal accuracy, though, and the other evidence presented throughout this paper supported the idea that, if there was any gender bias, it was small and likely not very significant.

Lastly, finding any clustering or groupings was examined, as it is hypothesized that with many tests taken, there is a more-prepared and a lesser-prepared cohort.

To begin with, the optimal number of clusters is explored. The available evidence suggests that two groups is the optimal number of clusters within this data.

```
Clustering Methods:
 hierarchical kmeans pam

Cluster sizes:
 2 3 4 5 6 7 8 9 10

Validation Measures:
                                  2         3         4         5         6         7         8         9        10

hierarchical Connectivity    13.9266  484.6500  518.0857  545.1353  545.1353  569.2056  612.0405  678.3687  699.0877
             Dunn             0.4472    0.2357    0.2357    0.2357    0.2357    0.2357    0.2357    0.2357    0.2357
             Silhouette       0.0914    0.0873    0.0656    0.0559    0.0445    0.0375    0.0340    0.0336    0.0252
kmeans       Connectivity   457.0774  740.8464  901.9635 1077.3306 1220.7290 1305.9948 1374.1373 1487.4575 1534.4048
             Dunn             0.2425    0.2425    0.2425    0.2500    0.2582    0.2500    0.2500    0.2500    0.2500
             Silhouette       0.1176    0.0823    0.0784    0.0746    0.0676    0.0656    0.0577    0.0591    0.0597
pam          Connectivity   602.8968  902.6421 1067.0393 1224.6663 1366.4762 1472.6262 1533.3306 1530.3683 1610.8679
             Dunn             0.2357    0.2425    0.2500    0.2500    0.2500    0.2500    0.2500    0.2582    0.2582
             Silhouette       0.1026    0.0728    0.0701    0.0569    0.0565    0.0561    0.0498    0.0454    0.0501

Optimal Scores:

             Score   Method       Clusters
Connectivity 13.9266 hierarchical 2
Dunn          0.4472 hierarchical 2
Silhouette    0.1176 kmeans       2
```

Figure 5.42: Number of Clusters Optimal with Selected Methods

Therefore, k-means clustering was undertaken with k=2, using 50 different random sets chosen. The graphical depiction is shown:

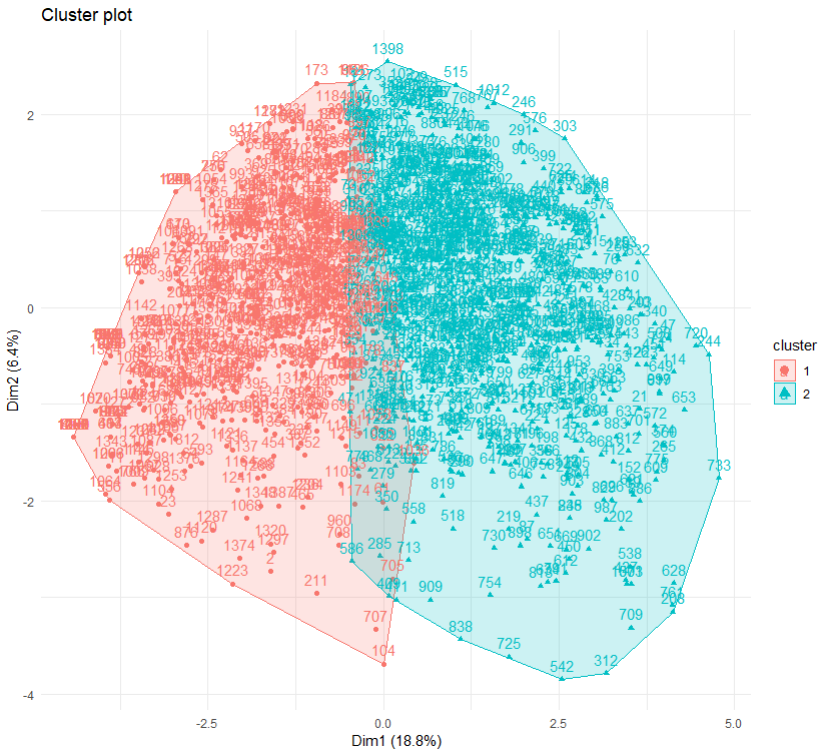


Figure 5.43: K-Means Clustering with k=2

There was clearly some separation between the two groups, so the clustering appeared fair; although, there was definitely room for improvement. A hierarchical approach to clustering was plotted with the dendogram shown:

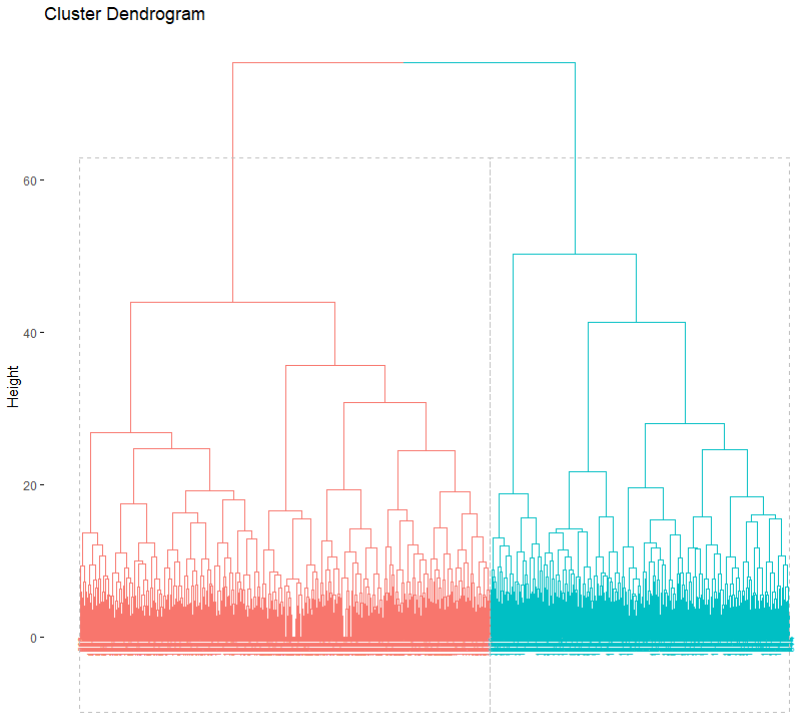


Figure 5.44: Hierarchical Clustering Dendrogram

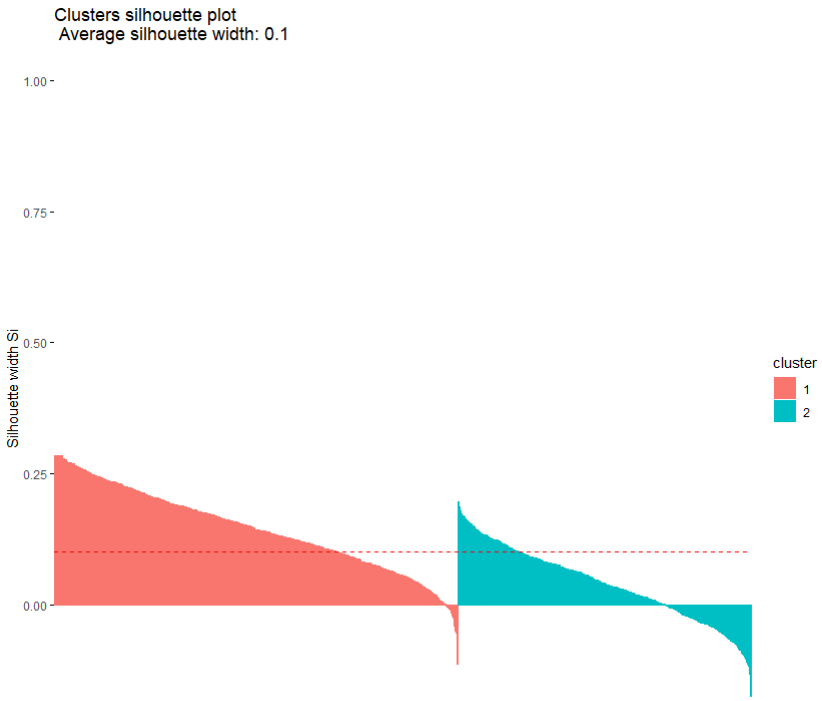


Figure 5.45: Silhouette Plot of the Two Clusters

```
Cluster means:
    Item49    Item27    Item41     Item7    Item38     Item28     Item9    Item47    Item75    Item17    Item76    Item10
1 0.960199 0.4212272 0.6865672 0.6633499 0.9270315 0.30182421 0.8756219 0.9751244 0.8175788 0.4643449 0.8208955 0.9452736
2 0.778607 0.1082090 0.1529851 0.2350746 0.5223881 0.09701493 0.5746269 0.8855721 0.3470149 0.1355721 0.3557214 0.7176617
     Item64     Item45    Item24     Item1    Item68    Item61    Item25     Item2    gender
1 0.6152570 0.8772803 0.6036484 0.7960199 0.8275290 0.9087894 0.7827529 0.4129353 0.6301824
2 0.3843284 0.5124378 0.1878109 0.5671642 0.5348259 0.6380597 0.4664179 0.1616915 0.6753731
```

Figure 5.46: Cluster Means with k=2

It definitely appeared that the k-means grouping was more prominent and it was chosen for a further check. The item comparison for the two groups was analyzed and shown:

As apparent, the first group performed significantly better on most items than the second group. To further check this, the mean of the entire first group's performance was checked, and this group average 14.68 correct questions out of the 20 total questions in the set. Meanwhile, the second cluster averaged as score of only 8.36 correct questions out of the 20 total questions in the set. If the Expected Total Score for the 3PL model were examined, these scores would be equivalent to approximately a +1.5 ability level and a -0.5 ability level, respectively. This implied that the two groupings, roughly speaking, corresponded to those that were prepared and did well versus those that were not prepared and did poorly.

As a small caveat here, it was desired to check and see if there was any clear gender bias between the two clusters, ie, were mostly males present in the upper group versus females in the lower group, or vice versa? The means between the groups with gender included are very small differences. For Group 1 (high group) females, the mean was 14.56, and for males 14.76. For Group 2 (the low group), the mean was 8.36 for both males and females. Therefore, there does not appear to be evidence of a gender bias among the two clusters.

We also wanted to look and see if there were any natural clusterings among the questions. To do this, we first removed about a dozen participants who scored all 1's, so that the analysis wouldn't produce matrices that wouldn't work. Then, the number of clusters was tested as before, with 2 being the number chosen. So, the k-means algorithm was run on the 20 questions, with the following plot produced showing clear separation into two groups:

Among these two groups, the mean item difficulty of Group 1 was 70.4% correct, while the mean item difficulty of Group 2 was 32.9% correct. This represented a huge difference in the items. Since the questions themselves are unknown, it is impossible to truly know why. However, these questions may have represented two distinct sub-categories within Biology

Figure 5.47: Cluster Among Questions (Variables) with k=2

(which the test was on). Or one group may have represented more obscure knowledge or a higher level of critical thinking than the other group. Or it could be that one group is an effectively performing set of questions and the other group, with only 32.9% of participants getting them right, is a poorly written or poorly performing group of questions and should be analyzed further. Should the questions have been known, this type of question analysis is exactly the type of investigation that should be conducted. With less than 1/3 of all examinees getting Group 2 questions correct, if the domain of Group 1's questions is not well known to an examinee, then inevitably they do badly on the test. So this type of analysis has proven fruitful, although the answers themselves remain obscured behind the question content.

Chapter 6

Conclusions

The field of Item Response Theory is an important and dynamic field in educational and psychological assessment. With a background intertwined in statistics, IRT makes it possible to administer standardized tests on a large scale basis and have the results be both interpretable and consistent throughout a large population of examinees. Before the underpinnings of IRT became widely used, Classical Test Theory attempted to use statistics to validly construct, administer, score, and interpret tests, but the limitations were huge in that the results were dependent on the examinees who took the test and on the test questions itself. IRT overcame these restrictions to become the most widely used paradigm in standardized testing today. However, as demonstrated in this paper, the ideas of statistical machine learning can readily compete with IRT in terms of classification and interpretation.

This paper delved into the basics of Item Response Theory and its paramater estimation techniques before analyzing a real data set using those techniques. Various graphs and curves were developed that showed the 3PL model fit the data the best. A subset of the data was then selected, using the full number of items as an item bank, to create a subtest based on a cutoff score. The diagnostics of this test were calculated, and its effectiveness discussed. Finally, statistical machine learning algorithms were used to compare with the IRT methods and demonstrate their potential usefulness in supplementing or replacing traditional IRT methods in testing theory and administration.

In their 2012 paper, Bergner et. al. [3] discussed that IRT parameter estimation is a technical and complicated iterative process and the goodness-of-fit analysis continues to be a

subject of research. They argue that a machine-learning approach to testing theory will be more powerful moving forward and provide more flexibility to estimate examinees ability accurately.

IRT also holds promise for working together with machine learning in unexpected ways. Martinez in 2016 [9] found that there are some interesting uses of IRT in machine learning models, particularly in IRT's methods of equating test sets for evaluation purposes. In the paper, Martinez applies IRT frameworks to a variety of machine-learning models and claims promise in using the two harmoniously in future endeavors.

IRT is also playing a big role in intelligent tutoring systems and computerized adapative testing (CAT), as mentioned by Desmarais and Baker [7] in their 2012 paper. Complicated algorithms such as Bayesian Knowledge Tracing and Deep Knowledge Tracing [10] (2015) use IRT and neural networks together to create a dynamic, responsive learning system that looks to the future in its important applications in educational and psychological testing. With the ability to adapt in real time to examinees responses, these complex systems are able to show a degree of flexibility and self-learning that is already proving useful.

This paper just scratched the surface on the deep and exciting field of Item Response Theory. However, even as more complex and sophisticated methods from machine and statistical learning become more widely available and used, the tenants of Item Response Theory still stand as the benchmark for standardized tests. Even if IRT is eventually supplanted by the advanced techniques of Deep Knowledge Tracing and Bayesian Knowledge Tracing with Computerized Adapative Testing, IRT will leave its legacy as the fingerprints inside the algorithms and will vicariously live on with its important contributions.

References

[1] Baker, F. B., & Kim, S. H. (2017). The basics of item response theory using R. New York, NY: Springer.

[2] Baker, Frank & Seock-Ho Kim. Item Response Theory: Parameter Estimation Techniques, 2nd edition, Revised and expanded. 2004.

[3] Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory. International Educational Data Mining Society.

[4] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores.

[5] Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1-29.

[6] Desjardins, C. D., & Bulut, O. (2017). Handbook of Educational Measurement and Psychometrics Using R. CRC Press.

[7] Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction, 22(1-2), 9-38.

[8] Drabinova, A. & Martinkova P. (2017). Detection of Differential Item Functioning with NonLinear Regression: Non-IRT Approach Accounting for Guessing. Journal of Educational Measurement, 54(4), 498-517

[9] Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2016, August). Making sense of item response theory in machine learning. In Proceedings of the Twenty-second European Conference on Artificial Intelligence (pp. 1140-1148). IOS Press.

[10] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In Advances in neural information processing systems (pp. 505-513).

Appendices

The R code used for this paper is provided here.

```
# include graphics in latex document

#graph a basic logistic curve
theta <- seq(-3,3,.1)
beta <- 0
alpha <- 1
P <- 1/(1+exp(-alpha*(theta-beta)))
plot(theta,P,type="l")

#load data set
library(difNLR)
View(MSATB)
write.csv(MSATB,"thesisdata.csv") # make a csv of the msat-b data

data1 <- MSATB
summary(data1) # overall upper end view of the data
sum(data1$gender) #number of females in the data set (coded as "1"), 923
nrow(data1)-sum(data1$gender) #number of males in the data set (coded as "0"), 484
AllScores <- NULL
for (i in 1:nrow(data1)){
examineeScore <- NULL
examineeScore <- sum(data1[i,1:20])/20
AllScores <- cbind(AllScores,examineeScore)
}
length(AllScores) # 1407 examinees
mean(AllScores) #Average score of all examinees, 0.5535892.
min(AllScores) #Minimum score, 0.1
max(AllScores) #Maximum score, 1
median(AllScores) #Median score, 0.55
hist(AllScores) #Mostly normally distributed it looks like
sd(AllScores) #Overall SD of 0.19

AllItems <- NULL
for (i in 1:(ncol(data1)-1)){ .
itemScore <- NULL
itemScore <- sum(data1[1:1407,i])/1407
AllItems <- cbind(AllItems,itemScore)
}
length(AllItems) # 20 items
mean(AllItems) #Average difficulty of all items, 0.5535892
min(AllItems) #Minimum difficulty (hardest item), 0.1848
max(AllItems) #Maximum difficulty (easiest item), 0.9240
median(AllItems) #Median difficulty, 0.5785
hist(AllItems)
sd(AllItems) #Item SD of 0.22
```

```
aggregate(data1,by=list(data1$gender),FUN=mean)
AllMaleScores <- NULL
AllFemaleScores <- NULL
for (i in 1:nrow(data1)){
maleScore <- NULL
if (data1[i,21] == 0) {
maleScore <-sum(data1[i,1:20])/20
AllMaleScores <- cbind(AllMaleScores,maleScore)
} else {
femaleScore <- NULL
femaleScore <- sum(data1[i,1:20])/20
AllFemaleScores <- cbind(AllFemaleScores,femaleScore)
}

}
length(AllMaleScores) #484 males, as it should be
length(AllFemaleScores) #923 females, as it should be
mean(AllMaleScores) #Males average a score of 0.56
mean(AllFemaleScores) #Females average a score of 0.55
range(AllMaleScores) #Scores range from 0.1 to 1.0 for males
range(AllFemaleScores) #Same for female scores
median(AllMaleScores) #0.55 for male median
median(AllFemaleScores) #0.55 for female median
sd(AllMaleScores) #SD of 0.19 for males
sd(AllFemaleScores)
hist(AllMaleScores)
hist(AllFemaleScores)

t.test(x=AllMaleScores,y=AllFemaleScores, alternative="two.sided",paired=F)


###################################
###################################

library(mirt)

#Rasch fit to the data
rasch_mod <- "F = 1 - 20"
rasch_fit <- mirt(data= data1[,1:20], model= rasch_mod, itemtype="Rasch", SE = T)
rasch_params <- coef(rasch_fit, IRTpars=T, simplify=T)
rasch_items <- rasch_params$items
rasch_items
plot(rasch_fit, type="trace") #ICC for each item
plot(rasch_fit, type="infotrace") # IIC for each item.
plot(rasch_fit, type="infoSE") # TIC and cSEM plot for the test
plot(rasch_fit, type="score") #Expected total score on the test
ability_mle <- fscores(rasch_fit, method="ML", full.scores=T, full.scores.SE=T)
head(ability_mle)
ability_finite <- ability_mle[is.finite(ability_mle[,1]),]
ability_summary <- apply(ability_finite,2,summary)
ability_summary
rasch_itemfit <- itemfit(rasch_fit, fit_stats = c("X2","G2","Zh"))
rasch_itemfit #data for the most part do not fit the Rasch model
#Three items chosen below to plot their empirical plot
itemfit(rasch_fit,empirical.plot=c(1))
itemfit(rasch_fit,empirical.plot=10)
itemfit(rasch_fit,empirical.plot=19)
sum(rasch_itemfit$X2) #616 total chi-square
rasch_personfit <- personfit(rasch_fit)
```

```
hist(rasch_personfit$Zh, xlab="Zh Statistic")
abline(v=-2, lwd=2, lty=2)

#2PL fit to the data
twopl_mod <- "F = 1 - 20"
twopl_fit <- mirt(data=data1[,1:20], model=twopl_mod, itemtype="2PL", SE=T)
twopl_params <- coef(twopl_fit, IRTpars = T, simplify = T)
twopl_items <- twopl_params$items
twopl_items
plot(twopl_fit, type="trace") #ICC for each item
plot(twopl_fit, type="infotrace") #IIC for each item
plot(twopl_fit, type="infoSE") #TIC and cSEM plot for the test
plot(twopl_fit, type="score") #Expected total score on the test
ability_mle2 <- fscores(twopl_fit, method="ML", full.scores=T, full.scores.SE=T)
head(ability_mle2)
ability_finite2 <- ability_mle2[is.finite(ability_mle2[,1]),]
ability_summary2 <- apply(ability_finite2,2,summary)
ability_summary2
twopl_itemfit <- itemfit(twopl_fit, fit_stats = c("X2","G2","Zh"))
twopl_itemfit #almost none of the data fits the 2PL model well
#Three items chosen below to plot their empirical plot
itemfit(twopl_fit,empirical.plot=c(1))
itemfit(twopl_fit,empirical.plot=10)
itemfit(twopl_fit,empirical.plot=19)
sum(twopl_itemfit$X2) #467.18 total chi-square
twopl_personfit <- personfit(twopl_fit)
hist(twopl_personfit$Zh, xlab="Zh Statistic")
abline(v=-2, lwd=2, lty=2)

#3PL fit to the data
threepl_mod <- "F = 1 - 20"
threepl_fit <- mirt(data=data1[,1:20], model=threepl_mod, itemtype="3PL", SE = T)
threepl_params <- coef(threepl_fit, IRTpars = T, simplify = T)
threepl_items <- threepl_params$items
threepl_items
plot(threepl_fit, type="trace") #ICC for each item
plot(threepl_fit, type="infotrace") #IIC for each item
plot(threepl_fit, type="infoSE") #TIC and cSEM plot for the test
plot(threepl_fit, type="score") #Expected total score on the test
ability_mle3 <- fscores(threepl_fit, method="ML", full.scores=T, full.scores.SE=T)
head(ability_mle3)
ability_finite3 <- ability_mle3[is.finite(ability_mle3[,1]),]
ability_summary3 <- apply(ability_finite3,2,summary)
ability_summary3
threepl_itemfit <- itemfit(threepl_fit, fit_stats = c("X2","G2","Zh"))
threepl_itemfit #data for the most part do not fit the 3PL model
#Three items chosen below to plot their empirical plot
itemfit(threepl_fit,empirical.plot=c(1))
itemfit(threepl_fit,empirical.plot=10)
itemfit(threepl_fit,empirical.plot=19)
sum(threepl_itemfit$X2) #433 total chi-square
threepl_personfit <- personfit(threepl_fit)
hist(threepl_personfit$Zh, xlab="Zh Statistic")
abline(v=-2, lwd=2, lty=2)

#4PL fit to the data
fourpl_mod <- "F = 1 - 20"
fourpl_fit <- mirt(data= data1[,1:20], model=fourpl_mod, itemtype="4PL", SE = T)
fourpl_params <- coef(fourpl_fit, IRTpars = T, simplify = T)
fourpl_items <- fourpl_params$items
```

```
fourpl_items
plot(fourpl_fit, type="trace") #ICC for each item
plot(fourpl_fit, type="infotrace") #IIC for each item
plot(fourpl_fit, type="infoSE") #TIC and cSEM plot for the test
plot(fourpl_fit, type="score") #Expected total score on the test
ability_mle4 <- fscores(fourpl_fit, method="ML", full.scores=T, full.scores.SE=T)
head(ability_mle4)
ability_finite4 <- ability_mle4[is.finite(ability_mle4[,1]),]
ability_summary4 <- apply(ability_finite4,2,summary)
ability_summary4
fourpl_itemfit <- itemfit(fourpl_fit, fit_stats = c("X2","G2", "Zh"))
fourpl_itemfit #data for the most part do not fit the Rasch model
#Three items chosen below to plot their empirical plot
itemfit(fourpl_fit,empirical.plot=c(1))
itemfit(fourpl_fit,empirical.plot=10)
itemfit(fourpl_fit,empirical.plot=19)
sum(fourpl_itemfit$X2) #382 total chi-square.
fourpl_personfit <- personfit(fourpl_fit)
hist(fourpl_personfit$Zh, xlab="Zh Statistic")
abline(v=-2, lwd=2, lty=2)

#model fit comparison
anova(rasch_fit,twopl_fit)
anova(twopl_fit, threepl_fit)
anova(threepl_fit,fourpl_fit)

#constructing a test using cutoff score of +1
# our model, as found above, will be the 3PL

threepl_items
subtest <- data1[,c(2,10,13,16,20)]
subtest_mod <- "F = 1 - 5"
subtest_fit <- mirt(data= subtest, model=subtest_mod, itemtype="3PL", SE = T)
subtest_params <- coef(subtest_fit, IRTpars = T, simplify = T)
subtest_items <- subtest_params$items
subtest_items
plot(subtest_fit, type="trace") #ICC for each item
plot(subtest_fit, type="infotrace") #IIC for each item
plot(subtest_fit, type="infoSE") #TIC and cSEM plot for the test
plot(subtest_fit, type="score") #Expected total score on the test
subtest2 <- data1[,c(4,8,10,12,17)]
subtest_mod2 <- "F = 1 - 5"
subtest_fit2 <- mirt(data= subtest2, model=subtest_mod2, itemtype="3PL", SE = T)
subtest_params2 <- coef(subtest_fit2, IRTpars = T, simplify = T)
subtest_items2 <- subtest_params2$items
subtest_items2
plot(subtest_fit2, type="trace")
plot(subtest_fit2, type="infotrace")
plot(subtest_fit2, type="infoSE")
plot(subtest_fit2, type="score")




subtestScores <- fscores(subtest_fit,method="ML", full.scores=T)
for (i in 1:length(subtestScores)){
if(!is.finite(subtestScores[i,1])){
if(subtestScores[i,1] <0){
subtestScores[i,1] <- min(subtestScores[is.finite(subtestScores[,1]),])
} else {
```

```
subtestScores[i,1] <- max(subtestScores[is.finite(subtestScores[,1]),])
}
}
}
PassScore <- table(subtestScores >=1.5)
PassScore
ResultVector <- NULL
for(i in 1:length(subtestScores)){
if(subtestScores[i,1]>=1.5){
res <- "PASS"
ResultVector <- rbind(ResultVector,res)
} else {
res <- "FAIL"
ResultVector <- rbind(ResultVector,res)
}
}
ScoreResults <- as.data.frame(cbind(ResultVector,subtestScores[,1]))
colnames(ScoreResults) <- c("Result", "Score")
rownames(ScoreResults)<-1:1407
rawData <- data1[,c(2,10,13,16,20)]
ScoreData <- as.data.frame(cbind(ScoreResults,rawData))
colnames(ScoreData) <- c("Result","Score","Q1","Q2","Q3","Q4","Q5")
trainingSet1 <- ScoreData[1:1000,]
testSet1 <- ScoreData[1001:1407,]
library(arm)
modelLog <- bayesglm(data=trainingSet1, Result ~ Q1+Q2+Q3+Q4+Q5,
 family="binomial", control=list(maxit=100), prior.df=10)
summary(modelLog)
predLog <- predict(modelLog, testSet1)
tab <- table(round(predLog),testSet1$Result)
tab

trainData <- data1[1:1000,]
testData <- data1[1001:1407,]

trainLog <- glm(data=trainData, trainData$gender~.,
 family="binomial", control=list(maxit=100))
summary(trainLog)
predtrainLog <- predict(trainLog,testData, type="response")
tab2 <- table(round(predtrainLog),testData$gender)
tab2
errorRateLog <- 1-sum(diag(tab2))/(sum(rowSums(tab2)))
errorRateLog

library(randomForest)
trainData$gender <- as.factor(trainData$gender)
testData$gender <- as.factor(testData$gender)
f <- as.formula(paste("trainData$gender~",
paste("trainData[,",1:20,"]",collapse="+")))
rf <- randomForest(data=trainData,x=trainData[,1:20],
y=trainData$gender, xtest = testData[,1:20], ytest=testData[,21], ntree=1000)
mean(rf$err.rate) #41% error rate

summary(prcomp(trainData[,1:20]))
biplot(prcomp(trainData[,1:20]))
screeplot(prcomp(trainData[,1:20]))

library(class)
k <- knn(train=trainData[,1:20],test=testData[,1:20], cl=trainData[,21],k=5, prob=T)
summary(k)
```

```
tab3<- table(k,testData[,21])
tab3
errRate <- 1 - sum(diag(tab3))/sum(rowSums(tab3))
errRate #error rate of almost 41%, mostly due to males.

library(neuralnet)
trainData$gender <- as.numeric(trainData$gender)
testData$gender <- as.numeric(testData$gender)
n <- neuralnet(formula=trainData$gender~., data=trainData,
hidden=c(5,2), stepmax=1e+06, algorithm="backprop",learningrate=0.001, linear.output=F)
summary(n)
computeN <- compute(n,testData[,1:20])
plot(testData[,21],computeN$net.result)
mean(abs(computeN$net.result-testData[,21])/testData[,21]) #31% error, best yet.



library(cluster)
library(factoextra)
library(NbClust)
fviz_nbclust(data1,kmeans,method="silhouette", k.max=20)
NbClust(data1,method="kmeans")
# Says 2 groups is the best, both methods
km.data <- kmeans(data1,2,nstart=50)
fviz_cluster(km.data, data=data1[,1:20], frame.type="convex")+ theme_minimal()
print(km.data)


library(clValid)
cl <- clValid(data1, nClust=2:10,
clMethods=c("hierarchical","kmeans","pam"),validation="internal", maxitems=1500)
summary(cl)

data.hc <- hcut(data1,k=2,stand=T)
fviz_dend(data.hc,rect=T,cex=0.5)

fviz_silhouette(data.hc) #not a very strong clustering


g<-get_clust_tendency(data1, n=300, gradient=list(low="blue",high="white"))
g$hopkins_stat #0.42, data is not that clusterable.
g$plot

Dat2Sum <- NULL
Dat2Sum2 <- NULL
data2 <- as.data.frame(cbind(data1,km.data$cluster))
for (i in 1:nrow(data2)){
if (data2[i,22]==1){
Dat2Sum <- cbind(Dat2Sum,sum(data2[i,1:20]))
} else {
Dat2Sum2 <- cbind(Dat2Sum2,sum(data2[i,1:20]))
}
}
mean(Dat2Sum) #mean score of 14.68 correct
mean(Dat2Sum2) # mean score of 8.36 correct

a <- aggregate(data2[,1:20],by=list(data2[,21],data2[,22]), mean)
rowSums(a[,3:22])
```

```
#We also want to look and see if there are natural clusterings among the questions.
tr <- t(data1)


tr <- tr[,-which(colSums(tr[1:20,])==20)]
fviz_nbclust(tr[1:20,],kmeans,method="silhouette", k.max=15)
NbClust(tr[1:20,],method="kmeans") #error

km.data2 <- kmeans(tr[1:20,],2,nstart=50)
fviz_cluster(km.data2, data=tr[1:20,], frame.type="convex")+ theme_minimal()
# Definite groupings between items.

#What is the mean of the two groups of questions?


tr <- t(data1)

data3 <- as.data.frame(cbind(tr[1:20,],km.data2$cluster))
a2 <- aggregate(data3[,1:1407],by=list(data3[,1408]),mean)
rowSums(a2[,1:1407])/1407
```