The University of Texas at TYLER The Graduate School

University of Texas at Tyler
Scholar Works at UT Tyler

Math Theses

Math

Spring 5-7-2012

# Item Response Models for Dichotomous and Polytomous Data in the Context of Generalized Linear Models with Applications

Grant Campbell

# Item Response Models for Dichotomous and Polytomous Data in the Context of Generalized Linear models with Applications

by

Grant Campbell

A thesis submitted in partial fulfillment
of the requirements for the degree of
Masters of Science
Department of Mathematics

Nathan Smith, Ph.D, Committee Chair

College of Arts and Sciences

The University of Texas at Tyler
May 2012

The University of Texas at Tyler
Tyler,Texas

This is to certify that the Master's Thesis of

GRANT CAMPBELL

has been approved for the thesis requirement on
April 10th, 2012
for the Masters of Mathematics degree

Approvals:

Nathan Smith

Thesis Chair: Nathan Smith, Ph.D.

Member: Reagan Beckham, Ph.D.

Ron Livingston

Member: Ron Livingston, Ph.D

Chair, Department of Mathematics

Martin Slaney

Dean, College of Arts and Sciences

# Contents

# List of Tables

## List of Figures

# Abstract

## Item Response Models for Dichotomous and Polytomous Data in the Context of Generalized Linear models with Applications

Grant Campbell

Thesis Chair: Nathan Smith, Ph.D.

The University of Texas at Tyler
May 2012

Item response theory is a test theory, in contrast to classical test theory, that focuses on the individual items of an exam in order to analyze test accuracy amd reliability, and evaluate examinee ability levels. Developed in the mid 20th century, item response theory, or IRT, is consider superior in a number of ways to many other test theory approaches. Provided here is an overview of basic IRT models using the a test theory approach, as well as the development of IRT models in the context of generalized linear models. Binary, correct/incorrect, response and polytomous, or mutiple, response items are considered for item response theory models developed here in the context of generalized linear models. Applications of IRT are also explored using the statistical software R.

# 1 Item Response Theory

## 1.1 Introduction

Item Response Theory is a mathematical way of creating and analyzing models mostly used in the field of psychometrics. Quite popular within the last twenty to thirty years due to improvements in computing power, the concepts of IRT have been around since the middle of the twentieth century. So let us first begin by describing, in a very broad sense, the idea of classical test theory so as to have something to compare with item response theory. As for any test theory classical test theory is used to evaluate test scores, evaluate the ability of a test to measure some latent trait, and determine the reliability of the test itself. To achieve these objectives classical test focuses on the observed scores that students produce on a given test. Classical test theory then assumes that these observed scores are equivalent to some true score plus some random error term that in most cases is assumed to be normally distributed. Therefore classical test theory is developed around this relationship between the observed scores and the true score and error, this is where item response theory will differ from classical test theory. Like classical test theory item response theory, or IRT, is a method used to create, analyze, and calculate scores or a persons "ability" on tests, exams, surveys, and questionnaires. Differing from classical test theory, item response theory focusses on individual items that appear on the exam or survey. Focussing on individual items rather than entire test scores allow the analyst to determine not just how well a person does on an evaluation but how well that evaluation assesses the latent trait being tested (i.e.

a persons math skills if the evaluation is a math test). IRT models also allow one to determine within a test which items or questions work better to evaluate the examinees ability or trait being tested. Because item response theory has the ability to test individual questions on exams it is used quite frequently for educational purposes to collect banks of test questions and for improvement of exams over time since poor questions can be quickly realized. More recently item response theory has been implemented in the use of computer based testing or computerized adaptive testing, in which the test can be adapted to an examinees ability by asking questions that are more closely related to the person's ability level. The benefit of IRT is that even though difficulties of exam questions might differ between examinees, one can evaluate or estimate all examinees ability level on the same continuum.

Let us first discuss the assumptions that must be met in order to accept information gained by IRT. There exist three assumptions when dealing with item response models:

1) The latent trait being studied, usually denoted $\theta$, is one-dimensional. The latent trait of an item response model is defined as the measurable ability level that is being tested which differs over persons within the tested group.

2) Items are locally independent, i.e. the probability of responding to item $i$ has no effect on the probability of responding to item $j$.

3) There exists a function , called the item response function, that relates the person's latent trait and the actual response the person makes on an item.

The assumptions for IRT models are considered stronger than the assumptions for classical test theory, which are essentially that there exists a raw test score for every individual which consists of an observed score and a random error and that random error is normally distributed with an expected value of zero. Logically one might note that the assumption of item independence in IRT might be a little misleading, since responding to one item might have an effect on a person's response to another item. It has been shown that "Monte Carlo work and experience with IRT programs suggests that minor violations of this assumption do not make much of a difference" [6].

## 1.2 Rasch Model

Let us now consider the item response function that will be spotlighted and built upon in our discussion throughout this article. The most commonly used statistical model in IRT to relate the probability of choosing the correct response to an item with some ability or latent trait level is the logistic model. The aim of our discussion is to express item response models as a subset of a larger group of models termed generalized linear models. Logistic models are a particular type of generalized linear models and will be further expanded on in our general discussion on generalized linear models. For now let us simply define what the logistic regression function is and in our discussion on generalized linear models we will further explore the entire logistic model. The logistic function is defined as

$$f\left(y\right) = \frac{e^{y}}{1+e^{y}} \tag{1.1}$$

where $y$ is a linear combination of predictor variables in a statistical model. We will explore three forms of this logistic function and how they relate to item response theory. The three logistic models discussed differ in how many parameters are used to explain the relationship between the latent trait being studied and the probability of choosing the correct response to an item. The simplest logistic model, and quite possibly the model most used in the field of IRT, is the one-parameter logistic model or Rasch model, named after Georg Rasch a Danish mathematician and statistician. Note here that the logistic function is termed the Rasch model in the field of item response theory. The Rasch model, which allows us to calculate the probability of choosing the correct response for a given latent trait and item parameter, takes the following form.

$$P\left(Y_{pi}=1|\theta_p,\beta_i\right)=\frac{\exp\left[\theta_p-\beta_i\right]}{1+\exp\left[\theta_p-\beta_i\right]}. \tag{1.2}$$

Here $Y_{pi}=1$ refers to choosing a "correct" response to item $i$ by person $p$ ($Y_{pi}=0$ would account for an incorrect response). Note a correct response might not necessarily be a correct answer to an exam question, but might simply be the studied response on a survey. Furthermore $\theta_p$ represents the latent trait for person $p$ (a person's skill level if the items were exam questions), and $\beta_i$ denotes the item parameter (if the item is on an exam, testing a certain skill level, $\beta_i$ would be the difficulty of that item for whatever skill being tested.) Although it might not be clear yet, one might notice the similarity of the Rasch model in 1.2 and the logistic model in 1.1. In fact the Rasch model is simply an application of the logistic model which we discuss in a later chapter. The Rasch model is mathematically nice in the sense that it has domain of $\mathbb{R}$ for the latent trait, $\theta_p$ and range of $(0,1)$ for the probability of choosing the "correct" response to the item. Another important aspect of the Rasch model is that the fixed item parameter, $\beta_i$, is scaled on the same continuum as the person's latent trait. The fact that the item parameters and the person parameter allows one to relate specific person abilities to specific difficulties of a test question. Relating unique ability parameters to unique item parameters gives a test creator the ability to develop specialized tests depending on the group or person being tested since ability levels are unique amongst groups or individuals. The curve representing the logistic IRT model is termed the item characteristic curve or ICC. As seen in Figure 1.1 the point at which the the

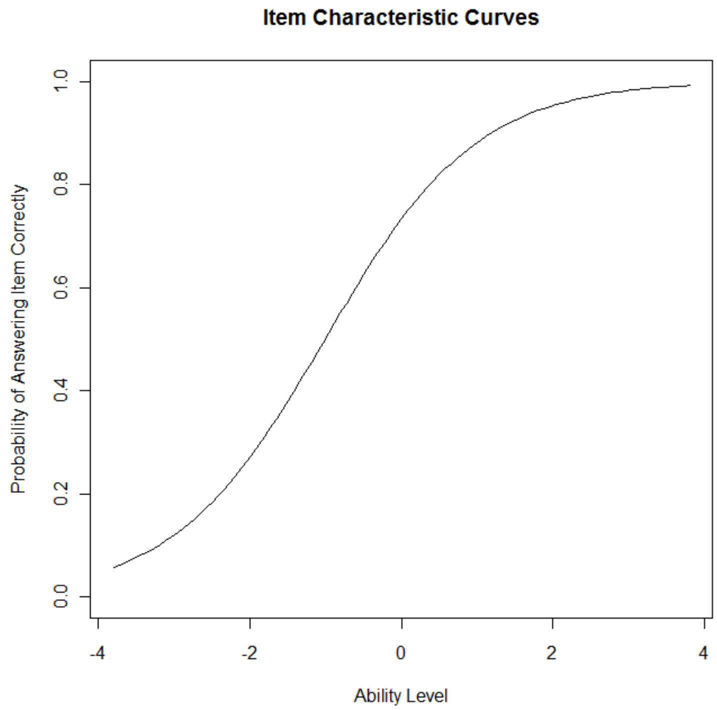


Figure 1.1: $P\left(Y_{pi} = 1 \middle| \, \beta_i = -1.009\right) = \frac{e^{\theta_p + 1.009}}{1 + e^{\theta_p + 1.009}}$

probability of answering the item correctly is 0.50 on the item characteristic curve is the position of the item difficulty. Therefore, as the item difficulty increases, the probability of answering that item correctly decreases for a fixed person latent trait, $\theta_p$. Figure 1.2 shows this idea by plotting several Rasch model curves with different item difficulty levels together The Rasch Model is essentially different to most statistical models for fitting data in the sense that the Rasch model requires the data to fit the model rather than trying to fit a model to the data. In order to use the properties of the Rasch model, then one must have data that approximately fits the model. Therefore when obtaining the data one wishes to examine using this Rasch model, the analyst is allowed to discard any data values that do not conform to the model. That is if the majority of observed response patterns, an examinees set of responses, follow a certain pattern, then outlier response patterns may be discarded as they will not affect the model. The next natural question to ask is how

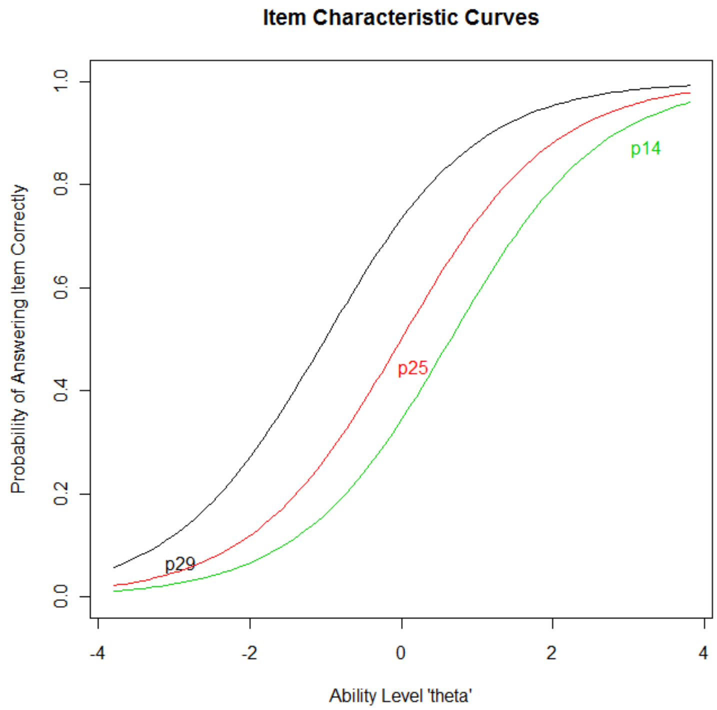


Figure 1.2: Item p29: $\beta_i = -1.009$, Item p25: $\beta_i = -0.001$, Item p14: $\beta_i = .648$

do the values discarded affect the measurements gained from using the Rasch model? This idea of "dropping" data values is considered admissible since these values are considered extreme as well as the information obtained from fitting the model is quite significant [3]. Keep in mind that the significance of the Rasch model allows an analyst to very accurately choose items that test the latent trait of an examinee. Therefore the idea of having data fit the model is not a far stretch for what the Rasch model is being used for. Now using the Rasch model an analyst can only address a valid latent trait explained by the data if the data conforms to the the model the analyst is using. Because of this approach to using the one-parameter logistic model, it is considered more of a confirmatory approach rather than an explanatory approach that allows an analyst to manipulate a model to fit data. The idea of only using data that conforms to the model requires very large data sets when applying a one-parameter Rasch model to explain a latent described by the data. As stated earlier, because of the need for very large data sets and, as will be seen later, the use of non-closed form integrals for parameter prediction, IRT and the use of the Rasch model have only become applicable in in recent decades due to increased computing power for performing numerical approximations.

## 1.3 Two-Parameter Logistic Model

The second basic IRT model studied is called the two-parameter logistic model. This model introduced a parameter, denoted $\alpha_i$, which allows for the discrimination an item might have on the latent trait being studied. The discrimination factor in the two-parameter logistic model changes the slope at the inflection point of the item characteristic curve, i.e. the slope at $P(Y_{pi} = 1) = .05$. The two-parameter logistic IRT model is given by the following,

$$P\left(Y_{pi} = 1 | \theta_p, \beta_i, \alpha_i\right) = \frac{\exp\left[\alpha_i\left(\theta_p - \beta_i\right)\right]}{1 + \exp\left[\alpha_i\left(\theta_p - \beta_i\right)\right]} \tag{1.3}$$

where $Y_{pi} = 1$, $\theta_p$, $\beta_i$ represent the same parameters as in the Rasch model, and $\alpha_i$ represents the item discrimination. Let us now discuss the physical interpretation of the item discrimination factor. As stated previously the item discrimination affects the slope at the point where the probability of choosing the correct response is .05, and one can see in Figure 1.3 the lower the value for $\alpha_i$ the flatter the item characteristic curve and the higher the $\alpha_i$ value the steeper the curve. The $\alpha_i$ parameter is called the discrimination factor since it describes how well an item distinguishes the probability of responding correctly among the examinees along different ability levels. Consider a low $\alpha_i$ value, i.e. a flatter item characteristic curve, then large changes in the ability level of the examinees leads to only small changes in the probability of that the examinees will respond to the item correctly. Large $\alpha_i$ values cause the ICC to be steep, which in turn means small changes in

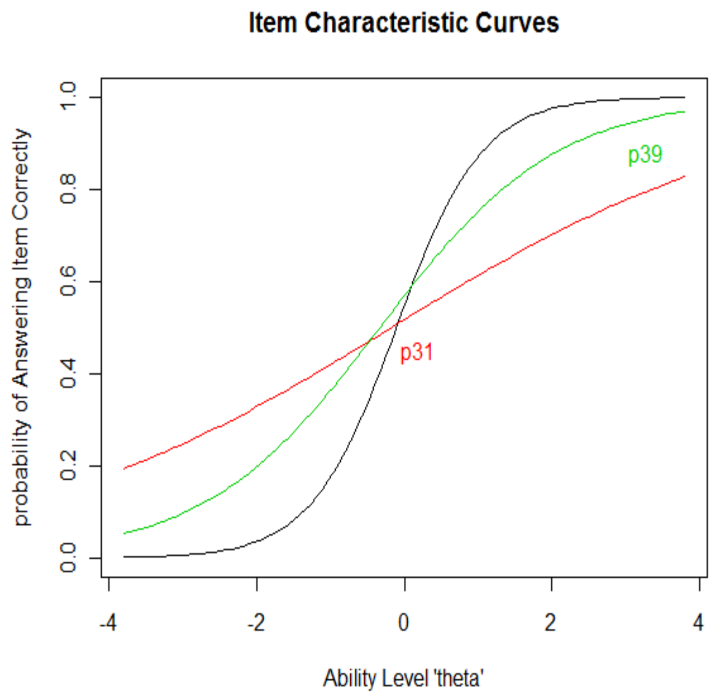


Figure 1.3: Item p31 $\beta_i = -1.77$, $\alpha_i = .393$; Item p39 $\beta_i = -.332$, $\alpha_i = .835$; Item p15: $\beta_i = -0.11$, $\alpha_i = 1.732$

ability lead to large changes in the probability of responding correctly. Thus, items with larger discrimination factors are more desirable to analysts who wish to use the items that describe an examinees ability the greatest. Notice how the discrimination factor plays a role in the item characteristic curve, consider a discrimination factor value of $\alpha_i = 0$, which would yield the item characteristic curve of

$$P\left(Y_{pi} = 1 | \theta_p, \beta_i, \alpha_i = 0\right) = \frac{\exp\left[\alpha_i\left(\theta_p - \beta_i\right)\right]}{1 + \exp\left[\alpha_i\left(\theta_p - \beta_i\right)\right]} = \frac{\exp\left[0\right]}{1 + \exp\left[0\right]} = \frac{1}{2}. \tag{1.4}$$

Therefore an alpha value of zero yields the line with slope 0 at the probability value of .5. In other words, if there is no discrimination for an item an examinee with any ability level will have a 50% of selecting the correct answer. This type of item would be worthless for an analyst looking for quality items for an examine or survey.

Consider the relationship between the two-parameter logistic IRT model and the

Rasch model. If the assumption is made that all items have equal discrimination, i.e. this is equivalent to setting all discrimination factors to one, the Rasch model falls out. This assumption leads to less information about the items in consideration, but yields a much simpler model and methods that can be used when approximating item parameters [6] [4].

## 1.4 Three-Parameter Logistic Model

A third basic IRT model is called, in conjunction with the previous model, the three-parameter IRT model, or three-parameter logistic model. The additional parameter included in this model takes into consideration the ability of an examinee to guess the correct response of an item. This "guessing" factor takes the form of a lower asymptote on the item characteristic curve. This lower asymptote correctly describes the allowance for guessing the correct response since it would allow for the probability of choosing the correct response to increase for those examinees with very low ability levels. Consider the item characteristic curve in Figure 1.4. It is clear that even those examinees with ability levels of -2.5 and lower will still have about a 30% chance of choosing the correct response, but the probability of choosing the correct response for high ability examinees has not been affected. Therefore the three-parameter logistic model takes the form,

$$P\left(Y_{pi}=1|\theta_p,\beta_i,\alpha_i,c_i\right)=c_i+\left(1-c_i\right)\frac{\exp\left[\alpha_i\left(\theta_p-\beta_i\right)\right]}{1+\exp\left[\alpha_i\left(\theta_p-\beta_i\right)\right]}. \tag{1.5}$$

There exist four and five-parameter logistic models, but there is a lack of literature applying these more complex logistic models to item response theory. Also, our discussion will mainly focus on the Rasch, or one-parameter logistic, model and extensions of it. As one might imagine as the number of parameters used in the model increases the ability to accurately approximate the values of these

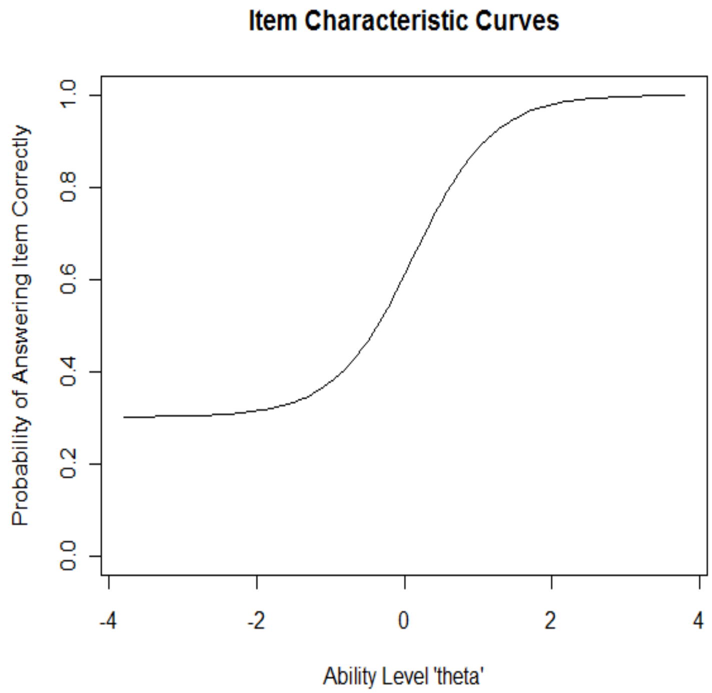


Figure 1.4: ICC with $\beta_i = 0.128$, $\alpha_i = 1.862$, $c_i = .302$

parameters, given a set of responses, would decrease. This fact leads to many using IRT models to ignore most other models besides the simplest, i.e. the Rasch model. Most IRT analysts consider the Rasch model sufficient in the sense that it best describes their data to the extent in which they need, although there are those who adamantly defend the need to use more complex models. It has also been confirmed that as the complexity of the model increases, much more data is needed to accurately approximate the item characteristic curves for items.

## 1.5 Item Information and Scoring

Let us not forget the primary purpose of implementing a well defined test theory, which is to accurately approximate an examinees latent trait or ability being tested given a set of responses to questions or items. The information function for a specific item is important in determining how accurate this approximation is along with other applications. We define the information function of an item as the reciprocal of the precision measurement of the item. That is, the precision of an item is how well that item accurately estimates the ability level, hence the variance of the ability estimates around the actual ability parameter. Therefore we define the information for a specific item on an exam as follows:

$$I_i(\theta) = \frac{1}{\sigma^2} \tag{1.6}$$

where $\sigma^2$ takes different forms depending on the type of model being used. For instance, in regards to the Rasch model and the 2-parameter model,

$$\sigma_i^2 = \frac{1}{a_i^2 P_i\left(\hat{\theta}\right) Q_i\left(\hat{\theta}\right)} \tag{1.7}$$

where $a_i$ is the discrimination factor of item $i$ (recall for the Rasch model we let $a_i = 1$), $\hat{\theta}$ is the estimated value of the ability parameter $\theta$, $P_i\left(\hat{\theta}\right)$ represents the probability of answering correctly to item $i$ give ability level $\hat{\theta}$, and $Q_i\left(\hat{\theta}\right) = 1 - P_i\left(\hat{\theta}\right)$ [4]. Therefore it is clear to see by Figure 1.5 that as the standard deviation (i.e. $\sigma$) increases (the precision of estimating the parameter is decreasing), the amount of information a particular item yields decreases for that particular ability level. So, the information function can tell us how well an item will accurately test a person's ability level, although an item with a large amount of information for a specific ability level will have very little information for other ability levels. Therefore you would not want to use items with large amounts of information for an ability level of 2 to accurately test a group of individuals with ability level of -1. The next step is to determine an information function for an entire test since we use entire tests to approximate an examinees ability level. We define the test information function in a very natural way, that is for a test with $N$

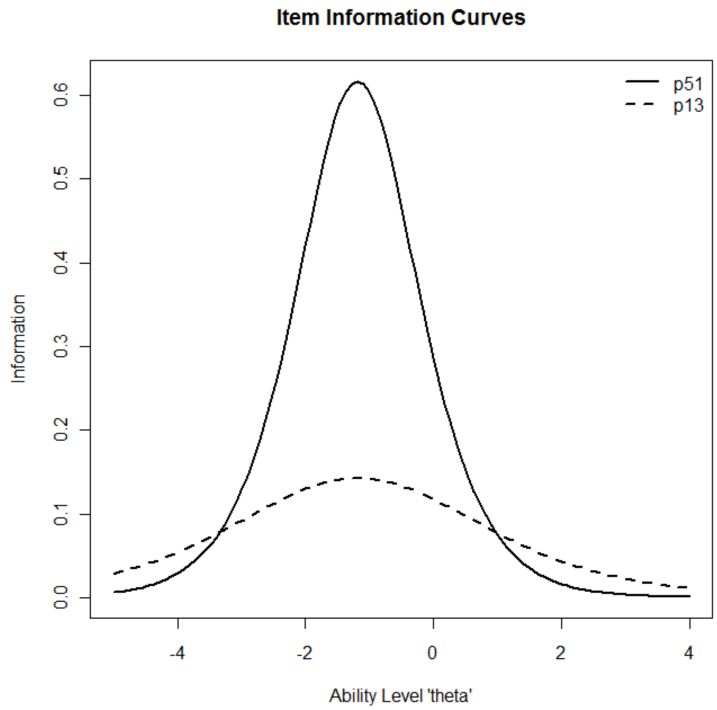


Figure 1.5: For $\theta \approx -1.18$, p13: $\sigma^2 \approx 7.0126$ and p51: $\sigma^2 \approx 1.6205$

items

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta). \tag{1.8}$$

Clearly the test information function will always be equal to or larger than any one item information function (it will only be equal to if the test consists of only one item, in which case the test would not be very strong). Theoretically a test maker would want to create a test with a test information function that is very high over a large range of ability levels. This can be achieved by using a number of questions that have high information functions over a wide range of ability levels.

Once a quality test has been created, approximating ability levels based on how an examinee responded is the final step. This is achieved by holding the item parameters stable and using methods to estimate the ability parameter, in the same manner that we held the ability parameters stable to estimate the item parameters, which was briefly discussed earlier. Note that the item parameter estimation methods will be explored in greater detail in a later section. We take an iterative approach when approximating the ability parameter $\theta$ using a maximum likelihood estimation method. For the Rasch and 2-parameter model we first estimate an ability of the examinee and denote it $\hat{\theta}_s$ then,

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} -a_i^2 \left[ y_i - P_i\left(\hat{\theta}_s\right) \right]}{\sum_{i=1}^{N} a_i^2 P_i\left(\hat{\theta}_s\right) Q_i\left(\hat{\theta}_s\right)} \tag{1.9}$$

where $a_i$ is the discrimination parameter of item $i$, $y_i$ is the response of the examinee ($y_i = 1, 0$), $P_i\left(\hat{\theta}_s\right)$ is the probability that the examinee with ability level $\hat{\theta}_s$ correctly responds to item $i$ and $Q_i\left(\hat{\theta}_s\right) = 1 - P_i\left(\hat{\theta}_s\right)$ [4]. This iterative process continues until the different from one iteration to the next is negligible. This process is valid for approximating the ability level since as the ability level becomes closer and closer to the real value of $\theta$ the numerator of the second term in the iteration formula will gradually converge to zero. And thus the ability has been accurately approximated. One note to make is that if an examinee has a response pattern of every item being answered correctly or incorrectly, this iterative method yields no ability estimate for the examinee since the iterations will not converge to a real value. Therefore the analyst must take this into consideration when using the previously discussed estimation procedure. There exist a number of other methods for accurately approximating an examinees ability level given that the item parameters have been accurately estimated. One should keep in mind that the ability parameters are scaled on the same continuum as the item parameters and usually range from values of -3 to 3, thus interpretation of scoring in item response theory is completely different than that of classical test theory which computes a score based off of total correct responses. In other words, two examinees that answered the same number of responses correctly could very possibly have two different scores using item response theory depending on which items they responded correctly to. Because of this weighted score that item response theory brings to test theory, many believe that IRT is a superior method for scoring an examinees ability, especially when considering high stakes exams.

## 2 Generalized Linear Models

To further explore item response theory in the context we wish, we will need a brief review of generalized linear models. Generalized linear models unify a number of different statistical modeling methods under a single theoretical approach. For example one can express everything from the basic general linear model to logit analysis and analysis of variance as simply special cases of a broader group of generalized linear models.

The first step in formalizing generalized linear models is to understand the exponential family of functions. The exponential family is simply a group of probability density functions that can be rewritten in a certain form for theoretical convenience. The exponential family was developed by R.A. Fisher, an English statistician and biologist, who discovered that many probability mass and density functions can be represented as a more general type of function [2]. Functions of the exponential family are probability density functions (PDFs) and probability mass functions (PMFs) of the following form:

$$\begin{aligned} f(x \mid \alpha) &= \exp\left[\phi(x)\,\psi(\alpha)\right]\eta(x)\,\delta(\alpha) \\ &= \exp\left[\phi(x)\,\psi(\alpha) + \log(\eta(x)) + \log(\delta(\alpha))\right], \end{aligned}$$

where $\phi$, $\psi$, $\eta$, and $\delta$ are real valued functions and $\eta$, $\delta > 0 \; \forall x, \alpha$. We also want to note that as long as the random vector $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is independent, identically distributed then the joint probability mass or density function belonging to the exponential family is as follows:

$$\begin{aligned} f(\mathbf{X} \mid \alpha) &= \prod_{i=1}^{n} f(x_i \mid \alpha) \\ &= \prod_{i=1}^{n} \exp\left[\phi(x_i)\psi(\alpha) + \log(\eta(x_i)) + \log(\delta(\alpha))\right] \\ &= \exp\left[\psi(\alpha)\sum_{i=1}^{n}\phi(x_i) + \sum_{i=1}^{n}\log(\eta(x_i)) + n\log(\delta(\alpha))\right]. \end{aligned}$$

Thus the joint probability density function or joint probability mass function for a set of independently, identically distributed random variables, each of which having equivalent PDFs or PMFs of the exponential family, is of the exponential family as well. This generalization allows for a natural extension into statistical analysis since one usually considers data sets with multiple variates.

Now let us consider a simplified version of the general form, i.e. the canonical form, for functions in the exponential family. The motivation for the following derivation will be clear when showing the relationship between the exponential family and what is called the link function for generalized linear models. Let $y = \phi(x)$ and $\theta = \psi(\alpha)$ be a one-to-one transformation of the components of the general form for functions in the exponential family ($\phi^{-1}$ and $\psi^{-1}$ exist). Making these transformations yields the following canonical form for the exponential family:

$$f(x \mid \alpha) = \exp\left[\phi(x)\psi(\alpha) + \log(\eta(x)) + \log(\delta(\alpha))\right]$$

which implies,

$$\begin{aligned} f(y \mid \theta) &= \exp\left[y\theta + \log\left(\eta\left(\phi^{-1}(y)\right)\right) + \log\left(\delta\left(\psi^{-1}(\theta)\right)\right)\right] \\ &= \exp\left[y\theta - b(\theta) + c(y)\right], \end{aligned}$$

where $\log(\eta(\phi^{-1}(y))) = c(y))$ and $\log(\delta(\psi^{-1}(\theta))) = -b(\theta)$, and we call $\theta$ the canonical parameter and the form $\theta$ takes the canonical link. Note that the $b(\theta)$ term in the canonical PDF or PMF is not a function of the unknown parameter but yet a function of the known parameter and therefore is the term, sometimes called the "normalizing constant", that can be manipulated so that the PDF or PMF integrates or sums, respectively, to one [2]. Similarly for the multivariate case assuming $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is independently, identically distributed, and by letting $y_i = \phi(x_i)$ be the one-to-one transformation of the PDF of the $i$th random variable then the joint PDF is as follows:

$$\begin{aligned}
f(\mathbf{X} \mid \alpha) &= \exp\left[\psi(\alpha)\sum_{i=1}^{n}\phi(x_i) + \sum_{i=1}^{n}\log(\eta(x_i)) + n\log(\delta(\alpha))\right] \\
&= \exp\left[\theta\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\log\left(\eta\left(\phi^{-1}(y_i)\right)\right) + n\log\left(\delta\left(\psi^{-1}(\theta)\right)\right)\right] \\
&= \exp\left[\theta\sum_{i=1}^{n} y_i - nb(\theta) + \sum_{i=1}^{n} c(y_i)\right],
\end{aligned}$$

where $c(y_i) = \log(\eta(\phi^{-1}(y_i)))$ and $-b(\theta) = \log(\delta(\psi^{-1}(\theta)))$. Notice that every derivation thus far has been for PDFs or PMFs of distributions of only one known parameter, $\theta$. Clearly in many instances this is not the case, for example in the normal distribution we consider two known parameters, $\mu$ and $\sigma^2$. Therefore for the multiparameter case we wish to have a similar formulation of the PDF or PMF in consideration. Let $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ be a vector of known parameters for a distribution. Then the canonical form for a exponential family function becomes

$$f(y \mid \theta) = \exp\left[\sum_{j=1}^{k} y\theta_j - b(\theta_j) + c(y)\right].$$

As an example of deriving the canonical form for the probability mass function of the exponential family let us consider the binomial distribution. This example will be handy when considering the translation to item response theory since the binomial distribution is used time and again to model counts of successes or failures [2] of exam or survey questions. Let $Y$ be the random variable such that $Y \sim Binomial\,(n,p)$ where $Y$ represents the number of "correct" responses to an exam, questionnaire, survey, etc., $n$ represents the known number of "questions" asked, and $p$ is the probability of answering with a "correct" response. Note here that since $n$, the number of "questions", is known we can disregard this value as a parameter and consider the binomial distribution as a one-parameter distribution. Parameters that can be disregarded are known as nuisance parameters. So for the PMF of the binomial distribution we can say, $f\,(y \mid n,p) = f\,(y \mid p)$. This disregard of the nuisance parameter allows us to describe the probability mass function of the binomial distribution in its canonical form as follows:

$$\begin{aligned}
f\,(y \mid p) &= \binom{n}{r} p^{y} (1-p)^{n-y} \\
&= \exp\left[\log \binom{n}{r} + y\log(p) + (n-y)\log(1-p)\right] \\
&= \exp\left[\log \binom{n}{r} + y\log(p) - y\log(1-p) + n\log(1-p)\right] \\
&= \exp\left[y\log\left(\frac{p}{1-p}\right) - (-n\log(1-p)) + \log\binom{n}{r}\right].
\end{aligned}$$

Thus $y\theta = y\log\left(\frac{p}{1-p}\right)$, $b\,(\theta) = -n\log(1-p)$, and $c\,(y) = \log\binom{n}{r}$ fits the canonical form. The canonical link $\theta = \log\left(\frac{p}{1-p}\right)$ plays an important role in modeling data using item response theory and we will call this the logit link.

We will now explore the relationship between the exponential family of functions and generalized linear models. There are three components that make up a generalized linear model and are as follows,

1) Random Component: The random or stochastic component consists of the vector $\mathbf{Y}$ of independent, identically distributed random variables that take on a distribution belonging to the exponential family with mean $\boldsymbol{\mu}$.

2) Systematic Component: The systematic component, denoted $\eta$, is a linear combination of predictors $\{x_1, x_2, \ldots, x_n\}$. that is

$$\boldsymbol{\eta} = \sum_{i=1}^{n} x_i \beta_i.$$

Note that the systematic component describes the observed data, $\mathbf{Y}$, through some set of linear predictors.

3) The Link Function: The link function shows the relationship of the observed data and the linear combination through a function of the means of the distributions. In other words the link function, denoted $g(\cdot)$, is described as follows:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \sum_{j=1}^{n} \mathbf{x}_j \beta_j = \boldsymbol{X}\boldsymbol{\beta}$$

such that

$$g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu} = E(\mathbf{Y}),$$

where $\mathbf{X}$ is the $m \times n$ model matrix for the observed data and $\boldsymbol{\beta}$ is the $n \times 1$ vector of coefficients for the linear predictors that will need to be approximated [5], [12].

It turns out that the link function for a specific distribution is simply the canonical link introduced earlier in the canonical form of the PDF or PMF of that particular distribution [5] [2]. The relaxation of assumptions is the goal in modeling a data set using generalized linear models. "The basic philosophy is to employ a function of the mean vector to link the normal theory environment with Gauss-Markov assumptions, to another environment that encompasses a wide class of outcome variables"[2]. Since many classes of outcome variables contain no solid assumptions that can be made, transforming these random variables into a linear combination of predictors allows one to treat the model as linear regression and thus a number of different fitting algorithms and prediction methods open up to the modeler.

In order to explore how a generalized linear model works consider the simple case of a general linear model, i.e. we will generalize the general linear model in terms of the three components discussed earlier. The general linear model consists of a normally distributed random variable $\mathbf{Y}$ such that the components are independently, identically distributed with mean $\boldsymbol{\mu}$ and an equal variance over all $Y_i$ of $\sigma^2$. In the general linear model the random variables are expressed as a linear combination of predictors and a random error, that is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

We can assume the Gauss-Markov assumptions for the general linear model, thus $E[\boldsymbol{\epsilon}] = 0$. Therefore,

$$E[\mathbf{Y}] = E[\mathbf{X}\boldsymbol{\beta}] + E[\boldsymbol{\epsilon}]$$

implies

$$\boldsymbol{\mu} = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \tag{2.1}$$

We now have the random and systematic components for the generalization of the general linear model, i.e. the normal random variable $\mathbf{Y}$ and the systematic component $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. It is also clear from 2.1 that the link function that correlates the systematic component to the mean of the probability distribution is

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{\mu}.$$

That is the link function for the general linear model is simply the identity function. So to sum up the general linear model can be observed as a generalized linear model using the following three components:

1) Random Component: The independent, identically distributed random variable of observations $\mathbf{Y}$ takes on a normal distribution with mean $\boldsymbol{\mu}$ and constant variance $\sigma^2$ for all $Y_i$.
2) Systematic Component: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is called the model matrix and $\boldsymbol{\beta}$ is a vector of coefficients.
3) The Link Function: The link between the systematic component and the mean of the random component is given by the identity function, that is $\boldsymbol{\eta} = \boldsymbol{\mu}$ or $\eta_i = \mu_i \ \forall i$ [5].

Another very interesting example of a generalized linear model is one where the observations in the data set take a binomial distribution. It is clear to see how observations of this type are quite natural when discussing test theories, i.e. a response to a question can either be correct or incorrect. Therefore the three components to this model would be as follows:

1) Random Component: The components of the random variable $\mathbf{Y}$ have independent, identically distributed binomial distributions.
2) Systematic Component: There exists a linear combination of covariates represented in a model matrix that produces the linear predictor

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

3) The Link Function: Recall that canonical link formulated from the canonical form of the PDF or PMF of the distribution is exactly the link function of the generalized linear model that uses that particular probability distributions.

Therefore the function $g(\mu_i) = \eta_i$ that satisfies this link between the systematic component and the mean of the probability distribution is

$$\eta_i = g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right)$$

We call this link function for the binomial distribution the logit link, which we will use throughout our discussion item response theory.

Recall that one motivation for using generalized linear models is to allow the modeler access to a number of more estimation procedures for best fitting observed data. A few of these methods are the Newton-Raphson Method, weighted least squares method, and perhaps the most used iteratively re-weighted least squares. These estimation methods will be derived in the discussions of item response theory in the context of generalized linear models [1].

# 3 IRT Models for Dichotomous Data

A test theory point of view was taken in the development of the Rasch, two-parameter, and three-parameter IRT models that were discussed in chapter 1. That is when considering an item on a test, we created models that would comply with possible variations on how that particular item might affect the probability of responding correctly. Therefore things like guessing factors were inserted into the model. Conversely a generalized linear model approach can also be taken when developing a item response model, which is how we will look at the Rasch model in this chapter. In other words we will show how the Rasch model is simply a subset of the larger group of models, called generalized linear models. Keep in mind any model constructed here will be an extension on the Rasch model, or one-parameter IRT model, although theoretically the same extensions could be applied to two- and three-parameter models as well. Four extensions on the Rasch model for dichotomous data will be derived along with the derivations of likelihood functions for the unknown parameters in each model which will be used for the means of predicting the values of these parameters. We will also establish and discuss methods for maximizing the likelihood functions developed for each model.
Consider a test item, question, that has two possible responses, i.e. the correct answer to the problem and the incorrect. For every model representing this binary data, we will let the correct solutions and incorrect solutions be represented by 1 and 0 respectively. Thus if person $p$ chooses the correct solution to item (question) $i$ then the random variable $Y_{pi} = 1$. Similarly the random variable $Y_{pi} = 0$ represents person $p$ answering incorrect on item $i$.

## 3.1 Rasch Model

Let $Y_{pi} \sim Binomial(1, \pi_{pi})$ such that $\pi_{pi}$ is the probability that $Y_{pi} = 1$, i.e. person $p$ gets item $i$ correct.

Link function:

$$\eta_{pi} = log\left(\frac{\pi}{1-\pi}\right) \tag{3.1}$$

and

$$\eta_{pi} = \theta_p - \beta_i \tag{3.2}$$

where $\theta_p \sim Normal(0, \sigma_\theta^2)$ is the random ability parameter of person $p$, and $\beta_i$ is the fixed weight of item predictor (can be thought of the item difficulty).

So, from (3.1) and (3.2) we get

$$log\left(\frac{\pi}{1-\pi}\right) = \theta_p - \beta_i$$

which yields

$$P(Y_{pi} = 1) = \pi_{pi} = \frac{e^{\theta_p - \beta_i}}{1 + e^{\theta_p - \beta_i}}.$$

Recall that a generalized linear model consists of three components; a random component that takes on a particular distribution, the systematic or linear component, and the link function that connect the mean of the random component with the systematic component. So simply defining a random variable, $Y_{pi} \sim Binomial(1, \pi_{pi}$, a particular systematic component, $\theta_p - \beta_i$, and a link, the logit function or $log\left(\frac{\pi}{1-\pi}\right)$, then it becomes clear that the Rasch model is an example of the larger group of generalized linear models. Being able to define the Rasch model in this way opens up numerous different methods for approximation and maximization of fitting the model to a set of data, that is any method that has been proven to be succesful for generalized linear models can now be applied to the

IRT Rasch model. Also, since we can now consider the Rasch model as a GLM, certain extensions can be made to the model in order to describe different situations. For instance certain hidden person or item traits within a set of data might have an affect on the how people respond to an item. The following extensions of the Rasch model take these possibilities into consideration.

## 3.2 Latent Regression Rasch Model

Let $Y_{pi} \sim Binomial(1, \pi_{pi})$ such that $\pi_{pi}$ is the probability that $Y_{pi} = 1$, i.e. person $p$ gets item $i$ correct.

Now let $\theta_p = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p$, we get the link function

$$\eta_{pi} = log\left(\frac{\pi}{1-\pi}\right) = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i,$$

where $Z_{pj}$ is the value of the person predictor of person $p$ on person property $j$, $\vartheta_j$ is the fixed weight of property $j$, and $\epsilon_p$ is the remaining person effect after the person property effect is accounted for. $\epsilon_p$ can be considered the random error that occurs, and $\epsilon_p \sim Normal(0, \sigma_\epsilon^2)$

So (5) yields

$$P(Y_{pi} = 1) = \pi_{pi} = \frac{e^{\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i}}{1 + e^{\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i}}$$

## 3.3 Linear Logistic Test Model (LLTM)

Let $Y_{pi} \sim Binomial(1, \pi_{pi})$ such that $\pi_{pi}$ is the probability that $Y_{pi} = 1$, i.e. person p gets item i correct.

Now consider $\beta_i = \sum_{k=0}^{K} \beta_k X_{ik}$ and working off the original Rasch model we get the link function,

$$\eta_{pi} = log(\frac{\pi}{1-\pi}) = \theta_p - \sum_{k=0}^{K} \beta_k X_{ik}$$

where $\theta_p \sim Normal(0, \sigma_\theta^2)$ is the random ability parameter of person p, $X_{ik}$ is the value of the item predictor of item $i$ on item property $k$, and $\beta_k$ is the regression weight of item property k.

The previous equation yields,

$$P(Y_{pi} = 1) = \pi_{pi} = \frac{e^{\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}}}{1 + e^{\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}}}$$

## 3.4 Latent Regression LLTM

Let $Y_{pi} \sim Binomial(1, \pi_{pi})$ such that $\pi_{pi}$ is the probability that $Y_{pi} = 1$, i.e. person p gets item i correct.

Again working off the original Rasch model consider, $\theta_p = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p$ and $\beta_i = \sum_{k=0}^{K} \beta_k X_{ik}$, in which we have the link function,

$$\eta_{pi} = log(\frac{\pi}{1 - \pi}) = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik},$$

where $Z_{pj}$ is the value of the person predictor of person $p$ on person property $j$, $\vartheta_j$ is the fixed weight of property $j$, and $\epsilon_p$ is the remaining person effect after the person property effect is accounted for. $\epsilon_p$ can be considered the random error that occurs, $\epsilon_p \sim Normal(0, \sigma_\epsilon^2)$, $X_{ik}$ is the value of the item predictor of item $i$ on item property $k$, and $\beta_k$ is the regression weight of item property k.

which yields,

$$P(Y_{pi} = 1) = \pi_{pi} = \frac{e^{\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}}}{1 + e^{\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}}}$$

# 4 Likelihood Functions for Item Parameters

The next section is aimed at creating a way to predict the fixed weights of the predictors, i.e. the $\beta$ values, so that for any particular item we will be able to measure the probability of answering that item correctly. First we need to develop the marginal maximum likelihood (MML) functions as a function of the fixed weights of the predictors for each of the four IRT models for dichotomous data presented above. The marginal maximum likelihood is used here simply for the fact that the software used in this paper implements this type of likelihood function, but note that there exist other types of likelihoods that have been applied to estimation inside of item response theory. Therefore given a random variable $x$ and two parameters $\psi$ and $\lambda$ the marginal maximum likelihood function for a parameter $\psi$ on random variable $x$ is denoted $L(\psi; x)$ and is given by the equation

$$L\left(\psi; x\right) = P\left(x \mid \psi\right) = \int P\left(x \mid \psi, \lambda\right) P\left(\lambda \mid \psi\right) d\lambda \tag{4.1}$$

## 4.1 MML for The Rasch Model

Towards developing the marginal maximum likelihood function for Rasch model developed in the previous section consider,

$$\begin{aligned} P\left(Y_{pi} = 1 \mid \beta_i, \theta_p\right) &= \pi_{pi} \\ &= \frac{e^{\theta_p - \beta_i}}{1 + e^{\theta_p - \beta_i}} \\ &= \frac{e^{y_{pi}(\theta_p - \beta_i)}}{1 + e^{\theta_p - \beta_i}} \end{aligned} \tag{4.2}$$

and

$$
\begin{aligned}
P\left(Y_{pi}=0 \mid \beta_i,\theta_p\right) &= 1-\pi_{pi} \\
&= 1-\frac{e^{\theta_p-\beta_i}}{1+e^{\theta_p-\beta_i}} \\
&= \frac{1}{1+e^{\theta_p-\beta_i}} \\
&= \frac{e^{y_{pi}(\theta_p-\beta_i)}}{1+e^{\theta_p-\beta_i}}.
\end{aligned} \tag{4.3}
$$

Thus, since $Y_{pi}$ can only take the values 0 or 1, equations 4.2 and 4.3 yield,

$$
P(Y_{pi}=y_{pi} \mid \beta_i,\theta_p)=\frac{e^{y_{pi}(\theta_p-\beta_i)}}{1+e^{\theta_p-\beta_i}}, \tag{4.4}
$$

where $y_{pi}$ is the observed value on item $i$ by person $p$, i.e. 0 if person $p$ chose the incorrect solution to item $i$ and 1 if person $p$ chose the correct solution to item $i$. Note that $\forall i,j$, the events $Y_{pi}=y_{pi}$ and $Y_{pj}=y_{pj}$ are independent of each other. Therefore the probability of choosing response $y_{pi}$ over all items, 1,....,$I$, for person $p$ becomes

$$
\begin{aligned}
P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p \mid \boldsymbol{\beta},\theta_p\right) &= P\left(Y_{p1}=y_{p1}\cap Y_{p2}=y_{p2}\cap\ldots\cap Y_{pI}=y_{pI} \mid \boldsymbol{\beta},\theta_p\right) \\
&= \prod_{i=1}^{I} P\left(Y_{pi}=y_{pi} \mid \beta_i,\theta_p\right) \\
&= \prod_{i=1}^{I}\frac{e^{y_{pi}(\theta_p-\beta_i)}}{1+e^{\theta_p-\beta_i}}.
\end{aligned} \tag{4.5}
$$

Therefore using 4.1 we can see that

$$
\begin{aligned}
L_p\left(\boldsymbol{\beta};\boldsymbol{Y}_p\right) &= P\left(\boldsymbol{Y}_p=y_p \mid \boldsymbol{\beta}\right) \\
&= \int_{-\infty}^{\infty}\prod_{i=1}^{I}\frac{e^{y_{pi}(\theta_p-\beta_i)}}{1+e^{\theta_p-\beta_i}} g\left(\theta_p \mid \mu_\theta,\sigma_\theta\right) d\theta_p,
\end{aligned} \tag{4.6}
$$

where $g(\theta_p|\mu_\theta, \sigma_\theta)$ represents the normal density function for $\theta_p$ such that $\theta_p$ has mean $\mu_\theta$ and standard deviation $\sigma_\theta$. Note that the previous equation represents the probability of choosing a particular set of responses for items $i = 1, \ldots, I$ independent of the person parameter $\theta_p$. Thus having the probability of responding a certain way to the items as a function of the fixed item weights $\beta_i$ will allow us to predict value of each particular item predictor weight so as to more fully understand how each item affects the person parameter.

## 4.2 MML for the Latent Regression Rasch Model

Now to form the MML in regards to the Latent Regression Rasch Model, recall that we will consider predictors on the person parameter. This will allow for different person properties to come in effect when considering a response to an item. Recall the link function for the Latent Regression Rasch Model is as follows

$$\begin{aligned} \eta_{pi} &= \log\left(\frac{\pi_{pi}}{1-\pi_{pi}}\right) \\ &= \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i. \end{aligned} \tag{4.7}$$

Therefore we wish to proceed in the same manner as for the Rasch Model, hence we can clearly see from (4.7)

$$\begin{aligned} \frac{\pi_{pi}}{1-\pi_{pi}} &= \exp\left[\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i\right] \\ \pi_{pi} &= \frac{\exp\left[\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i\right]}{1+\exp\left[\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i\right]} \\ P\left(Y_{pi} = 1 \mid \beta_i, \vartheta_j, \theta_p\right) &= \frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i\right)\right]}{1+\exp\left[\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i\right]} \end{aligned} \tag{4.8}$$

and

$$
\begin{aligned}
P\left(Y_{pi}=0 \mid \beta_i, \vartheta_j, \theta_p\right) &= 1-\pi_{pi} \\
&= 1-\frac{\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right]}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right]} \\
&= \frac{1}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right]} \\
&= \frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right)\right]}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right]}. \qquad (4.9)
\end{aligned}
$$

Therefore since (4.8) and (4.9) are equivalent then we can generalize to say that the probability of choosing either a correct response or an incorrect is as follows,

$$
P\left(Y_{pi}=y_{pi} \mid \beta_i, \vartheta_j, \theta_p\right)=\frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right)\right]}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right]}.
$$

Assuming that the response of one item is independent of the response of another, we can derive in the same manner as for the Rasch Model using (4.1), the marginal maximum likelihood for person $p$ is,

$$
\begin{aligned}
L_p\left(\boldsymbol{\beta}, \boldsymbol{\vartheta}; \boldsymbol{Y}_p\right) &= P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p \mid \boldsymbol{\beta}, \boldsymbol{\vartheta}, \theta_p\right) \qquad (4.10) \\
&= \int_{\infty}^{\infty}\prod_{i=1}^{I}\frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right)\right]}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right]} g\left(\epsilon_p \mid \mu_\epsilon, \sigma_\epsilon\right) d\epsilon_p,
\end{aligned}
$$

where $g$ is the normal density function of $\epsilon_p$. Therefore, this likelihood function is equivalent to the probability of choosing a specific pattern of responses for items $i=1,\ldots,I$ for person $p$ under a model that has item parameters $\beta_i$ and person parameters $\vartheta_j$.

## 4.3 MML for the Linear Logistic Test Model

For the Linear Logistic Test Model, recall that we consider the link function

$$\eta_{pi} = \log\left(\frac{\pi_{pi}}{1-\pi_{pi}}\right) = \theta_p - \sum_{k=0}^{K} \beta_k X_{ik} \tag{4.11}$$

where $\theta_p$ represents the random ability parameter of person $p$. Recall, this model allows us to take into consideration different properties that an item $i$ might take, i.e. one particular item might have a number of different properties that it can represent or test the person on. So again we will proceed in the same manner as was done in the previous models, hence find $P\left(Y_{pi}=1 \mid \beta_i, \theta_p\right) = \pi_{pi}$ and $P\left(Y_{pi}=0 \mid \beta_i, \theta_p\right) = 1-\pi_{pi}$ and create a general equation for $P\left(Y_{pi}=y_{pi} \mid \beta_i, \theta_p\right)$ which can be substituted into the MML equation to find the marginal maximum likelihood function of the unknown item property weights given the responses of person $p$ to item $i$. Therefore notice from the link function of the linear logistic test model that we get,

$$\begin{aligned}
\frac{\pi_{pi}}{1-\pi_{pi}} &= \exp\left[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}\right] \\
\pi_{pi} &= \frac{\exp\left[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]}{1+\exp\left[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]} \\
\pi_{pi} &= \frac{\exp\left[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]}{1+\exp\left[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]}
\end{aligned}$$

and

$$\begin{aligned}
1-\pi_{pi} &= 1 - \frac{\exp[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}]}{1+\exp[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}]} \\
&= \frac{1}{1+\exp[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}]}.
\end{aligned}$$

Therefore we can see that,

$$P\left(Y_{pi}=1 \mid \beta_k,\theta_p\right) = \frac{\exp\left[\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]}{1+\exp\left[\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]} \tag{4.12}$$

$$P\left(Y_{pi}=0 \mid \beta_k,\theta_p\right) = \frac{1}{1+\exp\left[\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]}. \tag{4.13}$$

Thus using (4.12) and (4.13) one can express the probability of person $p$ choosing response $y_{pi}$ to item $i$ as follows,

$$P\left(Y_{pi}=y_{pi} \mid \beta_k,\theta_p\right)=\frac{\exp\left[y_{pi}\left(\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right)\right]}{1+\exp\left[\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]}.$$

Now we will consider the collection of all responses for person $p$ over all items, $i=1,\ldots,I$. With the assumption that items are independent of each other we can now write the probability that person $p$ chooses a specific collection of responses as follows,

$$\begin{aligned}
P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p \mid \boldsymbol{\beta},\theta_p\right) &= P\left(Y_{p1}=y_{p1}\cap Y_{p2}=y_{p2}\cap\ldots\cap Y_{pI}=y_{pI} \mid \boldsymbol{\beta},\theta_p\right)\\
&= \prod_{i=1}^{I} P\left(Y_{pi}=y_{pi} \mid \beta_k,\theta_p\right)\\
&= \prod_{i=1}^{I}\frac{\exp\left[y_{pi}\left(\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right)\right]}{1+\exp\left[\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]}.
\end{aligned}$$

We wish to find a function that is independent of the random component $\theta_p$ so as to predict the fixed item property parameters, $\beta_k$, for a given collection of responses $y_p$. Therefore using the marginal maximum likelihood function from (4.1) to eliminate our unknown random component $\theta_p$ from the likelihood function, we get

$$\begin{aligned} L_p(\boldsymbol{\beta}; \boldsymbol{Y}_p) &= P\left(\boldsymbol{Y}_p = \boldsymbol{y}_p \mid \boldsymbol{\beta}, \theta_p\right) \\ &= \int_{\infty}^{\infty} \prod_{i=1}^{I} \frac{\exp\left[y_{pi}\left(\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}\right)\right]}{1 + \exp\left[\theta_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]} g\left(\theta_p \mid \mu_\theta, \sigma_\theta\right) d\theta \quad (4.14) \end{aligned}$$

such that $g$ is the normal density function of $\theta_p \sim Normal(0, \sigma_\theta^2)$ Therefore the LLTM likelihood function gives us a function of the fixed item property weights regardless of any random component.

## 4.4 MML for the Latent Regression LLTM

We now look at the marginal maximum likelihood function for the fourth model presented for dichotomous data, which is the Latent Regression Linear Logistic Test Model. Recall that this model is used to represent a situation where both a number of different person effects and different properties over the items might play a roll in how the person parameter and items interact. The link function for the Latent Regression LLTM is

$$\eta_{pi} = \log\left(\frac{\pi_{pi}}{1 - \pi_{pi}}\right) = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}, \quad (4.15)$$

where $\epsilon_p \sim Normal(0, \sigma_\epsilon^2)$ is the random component that varies over persons. With similar computations as were done in the previous three examples, one can see from (4.15) that we get

$$\begin{aligned} P\left(Y_{pi} = 1 \mid \beta_k, \vartheta_j, \theta_p\right) &= \pi_{pi} \\ &= \frac{\exp\left[\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]}{1 + \exp\left[\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]} \end{aligned}$$

and

$$
\begin{aligned}
P\left(Y_{pi}=0 \mid \beta_k, \vartheta_j, \theta_p\right) &= 1-\pi_{pi} \\
&= \frac{1}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]}.
\end{aligned}
$$

Therefore, to create a general equation for the probability of any response to item $i$, notice

$$
P\left(Y_{pi}=y_{pi} \mid \beta_k, \vartheta_j, \theta_p\right)=\frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ik}\right)\right]}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]}. \tag{4.16}
$$

Using the assumption of independence on the items and (4.16) we get

$$
P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p \mid \boldsymbol{\beta}, \boldsymbol{\vartheta}, \theta_p\right)=\prod_{i=1}^{I}\frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ik}\right)\right]}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]}. \tag{4.17}
$$

Now applying the the marginal maximum likelihood function (4.1) we come to the following function of just the fixed parameters,

$$
\begin{aligned}
L_p\left(\boldsymbol{\beta}, \boldsymbol{\vartheta}; \boldsymbol{Y}_p\right) &= P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p \mid \boldsymbol{\beta}, \boldsymbol{\vartheta}\right) \\
&= \int_{\infty}^{\infty} P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p \mid \boldsymbol{\beta}, \boldsymbol{\vartheta}, \theta_p\right) g\left(\epsilon_p \mid \mu_\epsilon, \sigma\epsilon\right) d\epsilon \\
&= \int_{\infty}^{\infty}\prod_{i=1}^{I}\frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ik}\right)\right]}{1+\exp\left[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]} g\left(\epsilon_p \mid \mu_\epsilon, \sigma\epsilon\right)
\end{aligned}
$$

where again $g$ is the normal density function for $\epsilon_p$.

To derive the complete marginal maximum likelihood functions for each model in turn, we note the assumption of independence of the responses over persons. This assumption allows us to create MML functions for each model in a general case, since as of now we only have a MML function for a single examinee given that persons set of responses. Notice that since person's responses are independent we have

$$L(\boldsymbol{\beta};\boldsymbol{Y}) = P(\boldsymbol{Y}=\boldsymbol{y}\mid\boldsymbol{\beta}) = \prod_{p=1}^{P} P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p\mid\boldsymbol{\beta}\right) = \prod_{p=1}^{P} L_p(\boldsymbol{\beta};\boldsymbol{Y}_p). \tag{4.18}$$

Thus applying (4.6) and (4.18), the complete likelihood function of the unknown parameters $\boldsymbol{\beta}$ is

$$\begin{aligned} L(\boldsymbol{\beta};\boldsymbol{Y}) &= \prod_{p=1}^{P} L_p(\boldsymbol{\beta};\boldsymbol{Y}_p) \\ &= \prod_{p=1}^{P} P\left(\boldsymbol{Y}_p=\boldsymbol{y}_p\mid\boldsymbol{\beta}\right) \\ &= \prod_{p=1}^{P}\int_{-\infty}^{\infty}\prod_{i=1}^{I}\frac{\exp\left[y_{pi}\left(\theta_p-\beta_i\right)\right]}{1+\exp\left[\theta_p-\beta_i\right]} g\left(\theta_p\mid\mu_\theta,\sigma_\theta\right)d\theta_p. \end{aligned}$$

Applying equations (4.10) and (4.18) the complete marginal maximum likelihood function for the latent Regression Rasch Model is

$$L(\boldsymbol{\beta},\boldsymbol{\vartheta};\boldsymbol{Y}) = \prod_{p=1}^{P}\int_{\infty}^{\infty}\prod_{i=1}^{I}\frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i\right)\right]}{1+\exp[\sum_{j=1}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i]} g\left(\epsilon_p\mid\mu_\epsilon,\sigma_\epsilon\right)d\epsilon_p.$$

Applying equations (4.14) and (4.18) the marginal maximum likelihood function for the complete set data set over all persons for the Linear Logistic Test Model is

$$L(\boldsymbol{\beta}_k;\boldsymbol{Y}) = \prod_{p=1}^{P}\int_{\infty}^{\infty}\prod_{i=1}^{I}\frac{\exp\left[y_{pi}\left(\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right)\right]}{1+\exp\left[\theta_p-\sum_{k=0}^{K}\beta_k X_{ik}\right]} g\left(\theta_p\mid\mu_\theta,\sigma_\theta\right)d\theta.$$

And lastly the complete marginal maximum likelihood function for the Latent Regression Linear Logistic Test Model is

$$L\left(\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_j; \boldsymbol{Y}\right) = \prod_{p=1}^{P} \int_{\infty}^{\infty} \prod_{i=1}^{I} \frac{\exp\left[y_{pi}\left(\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}\right)\right]}{1+\exp\left[\sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}\right]} g\left(\epsilon_p \mid \mu_\epsilon, \sigma\epsilon\right).$$

Thus, for the four Item Response Models we are exploring in this section we have have constructed functions which are independent of the random components, $\theta_p$ or $\epsilon_p$. This is vital for fitting our model since the item and person parameters $\beta_i$ and/or $\vartheta_j$ are the contributing factors to how the model should look. In other words the unknown random component, or latent trait, is only considered after a model has been sufficiently fit. Since these functions are now independent of $\theta_p$ we may maximize the values of the fixed components, that is find the $\beta_i$ values that best fit our model to the given observed data. Note the complexity of each likelihood function though. The integrals of each likelihood do not have closed form solutions, therefore simplification methods must be applied in order to have tractable functions in which to maximize. In other words the fitting procedure consists of two parts; first, one must approximate the integral of the likelihood function; secondly, once the integral has been approximated then a maximization method may be applied [9].

# 5 Approximating for Best Fit of the Model

## 5.1 Integral Approximation

In this section, we will discuss a number of procedures and algorithms that have been introduced to deal with approximating the marginal maximum likelihood functions introduced previously. For sake of simplicity on seeing how each procedure works, we will consider the Rasch model when presenting each approximation procedure. Generalizing each method is straightforward for more complex models.
To approximate the unknown fixed parameters of our model we will want to maximize the likelihood functions of each model that were derived previously.
Notice that each model has multiple unknown parameters, for the Rasch model the latent trait $\theta$ and the item difficulty parameters $\beta$ are both unknown. Therefore to fit the model we will hold the latent trait $\theta$ stable in order to approximate $\beta$. In a later section we will see that once the $\beta$ values have been estimated to best fit our model we can then approximate an individuals latent trait or ability level from a given set of responses, keep in mind this is our ultimate goal.
Again for simplicity sake let us consider the Rasch model. In order to approximate all values $\beta_i$ we will maximize the marginal maximum likelihood function,

$$L\left(\boldsymbol{\beta};\boldsymbol{Y}\right) = \prod_{p=1}^{P} \int_{-\infty}^{\infty} \prod_{i=1}^{I} \frac{e^{y_{pi}(\theta_p - \beta_i)}}{1 + e^{\theta_p - \beta_i}} g\left(\theta_p | \mu_\theta, \sigma_\theta\right) d\theta_p,$$

where $g$ is the normal density function. One problem with maximizing $L\left(\boldsymbol{\beta};\boldsymbol{Y}\right)$ is that the integral in the equation has no closed form solution. There are two groups of methods one can take to attempt maximizing this likelihood function: (1) Approximate the integral with numerical techniques then maximizing the function.

(2) Approximate the integrand such that the integral created from the approximation has a closed form solution, then solve the integral and maximize the solution. Although numerous procedures have been created to accomplish both approximating the integrand or the integral in it's entirety, we will focus on a couple methods for approximating the integral.

The most commonly used method for approximating the integral, and the method used in the R package that we will employ, is the Gauss-Hermite quadrature approximation. This approximation method states that given an integral of the form,

$$\int_{-\infty}^{\infty} e^{x^2} f(x)\, dx,$$

we can approximate the integral as such

$$\int_{-\infty}^{\infty} e^{x^2} f(x)\, dx \approx \sum_{i=1}^{n} f(x_i)\, w_i$$

where $x_i$ for $i = 1, \ldots, n$ are the nodes of the approximation and $w_i$ for $i = 1, \ldots, n$ are the weights that approximate the integral such that the nodes and weights have a predetermined form they take [8].

In order to demonstrate what the approximation looks like in terms of our likelihood function, let us consider the marginal maximum likelihood function for a single person $p$, that is

$$L_p\left(\boldsymbol{\beta}; \sigma_\theta^2\right) = \int_{-\infty}^{\infty} P\left(Y_p \mid \boldsymbol{\beta}, \theta_p\right) g\left(\theta_p \mid 0, \sigma_\theta^2\right) d\theta_p,$$

where is the normal density function of $\theta_p$ with $\mu_\theta = 0$. Notice that this is exactly the integral of the Rasch model that needs to be approximated. Since $g$ is the normal density function with $\mu_\theta = 0$ we have that

$$g\left(\theta_p \mid 0, \sigma_\theta^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\theta_p^2}{2\sigma^2}\right].$$

Towards achieving the form of the integral needed to use the Gauss-Hermite quadrature method, let

$$x = \frac{\theta_p}{\sigma_\theta\sqrt{2}} \Rightarrow \theta_p = \sqrt{2}\sigma_\theta x \Rightarrow d\theta_p = \sqrt{2}\sigma_\theta dx.$$

Therefore the derivation of the approximation of $L_p\left(\boldsymbol{\beta} \mid; \sigma_\theta^2\right)$ is as follows,

$$\begin{aligned}
L_p\left(\boldsymbol{\beta}; \sigma_\theta^2\right) &= \int_{-\infty}^{\infty} P\left(Y_p \mid \boldsymbol{\beta}, \theta_p\right) g\left(\theta_p \mid 0, \sigma_\theta^2\right) d\theta_p \\
&= \frac{1}{\sigma_\theta\sqrt{2\pi}} \int_{-\infty}^{\infty} P\left(Y_p \mid \boldsymbol{\beta}, \theta_p\right) \exp\left[\left(\frac{\theta_p}{\sqrt{2}\sigma_\theta}\right)^2\right] d\theta_p \\
&= \frac{1}{\sigma_\theta\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{x^2} P\left(Y_p \mid \boldsymbol{\beta}, \sqrt{2}\sigma_\theta x\right) \sqrt{2}\sigma_\theta dx \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{x^2} P\left(Y_p \mid \boldsymbol{\beta}, \sqrt{2}\sigma_\theta x\right) dx \\
&\approx \sum_{i=1}^{n} P\left(Y_p \mid \boldsymbol{\beta}, \sqrt{2}\sigma_\theta x_i\right) \frac{w_i}{\sqrt{\pi}}.
\end{aligned}$$

Furthermore the approximation becomes exact if the function $P\left(Y_p \mid \boldsymbol{\beta}, \theta_p\right)$ is a polynomial of degree $2n - 1$ or less [1]. Clearly these approximations are not computed by hand, but note that all statistical packages, that have been discussed in the literature, have a Gauss-Hermite quadrature method built in, and furthermore the package that will be used in this paper to show examples uses this method with a varying number of nodes depending on the model being fitted.
The Monte Carlo integration technique is also a method widely used to approximate such integrals that have been discussed thus far. This method involves using randomly chosen nodes rather than predetermined values, in order to approximate

the integral. Because of this random component one can consider the Monte Carlo method as the stochastic counterpart to the Gaussian quadrature method [1]. The Monte Carlo method can be applied because of the following fact, given $f(x)$ the pdf of a normal distribution then

$$E[u(x)] = \int_{-\infty}^{\infty} u(x) f(x).$$

Therefore the marginal maximum likelihood function for person $p$, the integral that needs to be approximated, can be written as the expected value of $P(Y_p \mid \boldsymbol{\beta}, \theta_p)$, that is

$$L_p\left(\boldsymbol{\beta}; \sigma_\theta^2\right) = \int_{-\infty}^{\infty} P(Y_p \mid \boldsymbol{\beta}, \theta_p)\, g\left(\theta_p \mid 0, \sigma_\theta^2\right) d\theta_p = E\left[P(Y_p \mid \boldsymbol{\beta}, \theta_p)\right]$$

since $g$ is a normal density function. Then, choosing at random $i = 1, \ldots, n$ nodes and finding $P(Y_p \mid \boldsymbol{\beta}, \theta_p)$ evaluated with at each value $\theta_p$ for the $i$th node and averaging these values we get an approximation for the likelihood function in question. In other words we get that

$$\begin{aligned} L_p\left(\boldsymbol{\beta}; \sigma_\theta^2\right) &= E\left[P(Y_p \mid \boldsymbol{\beta}, \theta_p)\right] \\ &\approx \frac{1}{n} \sum_{i=1}^{n} P\left(Y_p \mid \boldsymbol{\beta}, \theta_p^{(i)}\right) \end{aligned}$$

where $\theta_p^{(i)}$ is the value of $\theta_p$ evaluated at node $i$ [1]. Monte Carlo approximation is a consistent method in the sense as $n$, the number of nodes, gets sent to $\infty$, $\frac{1}{n} \sum_{i=1}^{n} P\left(Y_p \mid \boldsymbol{\beta}, \theta_p^{(i)}\right)$ converges to the expected value. Therefore the number of random nodes used in the approximation the better the approximation to the expected value of $P(Y_p \mid \boldsymbol{\beta}, \theta_p)$, hence the better approximation for the maximum marginal likelihood function in question. Once the approximated integral has been found, one can then choose from a number of different methods to maximize the newly found function in order to accurately estimate the unknown parameters, the difficulty parameter for the Rasch model. Lets now discuss a couple of these methods for maximization of dichotomous item response models, specifically the Rasch model.

## 5.2 Maximization Methods

There exist many iterative approaches used in maximizing maximum likelihood functions whether it be from a item response theory model or just a GLM in general. One such maximization procedure we will explore here is the expectation-maximization method, or the EM algorithm. the general premise of the EM algorithm is to maximize a lower bound of the log likelihood function for a particular parameter. Therefore, consider the logarithm of the marginal maximum likelihood for a general IRT model discussed previously,

$$\log L\left(\boldsymbol{\beta}\middle|\ \boldsymbol{Y}\right) = \log P\left(\boldsymbol{y}\middle|\ \boldsymbol{\beta}\right) = \log \prod_{p=1}^{P} \int P\left(\boldsymbol{Y}_p\middle|\ \boldsymbol{\beta}, \theta_p\right) g\left(\theta_p\right) d\theta_p,$$

where $g$ is the normal density function and the limits of integration have been dropped for simplification of notation. Then towards defining the EM algorithm notice

$$\begin{aligned}
\log\left[L\left(\boldsymbol{\beta}\mid\boldsymbol{Y}\right)\right] &= \log\left[\prod_{p=1}^{P}\int P\left(\boldsymbol{Y}_p\mid\boldsymbol{\beta},\theta_p\right)g\left(\theta_p\right)d\theta_p\right] \\
&= \sum_{p=1}^{P}\log\left[\int P\left(\boldsymbol{Y}_p\mid\boldsymbol{\beta},\theta_p\right)g\left(\theta_p\right)d\theta_p\right] \\
&\approx \sum_{p=1}^{P}\log\left[\sum_{i=1}^{n}P\left(\boldsymbol{Y}_p\mid\boldsymbol{\beta},\sqrt{2}\sigma_\theta x_i\right)\frac{w_i}{\sqrt{\pi}}\right] \\
&= \sum_{p=1}^{P}\log\left[\sum_{i=1}^{n}f\left(\sqrt{2}\sigma_\theta x_i\mid\boldsymbol{Y},\boldsymbol{\beta}\right)\frac{P\left(\boldsymbol{Y}_p\mid\boldsymbol{\beta},\sqrt{2}\sigma_\theta x_i\right)}{f\left(\sqrt{2}\sigma_\theta x_i\mid\boldsymbol{Y},\boldsymbol{\beta}\right)}\frac{w_i}{\sqrt{\pi}}\right] \\
&= \sum_{p=1}^{P}\log\left[\sum_{i=1}^{n}f\left(\sqrt{2}\sigma_\theta x_i\right)\frac{P\left(\boldsymbol{Y}_p\mid\boldsymbol{\beta},\sqrt{2}\sigma_\theta x_i\right)}{f\left(\sqrt{2}\sigma_\theta x_i\right)}\frac{w_i}{\sqrt{\pi}}\right] \\
&\geq \sum_{p=1}^{P}\sum_{i=1}^{n}f\left(\sqrt{2}\sigma_\theta x_i\right)\frac{w_i}{\sqrt{\pi}}\log\left[\frac{P\left(\boldsymbol{Y}_p\mid\boldsymbol{\beta},\sqrt{2}\sigma_\theta x_i\right)}{f\left(\sqrt{2}\sigma_\theta x_i\right)}\right] \\
&= F\left(f,\boldsymbol{\beta}\right)
\end{aligned}$$

where $f$ is the normal density function, but in general the EM algorithm states that any $f$ can be any arbitrary density function [11]. Note that the approximation comes from the Guass-Hermite quadrature approximation and the inequality comes from Jensen's inequality. Therefore we now have a lower bound for the log likelihood our models. By finding the $\beta_i$ values that maximize this lower bound, we can say these $\beta_i$ values maximize the original likelihood function. Each iteration in the EM algorithm is as follows

E-step: compute $f^{(k+1)} = argmax_f F\left(f,\boldsymbol{\beta^k}\right)$

M-step: compute $\boldsymbol{\beta}^{k+1} = argmax_{\boldsymbol{\beta}} F\left(f^{(k+1)},\boldsymbol{\beta}\right)$

where $argmax_f F\left(f,\boldsymbol{\beta}\right) = \left\{f \mid \forall f'\ F\left(f',\boldsymbol{\beta}\right)\leq F\left(f,\boldsymbol{\beta}\right)\right\}$ and $argmax_{\boldsymbol{\beta}} F\left(f,\boldsymbol{\beta}\right) = \left\{\boldsymbol{\beta} \mid \forall \boldsymbol{\beta}'\ F\left(f,\boldsymbol{\beta}'\right)\leq F\left(f,\boldsymbol{\beta}\right)\right\}$ [11].

Note that $f^{(k+1)} = P\left(\theta_p \mid \boldsymbol{\beta}^{(k)}\right)$. Therefore the E-step simplified by considering the following.

$$\begin{aligned}
F(f,\boldsymbol{\beta}) &= \sum_{p=1}^{P}\sum_{i=1}^{I} f\left(\sqrt{2}\sigma_\theta x_i\right)\frac{w_i}{\sqrt{\pi}}\log\left[P\left(\boldsymbol{Y}_p \mid \boldsymbol{\beta}, \sqrt{2}\sigma_\theta x_i\right)\right] \\
&\quad -\sum_{p=1}^{P}\sum_{i=1}^{I} f\left(\sqrt{2}\sigma_\theta x_i\right)\frac{w_i}{\sqrt{\pi}}\log\left[f\left(\sqrt{2}\sigma_\theta x_i\right)\right] \\
&= \sum_{p=1}^{P}\sum_{i=1}^{I} P\left(\theta_p \mid \boldsymbol{\beta}^{(k)}\right)\frac{w_i}{\sqrt{\pi}}\log\left[P\left(\boldsymbol{Y}_p \mid \boldsymbol{\beta}, \sqrt{2}\sigma_\theta x_i\right)\right] \\
&\quad -P\sum_{i=1}^{I} f\left(\sqrt{2}\sigma_\theta x_i\right)\frac{w_i}{\sqrt{\pi}}\log\left[f\left(\sqrt{2}\sigma_\theta x_i\right)\right] \\
&= G\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}\right) - H(f).
\end{aligned}$$

$H(f)$ is not a function of the unknown parameter $\boldsymbol{\beta}$, therfore can be ignored when trying to maximize $F$ since it is simply a constant function. Now notice

$$G\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}\right) = \sum_{p=1}^{P} E\left[\frac{w_i}{\sqrt{\pi}}\log\left[P\left(\boldsymbol{Y}_p \mid \boldsymbol{\beta}, \sqrt{2}\sigma_\theta x_i\right)\right]\right] \tag{5.-28}$$

where the expectation is over the probability distribution of $f$. Therefore the iteration steps become

E-step: compute $G\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}\right) = \sum_{p=1}^{P} E\left[\frac{w_i}{\sqrt{\pi}}\log\left[P\left(\boldsymbol{Y}_p \mid \boldsymbol{\beta}, \sqrt{2}\sigma_\theta x_i\right)\right]\right]$

M-step: compute $\boldsymbol{\beta}^{k+1} = argmax_{\boldsymbol{\beta}} F\left(f^{(k+1)}, \boldsymbol{\beta}\right)$

where in most cases normal maximization methods, i.e. setting derivatives to zero, can be applied to the M-step [10].

Note that each iteration consists of both the E-step and the M-step, thus the iteration is repeated until the $\beta_i$ values converge. The EM algorithm is applied to cases in which a data set is said to be incomplete, in otherwords there are missing values that describe the model. "Repeating the E and M steps, the algorithm is guaranteed to converge to a local maximum of the likelihood function with each iteration increasing the log-likelihood" [10]. Note that the EM algorithm guarantees a local extrema, therefore in terms of the models used in item response theory further study might be needed to guarantee global maxima are reached using the EM algorithm. But for applications purposes, even though method converges to a maximum value rather slowly, the EM algorithm is one of the preferred methods used in IRT modeling [1].

# 6 IRT models for Polytomous Data

In the previous section we discussed item response theory models for the case of having a correct and an incorrect response to an item, i.e. dichotomous response sets. A natural progression when modeling under item response theory would be to take into consider multiple response items. That is consider all possible responses for an item rather than just the correct response and an incorrect response. So one might want to further analyze an item in terms of how examinees are responding to all possible choices it might have. Sets of data where more than two possible responses are available we will call polytomous data sets or multicategorical data sets. There are two types of multicategorical data that will be examined here: 1) Ordinal data sets are sets of responses or categories that have order to them and 2) nominal data sets are sets of responses or categories that have no inherent order to them. The type of data one is looking at, be it ordinal or nominal, determines the type of model used. In other words, models that best describe the relationships among ordinal data sets are most likely not the best models to use to describes data sets of the nominal type and vice versa.

To start lets set up some notation and assumptions that will be used throughout our discussion of polytomous data sets. We will say that item $i$ has $M_i$ possible responses and we label these responses as $m$ such that $m = 0, \ldots, M_i - 1$. We will also assume that an examinee may only choose one response per item, but the number of responses per item can change over an exam. Note that since each item has $M_i$ possible response categories there exist $M_i$ probabilities within each item representing the probabilities of choosing each possible response. Therefore for person $p$ and item $i$ we say the probability of choosing response $m$ is $\pi_{pim}$. Also, note that $\sum_{m=0}^{M_i-1} \pi_{pim} = 1$ since we assume an examinee may not leave answers blank, without a response. Another assumption that will be made, as was made for the dichotomous response data sets, is that items will be considered multivariate independent, that is there are no correlations between responses to items. This assumption of independence of items will again greatly simplify the analysis.

## 6.1 Multivariate Generalized Linear Models

To perform IRT analysis for items that have multiple responses we will construct each model in terms of a multivariate generalized linear mixed model. The first step in doing this is to recode the set of responses into random vector of binary data. That is, say person $p$ on item $i$ chooses the response $m$, then we assign the random variable $Y$ the value $m$ for person $p$ on item $i$, i.e. $Y_{pi} = m$. Notice that this is simply an extension of the dichotomous responses discussed earlier since in the case where there are only two possible responses, correct and incorrect, then $M_i = 2$ which implies $m = 0, 1$. So to recode our data in terms of 0s and 1s we create a random vector $\boldsymbol{C}_{pi}$ such that $\boldsymbol{C}_{pi}$ has length $M_i - 1$ and we define the components of the possible vectors for $\boldsymbol{C}_{pi}$ as

$$c_{pim} = \begin{cases} 1 & \text{if } Y_{pi} = m \text{ where } m = 1, \ldots, M_i - 1 \\ 0 & \text{otherwise} \end{cases} .$$

Therefore we have

$$P\left(Y_{pi} = m\right) = P\left(c_{pim} = 1\right) = \pi_{pim}.$$

For visualization of how this construction of a random vector works lets consider the case where there are three possible responses to item $i$, hence $M_i = 3$. Since $M_i = 3$ we have that $m = 0, 1, 2$.

So suppose person $p$ chooses response $m = 0$ for item $i$, then we have that $Y_{pi} = 0$. Thus $c_{pi1} = 0$ and $c_{pi2} = 0$. Then the realization of the random vector $C_{pi}$ call it $c_{pi}$

is the vector $(0, 0)$. Thus the three possible realizations of the random vector $C_{pi}$ for the case of having an item having three responses is as follows:

If $Y_{pi} = m = 0$ then $c_{pi} = (c_{pi1}, c_{pi2}) = (0, 0)$

If $Y_{pi} = m = 1$ then $c_{pi} = (c_{pi1}, c_{pi2}) = (1, 0)$

If $Y_{pi} = m = 2$ then $c_{pi} = (c_{pi1}, c_{pi2}) = (0, 1)$.

Now that a sufficient recoding of the random variable has been constructed the next step is to develop a multivariate generalized linear model in order to develop item response model in this framework. Recall the components of a generalized linear model consist of 1) a random component which describes the probability distribution of the data, 2) a link function that relates the expected value of the distribution for the data in terms of some set of linear predictors, and 3) a set of linear predictors that will describe the data under question.

The Distribution: in the general case, we have created from a set of $M$ possible responses a random vector such that the realization of this random vector, $c_{pi}$, has length $M_i - 1$. But, each of these vectors will contain one component that takes the value 1 and the rest of the components taking the value 0. This yields exactly a multinomial distribution with a total count of 1, hence a multivariate Bernoulli distribution. Therefore we have that

$$P\left(Y_{pi} = m\right) = P\left(\boldsymbol{C}_{pi} = \boldsymbol{c}_{pi}\right) = \pi_{pi0}^{c_{pi0}} \pi_{pi1}^{c_{pi1}} \pi_{pi2}^{c_{pi2}} \ldots \pi_{piM_i-1}^{c_{piM_i-1}},$$

where $\pi_{pi0} = 1 - \pi_{pi1} - \pi_{pi2} - \ldots - \pi_{piM_i-1}$ and $c_{pi0} = 1 - c_{pi1} - c_{pi2} - \ldots - c_{piM_i-1}$.
Also notice that the expected value of this distribution will simply be the vector of marginal probabilities for each possible response, that is

$$E[\boldsymbol{C}_{pi}] = \boldsymbol{\pi}_{pi} = (\pi_{pi1}, \pi_{pi2}, \ldots, \pi_{piM_i-1})$$

The Link Function: Generalizing a link function for the multinomial generalized linear model comes from a natural extension of the binomial case. Recall that the link function used for a binomial distribution was as follows,

$$\eta\left(\pi_{pi}\right) = log\left(\frac{\pi_{pi}}{1-\pi_{pi}}\right).$$

Note how the link function is the logarithm of the ratio of the probability of responding correctly and responding incorrectly to item $i$, which happen to be the only two possible responses in the dichotomous case. Clearly the two events of responding correctly and incorrectly are mutually exclusive as well. This idea of looking at the ratio of two mutually exclusive events is how we will generalize the link function. Given that there are more than two possible responses to an item consider the mutually exclusive sets of possible responses $A_m$ and $B_m$. Then we define the $m$th component to the logit link function as follows

$$\eta_{pim} = f_{\text{link}m}\left(\boldsymbol{\pi}_{pi}\right) = log\left(\frac{\pi_{pi}\left(A_m\right)}{\pi_{pi}\left(B_m\right)}\right)$$

where $\pi_{pi}\left(A_m\right)$ and $\pi_{pi}\left(B_m\right)$ are the probabilities of responding to a possible response in the sets $A_m$ and $B_m$ respectively. This construction can be interpreted as the "attractiveness" of the subset $A_m$ of responses over the subset $B_m$ of responses. It is also important to note that because of the many ways the mutually exclusive sets $A_m$ and $B_m$ can be chosen, we will only explore four unique ways to do this in the upcoming sections. Also note that the vector link function has the same dimension as the mean, or expected value, of the multivariate bernoulli distribution under study. So the full link function for an item $i$ with $M_i$ possible responses takes the form

$$\boldsymbol{\eta}_{pi} = \boldsymbol{f}_{\text{link}}\left(\boldsymbol{\pi}_{pi}\right) = \begin{pmatrix} f_{\text{link}1}\left(\boldsymbol{\pi}_{pi}\right) \\ f_{\text{link}2}\left(\boldsymbol{\pi}_{pi}\right) \\ \vdots \\ f_{\text{link}M_i-1}\left(\boldsymbol{\pi}_{pi}\right) \end{pmatrix} = \begin{pmatrix} \eta_{pi1} \\ \eta_{pi2} \\ \vdots \\ \eta_{piM_i-1} \end{pmatrix}.$$

Keep in mind the point of creating this link function is to be able to predict the probability of responding a certain way to an item. So as seen in the dichotomous data case, there does exist an inverse to the logit link function therefore the probability of choosing each possible response within an item $i$, described in the vector $\boldsymbol{\pi}_{pi}$, is given by $\boldsymbol{f}_{\text{link}}^{-1}\left(\boldsymbol{\eta}_{pi}\right)$.

The Linear Predictor: suppose after considering an item we wish to create a model matrix such that certain factors, whether it be inherent to how the modeler thinks the data is related or simply factors that modeler wishes to focus on, the possible types of predictors that can be used are item, person, or logit predictors or any combination of the three. Depending on the type of predictor being used the model matrix will take different forms. An item or person predictor model matrix is used when one wants to predict the model by showing a relationship between items or persons respectively. That is, does changing from one item to another or one person to another affect the probability of responding to that item in a specific way. The logit predictor takes into consideration the different responses within an item, i.e. is there an affect to the probability of choosing reponse $m$ over response $n$ within one particular item. Note that we label this predictor as the logit predictor since we represent the different possible responses through the individual logistic function components within the link function in which case each component consists of a different pair subsets of possible responses. So for some notation label each individual observed value for person $p$, item $i$, and logit $m$ under variable $X_k$ as

$X_{pimk}$. We will then collect these predictors for person $p$ on item $i$ and logit $m$ into a vector of predictors, that is

$$\boldsymbol{X}_{pim}^T = \left(X_{pim1}\ X_{pim2}\ \ldots\ X_{pimK}\right).$$

Next, stack the logit predictor vectors to create a matrix of predictors for person $p$ on item $i$ taking into consideration every type of possible logit predictor, and label this matrix $\boldsymbol{X}_{pi}$, which will take the form

$$\boldsymbol{X}_{pi} = \begin{pmatrix} X_{pi11} & X_{pi12} & \ldots & X_{pi1K} \\ X_{pi21} & X_{pi22} & \ldots & X_{pi2K} \\ \vdots & \vdots & \ddots & \vdots \\ X_{pi(M_i-1)1} & X_{pi(M_i-1)1} & \ldots & X_{pi(M_i-1)K} \end{pmatrix}$$

such that item $i$ has $M_i$ possible responses. Thus $X_{pi}$ has dimensions $M_i - 1 \times K$. One can then stack all $\boldsymbol{X}_{pi}$ matrices over all items $i = 1, \ldots, I$ on top of one another to create the matrix $\boldsymbol{X}_p$ which has the dimensions $(M_i - 1)(I) \times K$ [1]. Next to construct a model matrix for an entire group of examinees over the total number of items on an exam, stack all $\boldsymbol{X}_p$ matrices for all persons $p = 1, \ldots, P$ on top of one another yielding a "super matrix" $\boldsymbol{X}$ which would have the dimensions $(M_i - 1)(I)(P) \times K$. Now that a model matrix has been constructed the next step in developing the linear predictor component to the generalized linear model is to create a vector of regressions coefficients. Lets label this vector of coefficients $\boldsymbol{\beta}$ which takes the form

$$\boldsymbol{\beta}^T = \left(\beta_1\ \beta_2\ \ldots\ \beta_K\right)$$

Therefore the linear predictor for a item response model in terms of a generalized linear models will be $\boldsymbol{X\beta}$, which is well defined since $\boldsymbol{X}$ has dimensions $(M_i - 1)(I)(P) \times K$ and $\boldsymbol{\beta}$ has dimensions $K \times 1$. For simplicity sake let us consider the link function for person $p$ on item $i$. Also, because we are building this linear predictor in the context of item response models we will let the linear predictor be negative in order to show how certain predictors, i.e. the difficulty parameter in the dichotomous case, affect the probability of choosing a specific response. Thus the link function for a specific person $p$ on item $i$ using logit $m$ will be as follows,

$$\begin{aligned}
\boldsymbol{\eta}_{pim} &= -\sum_{k=1}^{K} X_{pimk}\beta_k \\
&= -X_{pim1}\beta_1 - X_{pim2}\beta_2 - \ldots - X_{pimK}\beta_K \\
&= -\boldsymbol{X}_{pim}^T\boldsymbol{\beta}
\end{aligned}$$

Therefore the link function for person $p$ on item $i$ which has $M_i - 1$ possible responses is

$$\boldsymbol{\eta}_{pi} = \begin{pmatrix} \eta_{pi1} \\ \eta_{pi2} \\ \vdots \\ \eta_{piM_i-1} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{X}_{pi1}^T\boldsymbol{\beta} \\ -\boldsymbol{X}_{pi2}^T\boldsymbol{\beta} \\ \vdots \\ -\boldsymbol{X}_{piM_i-1}^T\boldsymbol{\beta} \end{pmatrix} = -\boldsymbol{X}_{pi}\boldsymbol{\beta}$$

## 6.2 Multivariate Generalized Linear Mixed Models

In order to fully develop our item response models in the framework of a generalized linear model, we must introduce the idea of a multivariate generalized linear mixed model. Notice that the multivariate GLM we have developed in the previous section only includes fixed effects to describe the probability of responding a certain way, therefore lets introduce a random effect into our model. In many cases, and almost always in item response theory, there exists a person specific effect that determines how that person will respond to an item. In the case of item response theory this person specific effect is the latent trait of ability level being tested. Lets call this latent ability level $\theta_p$ just as we did in the dichotomous data case. The norm in item response theory is to assume $\theta_p$ is normally distributed, i.e. $\theta_p \sim Normal\,(0, \sigma_\theta^2)$. So, to properly fit this random effect component into our multivariate generalized linear model let $\boldsymbol{Z}_{pi}^T = (Z_{pi1}\ Z_{pi2}\ \ldots\ Z_{piM_i-1})$ such that

$$\boldsymbol{Z}_{pi}\theta_p = \begin{pmatrix} Z_{pi1}\theta_p \\ Z_{pi2}\theta_p \\ \vdots \\ Z_{piM_i-1}\theta_p \end{pmatrix}.$$

In most cases the vector $\boldsymbol{Z}_{pi}$ is made to be a $M_i - 1$ vector of 1's in order to simplify the model. Once this random component is added to the already formed fixed component of our generalized linear model we have what is now called a generalized linear mixed model or GLMM. This particular GLMM takes the form

$$\begin{aligned}
\boldsymbol{\eta}_{pi} &= \boldsymbol{Z}_{pi}\theta_p - \boldsymbol{X}_{pi}\boldsymbol{\beta} \\
&= \begin{pmatrix} Z_{pi1}\theta_p - \boldsymbol{X}_{pi1}^T\boldsymbol{\beta} \\ Z_{pi2}\theta_p - \boldsymbol{X}_{pi2}^T\boldsymbol{\beta} \\ \vdots \\ Z_{piM_i-1}\theta_p - \boldsymbol{X}_{piM_i-1}^T\boldsymbol{\beta} \end{pmatrix} \\
&= \begin{pmatrix} Z_{pi1}\theta_p - \sum_{k=1}^K X_{pi1k}\beta_k \\ Z_{pi2}\theta_p - \sum_{k=1}^K X_{pi2k}\beta_k \\ \vdots \\ Z_{pi(M_i-1)}\theta_p - \sum_{k=1}^K X_{pi(M_i-1)k}\beta_k \end{pmatrix}
\end{aligned}$$

So depending on how the model matrix $X_{pi}$ is constructed different IRT models can fall out. In the following two sections we will develop two item response models in the framework of this generalized mixed model, and to do so we will first construct a model matrix that is used quite frequently in a number of different item response models. We will also explore how changing the type of logit link functions used, i.e. differing the subsets $A_m$ and $B_m$ of responses to use in the link function, changes the type of model.

## 6.3 Model Building and Predictor Model Matrices

The benefit of using this generalized linear model format to create an item response model is that the model matrices are unique to the person fitting the model [1]. That is an analyst using item response theory for polytomous data in the context of generalized linear mixed models can uniquely create a model matrix in order to fit the model specifically to their own needs. For simplicity purposes and for the sake of applying GLMM to item response theory we will assume the vector $\boldsymbol{Z}_{pi}$ is a vector of 1's and thus the random component of the generalized linear mixed model is simply the value of $\theta_p$ for each person $p$. Recall that the model matrix of predictors for the fixed effect is denoted $\boldsymbol{X}$ and has dimension $(M_i - 1)(I)(P) \times K$ where $P$ is the total number of examinees in the data set, $I$ is the total number items on the exam or survey and $M_i - 1$ is the number of possible responses on item $i$. For the sake of presenting a tractable model lets consider an item $i$ with 3 possible responses, that is $M_i = 3$. Therefore, when we consider the model matrix $\boldsymbol{X}_{pi}$ for a person $p$ and item $i$, the matrix will have a total of $M_i - 1 = 2$ rows. Also recall that a predictor variable can consist of predictors that represent the item, person, logit, or any combination thereof, i.e. item-by-person, item-by-logit, person-by-logit, or item-by-person-by-logit predictors. The model matrix that is implemented in a number of frequently used polytomous response IRT models is termed the item-by-logit matrix. This model provides information about each individual possible response within a particular item. Although the components within a model matrix can take any number of values, we will simplify our model by setting each non-zero element of the matrix to be 1. For our example lets consider this item-by-logit predictor matrix for 3 items each of which have 3 different possible responses. De Boeck and Wilson refer to Ramsey and Schafer 2001 when they introduce this "fairly general tentative model" [1]. Thus the complete model matrix for 3 items each of which have 3 possible responses that displays the item-by-logit relationship can be seen below [1]. Two important models that use

Table 6.1: Item-by-Logit Model Matrix

| Item | Logit | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|---|
| 1 | $L_1$ | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | $L_2$ | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | $L_1$ | 0 | 0 | 1 | 0 | 0 | 0 |
| 2 | $L_2$ | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | $L_1$ | 0 | 0 | 0 | 0 | 1 | 0 |
| 3 | $L_2$ | 0 | 0 | 0 | 0 | 0 | 1 |

this model matrix are Masters' *partial credit model* (PCM) and the Samejima's *graded response model* (GRM). Although there are a number of different models that have been constructed to fit polytomous data we will focus on these two models for our discussion. Applying this model matrix to the link function of the multivariate generalized linear model along with the vector of regression coefficients $\boldsymbol{\beta}$ and the assumption that the random component vector $\boldsymbol{Z}_p$ is a vector of 1s we have the link function

$$
\begin{aligned}
\boldsymbol{\eta}_p &= \boldsymbol{Z}_p\theta_p - \boldsymbol{X}_p\boldsymbol{\beta} \\
\begin{pmatrix} \eta_{p11} \\ \eta_{p12} \\ \eta_{p21} \\ \eta_{p22} \\ \eta_{p31} \\ \eta_{p32} \end{pmatrix} &= \begin{pmatrix} \theta_p \\ \theta_p \\ \theta_p \\ \theta_p \\ \theta_p \\ \theta_p \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} \\
\begin{pmatrix} \log\left(\frac{\pi_{p1}(A_1)}{\pi_{p1}(B_1)}\right) \\ \log\left(\frac{\pi_{p1}(A_2)}{\pi_{p1}(B_2)}\right) \\ \log\left(\frac{\pi_{p2}(A_1)}{\pi_{p2}(B_1)}\right) \\ \log\left(\frac{\pi_{p2}(A_2)}{\pi_{p2}(B_2)}\right) \\ \log\left(\frac{\pi_{p3}(A_1)}{\pi_{p3}(B_1)}\right) \\ \log\left(\frac{\pi_{p3}(A_2)}{\pi_{p3}(B_2)}\right) \end{pmatrix} &= \begin{pmatrix} \theta_p - \beta_1 \\ \theta_p - \beta_2 \\ \theta_p - \beta_3 \\ \theta_p - \beta_4 \\ \theta_p - \beta_5 \\ \theta_p - \beta_6 \end{pmatrix}.
\end{aligned}
$$

Note the change in subtext from one logit component to the next in the left hand side of the link function above. Recall that these represent the probabilities of responding to each particular subset of responses on each item $i = 1, 2, 3$. One might see the motivation in constructing a model matrix in this way since it yields a link function where each component takes the form of the Rasch model explored earlier for dichotomous data. Since the Rasch model is the simplest of item response models, and in most cases yields enough information to accurately analyze items, we have constructed an IRT model for polytomous data that is quite tractable which otherwise would be quite complex. Now that the model matrix has been constructed and we have a general link function using that model matrix, the next step is to determine what form the individual logit functions, i.e. the components of the link function, will take. Recall that the logit functions differ in what ratio of mutually exclusive subsets of responses are used. In other words we need to determine how to split the responses into subsets and why certain combinations might work better than others.

## 6.4 Constructing the Logit Functions for the Links

There are three general ways to divide up possible responses of items to create the different link functions needed in creating a model for polytomous data. We label these three methods as 1) the adjacent category logit, 2) the cumulative logit, and 3) the baseline category logit. The main focus of our discussion will be on the adjacent category logit and cumulative logit methods although we will mention the general construction of the baseline category logit.

The first model to consider is called the adjacent category, or response, logit. Using this adjacent category method one will examine the relationship of response $m$ with response $m-1$. Recall that each individual logit takes the form

$$\eta_{pim} = f_{\text{logit}m}\left(\boldsymbol{\pi}_{pi}\right) = log\left(\frac{\pi_{pi}\left(A_m\right)}{\pi_{pi}\left(B_m\right)}\right).$$

Therefore if an adjacent category model is used then subset $A_m$ is response $m$ and subset $B_m$ is response $m-1$, or $A_m = \{m\}$ and $B_m = \{m-1\}$. hence each individual logit becomes

$$\eta_{pim} = log\left(\frac{\pi_{pi}\left(m\right)}{\pi_{pi}\left(m-1\right)}\right) = log\left(\frac{\pi_{pim}}{\pi_{pi(m-1)}}\right)$$

Again to explore a tractable example consider an item $i$ that has three possible responses and using the item-by-logit model matrix constructed in the previous section, the complete link function for item $i$ takes the form

$$\boldsymbol{\eta}_{pi} = \begin{pmatrix} \eta_{pi1} \\ \\ \eta_{pi2} \end{pmatrix} = \begin{pmatrix} log\left(\frac{\pi_{pi1}}{\pi_{pi0}}\right) \\ \\ log\left(\frac{\pi_{pi2}}{\pi_{pi1}}\right) \end{pmatrix} = \begin{pmatrix} \theta_p - \boldsymbol{X}_{pi1}^T\boldsymbol{\beta} \\ \\ \theta_p - \boldsymbol{X}_{pi2}^T\boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \theta_p - \beta_1 \\ \\ \theta_p - \beta_2 \end{pmatrix}$$

Again one can see the motivation of the creation of this specific model since it can be looked at as multivariate form of the much simpler Rasch model. This construction using an item-by-logit model matrix and the adjacent category logit is referred to as the *partial credit model* [1]. The adjacent category logit approach lends itself to ordinal data, or data that has a specified order to it, since we are considering the relationship of a response and the response just prior. Therefore the *partial credit model* is highly applicable for educational purposes if one wishes to analyze exam questions where responses can be ordered from "most right" to "least right", although the *partial credit model* has many other possible applications in the fields of education, psychometrics and even econometrics. We will further explore the *partial credit model* in sections to come.

The cumulative logit is another method that lends itself to ordinal data. This model will represent the likelihood of choosing a response $m$ or higher in relation to choosing a response that is less than $m$. We will refer back to the original random variable $Y_{pi}$ to get a better grasp of just how the cumulative logit is constructed. Note that the likelihood, or probability, of choosing a response $m$ or greater can be denoted as $P\left(Y_{pi} \geq m\right)$. Therefore for the example of item $i$ that has a total of three possible responses we set the mutually exclusive sets as

$$\begin{aligned} A_1 &= Y_{pi} \geq 1 \\ B_1 &= Y_{pi} < 1 \\ A_2 &= Y_{pi} \geq 2 \\ B_2 &= Y_{pi} < 2 \end{aligned}$$

and therefore

$$\begin{aligned} \pi_{pi}(A_1) &= P(Y_{pi} \geq 1) = \pi_{pi1} + \pi_{pi2} \\ \pi_{pi}(B_1) &= P(Y_{pi} < 1) = \pi_{pi0} \\ \pi_{pi}(A_2) &= P(Y_{pi} \geq 2) = \pi_{pi2} \\ \pi_{pi}(B_2) &= P(y_{pi} < 2) = \pi_{pi0} + \pi_{pi1}. \end{aligned}$$

So using the item-by-logit predictor matrix and the cumulative logit construction that was just presented yields the following link function for the example of an item $i$ that has three possible responses.

$$\boldsymbol{\eta}_{pi} = \begin{pmatrix} \eta_{pi1} \\ \eta_{pi2} \end{pmatrix} = \begin{pmatrix} log\left(\frac{\pi_{pi1}+\pi_{pi2}}{\pi_{pi0}}\right) \\ log\left(\frac{\pi_{pi2}}{\pi_{pi0}+\pi_{pi1}}\right) \end{pmatrix} = \begin{pmatrix} \theta_p - \boldsymbol{X}_{pi1}^T\boldsymbol{\beta} \\ \theta_p - \boldsymbol{X}_{pi2}^T\boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \theta_p - \beta_1 \\ \theta_p - \beta_2 \end{pmatrix}.$$

Note that the fixed regression coefficients, $\beta_1$ and $\beta_2$, need not be, and most likely will not be, equivalent between the adjacent category logits and cumulative logits even though they both take the same "Rasch model" form. One can clearly see how a set of ordinal responses would fit into this cumulative logit approach nicely since again we are considering the ratio of responding to a set of responses that have a certain level and the responses that have a "lower" level. "In the item response literature, models of this type have been called *graded response models* (GRM, Samejima, 1969)" [1], or Samejima's *graded response model.* A further examination of the graded response model and the partial credit model will take place in a section to come, where we will derive the probability of choosing a specific response from the link functions presented above along with likelihood functions that may be used to approximate the fixed regression coefficients, $\beta_i$. But first let us introduce a third method for constructing the mutually exclusive sets of responses for the logit functions, in this case for nominal data.

The third type of logit that can used to model polytomous data is a method where a baseline category or response is used when considering the change from one response to another, quite naturally this method is called the baseline category logit. In this construction we consider each specific response that is possible for an item in relation to a fixed baseline response. Note that we have labeled the possible responses to an item $i$ as $m = 0, 1, \ldots, M_i - 1$, thus the norm is to use response 0 as the baseline response, although any response can be used as the baseline response, and look at the ratio $\frac{\pi_{pi}(A_m)}{\pi_{pi}(B_m)} = \frac{\pi_{pim}}{\pi_{pi0}}$ where $m \neq 0$. That is we let the set used in the denominator of each logit to be the fixed response $m = 0$. The link function will look similar to those of the graded response model and the partial credit model yet have different logits as each specific component of the link. Therefore the link function for the baseline category model using an item by logit model matrix is as follows

$$\boldsymbol{\eta}_{pi} = \begin{pmatrix} \eta_{pi1} \\ \eta_{pi2} \end{pmatrix} = \begin{pmatrix} log\left(\frac{\pi_{pi1}}{\pi_{pi0}}\right) \\ log\left(\frac{\pi_{pi2}}{\pi_{pi0}}\right) \end{pmatrix} = \begin{pmatrix} \theta_p - \boldsymbol{X}_{pi1}^T\boldsymbol{\beta} \\ \theta_p - \boldsymbol{X}_{pi2}^T\boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \theta_p - \beta_1 \\ \theta_p - \beta_2 \end{pmatrix}.$$

This type of model construction is more applicable towards data that is nominal, i.e. no apparent order, since we are more or less considering each response separately. Make a note that even though the linear components for each of the three models introduced are identical, except for the value of the $\beta_i$s, the most important part of each model is how the logit components of the link function are formed. How these logits are constructed determines what the probability of responding to each item separately will be, hence different logits yield different probability mass functions for $Y_{pi}$. That being said, let us now further develop the partial credit model and graded response model in turn, that is we will now derive the probability mass functions for each model as well as the marginal maximum likelihood functions.

## 6.5 Partial Credit and Graded Response Models

Lets continue to use the example of an item $i$ that has three possible responses, keep in mind any derivation that follows extends to an item with $M$ possible responses. Partial Credit Model: Recall from the link function for the partial credit model that we have the two logit components

$$log\left(\frac{\pi_{pi1}}{\pi_{pi0}}\right) = \theta_p - \beta_1 \qquad\qquad log\left(\frac{\pi_{pi2}}{\pi_{pi1}}\right) = \theta_p - \beta_2$$

Which yield that

$$\begin{aligned}
\pi_{pi1} &= \pi_{pi0}e^{\theta_p-\beta_1} & \pi_{pi2} &= \pi_{pi1}e^{\theta_p-\beta_2}\\
\pi_{pi1} &= (1-\pi_{pi1}-\pi_{pi2})\,e^{\theta_p-\beta_1} & &\\
\pi_{pi1} &= \left(1-\pi_{pi1}-\pi_{pi1}e^{\theta_p-\beta_2}\right)e^{\theta_p-\beta_1} & &\\
\pi_{pi1} &= e^{\theta_p-\beta_1}-\pi_{pi1}e^{\theta_p-\beta_1}-\pi_{pi1}e^{2\theta_p-\beta_1-\beta_2} & &\\
\pi_{pi1}&\left(1+e^{\theta_p-\beta_1}+e^{2\theta_p-\beta_1-\beta_2}\right) = e^{\theta_p-\beta_1} & &\\
\pi_{pi1} &= \frac{e^{\theta_p-\beta_1}}{1+e^{\theta_p-\beta_1}+e^{2\theta_p-\beta_1-\beta_2}} & \pi_{pi2} &= \frac{e^{2\theta_p-\beta_1-\beta_2}}{1+e^{\theta_p-\beta_1}+e^{2\theta_p-\beta_1-\beta_2}}.
\end{aligned}$$

So for the partial credit model the probabilities of responding to each individual response given a latent ability level of $\theta_p$ and for fixed item parameters $\beta_1$ and $\beta_2$

are as follows

$$P\left(Y_{pi}=0\right)=\pi_{pi0}=\frac{\pi_{pi1}}{e^{\theta_p-\beta_1}}=\frac{1}{1+e^{\theta_p-\beta_1}+e^{2\theta_p-\beta_1-\beta_2}}$$

$$P\left(Y_{pi}=1\right)=\pi_{pi1}=\frac{e^{\theta_p-\beta_1}}{1+e^{\theta_p-\beta_1}+e^{2\theta_p-\beta_1-\beta_2}}$$

$$P\left(Y_{pi}=2\right)=\pi_{pi2}=\frac{e^{2\theta_p-\beta_1-\beta_2}}{1+e^{\theta_p-\beta_1}+e^{2\theta_p-\beta_1-\beta_2}}.$$

A slight modification to notation will allow for a general formula for the probability of responding to a certain response of any item $i$, that is, we will let $\beta_{ij}$ be the regression coefficient for the $j$th logit of item $i$. And so one can see that

$$P\left(Y_{pi}=m_{pi}\right)=\pi_{pim}=\frac{\exp\left[\sum_{j=1}^{m}\left(\theta_p-\beta_{ij}\right)\right]}{1+\sum_{c=1}^{M_i-1}\exp\left[\sum_{j=1}^{c}\left(\theta_p-\beta_{ij}\right)\right]}$$

where item $i$ has $M_i$ possible responses and $m_{pi}$ represents person $p$ response to item $i$, will lead to the probability mass functions developed in our example of an item that has three possible responses $\left[\text{note that if } m=0 \text{ then} \sum_{j=1}^{m}\left(\theta_p-\beta_j\right)=0\right]$ [7]. Keep in mind the motivation of developing a general function for the probability of responding to a question in a certain way is that we need this function in order to approximate the fixed regression coefficients, i.e. the $\beta_{ij}$ values. In order to do this we need the likelihood function for the vector of regression coefficients $\boldsymbol{\beta}$ over all responses to all items for every examinee involved in the exam, survey, etc. To do this we will construct the likelihood function around the likelihood function for specific examinees. Assuming that items are independent over an exam, that is responding to one item has no affect on a person response to another item, we can notice that the likelihood function for $\boldsymbol{\beta}$ on person $p$ is

$$
\begin{aligned}
L_p(\boldsymbol{\beta}; \boldsymbol{m}_p) &= P(\boldsymbol{Y_p} = \boldsymbol{m}_p |\ \boldsymbol{\beta}) \\
&= P(Y_{p1} = m_{p1} \cap Y_{p2} = m_{p2} \cap \ldots \cap Y_{pI} = m_{pI} |\ \boldsymbol{\beta}) \\
&= \prod_{i=1}^{I} P(Y_{pi} = m_{pi} |\ \boldsymbol{\beta})
\end{aligned}
$$

where $\boldsymbol{m}_p$ is the vector of responses for all items for person $p$. Then it follows from the definition of a marginal maximum likelihood that we get

$$
\begin{aligned}
L_p(\boldsymbol{\beta}; \boldsymbol{m}_p) &= \int_{-\infty}^{\infty} P(\boldsymbol{Y}_p = \boldsymbol{m}_p |\ \boldsymbol{\beta}, \theta_p)\, g(\theta_p | \boldsymbol{\beta})\, d\theta_p \\
&= \int_{-\infty}^{\infty} \prod_{i=1}^{I} P(Y_{pi} = m_{pi} |\ \boldsymbol{\beta})\, g(\theta_p | \boldsymbol{\beta})\, d\theta_p \\
&= \int_{-\infty}^{\infty} \prod_{i=1}^{I} \frac{\exp\left[\sum_{j=1}^{m} (\theta_p - \beta_{ij})\right]}{1 + \sum_{c=1}^{M_i - 1} \exp\left[\sum_{j=1}^{c} (\theta_p - \beta_{ij})\right]} g(\theta_p | \boldsymbol{\beta})\, d\theta_p
\end{aligned}
$$

where $g$, since we assume $\theta_p$ is normally distributed, is the normal density function. The final step in developing the marginal maximum likelihood function for the partial credit model is to assume that persons are independent in which case the complete likelihood function follows quite simply

$$
\begin{aligned}
L(\boldsymbol{\beta}; \boldsymbol{Y}) &= \prod_{p=1}^{P} L_p(\boldsymbol{\beta}; \boldsymbol{m}_p) \\
&= \prod_{p=1}^{P} \int_{-\infty}^{\infty} \prod_{i=1}^{I} \frac{\exp\left[\sum_{j=1}^{m} (\theta_p - \beta_{ij})\right]}{1 + \sum_{c=1}^{M_i - 1} \exp\left[\sum_{j=1}^{c} (\theta_p - \beta_{ij})\right]} g(\theta_p | \boldsymbol{\beta})\, d\theta_p
\end{aligned}
$$

where $g$ is the normal density function. Once we have our maximum marginal likelihood function that is independent of the unknown random component and just a function of the fixed parameters an approximation method can then be implemented to simplify the integral which does not have a closed form solution.

After the integral is approximated a maximization procedure is used to estimate the $\beta$ values in order to properly fit our model to the data. Note that any integral approximation and maximization method discussed in Chapter 5 as well as a number of other numerical approximation and maximization methods can be used here.

Now that the partial credit model has been fully explored, let us consider the Samejima's graded response model and develop its marginal maximum likelihood function for the use of fitting a set of data. Recall from our discussion earlier that the logit components of the vector link function for an examine item with three possible responses using the graded response model are

$$log\left(\frac{\pi_{pi2}+\pi_{pi1}}{\pi_{pi0}}\right)=\theta_p-\beta_1 \qquad\qquad log\left(\frac{\pi_{pi2}}{\pi_{pi1}+\pi_{pi0}}\right)=\theta_p-\beta_2.$$

Therefore, in the same manner that was done prior, we solve for $\pi_{pi1}$ and $\pi_{pi2}$. So the two logit components yield the following

$$\begin{aligned}
\pi_{pi1}+\pi_{pi2} &= \pi_{pi0}e^{\theta_p-\beta_1} \\
\pi_{pi1}+\pi_{pi2} &= \left[1-\left(\pi_{pi1}+\pi_{pi2}\right)\right]e^{\theta_p-\beta_1} \\
\left(\pi_{pi1}+\pi_{pi2}\right)\left(1+e^{\theta_p-\beta_1}\right) &= e^{\theta_p-\beta_1} \\
\pi_{pi1}+\pi_{pi2} &= \frac{e^{\theta_p-\beta_1}}{1+e^{\theta_p-\beta_1}} \\
P\left(Y_{pi}\geq 1\right) &= \frac{e^{\theta_p-\beta_1}}{1+e^{\theta_p-\beta_1}}
\end{aligned}
\qquad
\begin{aligned}
\pi_{pi2} &= \left(\pi_{pi1}+\pi_{pi0}\right)e^{\theta_b-\beta_2} \\
\pi_{pi2} &= \left(\pi_{pi1}+1-\pi_{pi1}-\pi_{pi2}\right)e^{\theta_p-\beta_2} \\
\pi_{pi2} &= \left(1-\pi_{pi2}\right)e^{\theta_p-\beta_2} \\
\pi_{pi2} &= \frac{e^{\theta_p-\beta_2}}{1+e^{\theta_p-\beta_2}} \\
P\left(Y_{pi}\geq 2\right) &= \frac{e^{\theta_p-\beta_2}}{1+e^{\theta_p-\beta_2}}.
\end{aligned}$$

Thus the probabilities person $p$ responding a certain way to item $i$ under the graded responses model for fixed item parameters $\beta_1$ and $\beta_2$ are

$$P\left(Y_{pi}=0\right)=1-P\left(Y_{pi}\geq 1\right)=1-\frac{e^{\theta_p-\beta_1}}{1+e^{\theta_p-\beta_1}}$$

$$P\left(Y_{pi}=1\right)=P\left(Y_{pi}\geq 1\right)-P\left(Y_{pi}\geq 2\right)=\frac{e^{\theta_p-\beta_1}}{1+e^{\theta_p-\beta_1}}-\frac{e^{\theta_p-\beta_2}}{1+e^{\theta_p-\beta_2}}$$

$$P\left(Y_{pi}=2\right)=P\left(Y_{pi}\geq 2\right)=\frac{e^{\theta_p-\beta_2}}{1+e^{\theta_p-\beta_2}}.$$

A special assumption must be made to ensure that the graded response model is well defined in the sense that all probabilities are in fact values between 0 and 1. Notice that with regards to $P\left(Y_{pi}=1\right)$, if $P\left(Y_{pi}\geq 2\right)\geq P\left(Y_{pi}\geq 1\right)$ then we would have that $P\left(Y_{pi}=1\right)$ takes a negative value. A fix for this inconsistency can be made by simply allowing $\beta_1<\beta_2$ along the continuum under which the $\beta$ values and $\theta_p$ fall. Although this seems like a pretty strong assumption to be made, it allows us to properly fit a data set using this graded responses model. And even though we are making this assumption, important information about the items can still be taken from the model which is again the reason for fitting these models to begin with. Note that we are using an example of an item that has three possible responses, so to extend to a general item with $M_i$ possible responses we would have to make the assumption of $\beta_1<\beta_2<\ldots<\beta_{M_i-1}$. Therefore a well defined set of probabilities have been created for responding to each possible response for an item in question and thus a general formula must now be developed, that is a formula to represent the probability of responding to an arbitrary response $m$ for any item $i$. Like we have done for previous models, once this general formula has been found, it can be applied to the marginal maximum likelihood function in order to have a function that can be used to find the $\beta_i$ values that best fit our model to the data collected. To begin the development of this general item response function for Samejima's graded response model, notice that for item $i$ with $M_i$ possible responses $P\left(Y_{pi}\geq 0\right)=1$ and $P\left(Y_{pi}\geq M_i\right)=0$. That is since we denote the possible responses as $m=0,\ldots,M_i-1$ then clearly the probably of responding to any possible response is 1 and the probably of responding to a response greater than $M_i-1$ is 0. So using these two assumptions and noticing the pattern that is created from our example of an item using three possible responses we can come to the conclusion that the general function to represent the probability of responding a certain way is

$$P\left(Y_{pi}=m\right)=\pi_{pim}=P\left(Y_{pi}\geq m\right)-P\left(Y_{pi}\geq m+1\right)$$

Thus using the definition of a marginal maximum likelihood function and using the same argument as we did for the partial credit model we come to the marginal maximum likelihood function for our newly formed graded response function. Assuming items $i=1,\ldots,I$ are independent over an examine then the MML for a person $p$ is

$$\begin{aligned} L_p\left(\boldsymbol{\beta};\boldsymbol{m}_p\right) &= \int_{-\infty}^{\infty} P\left(\boldsymbol{Y}_p=\boldsymbol{m}_p \middle| \boldsymbol{\beta},\theta_p\right) g\left(\theta_p \middle| \boldsymbol{\beta}\right) d\theta_p \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^{I} P\left(Y_{pi}=m_{pi} \middle| \boldsymbol{\beta}\right) g\left(\theta_p \middle| \boldsymbol{\beta}\right) d\theta_p \end{aligned}$$

where $\boldsymbol{m}_p$ is the vector of responses for person $p$, $m_{pi}$ is the response for person $p$ on item $i$, $g$ is the normal density function, and $P\left(Y_{pi}=m_{pi}\right)=P\left(Y_{pi}\geq m_{pi}\right)-P\left(Y_{pi}\geq m_{pi}+1\right)$. And using the assumption that persons $p=1,\ldots,P$ are independent over the entire data set then the full marginal maximum likelihood function becomes

$$\begin{aligned} L\left(\boldsymbol{\beta};\boldsymbol{Y}\right) &= \prod_{p=1}^{P} L_p\left(\boldsymbol{\beta};\boldsymbol{m}_p\right) \\ &= \prod_{p=1}^{P} \int_{-\infty}^{\infty} \prod_{i=1}^{I} P\left(Y_{pi}=m_{pi} \middle| \boldsymbol{\beta}\right) g\left(\theta_p \middle| \boldsymbol{\beta}\right) d\theta_p \end{aligned}$$

such that $P\left(Y_{pi}=m_{pi}\right)=P\left(Y_{pi}\geq m_{pi}\right)-P\left(Y_{pi}\geq m_{pi}+1\right)$. So we have explored and developed two possible ways to construct a model using the item-by-logit model matrix, they are the partial credit model and Samejima's graded response model.

Note there exist numerous different ways to build a model matrix in which case each unique matrix will describe the data in a different manner. Looking at these different model matrices is quite possibly an area for further research. Another possibility for further research would be to explore different ways to select our mutually exclusive sets of responses $A_m$ and $B_m$ that are implemented into the logit components of the link function, which would lead to completely different models being formed. Also note that throughout the sections on generalized linear models and the Rasch model we explored extensions of the one parameter item response model, or the Rasch model. Implementing these same types of extensions to the two and three parameter models from chapter 1 could lead to very interesting results as well. Another area of interest is to explore models that take into consideration multiple latent traits, i.e. could one model data under IRT using a multi-dimension latent trait rather than the one dimension discussed through this paper. Now that we have developed a number of different models, let us insert real life observed data into our models so that we can see just how these models work are beneficial to a test theorist or psychometrician studying such exams, surveys, etc.

# 7 Application of IRT models in R

In this sections we will apply selected models developed throughout our discussion to data collected from TAKS testing from around the Tyler, Texas area. Although numerous packages have been created for multiple statistical based software, we will implement the item response theory package developed by Dimitris Rizopoulos for R titled *ltm*. The *ltm* package in R implements the Rasch model, two-parameter logistic, and three-parameter logistic models for dichotomous data introduced in the overview section on item response theory. The polytomous models that are available in the *ltm* package are the generalized partial credit model and the graded response model that we have just finished developing [13]. The motivation for using this package in R is that it exercises the use of the marginal maximum likelihood function in the process of finding the fixed item parameters and uses the Newton-Raphson method for maximizing the likelihood function both of which we have used to develop our models. We make a note of this because there are numerous other methods for approximating the parameter values.
The data we will be modeling first is a set of responses for tenth grade students at a local High School from the Tyler, Texas area for the Texas standardized test TAKS. The responses to the questions have been organized such that a correct response is denoted as 1 and an incorrect response is denoted as 0. Therefore we can consider this data set as being dichotomous and thus we will employ a Rasch model to analyze the items in question. The first command one can use to analyze the data is the *descript* command, which yields descriptive statistics. The descriptive data given are 1) proportions for each type of response for every item, 2) frequencies of total scores obtained, 3) a point biserial correlation with total scores, with both an included and excluded correlation, for every item. This value describes the

correlation of the observed responses to a particular item with the total score on the test, both including that particular item in the total test score and not including that particular item. 4) A Cronbach's alpha score for each item on the exam is also give, which measure consistency within the test. That is the Cronchbach alpha places a value on whether all items are truly measuring the same latent trait. And 5) a pairwise association table for pairs of items that are highly correlated [13]. The data set in question contains 56 questions, i.e. items, so to see an example of the output let us consider the first five items on the exam, then the descriptive statistics given would be

| Descriptive Statistics | | | | | | |
|---|---|---|---|---|---|---|
| | | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 |
| Proportions for each Response | 0 | .0478 | .0957 | .0718 | .0622 | .1005 |
| | 1 | .9522 | .9043 | .9282 | .9378 | .8995 |
| | logit | 2.9907 | 2.2460 | 2.5598 | 2.7132 | 2.1919 |
| Point Biserial Correlation with Total Score | Included | .4710 | .5488 | .5971 | .5220 | .5946 |
| | Excluded | .1961 | .1684 | .2877 | .2160 | .2182 |
| Cronbach's alpha | | .3798 | .4054 | .3091 | .3646 | .3641 |

Also a list of pairwise item correlations are given along with p-values and the frequencies of possible total scores for individuals are listed as well. The point biserial correlation given here is a measurement of the relationship between the binomial random variable $Y$ that takes values 0 or 1 depending on how the person responds to the item and the total scores on the test. Note that the random variables in our model are the binary variable $Y$ that takes the values 1 and 0. Point biserial correlations can be interpreted in a similar manner to any other correlation coefficient, that the point biserial correlation coefficient will be positive when high total test scores relate with $Y = 1$ and small test scores relate with $Y = 0$. This is the natural interpretation in terms of item response theory and a test theory in general since one would expect a positive correlation for "good" items since one would wish that answering more items correctly leads to higher test scores will have a higher chance of responding to an item correctly. The other measurement shown here, Cronbach's alpha value, might also be unfamiliar. Cronbach's alpha is a statistical measure of reliability within a test or a measure of internal consistency. In other words Cronbach's alpha gives a numerical value to how well all items within a group measure the same inherent trait. Clearly, especially in terms of items response theory, one would wish to have large Cronbach alpha values through the items on a test so that all items are essentially measuring the same skill or latent trait. Low alpha values imply that the items in questions are measuring different latent traits, in other words an item, say on an exam measuring mathematics skills, that has a very low Cronbach alpha value is not necessarily measuring the examinee's mathematics skills. Once the descriptive statistics have been thoroughly analyzed, one can then attempt to fit the data to a model, whether it be the Rasch model, two-parameter, or three-parameter logistics irt models. The *ltm* package will fit dichotomous data to any of these three models, therefore the analyst must determine which model best fits the data while still being tractable enough to be analyzed. Lets begin with fitting the Rasch model. An important note is to *ltm* does not by default apply a discrimination parameter of 1 to a Rasch model. Instead the package simply uses an equal discrimination value over all items. Therefore if one wishes to apply a discrimination parameter equal to 1 for their Rasch model, it must be written into the R code. The following will fit our data to a Rasch model with discrimination parameter equal to 1, *rasch1 <- rasch(data.2,constraint=cbind(length(data.2)+1,1))*. A *summary* command can then be used to see log-likelihood of the model along with the Akaike's information criteron (AIC), Bayesian information criterion (BIC) as well the approximated values of each items difficulty parameter, i.e. the $\beta_i$ values for each item on the exam. Lets consider the first five items fitted to the Rasch model with discrimination parameter equal to 1, then following would be the estimated difficulty parameters: Note how the R output even indicates the integration method

Table 7.1: Rasch Model for Items 1 through 5

| | value | std.err | z.vals |
|---|---|---|---|
| Dffclt.p1 | -3.4248 | .3389 | -10.1067 |
| Dffclt.p2 | -2.6367 | .2544 | -10.3660 |
| Dffclt.p3 | -2.9716 | .2853 | -10.4165 |
| Dffclt.p4 | -3.1345 | .3029 | -10.6497 |
| Dffclt.p5 | -2.5784 | .2496 | -10.3287 |
| Dscrmn | 1.0000 | NA | NA |

log.lik=-6095.851  AIC=12303.70  BIC=12490.87

Integration Method: Gauss-Hermite quadrature points 21

used to approximate the marginal maximum likelihood function of the model. If the discrimination parameter is not constrained to be 1 a different model can be fit, all items having the same discrimination parameter yet not necessarily 1. The following commands would give this new Rasch model and the summary information respectively, *rasch2 <- rasch(data.2)*, *summary(rasch2)*. This new model would yield the following values for the difficulty parameters for the first five items Now

Table 7.2: Rasch Model for Items 1 through 5

| | value | std.err | z.vals |
|---|---|---|---|
| Dffclt.p1 | -3.4382 | .3831 | -8.9756 |
| Dffclt.p2 | -2.6470 | .2877 | -9.1991 |
| Dffclt.p3 | -2.9836 | .3238 | -9.2136 |
| Dffclt.p4 | -3.1466 | .3436 | -9.1587 |
| Dffclt.p5 | -2.5887 | .2821 | -9.1770 |
| Dscrmn | .9955 | .0560 | 17.7635 |

log.lik=-6095.847  AIC=12305.69  BIC=12496.21

Integration Method: Gauss-Hermite quadrature points 21

that two models have been fitted to the data, the natural question is how well do the two models fit to the data in questions and which of the two models fit the data better? The *ltm* package has a built in goodness of fit test for Rasch models. The goodness of fit test is a bootstrap method using the Pearson chi-squared test. A bootstrap method involves constructing a number of sample data sets using the model in question and comparing the sample statistics of each sample in with the sample statistic from the original data set, in our case the sample statistics are that of the Pearson chi-squared type. So, the goodness of fit test assumes a null hypothesis that the model does in fact fit the observed data, then letting $\chi_b$ be the chi-squared test statistic of the $b$th sample where $b = 1, \ldots, B$ and $\chi_{obs}$ be the chi-squared test statistic of the observed data, the p-value of the hypothesis test is approximated by the following

$$\text{p-value} \approx \frac{(\#\text{ of times } \chi_b \geq \chi_{obs}) + 1}{B + 1}$$

[13]. The command to run this bootstrap goodness of fit test *GoF.rasch('model here')*. Therefore let us compare the goodness of fit tests between the two models we have constructed.

| | |
|---|---|
| Tobs | 3.341564e+16 |
| # data-sets | 50 |
| p-value | 0.3 |

Table 7.3: Goodness of Fit Test for *rasch1*

| | |
|---|---|
| Tobs | 3.339928e+16 |
| # data-sets | 50 |
| p-value | 0.28 |

Table 7.4: Goodness of Fit Test for *rasch2*

So, the p-values for both models are large, hence in both cases we would fail to reject the null hypothesis that the models do indeed fit the observed data. But notice that the p-value for our *rasch1* model is slightly higher than that of the *rasch2* model, thus we can interpret that *rasch1* fits the data slightly better than that of *rasch2*, the Rasch without the constraint of the discrimination being equal to 1. This conclusion can also be seen from the fact that the AIC and BIC values for *rasch1* are slightly lower than those of *rasch2* which implies a slightly betters fit, i.e.

Likelihood Ratio Table

| | AIC | BIC | log.lik | LRT | df | p.value |
|---|---|---|---|---|---|---|
| rasch1 | 12303.70 | 12490.87 | -6095.85 | | | |
| rasch2 | 12305.69 | 12496.21 | -6095.85 | 0.01 | 1 | .927 |

Table 7.5: Anova test for *rasch1* and *rasch2*

lower AIC, BIC values are desired citeRizo. Also an analysis of variance test can be ran to determine if two models, one nested inside the other, are equivalent by using the command *anova('model1','model2')* such that *'model1'* is nested inside *'model2'*. Since the p-value of the anova test is quite large we would fail to reject that the two models are equivalent, but from what the goodness of fit test implied, the fact that the AIC and BIC are smaller for the *rasch1* model, and since *rasch1* is a slightly simpler model with the discrimination parameter being 1, it would most likely be beneficial to implement *rasch1* as the model to describe our observed data.
Although we have not directly developed the two- and three-parameter models in terms of likelihood functions as we did for the Rasch model, data can very easily be fit to these two models by using the *ltm()* and *tpm()* commands. But since we have more fully developed the Rasch model in our discussions and since by our goodness of fit and anova tests show that the model *rasch1* sufficiently fits our observed data, let us more fully analyze this model. Once a model has been fit to a set of observed data, that is the item parameters have been sufficiently approximated, then there are a number of commands in the IRT package that will allow one to further analyze the data. one such command is *coeff('model',prob=TRUE)* which will output a list of all items along with the any parameters that have been approximated and the probability of answering that item correctly given an average ability level. The output for the first five items on the exam would be as follows The latent ability being studied on this exam is denoted $z$ here, and because in the norm in item response theory is to assume the mean ability level is $z = 0$ then the third row in the table is giving the probability of answering that particular item correctly

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 |
|---|---|---|---|---|---|
| Dffclt | -3.4247 | -2.6367 | -2.9716 | -3.1345 | -2.5784 |
| Dscrmn | 1 | 1 | 1 | 1 | 1 |
| P(x=1\| z=0) | .9684 | .9331 | .9512 | .9582 | .9294 |

Table 7.6: *coeff(rasch1,prob=TRUE)* for items 1 through 5

given an average ability level. Notice how each of these probabilities is quite high for only an average ability level. This can be related to the fact that each items difficulty parameters is very low hence these five items are not hard questions on this particular exam. Also note that the discrimination parameters for every item is 1, this is because we are analyzing the Rasch model that we artificially constrained this parameter to be 1. Another impressive aspect of the *ltm* package in R is that it has a built in command that will output, for any particular model, the approximated latent ability level given a set of possible responses. The command *factor.scores('model')* yields the ability level for all possible sets of responding to the items. The exam under study has 56 items, or questions, therefore we will not list all possible responses patterns with their ability level here but note that every possible response pattern will have a unique ability level attached to it. In other words two response patterns with equal number of correct responses will still have different ability level outputs since each individual item has a unique difficulty parameter. A simple plot command can also be used to express the item characteristic curves, item information curves, and test information curve. Recall that the item characteristic curve shows the relationship of the ability level of an individual against the probability of answering that an item correctly. Figure 7.1 shows the item characteristic curve for items 1, 14, 20, 30, and 45. The inflection point of each item characteristic curve corresponds with a 0.50 probability of answering that item correctly given the ability level along the x-axis. So for item 1 an examinee with roughly an ability level around -3 will have a 50% chance answering that item correctly and examinees with ability levels between 0 and 1 will have a 50% chance

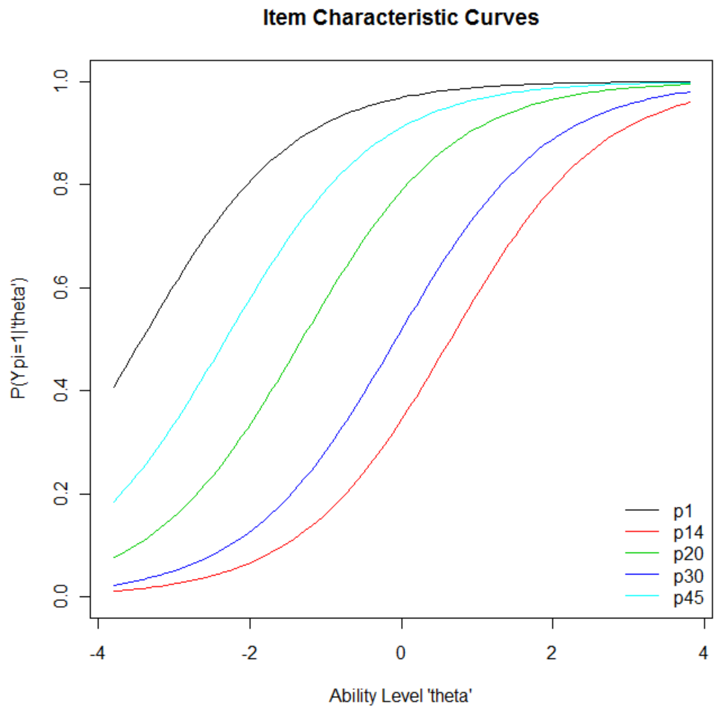


Figure 7.1: ICCs for five items from TAKS data set

of answering item 14 correctly. Being able to analyze each item individually using these item characteristic curves is quite beneficial in the sense items that are too "easy" or too "hard" in relation to all other items can be quickly identified and thrown out if one wishes to do so. Plotting all item characteristic curves can also be beneficial, as it can determine what type of ability levels the exam is testing. Figure 7.2 shows that the majority of the items on this exam are testing ability levels at or below an average ability level of 0. Because the exam under study is a standardized test to see if a student may move on to the next grade level, the fact that the items are generally testing to see if a student has the average ability of a student in that grade level seems to be quite natural. Notice also that this exam is using a number of items that have item characteristic curves that are spread over a wide range of ability levels. Being able to identify individual item characteristic curves allows test makers to uniquely construct exams based on how they would like the results to

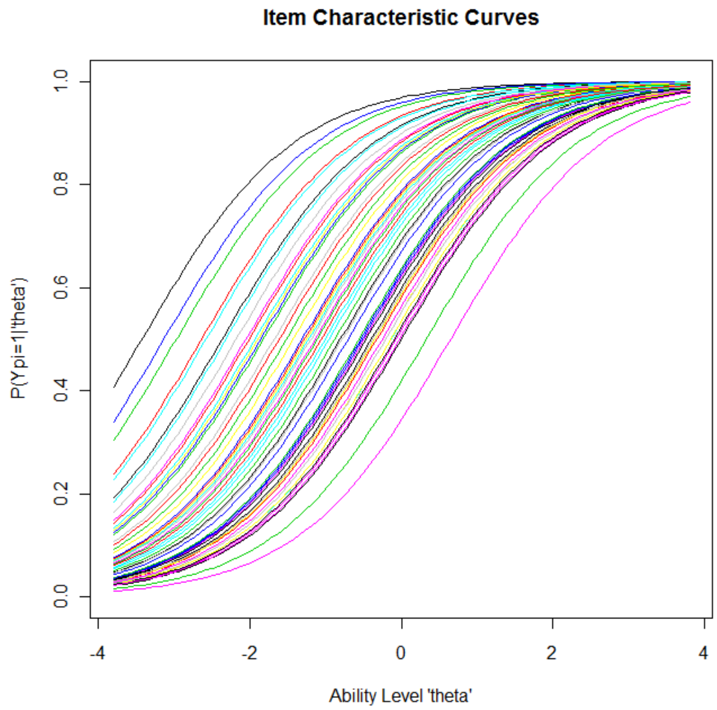


Figure 7.2: ICCs for all items from TAKS data set

look. For instance for very high stakes tests where the test maker neesd to know if an examinee has a particular ability level, the test maker can choose most of the items to be centered over that particular ability level. Item information curves can also be plotted quite easily using the R package *ltm* by identifying what type of curve to output. Item information curves show the relationship of how much information a particular item has over a certain range of ability levels, that is at what ability levels does a particular item measure the latent trait the 'best'. Recall that the information function of an item under the Rasch model is given by

$$I_i\left(\theta_p\right) = \frac{1}{\sigma^2} = P_i\left(\theta_p\right) Q_i\left(\theta_p\right)$$

where $P_i\left(\theta_p\right) = P\left(Y_{pi} = 1 |\ \theta_p\right)$ and $Q_i\left(\theta_p\right) = 1 - P_i\left(\theta_p\right)$, i.e. the probabilities of answering the item correctly and incorrectly respectively. So analyzation of Figure 7.3 shows us that item or problem thirty gives the gives the most information for

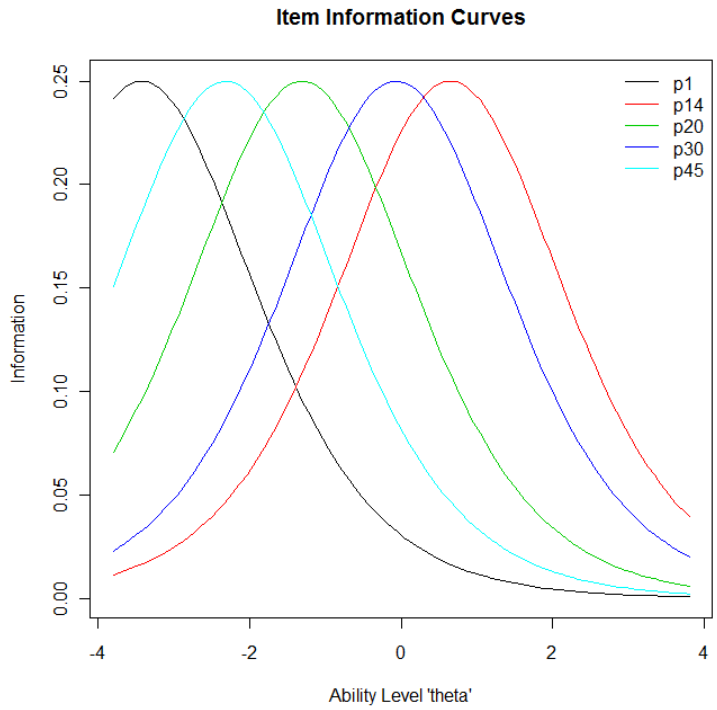


Figure 7.3: Information curves for five items from TAKS data set

examinees with average ability level and furthermore would probability not be a good question if one wishes to test someone with ability levels less than -1 or greater than 1. In comparing figures 7.1 and 7.3 one can see that the peak of the item information curve direectly relates to the inflection point of the item characteristic curves. Therefore for the same reason a test maker that wants to test students that have a wide range of ability levels would choose items that have item characteristic curves that cover a large range of ability levels, the test maker would choose items that have item information curves that cover a large range of ability levels. Intuitively though it might be more clear to use the item information curve to determine what items to choose for an exam, whether you want them to cover a large or small range of ability levels. Looking at all items information curves can be quite beneficial for reasons previously mentioned. Therefore from Figure 7.4 one can see that the test makers for this TAKS test wanted items that give the most

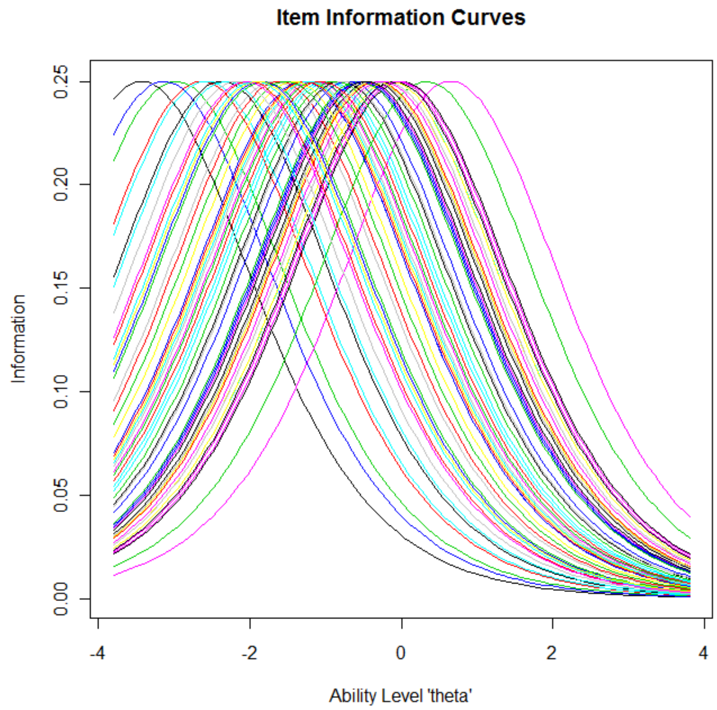


Figure 7.4: Information curves for all items from TAKS data set

information for ability levels that range from about -3 to just over 1. This might be anticipated since one assume that in a group of tenth graders there are going to be a wide range of abilities, thus to “cover all bases” the test maker made sure to have items that would measure abilities at all of those ability levels. On the other hand a test maker that was putting together questions for say a law school entry exam, he/she might only use items that give the most information for the ‘cut off’ ability level, in other words the test maker does not care to measure the ability levels at the extreme ends of the ability scale only whether they have the ability to succeed at their program. One can now see how the information function and item information curves can be quite beneficial for test making and analyzation of an already made exam. Also note that as mentioned in our general discussion on item response theory that we defined the test information curve as the sum of all item information functions, thus this test information curve can also be easily plotted and analyzed.

In a comparison of the item information curves and the test information curve in

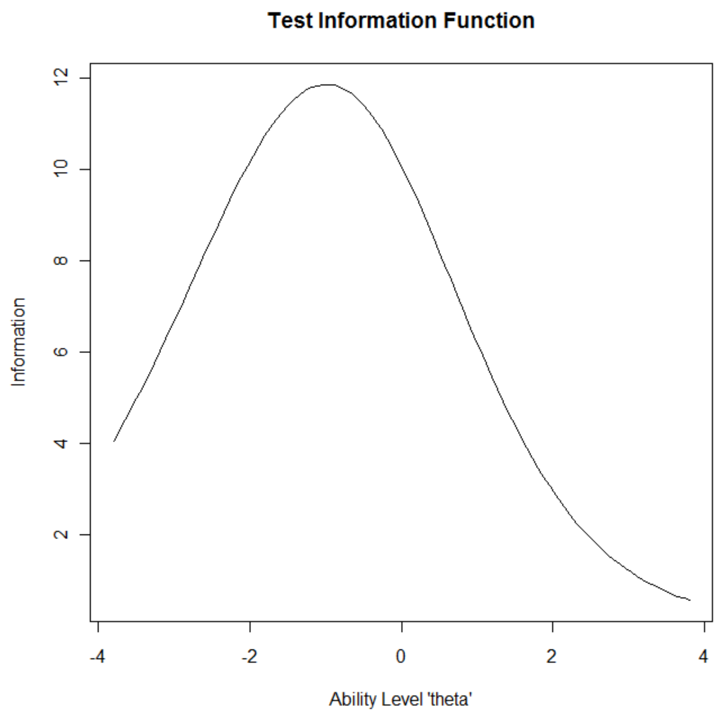


Figure 7.5: Test information curve for items from TAKS data set

Figure 7.5 one can see how the actual values on the y-axis, the information values, are much higher than on for the test information. This is natural from the way we define the test information function. The test information function for our TAKS test observed data essentially implies the same thing that we were able to conclude from Figure 7.4 of the individual item information functions, although having a single curve to represent the test information is a lot of times much easier to interpret and in general looks a little nicer graphically. So, to conclude the R package *ltm* is quite easy to navigate and although we only used a few applications in analyzing a data set with a Rasch model, there exist numerous other models and techniques that can be explored under this package including polytomous data models in particular the partial credit model and the graded response model [13].

# 8 Conclusion

Item response theory is considered by many as the future of psychological test theory and is being implemented more and more throughout the fields of psychometrics, econometrics, and education, as well as being implemented in many high stakes testing environments. Throughout this introduction to IRT we have discussed a test theory approach, a generalized linear model approach, and applied these methods to actual data obtained from a local highschool's standardized test responses. In our exploration of IRT using a test theory approach we developed three models to describe the relationship between a person's ability level and the probability of answering an item correctly. These three models were the Rasch model, or one-parameter logistic model which consists of an item difficulty parameter, the two-parameter model which introduces an item discrimination parameter, and the three-parameter model which introduces a guessing parameter to go along with difficulty and discrimination parameters from the previous models. We also discussed how the item information function can help with analyzation of items and what ability levels that an item is best suited for measuring. In the next few sections we explored the Rasch model in context of the larger group of statistical models termed generalized linear models. Four extensions of the Rasch model were constructed introducing "hidden" parameters into the models, that is person parameters, items parameters, or both were taken into consideration. Likelihood functions were then derived for each model under this context of GLMs which could then be applied to approximation and maximization methods for find parameter values that best fit the model to a set of observed data. We also explored a couple of popular integral approximation methods that are used for generalized linear models as well as a maximization process. The last two models developed were those for data sets that consisted of polytomous, or multiple, responses within an item. These two models were termed the partial credit model and the graded response model, each of which lend themselves to items that have ordinal responses. Lastly, we applied Dimitris Rizopoulos R package *ltm* to analyze TAKS, the Texas high school standardized test, responses for a group of tenth grade students. Using this package we explored how item characteristic curves and item information curves can be beneficial when interpreting how well a test measures a group of examinees ability levels. There are many area of further research that would be quite interesting to explore. For instance, considering the two- and three-parameters models in the context of generalized linear models and extending these models the same way we did with the Rasch model, introducing person and item parameters, could have some very interesting results. Another area of further research would be to consider models that take into account multi-dimensional latent trait parameters. Note that throughout our discussion the latent trait, $\theta_p$, we uni-dimensional , therefore each item was only measuring a single latent trait, or skill at a time. Extending this latent trait into multiple dimensions would most likely have some exciting results for item response theory.

## References


[1] P. De Boeck and M. Wilson. *Explanatory Item Response Models, a Generalized Linear and Nonlinear Approach*. New York: Springer, 2004.

[2] J. Gill. *Generalized Linear Models: A Unified Approach*. Thousand Oaks, California: Sage Publications, 2001.

[3] R. Smith. "Theory and Practice of Fit", American Dental Association. <http://rasch.org/rmt/rmt34b.htm>

[4] F.B. Baker. *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, 2001.

[5] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Boca Raton: Chapman & Hall/CRC, 1989.

[6] "Item Response Theory". M. Brannick. University of South Florida. 31 Jan 2012 <http://luna.cas.usf.edu/ mbrannic/files/pmet/irt.htm>.

[7] E. Muraki. "A Generalized Partial Credit Model: Application of an EM Algorithm". Sage Publications, 1992.

[8] W.J. Den Haan. "Numerical Integration." London School of Economics, London 3 June, 2011.

[9] M.S. Johnson. "Item Response Models and their use in Measuring Food Insecurity and Hunger." National Academies. Department of Statistics and

Computer Information Systems, Baruch College, City University of New York, 3 July, 2004.

[10] "EM Algorithm." S. Chang and H.J. Kim. University of Iowa, 9 Dec, 2007.

[11] "The EM Algorithm." A Singh. Carnegie Mellon University School of Computer Science, 20 Nov, 2005.

[12] J. Fox. *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks, California: Sage Publications, 2008.

[13] D. Rizopoulos. "Ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses." *Journal of Statistical Software, Issue 5*, 17 Nov, 2006.