Correcting for Rater Bias in the Presence of Non-Ignorable Missing Ratings

BY

ANDREW SWANLUND
B.A., Carleton College, 1999
M.S., University of Minnesota, 2005

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2016

Chicago, Illinois

Defense Committee:

George Karabatsos, Chair and Advisor
Everett Smith, Jr.
Yue Yin
Donald Hedeker, University of Chicago
Ryan Martin, North Carolina State University

ACKNOWLEDGMENTS

This thesis would not have been possible without the persistent help of family, friends, and colleagues. I want to thank Richard Smith and the staff at Data Recognition Corporation for inspiring me to return to graduate school, and Everett Smith, Jr. for facilitating that transition by connecting me to a job in Chicago and the University of Illinois at Chicago. I want to thank my advisor, George Karabatsos, for taking a chance on me as a graduate student and supporting me through my lengthy, part-time enrollment and studies.

I am especially indebted to my colleagues at American Institutes for Research (and formerly, Learning Point Associates) for providing a flexible and supportive environment to pursue my studies while simultaneously learning educational research and measurement on the job. In particular, my fellow employees (and previously, fellow graduate students) Kelly Hallberg, Chloe Gibbs, and Ryan Williams helped me hone my ideas and provided sympathetic ears as I struggled at times to move forward. My boss and mentor Lawrence Friedman (along with David Myers and Terri Pigott) applied proper pressure at key points to help me keep moving towards realization of my Ph.D.

Finally, I would like to thank my mom and dad for working hard to instill in me a love of learning, and for working hard to enable me to attend college and beyond.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF ABBREVIATIONS

GT	Generalizability Theory
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MFRM	Many-Facet Rasch Model
MNAR	Missing Not at Random
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
SEM	Standard Error of Measurement

SUMMARY

This thesis addresses the problem of non-ignorable missing ratings in judge rated data. A Bayesian bivariate probit ordinal missing data model implemented with Markov chain Monte Carlo (MCMC) was applied to simulated and real-world data sets to test the extent to which this proposed approach outperformed existing methods for analyzing judge rated data across a variety of evaluation criteria and data collection scenarios. The MCMC approach was compared to the many-facet Rasch model, generalizability theory (with a linear regression correction for rater effects), and the Rasch rating scale model. The objectives of the research were to test the extent to which the proposed methods could 1) calculate generalizability theory variance components when traditional methods could not be applied, and 2) produce more accurate latent trait measures than existing methods. The study used eight simulated data sets with varying numbers of examinees, raters, items, and distributional properties of examinee ability estimates. In addition a real-world data set consisting of classroom observations was used to test the applicability of the methods to non-simulated data.

The Bayesian bivariate missing data model produced variance component estimates (and D-study coefficients) that were quite accurate for measurement scenarios with only a single, randomly assigned rater. The MCMC approach yields confidence intervals with better coverage probabilities than traditional approaches, and this finding is consistent when raters are randomly or non-randomly assigned to examinees. This modeling approach more accurately models the uncertainty in examinee scores by taking into better account the error due to rater severity, and non-random assignment of raters.

I. INTRODUCTION

An important goal of educational measurement is to develop instruments and measures which provide researchers meaningful and valid information on educational phenomena, including the performance of examinees on a test or questionnaire. Of particular interest is the application of rigorous procedures to ensure that inferences are not based on statistical methods which do not account for potential issues with data collection and measurement. Failure to account for the possibility of systematic errors and biases during study design, data collection, and the subsequent statistical modeling can result in faulty conclusions about the effectiveness of educational programs and policies, as well as improper decisions about individual students, their teachers, or their schools.

This thesis will address the open problem of correcting for rater bias in the presence of non-ignorable missing ratings. In particular, single-rater and multi-rater designs, where raters form disjoint subsets in that they are totally uncrossed with examinees, pose a significant challenge to the development of fair, objective measures of examinee ability when raters are not interchangeable. Furthermore, measurement may even be problematic for multi-rater designs without disjoint subsets, when the missing data mechanism is non-ignorable. This thesis will draw upon the knowledge of several distinct lines of research. First, research on the sources and types of rater bias will help in understanding the nature of the problem, and suggest ways to locate such biases in the data. Next, the methods proposed here will build upon existing methods for analyzing and dealing with measurement error due to raters (such as Generalizability Theory, Many-Facet Rasch Measurement, and classical test theory correction methods). Finally, the fields of missing data analysis and Bayesian data analysis will be combined to develop a solution to this problem. This thesis will propose methods for correcting for rater bias in situations when

existing methods (which require partially crossed rating designs and no disjoint subsets in the data) cannot be applied. Furthermore, the proposed methods will handle measurement scenarios where missing data mechanisms cannot be considered ignorable.

There is a wealth of literature which has documented the presence, impact, and theory of rater effects. Rater effects have been shown to significantly decrease the reliability of measures, and remain persistent in practice, even with rigorous monitoring and training of raters. Research into rater effects has followed three main lines of research: rater cognition; contextual factors that relate to rater effects; and the identification of and correction of rater effects (Wolfe, 2004). This thesis will sit within the third line of research – identifying and correcting for rater effects. One particular example of a rater effect is severity/leniency bias. This type of bias occurs when judges have systematic biases that cause them to (on average) rate examinees higher (leniency) or lower (severity) than their true scores. While the models proposed in this thesis will focus on rater severity/leniency effects, this chapter provides detailed description of a wide range of rater effects.

To complement the literature on rater effects, there exists a body of research on determining the how rater effects manifest in a measurement system, the extent of those rater effects, how those rater effects are affecting examinee scores, and how to correct scores for those effects. The field of psychometrics focuses on developing, analyzing, and refining measurement systems across a wide-range of sciences including educational psychology. Generalizability Theory (GT), an important area of psychometrics, is the study of variance in scores attributable to distinct sources of measurement error (Brennan, 2001). These sources of error are typically referred to as facets. Essentially a modified version of traditional analysis of variance combined with classical test theory, GT seeks to estimate the level of error in a given observation that can

be allocated to each facet of measurement. Examples of facets include test or observation protocol items, raters, writing prompts, testing locations, and the occasion of testing. An alternative to GT which also examines the impact of the facets of measurement is the Many-Facet Rasch Model or MFRM (Linacre & Wright, 2004). Both of these approaches were developed to partition and account for variability due to facets of measurement for a given measurement scenario. For example, if classrooms are observed on multiple days by multiple raters, with respect to multiple items or tasks, GT and MFRM could produce estimates of measurement error attributable to items, raters, and occasion. However, for GT and MFRM to be applicable and useful, careful consideration must be given to the design of the data collection process (e.g., the number of raters per observation).

If data collection is not conducted in a particular fashion, standard GT and MFRM cannot be used. In particular, it is important that the data are sufficiently complete to allow for the estimation of random (for GT) or fixed (for MFRM) effects for the raters (or other facets of interest). Sufficiently complete is meant to imply that there are no disjoint subsets within the data. For example, if only one judge observes each person, then no model will be able to disentangle the judge effect from the person effect – that is, the two effects will be confounded (without adding additional assumptions or information to the model). For GT, this will result in the inability to separate out the variance due to persons from the variance due to persons-judge interactions. For MFRM, the model will be unable to determine the fixed effects (judge severity measures) associated with each judge. When data collection results in complete separation (i.e., the presence of disjoint subsets), traditional GT and MFRM cannot be used to analyze and deal with the measurement error attributable to the facets of measurement. For this scenario, it

becomes necessary to construct a method for dealing with the error introduced into the measures by the facets of measurement.

The situation where data do not exist to separate the effects attributable to the facets of measurement from the effects attributable to the objects of measurement can be treated as a missing data problem. One can imagine a measurement scenario with five judges rating examinees. If each examinee is rated by only one judge of the five, then the hypothetical and/or unknown ratings of the remaining four judges could be treated as missing data (which could have been collected under a more complete data collection process). This missing data scenario can be defined as missing response data (dependent variable missingness), with complete covariate data (i.e., the design matrix of judges, items, and persons). Assuming these missing data have no effect on the scores (or ability estimates) of the examinees would require at least an assumption that data were missing at random (MAR), meaning that the information from the observed ratings can be used to predict the values of the missing ratings, via the measurement (psychometric) model for the judge ratings. Indeed, if all the missing ratings are MAR, then a complete-case analysis can be used with the measurement model (Graham & Donaldson, 1993). However, MAR is arguably a very strong assumption, as almost always in practice, missing rating data results in the analyst's inability to determine the effect of the judges (or any other facet of measurement) on the observed scores, even before data collection. In turn, this can lead to biased estimates of judge rating severity parameters in a measurement model, when the model is used in a complete case analysis (Schafer & Graham, 2002).

When latent trait measures for individuals depend on rater perceptions (and thereby include error attributable to raters), it becomes necessary to consider methods for correcting those measures to account for error and bias induced by the raters. While GT focuses on

estimating the extent of variability due to rater effects, other methods have been applied to adjust scores based on inference about the effects due to judges (and other facets of measurement). These approaches include ordinary least squares (OLS) regression methods, MFRM and other latent trait models, linear scaling approaches, and data imputation (MacMillan, 2000). As with GT and MFRM, these methods have focused on cases where multiple raters (at least two) have rated each examinee. This thesis will discuss existing methods and then propose a correction model (a bivariate probit multiple imputation Markov chain Monte Carlo item response model) which applies to cases where only a single rater has rated each examinee. Additional scenarios where this model will be applicable will be for judge rating designs where the missing data mechanism is non-ignorable. Furthermore, this correction model will allow for the estimation of traditional GT statistics including reliability and the extent to which measurement error is partitioned among the facets.

This thesis outlines, develops, and tests a method for applying Markov chain Monte Carlo (MCMC) and missing data analysis techniques to a sparse data set of judge rated examinees (i.e., sparse to the point of only a single judge rating each examinee). The proposed methods are applied with the intent of generating estimates of reliability and error similar to traditional GT and correcting for rater bias (in a situation not previously addressed in the literature) including instances when missing data are non-ignorable. This thesis addresses the following research objectives:

- **Research Objective 1:** To develop statistical methods that allow one to investigate the extent to which measurement error (defined as the difference between one's true latent ability and the measure of that ability) is partitioned among the facets of measurement

(with an emphasis on rater bias) for data collection scenarios where traditional approaches in the literature cannot be applied.

- **Research Objective 2:** To develop methods which produce latent trait measure estimates (which are closer to true generating parameters) that account for rater bias and error due to the other facets of measurement, compared to existing methods which either cannot account for rater bias, or assume non-ignorable missing data.

This chapter discusses the key topics encountered in the discussion above. In particular, the following sections describe sources of bias and measurement error (i.e., that due to judge ratings), existing methods for addressing that measurement error (including outlining, comparing, and contrasting GT and MFRM), and existing methods used to correct individual measures of the latent trait. The topic of missing data is explored and includes foundational definitions of different missing data types and methods for handling missing data. Bayesian inference and MCMC methods are presented, keying in on how they can be used for polytomous multi-rater item response models. The remainder of this chapter serves only as an introduction and overview of these topics. A more technical, detailed discussion of how the models are developed and implemented is provided in the second chapter.

A. Sources of Bias and Measurement Error

Generalizability theory and MFRM are both concerned with the issue of external bias in observed scores for the objects of measurement. It is not difficult to imagine how any given measurement scenario has multiple competing and interacting sources of bias external to the trait of interest which can influence an individual's overall score. For example, when raters are involved in determining the score for an individual on a test, both the characteristics of the rater and the individual being rated (interacting with the rater's biases) can introduce considerable bias

into a measurement situation. While the effects of many different sources of measurement error can and should be considered when developing measures (e.g., raters, occasions, items), this thesis focuses on measurement error due to raters (also referred to as judges). Scenarios that involve judge rated scoring include classroom observations, college acceptance decisions, written essay scoring, partial credit item scoring, oral examinations, and job applicant decisions (to name a few).

Rater effects can be described as systematic patterns in the behavior of the raters that result in inaccurate or biased scoring of the objects of measurement (Wolfe, 2004). Despite rigorous training of raters aimed at minimizing or eliminating bias due to raters, many studies have shown that rater effects continue to be prevalent in the data (e.g., Blok, 1985; Braun, 1988; Englehard, 1992; Lane & Sabers, 1989; Luntz, Wright, & Linarcre, 1990). The reliability of judge ratings can be quite low. King, Hunter, & Schmidt (1980) found that reliabilities for single rater designs were often below 0.60. Similar to adding items to an assessment, adding additional raters (i.e., multiple raters for each object of measurement) can reduce the measurement error attributable to raters, thereby increasing reliability. Determining the extent to which additional raters may increase reliability is a major focus of GT. However, reliability is not the only issue when dealing with rater error in measurement. Bias introduced into examinee scores can affect inference based on those scores in a substantial manner. Examinees may be rewarded or penalized unjustifiably depending on the particular rater (or subset of raters) they are judged by (Guilford, 1954; Raymond, Webb, & Houston, 1991). That is, if an examinee is assigned a particularly lenient rater or severe rater, his or her score may be artificially higher or lower than his or her true score, respectively.

It is important to clarify what is meant by rater bias (or rater effects) in contrast to rater error (and random error in general), in terms of a bias-variance decomposition. Consider the following item-response model for a dichotomous item rating y_{nij}, with underlying (real-valued) latent response y^*_{nij}.

$$y_{nij} = \mathrm{I}\left(y^*_{nij} > 0\right) \tag{1}$$

$$y^*_{nij} = \mathbf{x}^{\mathrm{T}}_{nij}\boldsymbol{\beta} + \varepsilon_{nij} \tag{2}$$

In this model, $\mathrm{I}(\cdot)$ denotes the indicator function, each $\mathbf{x}^{\mathrm{T}}_{\mathrm{nij}}$ is the vector of person, item, and judge indicator (0/1) variables corresponding to person *n*, item *i*, and rater *j* (i.e., the *n*-th row of the design matrix), with corresponding coefficient parameter vector $\boldsymbol{\beta}$ having sample point-estimate $\widehat{\boldsymbol{\beta}}$. The associated error term is ε_{nij}, assuming $\mathrm{E}\left(\varepsilon_{nij}\right) = 0$ and $\mathrm{Var}\left(\varepsilon_{nij}\right) = \sigma^2_\varepsilon$. Specifically, $\sigma^2_\varepsilon = 1$ in the case of a probit model; for a logit model, σ^2_ε follows a Kolmogorov-Smirnov distribution (Holmes & Held, 2006). The prediction error for a matrix of predictors variables $\mathbf{X}$ can be decomposed as follows (Hastie, Tibshirani, & Friedman, 2009):

$$\begin{aligned}
\mathrm{Err}(\mathbf{X}) &= \mathrm{E}\left[\left(\mathbf{Y} - \mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\beta}}\right)^{\mathbf{2}}\right] && (3)\\
&= \sigma^2_\varepsilon + \left[\mathrm{E}\left(\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\beta}}\right) - \mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\beta}}\right] + E\left[\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\beta}} - \mathrm{E}\left(\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\beta}}\right)\right]^2 && (4)\\
&= \sigma^2_\varepsilon + \mathrm{Bias}^2\left(\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\beta}}\right) + \mathrm{Var}\left(\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\beta}}\right) && (5)\\
&= \text{irreducible random error} + \mathrm{Bias}^2 + \mathrm{Variance} && (6)
\end{aligned}$$

This decomposition shows that the prediction error for a particular item rating (as modeled by the regression model) can be decomposed into three parts: the stable random error of

observations around predicted values, estimator bias, and estimator variance. This framework may be helpful for thinking about measurement error and how it relates to rater bias. Part of the error term is stable. For example, in the above model, predicted values for the observable ratings will fall in the range of values between 0 and 1, whereas observed values will be either 1 or 0. If there was no systematic bias due to raters, persons, or items, and the variances attributable to these facets were 0, the term σ_{ε}^{2} would fully determine the error of prediction of the observable ratings. The discussion of rater effects that follows describes these effects in terms of how they impact this bias-variance decomposition.

Wolfe (2004) describes three prominent lines of research with respect to the study of rater effects. The first line of research has been focused on the area of rater cognition. That is, the extent to which there exists a relationship between a rater's cognitive processing abilities and his or her rating proficiency (e.g., Breland & Jones, 1984; Freedman, 1979; Freedman & Calfee, 1983; Wolfe, 1997; Wolfe, Kao, & Ramney, 1998). Similarities exist between this line of research and general research in the performance of experts and novices. The second line of research focuses on contextual factors that relate to rater effects such as the characteristics of raters, the tasks, interactions among raters and the targets (e.g., halo effects), and the rating environment (e.g., Dean, 1980; Hoyt, 1999; Hoyt, 2000; McIntyre, Smith, & Hassett, 1984; Murphy & Anhalt, 1992; Murphy & Jako, 1990). The third line of research described by Wolfe focuses on the impact of rater effects on the ratings themselves (i.e., reliability and bias), and the development of methods for statistically correcting these scores. This thesis sits within this third line of research, with a focus on estimating the size of rater effects and correcting for them in the estimation of examinee ability on the test.

Several types of rater effects exist and have been studied for their impact on reliability and examinee scores (Hoyt, 2000; Wolfe, 2004). These types of rater effect speak to characteristics of the rater (accuracy/inaccuracy, severity/leniency), how they use the scoring scale (centrality/extremism, restriction of range), and how they interact with examinees (dyadic variance, halo effect). Some terminology is necessary to describe these characteristics within a statistical framework. Each of these characteristics can be examined by describing the relationship between the actual (or observed) ratings and the conditional expected rating for an object of measurement. The conditional expected rating is the average rating that an individual would receive over an infinite number of hypothetical tests, conditional on a particular item, judge and specific levels of other facets of measurement. Within classical test theory, this is often referred to as the true score. Within latent trait modeling, this is the expected value of a rating given the generating parameters of the latent trait model.

Sources of rater effects due to characteristics of the rater include accuracy/inaccuracy and severity/leniency. A rater is considered to be accurate when his or her ratings are close to the true conditional expected value for the examinee. Accuracy is reflected by a low residual standard deviation (where the residuals are calculated as the observed score minus the expected score) and no significant correlation between the residuals and the expected score. Inaccurate raters have greater variation in their ratings around the expected score. This increase in variation manifests itself as a higher standard deviation of the residuals. However, for inaccurate raters, the correlation between the residuals and expected scores is still likely to be non-significant (assuming the inaccuracy is not biased in one direction or another). Rater severity (or harshness) refers to the general tendency of some raters to assign lower scores to examinees. In contrast, lenient raters generally assign higher scores to examinees. For latent trait models, severe and

lenient raters would have standard deviations of the residuals which were near to zero. For these models (such as MFRM), harshness and leniency would be exhibited by positive or negative rater parameters, respectively. For classical test theory models (utilizing raw scores), rater severity and leniency would show up as greater residual variance around the expected scores. Lenient judges would tend to have positive residuals across the scoring scale. Severe judges would tend to have negative residuals across the scoring scale. If leniency or severity are a consistent characteristic of a rater, there will tend to be a non-significant correlation between the residuals and expected score.

Table I describes the accuracy/inaccuracy and severity/leniency characteristics in the bias-variance decomposition terms described earlier. While it may seem counter-intuitive that raters can be both accurate and severe (or lenient and severe), this table is provided mostly to show how accuracy/inaccuracy is related to the variance of the ratings, and severity/leniency is related to the bias of the ratings. Accuracy in this sense refers to the low variance of the rater's scores (i.e., on hypothetical repeated ratings, the rater would assign scores which were quite similar to one another), rather than the correctness of the scores. Stated another way, accurate/severe or accurate/lenient raters assign incorrect, biased scores, which exhibit high precision.

TABLE I

ACCURACY VERSUS SEVERITY

	Accuracy	
Severity	Accurate	Inaccurate
Severe	Negative bias/low variance	Negative bias/high variance
Lenient	Positive bias/low variance	Positive bias/high variance

In addition to effects stemming from the characteristics of the raters, there can be effects due to interactions among the raters and the scoring scale. These interactions are the result of different raters interpreting the scoring scale in different manners. Raters who suffer from centrality effect (Wolfe, 2004) do not assign many scores at the high or low end of the score range. Their ratings tend to be near the middle of the scale. Conversely, raters who suffer from extremism (Wolfe, 2004) overuse the scores near the high or low end of the score range. Centrality can be seen statistically as large positive residuals for lower expected scores, and large negative residuals for higher expected scores. Or, statistically, the correlation between residuals and expected values for raters with centrality error will tend towards -1. In contrast, raters suffering from extremism will have correlations between residuals and expected values which approach +1. An additional form of interaction between raters and the scoring scale is referred to as restriction of range (Wolfe, 2004), which results from a combination of centrality and severity or leniency. However, if one were to correctly specify a latent trait model (such as the MFRM) with appropriate parameters describing the interactions between rating categories (thresholds) and judges, these correlations would become zero.

A rater effect can also occur as a result of an interaction between a rater and examinee. Hoyt (2000) refers to these effects as dyad specific bias – that is, a rater's unique interpretation of particular targets of measurement (e.g., examinees). An example of dyad specific bias is the halo effect. A halo effect (or dyadic covariance) occurs as a correlation among item scores for an individual examinee (conditional on the latent trait). In simpler terms, a halo effect occurs when a rater's general impression of an examinee influences the ratings for that person, regardless of the individual characteristic under consideration (Linn & Miller, 2005). If the rater has a favorable view of the individual, then ratings for all items are likely to be higher as a result – that is, item ratings will not be conditionally independent as would be assumed in either GT or MFRM. In MFRM, halo effects (or their opposite for unfavorable ratings) will manifest through over fit of the persons to the expectations of the model. Person over fit indices alert the data analyst to situations where item responses by the raters are too predictable for an individual (Wolfe & Smith, 2007).

B. Existing Approaches for Characterizing Measurement Error

Several approaches exist for characterizing and accounting for measurement error. This thesis first discusses two popular approaches (generalizability theory and the many-facet Rasch model), providing the general details of each and then comparing and contrasting their features. Following the description of generalizability theory and the many-facet Rasch model will be a review of various correction methods used to adjust examinee scores to better account for rater effects.

1. Generalizability Theory

Generalizability theory is an extension of classical test theory. GT seeks to understand the reliability of measures based on multiple facets of measurement, with raw scores

on items as the primary unit of analysis. This line of analysis combines traditional theories of reliability (taken from classical test theory) with the statistical tools of analysis of variance (ANOVA) techniques (Brennan, 2001). Cronbach, Rajaratnam, and Gleser (1963) produced the first work in this area and it has seen many developments over the subsequent decades (Brennan, 2001). While classical test theory viewed measurement error as one large pool of undifferentiated measurement error, GT seeks to disentangle the sources of measurement error from one another, evaluating them both as individual components of error, and as a whole (Shavelson & Webb, 2006). In measurement scenarios with persons, items, raters, and potentially other sources of variation, GT can be capable of attributing each facet its specific portion of the total variance. GT employs ANOVA to fit a linear model to the data, which can then be used to estimate the variance component associated with each measurement facet. Examples of these models will be shown below.

Test theory and measurement theory are based upon the assumption that an object of measurement has a "true score" on some underlying latent trait of interest (e.g., intelligence, satisfaction, mathematics ability, etc.). Instruments designed to measure that true score invariably result in scores that contain measurement error. Thus, a person's observed score can be defined as $X = T + E$ where X is the person's observed score, T is the person's true score, and E is the error of measurement (Webb, Shavelson, & Haertel, 2006). Classical reliability theory focuses on how the error of measurement affects the reliability (or consistency) of scores. GT goes further to partition this measurement error among the facets of measurement.

A key concept of GT is the "universe of admissible observations" (Shavelson & Webb, 2006). This "universe" is comprised of all observations which decision-makers would view as interchangeable with the current measurement. For example, if a student received a particular

score on a writing prompt scored by two judges, a decision maker would deem a score to be comparable or interchangeable with that score, if that score was based on a combination of a writing prompt and judges that were part of the universe of admissible observations. GT further distinguishes between generalizability studies (G-studies) and decision studies (D-Studies). G-studies classify the universe of admissible observations to be as all-encompassing as possible (e.g., all possible items, raters, occasions, etc.). By this method, the estimates of variance and reliability are appropriate for a variety of uses and needs of decision-makers. D-studies, in contrast, typically analyze a particular set of facets that are pertinent to a specific decision or measurement scenario (Brennan, 2001; Shavelson & Webb, 2006). The estimates of the variance components from a G-study can be used to inform the estimates of reliability of the person measures given the specifics of a measurement situation (i.e., the specific facets of measurement used in that setting). G-studies and D-studies sometimes are referred to under different names in the literature. For example, Raudenbush, Martinez, Bloom, Zhu, and Lin (2007) recommend conducting a "reliability study" before conducting an impact study. In that reliability study, the variance due to the facets of measurement are estimated (a G-study), and then those estimates are used to develop a data collection process (D-study) which will maximize statistical power for the impact study.

a. The Single Facet Case

The simplest case of a GT model is the single-facet case where each person (or object of measurement) is administered the same sample of items. In the GT literature, this design is usually denoted as $p \times i$ to signify that persons are crossed with items. Another single-facet design, denoted $i{:}\,p$, represents the situation where different sets of items are presented to

different examinees – that is, items are nested within persons. For the single-facet crossed design, the person item score can be represented by the linear model (Brennan, 2001):

$$X_{pi} = \mu + \nu_p + \nu_i + \nu_{pi} \quad (7)$$

where X_{pi} is the person-item score, μ is the grand mean, ν_p is the random person effect, ν_i is the random item effect, and ν_{pi} is the residual effect. GT is concerned with estimating the variance components for each of the facets of measurement. Calculations of the variance of the random effects associated with each facet of measurement can be used to determine the extent to which measurement error affects the total scores for the objects of measurement (presented generally as a G-coefficient) and in particular, how much of that error is attributable to different facets of measurement.

b. <u>The Multi-Facet Case</u>

A more general version of a GT model is the multi-facet GT model. One such multi-facet design can be denoted $p \times i \times r$ where persons p are crossed with items i and raters r. In this scenario, the items could represent constructed response test items, writing prompts, or even ratings relative to some indicator for a classroom observation. For the multi-facet design, the person-item score can be represented by the linear model (Brennan, 2001):

$$X_{pir} = \mu + \nu_p + \nu_i + \nu_r + \nu_{pi} + \nu_{pr} + \nu_{ir} + \nu_{pir} \quad (8)$$

For this more complete model, it becomes apparent that fully crossed data is necessary to completely estimate all of the variance components in the GT model. For example, if all raters do

not rate all items, $\sigma^2(ir)$ – i.e., the variance attributable to the interaction between raters and items – could not be calculated.

There are two main reliability coefficients that are calculated within generalizability theory (Shavelson & Webb, 2006). The first coefficient, typically called the generalizability coefficient (or G-coefficient) is similar to the reliability coefficient calculated in classical test theory. That is, the G-coefficient estimates the ratio of true score variance to total variance. The G-coefficient is useful for determining the reliability of the scores produced from a measurement system for the purpose of making relative decisions (i.e., comparing examinees to other examinees). A second coefficient is called the dependability index (Kane & Brennan, 1977). This coefficient provides an estimate of the reliability of absolute decisions (i.e., the error of measurement when comparing examinees to a particular criterion or cut score). The dependability index is the appropriate generalizability theory reliability index for criterion referenced decisions.

2. Many-Facet Rasch Model

The Many-Facet Rasch Model (MFRM) is an extension of the dichotomous Rasch model (Rasch, 1960; Wright & Stone, 1979), the Rasch Rating Scale Model (Andrich, 1978; Wright & Masters, 1982), and the Rasch Partial Credit Model (Wright & Masters, 1982). Like the family of Rasch models, the MFRM is a generalized linear model which incorporates a logistic link function to model the ordinal response function as predicted by fixed and random effects associated with the persons and facets of measurement. The MFRM adds parameters (to the traditional item and person parameters) for the additional facets of measurement considered here (e.g., rater severity). The MFRM allows one to model the impact of the variance due to the

facets of measurement on each measured person's score. In that sense, MFRM goes beyond the capabilities of traditional GT in that GT does not incorporate a model-based correction.

The basic MFRM (Linacre, 1989; Linacre & Wright, 2004) is commonly denoted in the Rasch measurement literature as follows (in model below, it is assumed that all items share a common rating scale, something that which can be modified to be more in line with a partial credit model formulation):

$$\log\left(P_{nijk}/P_{nij(k-1)}\right) = \theta_n - \delta_i - \gamma_j - \tau_k \tag{9}$$

where P_{nijk} is the probability of person (or object of measurement) n receiving a score of k on item i as rated by judge j. $P_{nij(k-1)}$ is the probability that person receives a score of $k-1$ in the same scenario. On the right side of the equation, θ_n represents the ability of person n, δ_i represents the difficulty of item i, γ_j represents the severity of judge j, and τ_k represents the Rasch threshold parameter for step k (i.e., the location on the latent continuum where there is equal probability of receiving a score of k relative to that of receiving a score of $k-1$). The MFRM is an extension of the Rasch rating scale model (Andrich, 1978; Wright & Masters, 1982). Including a parameter for rater severity allows the estimate of the latent trait to be adjusted based on the particular severity of the raters assigned to each examinee.

The model shown above can be generalized to allow for myriad measurement scenarios. For example, the model could be adjusted to account for different rating scales across items (a Partial Credit Model), rather than the single rating scale case shown above. This could be accomplished by replacing the τ_k term with τ_{ik} (which includes a subscript for each item i). Furthermore, the model could be adjusted to incorporate a variety of interactions among the main

effects (facets), including rater-person, rater-item, and rater-rating scale interactions (Kim & Wilson, 2009; Muckle & Karabatsos, 2009).

A general version of the Rasch model which incorporates many existing Rasch models including the MFRM is the multidimensional random coefficients multinomial logit model (Adams & Wilson, 1996; Adams, Wilson, & Wang, 1997). The item response probability model can be formulated as follows:

$$\Pr[X_{ik} = 1; \mathbf{A},\mathbf{b},\boldsymbol{\xi}|\theta] = \frac{\exp(b_{ik}\theta + \mathbf{a}_{ik}^{T}\boldsymbol{\xi})}{\sum_{k=1}^{K_i}\exp(b_{ik}\theta + \mathbf{a}_{ik}^{T}\boldsymbol{\xi})} \tag{10}$$

with response vector probability model

$$\Pr[\mathbf{X} = \mathbf{x}|\theta] = \boldsymbol{\Psi}(\theta,\boldsymbol{\xi})\exp[\mathbf{x}^{T}(\mathbf{b}\theta + \mathbf{A}\boldsymbol{\xi})] \tag{11}$$

where $\boldsymbol{\Psi}$ is defined as

$$\boldsymbol{\Psi}(\theta,\boldsymbol{\xi}) = \left\{\sum_{z\in\Omega}\exp[\mathbf{z}^{T}(\mathbf{b}\theta + \mathbf{A}\boldsymbol{\xi})]\right\}^{-1} \tag{12}$$

and $\mathbf{X}_i^{T} = (X_{i1}, X_{i2}, \ldots, X_{ik})$ is a vector-valued random variable with $X_{ik} = 1$ if the response to item i is in category k. The vector $\boldsymbol{\xi}^{T} = (\xi_1, \xi_2, \ldots, \xi_p)$ describes the p parameters for the items. The characteristics of the items and their response categories can be described through linear combinations of these parameters defined by the design vector $\mathbf{a}_{ik}$ ($i = 1, \ldots, I$ and $k = 1, \ldots, K_i$)

that can be assembled into a design matrix **A.** The vector $\mathbf{b}^T = (b_{11}, b_{12}, \ldots)$ defines the scoring function that assigns the performance level to each potential item response. θ is the (possibly multidimensional/vector-valued) latent variable of interest, and Ω is the set of all possible response vectors.

The MFRM and the more general multidimensional random coefficients multinomial logit model are members of a class of ordinal models referred to as adjacent categories models. Another class of ordinal models are called cumulative logit models and include item response theory models such as the modified Graded Response Model (Muraki, 1990). The MFRM and other adjacent category models allow for disordered estimates of the threshold parameters (the τ_k). Cumulative logit models, on the other hand, assume ordered thresholds (i.e., $\tau_0 < \tau_1 < \cdots < \tau_k$. The standard form of the modified Graded Response Model can be further modified to incorporate rater severity effects. Such a model can be defined as

$$P_{nijk} = \Pr\left[X_{nij} = k|\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\gamma}\right] = \int_{\tau_{k-1}}^{\tau_k} h\left(x_{nij}^*|\theta_n - \delta_i - \gamma_j, 1\right) dx_{nij}^* \tag{13}$$

where P_{nijk} is the probability of person (or object of measurement) n receiving a score of k on item i as rated by judge j. Furthermore, x_{nij}^* is the latent variable, and $h(\cdot)$is the density of a logistic distribution. On the right side of the equation, θ_n represents the ability of person n, δ_i represents the difficulty of item i, γ_j represents the severity of judge j, and τ_k represents the threshold parameter for step k (i.e., the point on the latent continuum that represents the transition point in probability from a score of k relative to that of receiving a score of $k-1$).

3. Contrasting Generalizability Theory and the Many-Facet Rasch Model

The two approaches (GT and MFRM) differ in the way they treat data and how the specific models are implemented. GT is typically implemented through a linear model (Brennan, 2001; Hocking, 1996). In contrast, MFRM fits a generalized linear model with a logistic link function to account for the ordinal response data (Kim & Wilson, 2009; Linacre, 1989; Linacre, 1995; Linacre & Wright, 2004). GT models either individual observed responses or total scores as the outcome variable and fits random effects for each of the facets of measurement and their interactions – if possible given the data collection structure (Brennan, 2001; Shavelson & Webb, 2006; Webb, Shavelson, & Haertel, 2006). Under such a model, the response variable is assumed to be continuous with interval properties – however, this assumption is troubling given that item responses and observed scores (especially in the case of rating items) are distinctly ordinal in nature (Wright & Lincare, 1989). Embretson (2006) has described problems that result from using standard parametric statistical routines applied to ordinal data. Maxwell and Delaney (1985) showed that differences in mean scores could be found using ordinal data when the true scores of the groups were the same. Embretson (1996) showed that improper use of ordinal data could result in the finding of significant interactions for experimental studies when no such interaction existed relative to the true scores. When data fit the Rasch model (and the MFRM), person measures approximate an interval scale.

In addition to assumptions about the nature of the outcome variable and how it should be modeled, GT and MFRM also model person, item, rater, and other facets in a different manner. In GT, all such effects are typically modeled as random effects, and are treated as nuisance parameters that are assumed to be normally distributed with mean zero. The specific effect for a particular rater or item are not of interest, rather the total variance and its effect on the precision

of measurement is considered most important. The targets of measurement in GT are the person scores and their universe score variance (Kim & Wilson, 2009; Linacre, 1995). In contrast, the MFRM fits a model with fixed effects for persons (depending on the estimation algorithm), items, and raters. The targets of measurement for the MFRM are all measurement facets. MFRM seeks to adjust person measures in a way to approximate each examinee's true score as accurately as possible.

Sudweeks, Reeve, and Bradshaw (2005) identify three key differences in terminology between GT and MFRM relating to the definition of facets, interactions among the facets, and the definition and interpretation of reliability. MFRM refers to any factor in the model as a facet of measurement. In particular, MFRM refers to persons (the object of measurement) as a measurement facet along with things like items and raters. In contrast, GT does not include persons in a listing of the measurement facets. GT and MFRM also differ in their treatment of interactions. In GT, interactions are analyzed through factorial analysis of variance. For example, a significant variance attributable to the interaction of raters and persons would imply that the rank ordering of persons varies substantially across raters. MFRM considers these interactions as phenomena which induce bias in the measurement system and are examined at the individual rater level (i.e., through differential rater functioning analysis, similar to differential item functioning analysis). Finally, the definition and interpretation of reliability varies slightly between GT and MFRM. GT reliability coefficients measure the precision of the mean score as a predictor of the true score (or universe score). For a single-facet model, this treatment of reliability is the same as that of Cronbach's alpha. In contrast, MFRM analyzes the spread of the data (standard deviation) relative to the amount of measurement error within the measurement system.

The ultimate goals of GT and MFRM also differ from one another. In GT, the goal is interchangeable items, raters, and all other facets of measurement. In GT, the desire is to develop and implement a set of items with similar difficulties. The case is similar for judges. By reducing the item and judge variance (i.e., having equally lenient/severe "machine-like" judges, and items of near equal difficulty), the precision of measurement can be greatly increased. The goal of MFRM on the other hand is to estimate measures for all facets of measurement, including examinees, on the same latent continuum.

The literature provides several examples of research comparing the utility and efficiency of GT and MFRM approaches (e.g., MacMillian, 2000; Smith & Kulikovich, 2004). Briggs and Wilson (2007) propose a model (which they refer to as Generalizability in Item Response Modeling – GIRM) that joins the two approaches under a common modeling framework. Their approach utilizes MCMC estimation methods to fit Rasch models with random item parameters (as opposed to the traditional fixed parameters). Fitting random item effects allows for a direct comparison of GIRM variance components to the variance components used in GT (while simultaneously applying a latent trait model to the ordinal response data).

4. Methods for Correcting Scores

Many methods of correcting scores for rater bias have been proposed and studied in the literature. The main approaches include regression based approaches, the use of the Many-facet Rasch Model, and the imputation of missing data. Regression approaches (e.g., de Gruijter, 1984; Houston, Raymond, & Svec, 1991; Raymond & Roberts, 1987; Raymond & Viswesvaran, 1983; Raymond, Webb, & Houston, 1991; Wilson, 1988), the approach most commonly applied in classical test theory (MacMillan, 2000), adjust observed scores (raw ordinal scores) to account for rater bias (or other bias due to measurement facets). These approaches fit linear models in the

observed score metric and add parameters for rater effects to the model. The adjusted scores are the conditional ability estimates (again, in the observed score metric), adjusted for rater fixed effects. The approach is similar to that of the MFRM (with the notable exception that these approaches do not use a logistic link function to account for the ordinal nature of the dependent variable).

The MFRM approach has seen wide use for detecting and correcting rater bias due to rater severity/leniency (e.g., Englehard, 1994; Englehard, 1996; Englehard & Myford, 2009; Gyagenda & Englehard, 2009; Kim & Wilson, 2009; Lang & Wilkerson, 2005; Looney, 2004; Myford & Wolfe, 2003; Myford & Wolfe, 2004; Wolfe, 2009; Wolfe, Moulder, & Myford, 2001). Corrected scores are available both in the logit metric (as "fair" scores accounting for differences in rater severity), and in the observed score metric (calculated by estimating the expected observed score conditional on item difficulty and rater severity). The details of the MFRM are provided earlier in this chapter.

Studies using the imputation approach have utilized the EM algorithm (Beale & Little, 1975; Dempster, Laird, & Rubin, 1977) to generate missing data (i.e., values of the missing ratings) based on values of the known and measured variables for raters and examinees (Houston, Raymond, & Svec, 1991). For these studies, imputation assumed an ignorable missing data pattern (discussed in greater detail in Section I.C). Imputation was conducted in the original raw score metric, with a similar model to that described above for GT analysis.

The model proposed in this study will draw on all three methods (regression, MFRM, and imputation) in combination with Bayesian inference to adjust for error due to raters. The proposed method will expand upon the previous research by imputing ratings using a more modern Markov Chain Monte Carlo approach for iterative multiple imputation (Gelman, Carlin,

Stern, & Rubin, 2009). Furthermore, the method will incorporate a latent trait model (which accounts for the ordinal outcome), will be designed to handle single-rater designs, and will not require the assumption of an ignorable missing data mechanism.

C. Methods for Categorizing and Analyzing Missing Data

The missing data pattern considered in this thesis is referred to as univariate non-response (Little & Rubin, 2002). That is, the design matrix is known and complete for all cases (one knows all possible combinations of items/tasks, examinees, and raters), but the outcome variable (ordinal ratings) can be missing. In the case of judge ratings of examinees, this missingness is considered planned missingness (i.e., not all judges were assigned to rate all examinees by design). In statistical data analysis, the pattern of missing data (or missingness) is often related to the data itself. Rubin (1976) proposed a framework for characterizing the relationship between the data and the pattern of missingness. I will limit the definition of each typology to the univariate non-response scenario relevant to this thesis. Missing data distributions are classified into two categories: missing at random (MAR) and missing not at random (MNAR). A third category, referred to as missing completely at random (MCAR) is a special case of MAR.

Let $\mathbf{y} = (y_1, \dots, y_n)^T$ represent the vector of ordinal judge ratings for a set of examinees to a set of items or tasks, where $y_i \in \{0,1, \dots, c\}$. To allow for missing data (e.g., planned missingness of rater responses), let this column vector of ordinal responses be comprised of both the observed data elements y_i^{obs} and the missing data elements y_i^{mis}. Furthermore, let $\mathbf{X} = (\mathbf{x}_i^T)_{n \times p}$ represent the design matrix consisting of n row vectors which describe the combination of examinee, item or task, and rater which produced the ordinal (observed or missing) response y_i. Then, the data are considered missing completely at random (MCAR) if

$$\Pr(m_i = 1|\mathbf{y},\mathbf{X}) = \Pr(m_i = 1) \quad (14)$$

where $m_i = 1$ when y_i is unobserved, and 0 otherwise.

For MCAR data, the probability of missingness is constant and not related to the variables in the design matrix. Stated another way, the missing data pattern is independent of both the observed data and the unobserved data. MCAR data is rare in practice, but a standard example is that of matrix sampling. Matrix sampling is a design of administering items to test takers where each test taker takes only a subset of the total items that are being administered. If items are randomly assigned to individual examinees, then the missing item responses can be considered MCAR. This situation is analogous to the assignment of raters to individual examinees (when that assignment is conducted randomly).

Data are considered missing at random (MAR) if

$$\Pr(m_i = 1|\mathbf{y},\mathbf{X}) = \Pr(m_i = 1|\mathbf{X}) \quad (15)$$

That is, the pattern of missing responses can depend on the observed data, but not on the unobserved data. An example of MAR data would be the case where responses were missing at a higher rate for girls than boys on a student survey, but missingness was not related to the unmeasured response itself. Together, the categories of MAR and MCAR are often referred to as ignorable non-response (Schafer & Graham, 2002).

The final missing data distribution pattern is missing not at random (MNAR), and occurs when missingness depends not only on the observed data, but also on the missing data themselves. MNAR data occur when there remains a residual relationship between the missing

data and the pattern of missingness after conditioning out the relationship between the observed data and missingness. An example of an MNAR missingness pattern would be missing outcome measures for patients in a longitudinal drug treatment study. If patients leave the study (and therefore stop being measured on the outcome) due to improvement or decline in their condition, the missing data pattern is related explicitly to the values of the missing data itself. Schafer & Graham (2002) provide a visual depiction of the difference between MCAR, MAR, and MNAR. This visual depiction has been replicated and modified to accurately reflect the data collection scenario described here. In Figure 1, **X**, **y**, and **m** are defined as above, and **Z** represents causes of missingness unrelated to **X**, and **y**. Lines connecting the different elements indicate a relationship between the two elements.

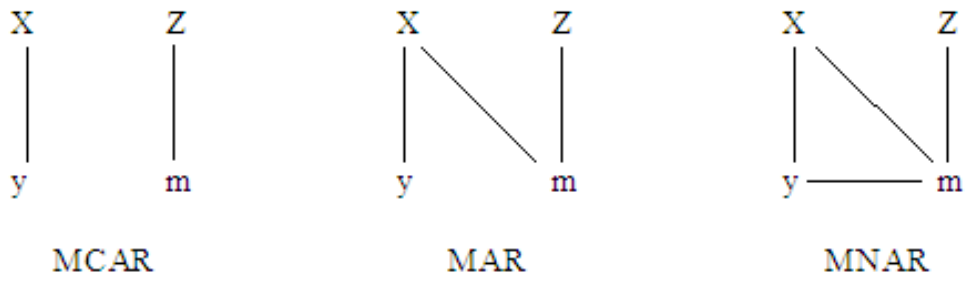

Figure 1. MCAR, MAR, and MNAR.

1. Missing Data Patterns in Judge Rated Data

While in many instances, missing data occurs for a variety of factors outside the control of the investigators, for the missing data scenario discussed here, data are missing by design of the study (i.e., planned missingness). Planned missingness is quite routine in statistical analysis, especially in experimental research. However, when the data collection method does not permit the estimation of variance components or fixed-effect parameters (for the MFRM), it becomes difficult to examine the extent to which person scores vary (and contain bias) as a result of the various facets of measurement. Block designs (Oehlert, 2000) are an example of how missing data can be planned and incorporated into a study. It will be useful for this thesis to describe three block designs and how they relate to data collection scenarios pertinent to the analysis of rater effects. The three designs are a complete block design, a balanced incomplete block design, and a partially balanced incomplete block design.

The ideal data collection case for GT or MFRM is a completely-crossed design (or complete block design). This design is "ideal" in the sense that it provides the most information possible about the facets, their interactions, and the variance attributable to them; however, this design is not ideal in the sense that it represents a very costly data collection plan. A complete block design is a data collection procedure which matches all levels of the treatment (or manipulated variables) within each block (grouping of outcome units). With respect to judge rated data, a complete block design simply describes the case where each judge rates all examinees on all items or tasks – that is, the relative effect of each judge (in relation to all other judges) can be calculated for each block (in this case, an individual examinee). For a small number of judges and examinees, this design may be feasible. However, as sample size increases, it can be especially costly to implement. Figure 2 shows a completely crossed design.

For this example, assume that three raters were used to rate three examinees and that each examinee was rated by all three raters.

	Rater 1	Rater 2	Rater 3
Examinee 1	X	X	X
Examinee 2	X	X	X
Examinee 3	X	X	X

Figure 2. A fully crossed design (complete block)

The data collection design shown in Figure 2 allows for the estimation of the variance due to the examinees, the raters, and the interaction among raters and examinees (or parameters for the main effects in the MFRM case). While complete, and beneficial for estimating all effects (including interactions), this design can be quite costly to implement.

While the complete block design represents the ideal for examining the reliability of scores, less complete data designs can still yield valuable information. A balanced incomplete block design matches all levels of treatment together an equal number of times, but not all treatments are applied to each block. An example of a balanced incomplete block design for judge rated data would be the following: each examinee is rated by two judges across all items or tasks. Each possible pairing of two judges appears in the data at least once, and all pairs of raters occur an equal number of times in the data set. Figure 3 shows a reduced data collection design through which the MFRM is capable of estimating measures for rater severity. For GT, rater and examinee effects could be estimated, but no interaction effect could be examined.

	Rater 1	Rater 2	Rater 3	Rater 4
Examinee 1	X	X		
Examinee 2	X		X	
Examinee 3	X			X
Examinee 4		X	X	
Examinee 5		X		X
Examinee 6			X	X

Figure 3. A balanced incomplete block design.

Finally, a partially balanced incomplete block design would not require that all judges pair together an equal number of times (or at all), only that there exists a pathway in the data to place all rater effects onto a similar metric (i.e., no disjoint subsets of raters). Figure 4 shows a partially balanced incomplete block design where each examinee is judged by two raters, but not all raters are paired with one another (e.g., rater 2 and rater 4 do not rate a common examinee). Despite the fact that not all rater pairs exist in the data, rater severity effects can still be calculated as long as no disjoint subsets of raters exist in the data (Bayesian approaches that specify prior distributions for rater severity can provide inference even in the case of disjoint rater subsets). The MFRM can also be implemented with disjoint subsets as long as some assumption is made about the average ability of the groups of examinees corresponding to each rater (or group of raters).

	Rater 1	Rater 2	Rater 3	Rater 4
Examinee 1	X	X		
Examinee 2		X	X	
Examinee 3			X	X
Examinee 4	X			X

Figure 4. A partially balanced incomplete block design.

Figure 5 shows a data collection design that utilizes only a single rater per examinee. With this data collection design, raters cannot be directly compared to one another as no two raters have provided a score for the same examinee. That is, examinee measures are completely confounded with rater severity. However, in a Bayesian framework, the specification of a prior distribution for the rater severities will allow for the generation of posterior distributions for each rater's severity parameter (despite the presence of disjoint rater subsets – i.e., one disjoint subset of data for each rater in the sample).

	Rater 1	Rater 2	Rater 3	Rater 4
Examinee 1	X			
Examinee 2		X		
Examinee 3			X	
Examinee 4				X

Figure 5. A single-rater (uncrossed/disjointed rater subsets) design.

2. Approaches to Analyzing Missing Data

Over the years, many approaches to dealing with the missing data problem have been developed and tested. This section will highlight a few of the major methods that have been popular at times. Methods range from simple deletion methods, to averaging of items (in the case of surveys or other psychometric measures), to single imputation methods (such as mean substitution, regression based substitution, and imputing from a conditional distribution), to multiple imputation. Each of these approaches will be briefly described, with a discussion of their assumptions and limitations.

Simple deletion methods are among the oldest of missing data analysis approaches. Case deletion (also referred to as list wise deletion) simply deletes all cases (rows of data) with missing observations for one or more variables. This approach is the default setting for many statistical software packages (i.e., when performing an analysis, the software automatically deletes cases with missing observations). Available case deletion (also referred to as pair-wise deletion) is different from case deletion in that it uses different subsets of the data for estimating different parameters. These deletion methods are generally valid only under MCAR (Schafer & Graham, 2002). When data are not MCAR, estimates can be biased; and, even in the situation where MCAR holds, case deletion is quite inefficient (as much of the sample may be discarded during the deletion process). If one conceptualizes the ratings for judges not assigned to rate an examinee as missing data, the traditional approach to analyzing these data is analogous to case deletion. That is, one analyzes the ratings from the judges who rated the examinees, and ignores the other judges. For this scenario, estimates of the latent trait measures are now dependent on the particular combination of judges that rated each individual in the given data set (i.e., the estimates are sample dependent). However, case-deletion can work in a regression (or

psychometric) modeling scenario, when the dependent variable data are only missing at random (MAR) – that is, the missingness of the response variable (the ratings) depend on the values of the observed data, including the observed covariates (Graham & Donaldson, 1993; Ibrahim, Chen, Lipsitz, & Herring, 2005). In particular, when missingness is univariate missingness (or dependent variable missingness, with complete covariate information), complete case methods can yield unbiased estimates. However, in the event that data are missing conditional on the unobserved examinee latent abilities (MNAR), complete case methods are inadequate. The models proposed in this thesis will not rely on the assumption of a MAR missing data mechanism.

For surveys or assessments with item non-response, one approach is to standardize the scores for each item and then use the average item response for missing data values. This approach is motivated by an assumption that all items are exchangeable. This approach may induce bias even under MCAR (Schafer & Graham, 2002), and reduces reliability.

Single imputation is another approach to handling missing data. Single imputation involves replacing missing data with values that are based somehow on the observed data. Methods for calculating the imputed values include (but are not limited to) mean substitution, regression predicted values, and sampling from a conditional distribution. Mean substitution is considered to be quite a poor approach to missing data as it can greatly bias both coefficient estimates as well as the of the standard error estimates of those coefficients. Substituting the mean for missing data will result in a downward biased estimate of the sample variance, while simultaneously inflating the sample size (Little & Rubin, 2002). Regression methods and conditional distribution sampling methods are an improvement on mean substitution, but they

still suffer from under-coverage (i.e., the confidence intervals are too narrow) or understated uncertainty (Rubin, 1987). Regression and conditional methods require the assumption of MAR.

Multiple imputation (Little & Rubin, 2002) is the newest of the major approaches to missing data analysis and is currently the most widely used. In multiple imputation, each missing data point is replaced by multiple imputed values. Each imputed data set is then analyzed separately and the results from each analysis are combined to produce an overall estimate of the parameters and their associated uncertainty. This approach eliminates the problem of understated uncertainty that plague single imputation methods. Multiple imputation relies on Bayesian methods, and therefore shares similarities with a likelihood based approach (another approach to missing data analysis not discussed here). Assumptions about the pattern of missing data matter for MI, with most applications assuming MAR data; however, there have been a few applications of MI when missingness was assumed to be MNAR (Schafer & Graham, 2002).

D. Open Problems in the Analysis of Judge Ratings

This chapter has served as a review of the extensive research that has been conducted on rater effects. Researchers have studied the nature of rater effects, the cognitive causes for those effects, and how contextual factors may explain variation in those effects. Statisticians have developed methods such as GT for estimating the measurement error attributable to raters. That information can then be used to improve the measurement system to improve measure reliability (e.g., by reducing rater variance, or by increasing the number of raters per examinee). Measurement researchers have used correction methods such as the MFRM that adjust scores to account for rater effects, thereby more accurately measuring the underlying latent examinee ability.

In general, existing methods for examining rater effects require two components. First, data collection must align with the needs of different statistical routines. In particular, disjoint subsets in the data and a failure to have more than one rater per examinee can result in non-identification of the model and confounding of variance components, respectively. Second, the methods for analyzing judge-rated data assume a MAR missing data mechanism. The implication is that existing methods make the strong assumption that raters are assigned to examinees in a random fashion and thus after conditioning on design matrix covariates, no residual relationship exists between missingness and examinee ability. However, if raters are not randomly assigned (as is often the case in educational research – e.g., assigning an observer who happens to be available to conduct a given classroom observations), MAR is an assumption that seems unrealistic. Dealing with judge rated data when missing responses are non-ignorable requires the development of a new set of methods.

This chapter has reviewed research from across a wide domain of psychometric and statistical research into rater effects, correction methods, missing data mechanisms, and the treatment of missing data. The methods proposed in this thesis will draw upon one additional statistical tool (Bayesian data analysis using MCMC) to address the open problem. The following section will provide background on the theory and applications of Bayesian inference along with examples of how it has been applied to other psychometric models. The infrastructure of Bayesian inference and Bayesian item response theory will be combined with the topics covered to this point to propose a model which simultaneously handles the problem of rater designs with disjoint subsets and non-ignorable missing data mechanisms.

E. Bayesian Data Analysis and a Proposed Solution

The modeling framework proposed as a solution to the open problem combines Bayesian analysis, missing data approaches, GT, and latent trait modeling. With that in mind, it makes sense to provide an overview of Bayesian inference, its rationale, implementation, and how it has been applied previously to item response theory models. For the purposes of this thesis, the Bayesian approach is desirable over a frequentist approach for several reasons. First, it allows for the researcher to establish prior beliefs about the extent and type of rater bias (along with priors for other model parameters as well). Second, by specifying prior distributions for the rater effects, the Bayesian approach allows one to coherently analyze data that could not be analyzed under a frequentist framework (e.g., data with disjoint subsets of raters like the single rater case). That is, establishing prior distributions for the rater effects provides a mechanism for ensuring the model is fully identified. Third, Bayesian methods allow for coherent and logical decisions in data analysis. In particular, Bayesian inference

> is fundamentally sound, very flexible, produces clear and direct inferences and makes use of all the available information. In contrast, the classical approach suffers from some philosophical flaws, has a restrictive range of inferences with rather indirect meanings and ignores prior information (O'Hagan & Forster, 2004, pp. 16-17).

Bayesian inference is based on Bayes theorem, which can be represented as follows:

$$f(\boldsymbol{\theta}|\mathbf{y}_n) = \frac{f(\mathbf{y}_n \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int f(\mathbf{y}_n \mid \boldsymbol{\theta})\, d\Pi(\boldsymbol{\theta})} \tag{16}$$

where $\mathbf{y}_n = \{y_i\}_{i=1}^{n}$ represents the collected data, $f(\mathbf{y}_n \mid \boldsymbol{\theta})$ is the data likelihood under the assumed model $f(\cdot \mid \boldsymbol{\theta})$ and a given value $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ of the model's parameter, assigned a prior density $\pi(\boldsymbol{\theta})$ corresponding to cumulative distribution function (c.d.f.) $\Pi(\cdot)$. The prior

density reflects pre-experimental beliefs about the plausible values of the parameter $\boldsymbol{\theta}$. Given the data, this prior is updated via Bayes' theorem to a posterior density, $f(\boldsymbol{\theta} \mid \mathbf{y}_n)$, that reflects plausible values of $\boldsymbol{\theta}$ given the data and prior. In essence, the two pieces of information (the prior distribution and the data) both provide information towards our inference about the distribution of the parameter (O'Hagan & Forster, 2004). Moreover, using standard arguments of probability theory, predictions of the model can be made on the basis of the posterior predictive density of a new data point y, given by:

$$f(y \mid \mathbf{y}_n) = \int f(y \mid \boldsymbol{\theta}) d\Pi(\boldsymbol{\theta} \mid \mathbf{y}_n) \tag{17}$$

The results of a Bayesian data analysis are usually summarized by the distributional properties of random samples drawn from the posterior distribution (which can be used to facilitate inference about the posterior predictive distribution). These summaries include statistics such as the distribution quantiles, means and modes, as well as the 95 percent posterior intervals. In some instances – generally single parameter models, or models with conjugate prior distributions – direct simulation from the posterior distribution is possible. However, when the posterior distribution does not have a familiar form, such direct simulation is not possible – or computationally efficient (Gelman, Carlin, Stern, & Rubin, 2009). This is especially true for more complex, multi-parameter problems (such as the one considered by this thesis).

Markov Chain Monte Carlo (MCMC) methods are typically used when it is not possible to directly simulate or derive the full joint posterior distribution of the model parameter, $\boldsymbol{\theta}$. MCMC sampling uses a Markov chain to iteratively draw values from the full conditional

posterior distributions of each of the subcomponents of $\boldsymbol{\theta}$; repeating this process a sufficiently large number of times leads to samples of $\boldsymbol{\theta}$ that converge to the full joint posterior density $f(\boldsymbol{\theta} \mid \mathbf{y}_n)$ of the model. The development and application of MCMC methods has been invaluable to Bayesian data analysis. Standard algorithms for the implementation of MCMC include the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) the Metropolis-Hastings algorithm (Hastings, 1970), and the Gibbs sampler (Gelfand & Smith, 1990).

Bayesian methods have been applied to measurement and item response theory models. Patz and Junker (1999a; 1999b) applied MCMC methods to item response models for both dichotomous and polytomous items using a Metropolis-Hastings within Gibbs sampling approach. Holmes and Held (2006) use a Bayesian auxiliary variable modeling approach to logistic regression that applies to cumulative logit models (e.g., the modified Graded Response Model), such that if multivariate normal priors are assumed for the model parameters, the easier to implement Gibbs sampling approach can be applied.

MCMC has also been applied to generalizability theory and the MFRM. In particular, Briggs and Wilson (2007) propose an approach called generalizability in item response modeling (GIRM) that applies MCMC methods to a random effects measurement model. This modified measurement model uses random item parameters to allow for the calculation of variance components analogous to those used in GT (while simultaneously estimating fixed item parameters typical of item response theory models). The Bayesian modeling approach proposed in this thesis will allow for estimation of variance components and generalizability coefficients by employing summary statistics for the posterior distributions of the model parameters.

3. MCMC, Polytomous Latent Trait Models, and Missing Data

This discussion now returns to the open problem of handling non-ignorable missing judge ratings, potentially in the situation where there are disjoint subsets in the data (due to the lack of a crossed-rater design). The goal of this thesis is to propose a model which accounts for the missing data mechanism (which may be MNAR), imputes values for those missing data within a full MCMC Bayesian framework, calculates rater severity parameters, and produces examinee ability parameters which are free from bias. To accomplish this task, this thesis will evaluate the performance of a bivariate regression model that is introduced briefly here, and then fully described in the methods section of this thesis.

Suppose there is a data set containing judge ratings of individual examinees on a set of tasks or items. Each examinee is rated by at least one judge on each item (although the proposed modeling framework could handle missing item data in addition to the missing rater data proposed here). The judge ratings for each item/examinee combination are ordinal responses. Let $\mathbf{y} = (y_1, \dots, y_n)^T$ represent the vector of ordinal judge ratings for a set of examinees to a set of items or tasks, where $y_i \in \{0,1, \dots, c\}$. To allow for missing data (e.g., planned missingness of rater responses), let this column vector of ordinal responses be comprised of both the observed data elements y_i^{obs} and the missing data elements y_i^{mis}. Furthermore, let $\mathbf{X} = (\mathbf{x}_i^T)_{n \times p}$ represent the design matrix consisting of n row vectors which describe the combination of examinee, item or task, and rater which produced the ordinal (observed or missing) response y_i.

For a complete data set (with imputed values for the missing ratings) denoted as $\{(y_i, \mathbf{x}_i, m_i)\}_{i=1}^n$, the joint likelihood for the general bivariate regression is as follows:

$$\Pr[\mathbf{y},\mathbf{X},\mathbf{m} \mid \boldsymbol{\beta},\boldsymbol{\phi},\sigma_Y,\sigma_m] = \prod_{i=1}^{n} \Pr[y_i \mid \mathbf{x}_i,\boldsymbol{\beta},\mathbf{w},\sigma_Y] Pr[m_i \mid \mathbf{x}_i, y_i,\boldsymbol{\phi},\sigma_m], \quad (18)$$

where the vector $\mathbf{w} = (w_k \mid k = 0, \dots, c;\ w_k \leq w_{k+1};\ w_0 = -\infty, w_c = \infty)$ contains the threshold parameters for the ordinal regression.

For the above model, there are two dependent variables. One dependent variable will be the ordinal ratings of the judges (which may or may not be observed), and the other dependent variable will be a binary indicator variable which denotes the presence or absence of missing ratings. Within a Bayesian MCMC sampling framework, this model will multiply-impute missing judge ratings and produce estimates of the joint posterior distribution of examinee, judge, and item/task parameters. Distributional properties of the posterior distribution can be used to estimate variance components and calculate traditional generalizability theory coefficients. As stated previously, this model can be fully identified under conditions when there are disjoint subsets within the data by establishing prior densities for the person, item, and rater parameters (along with priors for the threshold parameters). Full details of the model, the potential prior distributions, and the MCMC sampling framework are provided in Chapter 2.

F. Conclusions

Accounting for the presence of disjoint subsets, single rater designs, and non-ignorable missing data is an open problem within the research studying rater effects. Methods exist for dealing with rater effects when data has been collected in a highly rigorous manner (random assignment of raters to examinees, multiple raters per examinee). However, when randomization of raters cannot be accomplished due to practical requirements, or multiple raters cannot be assigned due to financial constraints, existing methods prove unable to adequately model the data

in such a manner as to provide unbiased estimates of examinee ability on the underlying latent trait.

This chapter has provided an introduction to the research basis for the methods proposed in this thesis. The statistical methods proposed here grow out of the literature across several lines of research including Generalizability Theory, Rasch measurement, score correction methods, Bayesian inference, and missing data analysis. Iterative multiple imputation procedures combined with the use of Markov Chain Monte Carlo logistic item response models demonstrate potential for dealing with the issue of bias induced into scores by the mishandling of rater effects under measurement scenarios not addressable by existing methods. The proceeding argument has established the rationale for pursuing this line of study given that statistical inference which does not account for the facets of measurement will produce estimates of individual measures on the latent trait that contain hidden, systematic bias above and beyond that due to random error. In the case where data collection allows for the estimation of the variance attributable to these facets and their effect on the reliability of the scores, traditional GT and MFRM that utilizes existing score correction methods may be sufficient. The modeling framework described in this thesis extends previous research to the scenario of single rater designs and non-ignorable missing data mechanisms, while simultaneously incorporating a more coherent Bayesian approach to statistical inference.

The methods introduced briefly in this first chapter are fleshed out in Chapter II of this thesis. Chapter III illustrates the use of the proposed statistical methods in the analysis of simulated data sets (for which generating parameters and rater characteristics are known). The use of simulated data sets will allow for the manipulation of missing data patterns, the number of raters per examinee, the number of items, sample size, and the extent and type of rater effects

present. Chapter IV applies the model to a real data set consisting of Pre-K classroom observations that was collected using a single rater design.

II. METHODS

This chapter describes the Bayesian bivariate probit ordinal missing data model, the algorithm for implementing it, and the research methods used to test the efficacy of this approach versus existing approaches for analyzing judge rated response data. The previous chapter established the open problem this study is addressing – namely, how to account for rater effects in data collection scenarios where either rater effects cannot be estimated, or the raters are assigned in a non-random manner (possibly leading to non-ignorable missing data). As stated previously, methods such as GT and MFRM have been developed to estimate the variance due to rater effects (GT and MFRM), correct for those effects (MFRM and linear regression corrections), and estimate the reliability and generalizability coefficients, along with the standard error of measurement of those scores. However, GT and MFRM do not readily apply to situations where multiple raters have not rated the same respondents in at least some of the observations. With single rater designs, there exist disjoint subsets in the data, which impedes the estimation of rater effects (in MFRM), and confounds the error variance due to raters with the variance due to differences among the examinees (in GT). MFRM can estimate rater effects in such a scenario if one makes a set of assumptions about the average ability of the group of examinees assigned to each rater. Furthermore, this thesis characterizes the problem in a missing data framework. In particular, it treats the responses for examinees not rated by a particular rater as missing data. While these types of missing-data-by-design approaches are common, they are often not analyzed as missing data problems. Without considering the implications of missing data, there is an implicit assumption that data are missing completely at random. However, if raters are not randomly assigned to examinees, the missing data pattern could be either MAR or MNAR. Under those patterns, the typical complete case analysis may not be appropriate

(especially if there is no attempt to correct for rater effects). The methods proposed in this thesis are attempting to reduce the bias due to the non-treatment of rater effects (in single rater designs), and the bias due to mistreatment of missing data.

This thesis has two major research objectives. The first objective is to develop and analyze a method for estimating the variance due to raters (or potentially other facets of measurement) when data do not support the use of traditional methods such as Generalizability Theory. The second objective is to use that method to reduce the bias in the examinee scores produced from the data analysis – accounting for rater effects in situations where rater effects cannot be estimated with traditional methods. To explore these objectives, this study uses a combination of simulated and real data sets. The simulated data sets serve as a basis for determining the efficacy of the new and existing methods when the "truth" is known. That is, simulating data based on "true" values of the underlying latent trait one is trying to estimate enables the evaluation of each approach's ability to return estimates of the parameters that mimic the generating parameters as closely as possible. In addition to simulated data, a real world data set is analyzed with each of the approaches to compare and contrast the results, as well as establish the new method as applicable to actual judge-rated data.

To accomplish these objectives, the Bayesian bivariate probit ordinal missing data model, MFRM, GT (with linear regression correction), and Rasch rating scale models are applied to real and simulated data sets. The simulated data sets are designed to compare these models to one another across a variety of data collection scenarios. To establish the robustness of the various approaches, data sets are considered where the number of raters, examinees, and items are varied. Furthermore, the methods are compared on their ability to handle both normal and non-normal (bimodal mixture normal) distributions of examinee abilities. Two rater designs are

explored (double rater and single rater). Finally, missing data patterns for the double and single rater designs are generated as both MCAR and MNAR.

The new method proposed in this thesis is the Bayesian bivariate probit ordinal missing data model. This model emulates a standard probit model (with variance parameter fixed to approximate a logistic model which is common in psychometric applications). It is contrasted with a typical Rasch rating scale model in that it is defined as a cumulative logit model as opposed to the rating scale model's adjacent categories form (Embretson & Reise, 2000). The cumulative logit model is chosen here because it allows for a more mathematically tractable Bayesian solution, although fitting an adjacent categories model is possible. The formulation of this model (which is described in mathematical detail below) uses a bivariate outcome, modeling the responses and the missing data indicator jointly as dependent variables. The predictor variables include fixed effects for the examinees, items, and raters (and the observed and unobserved ratings for the missing data indicator). The establishment of prior distributions for the independent fixed effect parameters, and the iterative imputation of missing responses identifies the model and enables it to both calculate rater effects, and determine their variance, even with disjoint subsets in the data. The model can be fit via Markov Chain Monte Carlo (MCMC) methods, using a hybrid Gibbs sampler employing Metropolis-Hastings sampling steps. This particular model implementation emulates a Bayesian “selection model” (Little & Rubin, 2002). The identifiability of such a model is key and is addressed after I specify the model.

The remainder of this chapter discusses the proposed new and existing methods, the data sets used for analyzing these methods, the research design, and the evaluation criteria used to judge their effectiveness and applicability. The first major section deals with the formulation and

explanation of the Bayesian bivariate probit ordinal missing data model and the algorithm that can be used to fit the model. Then, the details of the existing comparison methods are explained (i.e., the Rasch rating scale model, MFRM, and GT with the linear regression correction). Following that is a discussion of how variance components will be estimated in the different approaches and how those variance components are used to generate reliability and generalizability coefficients. The rest of the chapter deals with the design of the simulated data, the specifics of the real data sets, the hypotheses that will be examined to determine the extent to which the research objectives are met, and the evaluative criteria that will be used to test those hypotheses.

A. Bayesian Bivariate Probit Ordinal Missing Data Model

This section outlines the theoretical foundations and implementation of a Bayesian bivariate probit ordinal selection model for non-ignorable missing data. In particular, Bayesian likelihood theory is introduced, along with its applications to missing data problems. Following that is a description of the data structure that will support inference with non-ignorable missing ratings. Then, a psychometric model for ordinal responses is developed. Finally, the missing data solution, data structure, and psychometric model are combined to form the full model implemented in this thesis.

As provided earlier, Bayes theorem in its general form can be written as

$$f(\boldsymbol{\theta}|\mathbf{y}_n) = \frac{f(\mathbf{y}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}_n|\boldsymbol{\theta})\, \mathrm{d}\Pi(\boldsymbol{\theta})} \tag{19}$$

To implement Bayes theorem, one must specify both a likelihood for the data, and a prior distribution for the model parameters. Bayesian methods are very useful for the analysis of

missing data problems, and are particularly applicable to the problem addressed in this thesis. I first provide a general overview of how the data likelihood is specified for missing data problems following the guidance of Little and Rubin (2002). I then specify the likelihood used for the model proposed in this thesis by first deriving the psychometric model for the ratings, and then deriving the model for the missing data indicator.

Assume a vector $\mathbf{y}$ of ordinal responses (the dependent variable), and a matrix $\mathbf{X}$ of predictors (independent variables). Furthermore, assume that $\mathbf{X}$ is fully observed, and that $\mathbf{y}$ is only partially observed. Note that this scenario makes sense given that predictor variables in the rater response model described here will be indicator variables for persons, items, and raters (a design matrix) which is fully known. For simplicity, ignore the $\mathbf{X}$ for a description of the general case of missing data analysis. Now, $\mathbf{y}$ can be represented as a set of two vectors $\mathbf{y}_{obs}$ and $\mathbf{y}_{mis}$ which refer to the observed and missing portions of $\mathbf{y}$, respectively.

$$\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis}) \tag{20}$$

Next, assume there exists a vector $\mathbf{m}$ which reflects whether or not the corresponding element of $\mathbf{y}$ is missing or observed. To formulate a Bayesian selection model, represent the joint likelihood of $\mathbf{m}$ and $\mathbf{y}$ as

$$f(\mathbf{y},\mathbf{m}|\boldsymbol{\beta},\boldsymbol{\phi},\sigma_y,\sigma_m) = f(\mathbf{y}|\boldsymbol{\beta},\sigma_y) f(\mathbf{m}|\mathbf{y},\boldsymbol{\phi},\sigma_m) \tag{21}$$

For this data specification, note that the observed data are $(\mathbf{y}_{obs},\mathbf{m})$, with $\mathbf{y}_{mis}$ not observed. It follows that the joint distribution of the observed data is

$$f(\mathbf{y}_{obs},\mathbf{m}|\boldsymbol{\beta},\boldsymbol{\phi},\sigma_y,\sigma_m) = \int f(\mathbf{y}_{obs},\mathbf{y}_{mis}|\boldsymbol{\beta},\sigma_y) f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis},\boldsymbol{\phi},\sigma_m)\, d\boldsymbol{y}_{mis} \tag{22}$$

If the responses in $\mathbf{y}$ are missing at random, then the joint distribution simplifies as follows:

$$f(\mathbf{y}_{obs},\mathrm{m}|\boldsymbol{\beta},\boldsymbol{\phi},\sigma_y,\sigma_m) = f(\mathbf{m}|\mathbf{y}_{obs},\boldsymbol{\phi},\sigma_m) \int f(\mathbf{y}_{obs},\mathbf{y}_{mis}|\boldsymbol{\beta},\sigma_y)\, d\boldsymbol{y}_{mis} \tag{23}$$

$$= \boldsymbol{f}(\mathbf{m}|\mathbf{y}_{\boldsymbol{obs}},\boldsymbol{\phi},\sigma_m)\boldsymbol{f}(\mathbf{y}_{\boldsymbol{obs}}|\boldsymbol{\beta},\sigma_y) \tag{24}$$

This simplification is possible because if data are missing at random, then the distribution of the missing data $\mathbf{m}$ is independent of the missing ratings $\mathbf{y}_{\boldsymbol{mis}}$. Therefore,

$$f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis},\boldsymbol{\phi},\sigma_m) = f(\mathbf{m}|\mathbf{y}_{obs},\boldsymbol{\phi},\sigma_m) \tag{25}$$

and inferences can be based on

$$L(\boldsymbol{\beta}|\mathbf{y}_{obs}) \propto f(\mathbf{y}_{obs}|\boldsymbol{\beta}) \tag{26}$$

However, for the data problems posed in this thesis, the missing data mechanism is not assumed to be ignorable (and is simulated to be otherwise in some cases). Therefore, one must

deal with the likelihood for the complete data, and not just the observed data. For non-ignorable missing data, the above simplification is not possible, and therefore, one bases inference on

$$L(\boldsymbol{\beta},\boldsymbol{\phi}|\mathbf{y}_{obs},\mathbf{m}) \propto f(\mathbf{y}_{obs},\mathbf{m}|\boldsymbol{\beta},\boldsymbol{\phi}) \tag{27}$$

where $f(\mathbf{y}_{obs},\mathbf{m}|\boldsymbol{\beta},\boldsymbol{\phi})$ is defined in Equation 22.

To apply the likelihood theory just described to this study, first, consider the situation where some number of examinees are rated across a number of ordinal response items (assumed here to have a similar rating scale). Let y_{nir} denote the ordinal rating received by examinee n, from rater r, on item i, (where $n = 1, \dots, N$, $i = 1, \dots, I$, and $r = 1, \dots, R$). Each rating y_{nir} is drawn from the sample space $\{c = 1, \dots, C\}$. In practice, each item could have a varying number of response categories, but for simplicity, the model proposed here will assume all items have the same number of possible points. By design (or due to some other reason), the rating y_{nir} could be missing. Let m_{nir} denote a missing rating indicator, defined by

$$m_{nir} = \begin{cases} 1 \text{ if rating } y_{nir} \text{ is missing,} \\ 0 \text{ if rating } y_{nir} \text{ is nonmissing} \end{cases} \tag{28}$$

for all examinees, raters, and test items.

Given that definition of m_{nir}, the rating (observed or unobserved) can be represented as follows:

$$y_{nir} = y_{nir}^{mis} m_{nir} + y_{nir}^{obs}(1 - m_{nir}) \tag{29}$$

Now, define $\mathbf{X}$ as a design matrix, such that

$$\mathbf{X} = (\mathbf{x}_{nir}^{T})_{p\times(N+I+R)} \tag{30}$$

where $p = N \times I \times R$. More specifically, the rows of $\mathbf{X}$ contain values for a set of p covariates corresponding to the examinee, item, and rater for a particular response in $\mathbf{y}$. These covariates are indicator variables taking the value 1 for the column corresponding to examinee n, item i, and rater r, and 0 for all other columns.

Finally, let $\mathbf{Z} = (\mathbf{X},\mathbf{y})$ denote a side-by-side concatenation of $\mathbf{X}$ and $\mathbf{y}$, with rows $\mathbf{z}_{nir}$, such that

$$\mathbf{Z} = (\mathbf{z}_{nir}^{T})_{(p+1)\times(N+I+R+1)} \tag{31}$$

To begin to specify the Bayesian bivariate probit missing data model, one needs to first define the psychometric model used to model the ordinal response. For mathematical and algorithmic tractability, I will use a cumulative probit model. This model specification allows for a simple Gibbs sampling routine. Also, by specifying a scale parameter of 1.6, the probit model can emulate the more traditional cumulative logit model used in psychometric applications.

To develop a model for an ordinal outcome, it is natural to hypothesize that there exists an underlying latent variable measure of the trait of interest (e.g., the ability of the examinee which drives the ratings for that individual). This latent variable is assumed to be drawn from a normal distribution that is centered on a value specific to each response in the data (dependent on

the value of the covariates – i.e., the examinee, item, and rater indicator variables in the design matrix $\mathbf{X}$). Define this latent variable, y^*_{nir}, such that

$$y^*_{nir} \sim \text{Normal}\left(\mathbf{x}^T_{nir}\boldsymbol{\beta}, \sigma_{y^*}\right)\text{I}\left(w_{y_{nir}-1} < y^*_{nir} < w_{y_{nir}}\right) \tag{32}$$

Therefore, the auxiliary latent variable y^*_{nir} is drawn from a truncated normal distribution centered at the linear predictor for the response. To continue to specify the model, it is assumed that the latent metric is divided into categories that can map the latent variable into the ordinal response metric. If one assumes responses can take the values $1,2,\ldots,C$, then the model requires $C+1$ category thresholds w_c, defined such that

$$-\infty = w_0 < w_1 \leq w_2 \leq \cdots \leq w_C = \infty \tag{33}$$

The cumulative probit model defines probabilities such that

$$P\left(y_{nir} = c|\mathbf{X},\boldsymbol{\beta},\sigma_y\right) = P\left(y_{nir} \leq c|\mathbf{X},\boldsymbol{\beta},\sigma_y\right) - P\left(y_{nir} \leq c-1|\mathbf{X},\boldsymbol{\beta},\sigma_y\right) \tag{34}$$

Using the above relationship, the latent variable y^*_{nir}, the threshold w_c, and the normal distribution, I can then write the probability of a response c as

$$P\left(y_{nir} = \boldsymbol{c}|\mathbf{X},\boldsymbol{\beta},\sigma_y\right) = P\left(y^*_{nir} \leq w_c|\mathbf{X},\boldsymbol{\beta},\sigma_{y^*}\right) - P\left(y^*_{nir} \leq w_{c-1}|\mathbf{X},\boldsymbol{\beta},\sigma_{y^*}\right) \tag{35}$$

$$= P\left(w_{c-1} < y^*_{nir} \leq w_c\right) \tag{36}$$

$$= \int_{w_{c-1}}^{w_c} \mathrm{n}\left(y_{nir}^{*} | \mathbf{x}_{nir}^{T}\boldsymbol{\beta}, \sigma_{y^*}\right) dy^{*} \tag{37}$$

where $\mathrm{n}\left(y_{nir}^{*} | \mathbf{x}_{nir}^{T}\boldsymbol{\beta}, \sigma_{y^*}\right)$ is the density for the normal distribution centered at $\mathbf{x}_{nir}^{T}\boldsymbol{\beta}$, with scale parameter σ_{y^*}.

I now turn my attention to specifying a model for $\mathbf{m}$, the vector of missing data indicators. The intent is to formulate a model for the probability that a response y_{nir} is not observed. To account for non-ignorable missing data, the probability that a response is missing is allowed to depend not only on the covariates in $\mathbf{X}$, but also on the observed and unobserved responses themselves (i.e., $\mathbf{y}$). To model the probability (which is bounded by 0 and 1) when one has a binary outcome, a standard choice is a cumulative normal distribution. Similar to how I defined a latent variable y_{nir}^{*} for the response variable y_{nir}, I can define a latent variable, m_{nir}^{*}, such that

$$m_{nir}^{*} \sim \begin{cases} \mathrm{n}(m_{nir}^{*} | \mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*})\mathrm{I}(m_{nir}^{*} \geq 0) \text{ if } m_{nir} = 1 \\ \mathrm{n}(m_{nir}^{*} | \mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*})\mathrm{I}(m_{nir}^{*} < 0) \text{ if } m_{nir} = 0 \end{cases} \tag{38}$$

Then, I can define the probability that a response in $\mathbf{y}$ is missing as

$$P(m_{nir} = 1 | \mathbf{Z}, \boldsymbol{\phi}, \sigma_{m^*}) = \mathrm{N}(m_{nir}^{*} | \mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*})^{m_{nir}}[1 - \mathrm{N}(m_{nir}^{*} | \mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*})]^{1-m_{nir}} \tag{39}$$

where $\mathrm{N}(m_{nir}^{*} | \mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*})$ is the cumulative distribution function for the normal distribution centered at $\mathbf{z}_{nir}^{T}\boldsymbol{\phi}$ with scale parameter σ_{m^*}.

The next step is to write out the likelihood for our data $(\mathbf{y},\mathbf{m},\mathbf{X})$. First, note that the likelihood can be written in simple terms as follows:

$$P(\mathbf{y},\mathbf{m},\mathbf{X}|\boldsymbol{\beta},\boldsymbol{\phi},\sigma_{y^*},\sigma_m) = P(\mathbf{y}|\boldsymbol{\beta},\sigma_{y^*},\mathbf{X})P(\mathbf{m}|\boldsymbol{\phi},\sigma_{m^*},\mathbf{y},\mathbf{X}) \tag{40}$$

If I substitute the models derived above, this likelihood can be written as

$$\begin{aligned} &P(\mathbf{y},\mathbf{m},\mathbf{X}|\boldsymbol{\beta},\boldsymbol{\phi},\sigma_{y^*},\sigma_{m^*}) \\ &= \prod_{n=1}^{N}\prod_{i=1}^{I}\prod_{r}^{R}\left(\int_{w_{c-1}}^{w_c} \mathrm{n}(y^*_{nir}|\mathbf{x}^T_{nir}\boldsymbol{\beta},\sigma_{y^*})\,dy^*\,\{\mathrm{N}(m^*_{nir}|\mathbf{z}^T_{nir}\boldsymbol{\phi},\sigma_{m^*})^{m_{nir}}[1\right. \\ &\left. - \mathrm{N}(m^*_{nir}|\mathbf{z}^T_{nir}\boldsymbol{\phi},\sigma_{m^*})]^{1-m_{nir}}\}\right) \end{aligned} \tag{41}$$

where $\mathrm{n}(y^*_{nir}|\mathbf{x}^T_{nir}\boldsymbol{\beta},\sigma_{y^*})$ is the normal density of the auxiliary latent variable y^*_{nir} underlying the rating y_{nir} for person n on item i by rater r. $\mathrm{N}(m^*_{nir}|\mathbf{z}^T_{nir}\boldsymbol{\phi},\sigma_{m^*})^{m_{nir}}$ represents the normal cumulative distribution function evaluated at the auxiliary latent variable m^*_{nir} underlying the missing data indicator m_{nir}. The models above are defined as probit ordinal regression models. This lies in contrast to the logistic models typically used in psychometric applications (e.g., the Rasch rating scale model or MFRM). However, this model is flexible and can support either a probit or logit interpretation. To fit a bivariate probit model for $(\mathbf{y},\mathbf{m})$, one simply uses the choices of $\sigma_{y^*} = 1$ and $\sigma_{m^*} = 1$ for the variance parameter. The choices $\sigma_{y^*} = 1.6$ and $\sigma_{m^*} = 1.6$ lead to an approximate, bivariate logit model for $(\mathbf{y},\mathbf{m})$.

The next step is to define the joint prior density for $(\boldsymbol{\beta},\boldsymbol{\phi},\mathbf{w})$. The regression parameters are assumed to be multivariate normal and independent from one another, such that

$$P(\boldsymbol{\beta},\boldsymbol{\phi}) = |(2\pi)^{k}\Sigma|^{-\frac{1}{2}}\exp^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})} \tag{42}$$

where

$$\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\phi}) \tag{43}$$

is the vector of regression coefficients for the response vector $\mathbf{y}$ predicted by $\mathbf{X}$ (represented as $\boldsymbol{\beta}$) and the missing data indicator vector $\mathbf{m}$ predicted by $\mathbf{Z}$ (represented by $\boldsymbol{\phi}$). In addition, the prior mean for the parameters is

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_{\beta}, \boldsymbol{\mu}_{\phi}) \tag{44}$$

and the block-diagonal variance covariance matrix is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{V}_{\beta} & 0 \\ 0 & \mathbf{V}_{\phi} \end{bmatrix} \tag{45}$$

Therefore, the joint prior for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ is

$$(\boldsymbol{\beta},\boldsymbol{\phi})|\boldsymbol{\mu}_\beta,\boldsymbol{\mu}_\phi,\mathbf{V}_\beta,\mathbf{V}_\phi \sim \text{Normal}\left(\boldsymbol{\beta},\boldsymbol{\phi}|[\boldsymbol{\mu}_\beta,\boldsymbol{\mu}_\phi],\begin{bmatrix}\mathbf{V}_\beta & 0\\ 0 & \mathbf{V}_\phi\end{bmatrix}\right) \quad (46)$$

The variance parameters for the joint prior for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are modeled based on the hyper-priors

$$v_\beta \sim \text{Unif}(0,1000) \quad (47)$$

$$v_\phi \sim \text{Unif}(0,1000) \quad (48)$$

where $\mathbf{V}_\beta$ and $\mathbf{V}_\phi$ are diagonal matrices with equal values of v_β and v_ϕ on the diagonal (and all other entries as 0), respectively. These priors were chosen to keep these variance finite while not imposing too strong a restriction on their values. In addition, the prior for the thresholds is expressed as

$$P(\mathbf{w}) = \delta_{-\infty}(w_0)\delta_0(w_1)\delta_\infty(w_{C+1})\prod_{c=2}^{C}\text{unif}(w_c|w_{c-1},w_{c+1}) \quad (49)$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_{C+1})$ are the $C+2$ threshold parameters for the category divisions respective to the latent metric. To identify the model, w_0, w_1, and w_{C+1} are set to $-\infty$, 0, and ∞, respectively. Now that the priors are specified, it is possible to specify the full posterior distribution of the parameters as

$$P(\boldsymbol{\beta},\boldsymbol{\phi},\mathbf{w}|\mathbf{y},\mathbf{m},\mathbf{X}) \propto P\left(\mathbf{y}|\boldsymbol{\beta},\sigma_{y^*},\mathbf{X}\right)P(\mathbf{m}|\boldsymbol{\phi},\sigma_{m^*},\mathbf{y},\mathbf{X})P(\boldsymbol{\beta},\boldsymbol{\phi})P(\boldsymbol{w}) \tag{50}$$

$$\propto \prod_{n=1}^{N}\prod_{i=1}^{I}\prod_{r=1}^{R}\left(\int_{w_{y_{nir}-1}}^{w_{y_{nir}}} \mathrm{n}\left(\mathbf{y}^*_{nir}|\mathbf{x}^T_{nir}\boldsymbol{\beta},\sigma_{y^*}\right)dy^*\{\mathrm{N}(\mathbf{m}^*_{nir}|\mathbf{z}^T_{nir}\boldsymbol{\phi},\sigma_{m^*})^{m_{nir}}[1\right.$$

$$\left.-\mathrm{N}(\mathbf{m}^*_{nir}|\mathbf{z}^T_{nir}\boldsymbol{\phi},\sigma_{m^*})]^{1-m_{nir}}\}\right)\times|(2\pi)^k\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}$$

$$\times\,\delta_{-\infty}(w_0)\delta_0(w_1)\delta_{\infty}(w_{C+1})\prod_{c=2}^{C}\mathrm{unif}(w_c|w_{c-1},w_{c+1})$$

This posterior distribution cannot be solved analytically, and therefore requires that I implement a sampling algorithm to determine the distributions of the parameters. The next section of this chapter describes the hybrid Gibbs sampler with Metropolis-Hastings steps used to evaluate this model.

The final definition in the above model specifies a bivariate normal joint prior distribution for $(\boldsymbol{\beta},\boldsymbol{\phi})$. In theory, any mean vector and covariance matrix could be used for the prior specification. Lacking any prior information on examinee ability, rater severity, or item difficulty, one might choose to use a relatively flat prior centered around 0. However, if pre-existing information on the values of these parameters exists, that information could be added to the prior specification. For example, if there exist estimates of rater severity (along with some estimate of the uncertainty of those estimates), the prior distribution for the rater severity parameters need not be 0 for all raters. That is, the specific rater parameters and variances could be used in the prior mean vector and variance-covariance matrix, respectively.

The specification of prior distributions for the regression parameters is of key importance for this model. In particular, without thoughtful definition of the prior distributions, the models are not identifiable. Huang, Chen, and Ibrahim (2005) showed that identification of generalized linear models with non-ignorable missing data is possible when prior distributions for the model parameters are proper (e.g., a normal distribution with finite variance). In the model description that follows, all prior and hyper-prior distributions are proper (either finite variance normal distribution, truncated normal distribution, or bounded uniform distribution).

B. Markov Chain Monte Carlo Estimation Methods

The implementation of this model requires a Markov Chain Monte Carlo procedure implementing a hybrid Gibbs and Metropolis sampling routine. To fit this model, I implement the following iterative algorithm. During each major iteration loop, there are eight steps, each of which samples from the full conditional posterior densities of the parameters of interest.

1. The first step is to sample v_β, the value of the variance parameter for the prior distribution for $\boldsymbol{\beta}$ using random-walk Metropolis sampling. For the first iteration, the value of v_β is set to 1, with all subsequent values drawn as follows: Sample $\log(v_\beta)$ from the normal proposal distribution $\log(v_\beta^*) \sim \text{Normal}(v_\beta^{s-1}, 1)$. I then accept or reject the proposed value with probability

$$\min\left\{1, \exp\left[\sum_{\beta} \log\{n(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \mathbf{V}_\beta^*)\} - \sum_{\beta} \log\{n(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \mathbf{V}_\beta^{s-1})\}\right]\right\} \quad (51)$$

subject to the constraint that the proposed variance parameter is less than 1000.

2. The next step is to sample v_ϕ, the value of the variance parameter for the prior distribution for beta using random-walk Metroplis sampling. For the first iteration, the value of v_ϕ is set to 1, with all subsequent values drawn as follows: Sample $\log(v_\phi)$ from the normal proposal distribution $\log(v_\phi^*) \sim \text{Normal}(v_\phi^{s-1}, 1)$. I then accept or reject the proposed value with probability

$$\min\left\{1, \exp\left[\sum_{\phi} \log\{\text{n}(\boldsymbol{\phi}|\boldsymbol{\mu}_{\boldsymbol{\phi}}, \mathbf{V}_{\boldsymbol{\phi}}^{*})\} - \sum_{\phi} \log\{\text{n}(\boldsymbol{\phi}|\boldsymbol{\mu}_{\boldsymbol{\phi}}, \mathbf{V}_{\boldsymbol{\phi}}^{\mathbf{s-1}})\}\right]\right\} \tag{52}$$

subject to the constraint that the proposed variance parameter is less than 1000.

3. Next, sample y_{nir}^* (the latent variable underlying the ordinal responses) from its conditional posterior density, denoted as

$$\pi(y_{nir}^*|\cdots) \propto \text{n}(y^*|\mathbf{x}_{nir}^T\boldsymbol{\beta}, \sigma_{y^*})\text{I}(w_{y_{nir}} < y^* < w_{y_{nir}+1}) \tag{53}$$

for all i, j, and r, with I(·) as the indicator function. This density is that of a truncated normal distribution and can be sampled using the inverse cumulative distribution function (c.d.f.) method (Devroye, 1986).

4. Next, sample latent variables m_{nir}^* from the conditional posterior density, denoted as

$$\pi(m_{nir}^*|\cdots) \propto \begin{cases} \text{n}(m_{nir}^*|\mathbf{z}_{nir}^T\boldsymbol{\phi}, \sigma_{m^*})\text{I}(m_{nir}^* \geq 0) \text{ if } m_{nir} = 1 \\ \text{n}(m_{nir}^*|\mathbf{z}_{nir}^T\boldsymbol{\phi}, \sigma_{m^*})\text{I}(m_{nir}^* < 0) \text{ if } m_{nir} = 0 \end{cases} \tag{54}$$

for all i, j, and r. This density is also a truncated normal distribution and can be sampled via inverse cdf.

5. Draw values of the threshold parameters from their conditional posterior via random-walk Metropolis sampling, denoted as

$$\pi(w_2, \dots, w_C) \propto \prod_{n=1}^{N}\prod_{i=1}^{I}\sum_{r=1}^{R}\left[\mathrm{N}\left(w_{y_{nir}+1}|\mathbf{x}_{nir}^T\boldsymbol{\beta}, \sigma_{y^*}\right) - \mathrm{N}\left(w_{y_{nir}}|\mathbf{x}_{nir}^T\boldsymbol{\beta}, \sigma_{y^*}\right)\right] \times \mathrm{I}(-\infty \equiv w_0 \leq w_1 \equiv 0 \leq w_2 \leq \cdots \leq w_C < w_{C+1} \equiv \infty) \tag{55}$$

To accomplish this, at each MCMC iteration, I sample a proposal value for the threshold parameter. By definition, $w_0 = -\infty$, $w_1 = 0$, and $w_{C+1} = \infty$. Therefore I only need to update the sample at each iteration for the thresholds w_2 through w_C. Starting with threshold w_2, I draw a proposal w_2^* from a $\mathrm{Normal}\left(w_2^{s-1}, \sigma_{w_2}^2\right)$ distribution and draw a uniform random variable (bounded by 0 and 1). I then use the uniform random variable to accept the proposal w_2^* (set $w_2^s \equiv w_2^*$) with probability

$$\min\left\{1,\exp\left[\sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\log\{\left[\mathrm{N}\left(\mathbf{w}_{y_{nir}+1}^{*}|\mathbf{x}_{nir}^{T}\boldsymbol{\beta},\sigma_{y^{*}}\right)-\mathrm{N}\left(\mathbf{w}_{y_{nir}}^{*}|\mathbf{x}_{nir}^{T}\boldsymbol{\beta},\sigma_{y^{*}}\right)\right]\right.\right. \tag{56}$$

$$\times \mathrm{I}(-\infty\equiv w_0\le w_1\equiv 0\le w_2\le\cdots\le w_C<w_{C+1}\equiv\infty)\}$$

$$-\sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\log\left\{\left[\mathrm{N}\left(\mathbf{w}_{y_{nir}+1}^{(s-1)}|\mathbf{x}_{nir}^{T}\boldsymbol{\beta},\sigma_{y^{*}}\right)-\mathrm{N}\left(\mathbf{w}_{y_{nir}}^{(s-1)}|\mathbf{x}_{nir}^{T}\boldsymbol{\beta},\sigma_{y^{*}}\right)\right]\right.$$

$$\left.\left.\left.\times \mathrm{I}(-\infty\equiv w_0\le w_1\equiv 0\le w_2^{s-1}\le\cdots\le w_C^{s-1}<w_{C+1}\equiv\infty)\right\}\right]\right\}$$

If I fail to accept (or rather, reject) the proposal threshold, I then set $w_2^s\equiv w_2^{s-1}$. This step is then repeated for the remaining threshold values $(w_3,\ldots,w_C)$, within the same MCMC major iteration.

6. Sample the coefficients for the examinee, item, and rater parameters (which are predictive of the rating variable) from the full posterior density, denoted as

$$\pi(\boldsymbol{\beta}|\cdots)=\mathrm{n}\left(\boldsymbol{\beta}|\boldsymbol{\mu}_{\beta}^{*},\sigma_{y^{*}}^{2}\mathbf{V}_{\beta}^{*}\right) \tag{57}$$

With $\mathbf{V}_{\beta}^{*}=\left(\mathbf{V}_{\beta}^{-1}+\mathbf{X}^{T}\mathbf{X}\right)^{-1}$ and $\boldsymbol{\mu}_{\beta}^{*}=\mathbf{V}_{\beta}^{*}\left(\mathbf{V}_{\beta}^{-1}\boldsymbol{\mu}_{\beta}+\mathbf{X}^{T}\mathbf{y}^{*}\right)$. This can be sampled directly from the multivariate normal distribution.

7. Sample the coefficients for the examinee, item, rater, and rating parameters (which are predictive of the missing rating indicator) from the full conditional posterior density, denoted as

$$\pi(\boldsymbol{\phi}|\cdots) = \mathrm{n}\left(\boldsymbol{\phi}|\boldsymbol{\mu}_{\phi}^{*}, \sigma_{m^*}^{2}\mathbf{V}_{\phi}^{*}\right) \tag{58}$$

With $\mathbf{V}_{\phi}^{*} = \left(\mathbf{V}_{\phi}^{-1} + \mathbf{Z}^{T}\mathbf{Z}\right)^{-1}$ and $\boldsymbol{\mu}_{\phi}^{*} = \mathbf{V}_{\phi}^{*}\left(\mathbf{V}_{\phi}^{-1}\boldsymbol{\mu}_{\phi} + \mathbf{Z}^{T}\mathbf{m}^{*}\right)$. This can be sampled directly from the multivariate normal distribution.

8. Finally, sample imputed values for the missing ratings. The conditional posterior distribution is denoted as

$$\pi(y_{nir}^{*}|\cdots) \propto n\left(y_{nir}^{*}|\mathbf{x}_{nir}^{T}\boldsymbol{\beta}, \sigma_{y^*}\right)\mathrm{n}(m_{nir}^{*}|\mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*}) \tag{59}$$

This is accomplished via random walk Metropolis. Since the y_{nir} are conditionally independent, they can be sampled simultaneously. For each missing rating (where $m_{nir} = 1$), draw a proposed $y_{nir}^{*} \sim Normal\left(y_{nir}^{s-1}, \sigma_{y_{nir}^{*}}^{2}\right)$. The values of the non-missing ratings do not change from iteration to iteration. Again, draw a vector of uniform random variables (bounded by 0 and 1) and use that random vector to accept the individual elements of the proposal vector of $y_{nir}^{*}$with probability:

$$\min\left\{1, \frac{n\left(y_{nir}^{*proposal}|\mathbf{x}_{nir}^{T}\boldsymbol{\beta}, \sigma_{y^*}\right)\mathrm{n}(m_{nir}^{*}|\mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*})}{n\left(y_{nir}^{*(s-1)}|\mathbf{x}_{nir}^{T}\boldsymbol{\beta}, \sigma_{y^*}\right)\mathrm{n}(m_{nir}^{*}|\mathbf{z}_{nir}^{T}\boldsymbol{\phi}, \sigma_{m^*})}\right\} \tag{60}$$

If the proposal vector of imputed latent variables (underlying the ratings) is not accepted, the values from the previous iteration are retained. Then, regardless of acceptance or

rejection, the ordinal ratings (where the original data were missing) are updated such that $y_{nir} = k$ if and only if $w_{y_{nir}} < y^*_{nir} < w_{y_{nir}+1}$.

C. Comparison Methods

A number of other methodological approaches are applied to the data sets in addition to the bivariate probit ordinal Bayesian MCMC approach. In particular, data are analyzed using a standard Rasch rating scale model (fit using the WINSTEPS software), a many-facet Rasch model (fit using the FACETS software), and true score theory (i.e., generalizability with a linear regression adjustment). Not all methods are appropriate for all scenarios. For example, GT (using linear regression adjustment) is not appropriate for scenarios with only single rated data sets. With single rater designs, disjoint subsets in the data prevent these models from being fit. In practice, the Rasch rating scale model can be applied to any approach (as it ignores the rater facet entirely). The Bayesian approach is applicable to all situations as the establishment of prior distributions on the parameters allows for the calculation of all model parameters, even when there exists complete separation in the data.

1. The Rasch Rating Scale Model

The Rasch rating scale model (Andrich, 1978; Wright & Masters, 1982) is applicable to data where the substantive meaning of each ordinal category (or rating) is defined to be the same across all items. For example, the rating scale model is often used to analyze survey data with Likert scale type response options (e.g., Strongly disagree, Disagree, Agree, Strongly agree). The rating scale model is formulated as an adjacent categories model for ordinal data with the probability of response dependent on three types of parameters: 1) the ability (or latent trait level) of the object of measurement; 2) the difficulty of the item (typically defined as the average of the category thresholds); and 3) the category thresholds (which define the transition

point from one category to the next). The Rasch rating scale model imposes no restriction on the ordering of these thresholds; they are defined simply as the point along the latent continuum where there is equal probability of being in two adjacent categories. That is, for a four-point rating scale, the threshold dividing categories 1 and 2 is the point on the latent continuum where there is a .5 probability that the respondent would receive a rating of a 1 and a .5 probability that the respondent would receive a rating of 2, conditional on the rating being either a 1 or 2. The specification of the Rasch rating scale model is often denoted as follows:

$$P(X = x|\theta_n, \delta_i, \tau_k) = \frac{\exp \sum_{k=0}^{x} \left(\theta_n - (\delta_i + \tau_k)\right)}{\sum_{c=0}^{C} \exp \sum_{k=0}^{c} \left(\theta_n - (\delta_i + \tau_k)\right)} \quad (61)$$

where θ_n is the ability of person n, δ_i is the difficulty of item i, τ_k is the k-th step threshold, and C is the highest rating one can receive ($x \in \{0,1, \dots, C\}$). The rating scale model is formulated such that the thresholds are the same for all items (similar to how they will be defined in the Bayesian bivariate model and the MFRM). The Rasch partial credit model (Wright & Masters, Rating scale analysis, 1982) is a generalization of the rating scale model that allows the thresholds to vary across items. The partial credit model is not considered in this study, without loss of generalizability. While the rating scale model does not account for the rater facet, it is applied to all single-rater data sets in this study. The rationale for this is that it provides estimates of the latent variables associated with the items and persons, in a modern psychometrics sense (as opposed to true score theory), ignoring the rater variable. In that sense, it will provide a reference for comparing the quality of the parameter estimates and error estimates generated from the more favored approaches of MFRM, GT, and the bivariate Bayesian method.

2. The Many-Facet Rasch Model

The rating scale model described above is typically applied to data where there are not raters (or each respondent is rated by a single rater). It cannot account for variance among the raters. The MFRM is a generalization of the rating scale model which adds additional facets to the linear term in the model. A full description and discussion of the MFRM is provided in Chapter 1. For this study, the MFRM is fit using the FACETS software, which provides fixed effects estimates of person, item, and rater parameters. In addition, it provides estimates of the amount of error variance that is attributable to each facet. The MFRM model is applied to data sets where there are multiple raters that provide ratings for at least some of the individuals. It is not necessary that each respondent is rated by multiple individuals; however, there must be a path through the data in order to generate rater parameter estimates in the same metric. For example, if rater A and rater B rate student 1 and rater B and rater C rate student 2, one can connect raters A and C through their mutual co-rating with rater B. The MFRM model is also applied to the single rater data sub sets. Rater severity estimates can be calculated by the software via group mean ability anchoring (e.g., setting the average ability for each disjoint subset of examinees equal to 0).

3. Generalizability Theory and Linear Regression Adjustment

The three data collection scenarios (rated by all raters, rated by two raters, and rated by one rater) fall into two generalizability theory frameworks. For GT analysis, the first rating scenario exhibits the properties of a balanced, complete, crossed design. The second and third rating scenarios are nested designs, with persons nested in raters, crossed with items. I provide details on the fully-crossed, balanced design below.

When individuals are rated by all raters, the GT design is a fully crossed, balanced, complete $p \times i \times r$ design. For the fully crossed design, main effects can be calculated for persons, items, and raters. In addition, effects can also be calculated for all two-way and three-way interactions; however, the three-way interaction of persons, items, and raters is confounded with the residual error variance. The GT model for the fully crossed design is shown in Table II:

TABLE II

A FULLY CROSSED TWO-FACET GENERALIZABILITY THEORY MODEL

Residual Representation	Facet Effect Representation
$\boldsymbol{X_{pir} = \mu}$	$X_{pir} = \mu$
$+\boldsymbol{(\mu_p - \mu)}$	$+\nu_p$
$+\boldsymbol{(\mu_i - \mu)}$	$+\nu_i$
$+\boldsymbol{(\mu_r - \mu)}$	$+\nu_r$
$+\boldsymbol{(\mu_{pi} - \mu_p - \mu_i + \mu)}$	$+\nu_{pi}$
$+\boldsymbol{(\mu_{pr} - \mu_p - \mu_r + \mu)}$	$+\nu_{pr}$
$+\boldsymbol{(\mu_{ir} - \mu_i - \mu_r + \mu)}$	$+\nu_{ir}$
$+\boldsymbol{(X_{pir} - \mu_{pi} - \mu_{pr} - \mu_{ir} + \mu_p + \mu_i + \mu_r + \mu)}$	$+\nu_{pir}$

This additive model specification defines how each rating X_{pir} can be written as a sum of main effects and interaction term deviations from an overall mean μ. Going further, one can define variance components for each of the main effects and interactions as follows:

$$\sigma^2(p) = [MS(p) - MS(pi) - MS(pr) + MS(pir)]/n_i n_r \quad (62)$$

$$\sigma^2(i) = [MS(i) - MS(pi) - MS(ir) + MS(pir)]/n_p n_r \quad (63)$$

$$\sigma^2(r) = [MS(r) - MS(pr) - MS(ir) + MS(pir)]/n_p n_i \quad (64)$$

$$\sigma^2(pi) = [MS(pi) - MS(pir)]/n_r \quad (65)$$

$$\sigma^2(pr) = [MS(pr) - MS(pir)]/n_i \quad (66)$$

$$\sigma^2(ir) = [MS(ir) - MS(pir)]/n_p \quad (67)$$

$$\sigma^2(pir) = MS(pir) \quad (68)$$

where, for example, $MS(p)$ represents the mean square statistic for the main effect associated with persons. The above equations can be rewritten as follows:

$$MS(p) = \sigma^2(pir) + n_i\sigma^2(pr) + n_r\sigma^2(pi) + n_i n_r\sigma^2(p) \quad (69)$$

$$MS(i) = \sigma^2(pir) + n_p\sigma^2(ir) + n_r\sigma^2(pi) + n_p n_r\sigma^2(i) \quad (70)$$

$$MS(r) = \sigma^2(pir) + n_p\sigma^2(ir) + n_i\sigma^2(pr) + n_p n_i\sigma^2(r) \quad (71)$$

$$MS(pi) = \sigma^2(pir) + n_r\sigma^2(pi) \quad (72)$$

$$MS(pr) = \sigma^2(pir) + i\sigma^2(pr) \quad (73)$$

$$MS(ir) = \sigma^2(pir) + n_p\sigma^2(ir) \quad (74)$$

$$MS(pir) = \sigma^2(pir) \quad (75)$$

Starting with an estimate of $MS(pir)$, the system of equations can be solved starting with the bottom equation and moving upwards. The mean square statistic is defined as the sum of squares divided by the degrees of freedom. Table III shows the necessary calculations for

generating the sum of squares and the subsequent mean squares that can be used in these equations (note that $T(\mu) = n_p n_i n_r \bar{X}^2$).

TABLE III

DEGREES OF FREEDOM AND SUMS OF SQUARES FOR A FULLY-CROSSED DESIGN.

$\boldsymbol{\alpha}$	$\boldsymbol{df(\alpha)}$	$\boldsymbol{T(\alpha)}$	$\boldsymbol{SS(\alpha)}$
$\boldsymbol{p}$	$n_p - 1$	$n_i n_r \sum \bar{X}_p^2$	$T(p) - T(\pi)$
$\boldsymbol{i}$	$n_i - 1$	$n_p n_r \sum \bar{X}_i^2$	$T(i) - T(\pi)$
$\boldsymbol{r}$	$n_r - 1$	$n_p n_i \sum \bar{X}_r^2$	$T(r) - T(\pi)$
$\boldsymbol{pi}$	$(n_p - 1)(n_i - 1)$	$n_r \sum\sum \bar{X}_{pi}^2$	$T(pi) - T(p) - T(i) + T(\mu)$
$\boldsymbol{pr}$	$(n_p - 1)(n_r - 1)$	$n_i \sum\sum \bar{X}_{pr}^2$	$T(pr) - T(p) - T(r) + T(\mu)$
$\boldsymbol{ir}$	$(n_i - 1)(n_r - 1)$	$n_p \sum\sum \bar{X}_{ir}^2$	$T(ir) - T(i) - T(r) + T(\mu)$
$\boldsymbol{pir}$	$(n_p - 1)(n_i - 1)(n_r - 1)$	$\sum\sum\sum X_{pir}^2$	$T(pir) - T(pi) - T(pr) - T(ir) + T(p) + T(i) + T(r) - T(\mu)$

Using the above systems of equations for the $p \times i \times r$ design, it is a simple matter to calculate the variance components associated with each of the main effects and interactions. These variance components can then be used to calculate a generalizability coefficient (analogous to reliability) for the data collection scenario. The generalizability coefficient (a measure of reliability for relative or normative decisions) is formulated as

$$E\rho^2 = \frac{E_p(\mu_p - \mu)^2}{E_p E_I E_R(X_{pIR} - \mu_{IR})^2} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \tag{76}$$

where δ indexes all facets that do not include p (unless some of the facets are fixed). The term σ_δ^2 denotes the relative error variance attributable to the facets of measurement (in our case, items and raters). The above G-theory framework will be implemented for the simulated, complete data sets (all raters rate all examinees), and the imputed data sets generated at each iteration of the Bayesian bivariate probit ordinal regression model. That is, for each iteration of the MCMC algorithm, there will be a set of observed and imputed ratings for all combinations of persons, items, and raters. The variance components can be calculated for each of these data sets (each iteration), and then the distribution of the variance components can be described similar to the other parameters estimated in the MCMC process.

In addition to the GT model applied to the raw scores for each item/rating, I fit a linear regression model that adjusts for rater leniency/severity (Wilson, 1988). This model is similar conceptually to the MFRM, but does not use logistic regression. This method is applied to demonstrate a traditional (non-IRT based) approach to correcting for rater effects in the observed score metric. The model for the linear regression correction can be formulated as follows:

$$y_{nir} = \alpha_n + \beta_r + \delta_i + \varepsilon_{nir} \tag{77}$$

where y_{nir} is the observed score for examinee n rated by rater r on item i, α_n is the true score for examinee n, β_r is the leniency (or scoring bias) for rater r, δ_i is the easiness for item i and ε_{nir} is the random error term. Note that positive values for the leniency parameter indicate raters

who tended to give more favorable ratings on average (which is the opposite relationship seen in the MFRM, where positive parameter values indicate severity). If one assumes that there are $p = p_n + p_i + p_r$ parameters to be estimated, the model can be rewritten in matrix algebra as follows:

$$\mathbf{y} = \mathbf{X}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\\ \boldsymbol{\delta}\end{bmatrix} + \boldsymbol{\varepsilon} \tag{78}$$

where $\mathbf{y}$ is a vector containing the observed ratings, $\mathbf{X}$ is a design matrix with column vectors representing the examinees, judges, and items, $\boldsymbol{\alpha}$ is the vector of examinee abilities, $\boldsymbol{\beta}$ is the vector of rater leniency, and $\boldsymbol{\delta}$ is the vector of item easiness. But, under this definition, the matrix $\mathbf{X}$ is not full rank, and therefore has no inverse (which will impede OLS estimation). To identify the model, assume that $\sum \beta_r = 0$ and $\sum \delta_i = 0$. These assumptions are the equivalent of dropping a rater and item column vector from the design matrix, which for convenience will be the last rater and item. With these final assumptions, I can estimate our true score, item easiness, and leniency parameters using ordinary least squares (OLS) as follows:

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\\ \mathbf{d}\end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{79}$$

where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{d}$ are the estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$, respectively. The parameter estimate for the final rater can be computed as follows:

$$\beta_{p_r} = \sum_{k<p_r} -\hat{\beta}_k \tag{80}$$

Furthermore, the covariance matrix for the parameters is calculated as

$$s^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{81}$$

where

$$s^2 = \mathbf{y}^T\mathbf{y} - [\mathbf{a}^T\mathbf{b}^T\mathbf{d}^T]\mathbf{X}^T\mathbf{y}/(p_n p_i p_r - p_n - p_i - p_r + 2) \tag{82}$$

and the variance for the last rater parameter is

$$s_{p_r p_r} = \sum_{n=1}^{p_n} \sum_{i=1}^{p_i-1} \sum_{r=1}^{p_r-1} s_{nir} \tag{83}$$

Wilson (1988) goes beyond this basic OLS algorithm to use information on rater consistency to refine his model (employing consistency estimates as weights in a generalized least squares model). However, as the data for this study are simulated such that there is no systematic inconsistency across raters, this step is omitted for this method.

D. Study Design Parameters

The majority of the study utilizes simulated data sets to answer a variety of empirical questions and hypotheses. As an overarching theme, the study addresses the effectiveness of each of the methods in handling different data scenarios. A total of four methods are applied to

each simulated data set (with the exception of methods that don't apply to particular cases). The four methods are the true score theory/GT approach using a linear regression correction, the Rasch rating scale model fit with WINSTEPS, the MFRM fit with FACETS, and the bivariate Bayesian model (fit using MATLAB; the code is provided in the appendix). These methods are compared with one another on how well they handle different rating scenarios, sample sizes (items, raters, and examinees), data distributions, and different missing data patterns. In particular, three rating designs are considered. Under the first design, all examinees are rated by all raters on all items. This design will generally be referred to as the complete data case. Under the second design, each examinee is rated by two raters, with all other ratings from other raters treated as missing data. The third design utilizes a single rater approach, with each examinee being rated on the items by a single rater, and all of the ratings from the remaining raters treated as missing data. This study is proposing methods for analyzing single and multi-rater data as a missing data problem. To that end, each simulated data set is generated with complete data (all ratings from all raters on all examinees). In examining the second and third rating design, ratings are deleted under different missing data scenarios (MCAR and MNAR), for the purpose of examining the impact of ignorable and non-ignorable missingness.

A total of four main data sets are simulated to examine the sensitivity of model comparisons to various aspects of the data collection scenario. These aspects include the number of raters (5 versus 10), the number of items (5 versus 20 four-point rating scale items), the number of examinees (50 versus 200), and the distribution of examinee abilities (normal versus bimodal). Rather than examine the factorial combination of all approaches (along with the four missing data scenarios for each), I simulated data sets to address specific comparisons. The four data sets are as follows:

1. 50 examinees, 5 raters, 5 items, normal distribution of true abilities
2. 50 examinees, 5 raters, 20 items, normal distribution of true abilities
3. 200 examinees, 10 raters, 5 items, normal distribution of true abilities
4. 200 examinees, 10 raters, 5 items, bimodal distribution of true abilities

These four data sets were strategically chosen to minimize the number of data sets while maximizing the number of comparisons across the aspects of data collection. In addition, during the course of the thesis work, the model was adjusted to model the variance term of the prior distributions for the model parameters (which eliminated the need for some of the original comparisons proposed as part of this thesis). The comparisons of note are as follows:

- Data set 1 versus data set 2: The impact of more items on the estimates of examinee abilities.
- Data sets 1/2 versus data sets 3/4: The impact of more raters on the estimates of examinee abilities.
- Data set 3 versus data set 4: The sensitivity of the different approaches to non-normal ability distributions

All response data used are based on simulated true parameter values and the MFRM (to generate data based on person ability, item difficulty, step thresholds, and rater severity parameters). The initial simulated data sets serve as the data for the complete case analysis. The single and double rater designs have data deleted (response data, not design parameters) in accordance with either an MCAR or MNAR missing data pattern. The MCAR data are created by randomly deleting the ratings for the items from all but one or two judges for each examinee, depending on the rating design. The judges that were kept for each examinee were selected via simple random sampling within examinee. Simulating the MNAR data from the complete data is

slightly more complicated. As described in the first chapter, missing data are considered MNAR when the missingness is related to the values of the missing data themselves. In our case (judge rated data), data will be MNAR when there is a systematic relationship between a rating being missing, and the unmeasured value of that rating. When working with real data, one can never know if missing data are MNAR (except in rare circumstances – e.g., the data collection process is set up to be MNAR). In generating the MNAR data from the complete simulated data, ratings from more severe judges were kept with higher probability than for more lenient judges. The probability of missingness for a particular judges ratings for a given individual was based on the value of the ratings received for that judge. Higher ratings were more likely to be missing, which results in more ratings from more severe raters in the data set. The missing ratings for some individuals are more likely to be their higher ratings (given that lenient judge ratings are more likely to be missing). This induces bias into the ratings, especially for individuals who are "unlucky" in that they only receive ratings from more severe raters. Each of the four initial data sets yielded five analytic data sets: 1) the complete case; 2) single rater MCAR (1MCAR); 3) single rater MNAR (1MNAR); 4) double rater MCAR (2MCAR); and 5) double rater MNAR (2MNAR). Not all data collection scenarios are analyzed with each method (as not all methods are appropriate for each scenario – e.g., the Rasch rating scale model will not produce a single ability for each examinee when there are multiple raters). The different methods are be applied to the data collection scenarios as follows:

- The Rasch rating scale model (WINSTEPS): 1MCAR, and 1MNAR
- The Many-Facet Rasch Model (FACETS): 1MCAR, 1MNAR, 2MCAR, and 2MNAR
- The Bayesian bivariate probit model (MCMC): 1MCAR, 1MNAR, 2MCAR, and 2MNAR

- Generalizability Theory: Complete Case
- Linear Regression Adjustment: 2MCAR, and 2MNAR

In addition to the simulated data sets, a real data set of classroom observation data were analyzed to test the applicability and effectiveness of the four methods under real world conditions. This data set includes observations of 255 early childhood classrooms using the CLASS Pre-K observation protocol. For this data set, raters were not randomly assigned, although there was no systematic method for assignment either; rather, raters were assigned based on availability and scheduling issues. For the CLASS data, only a single rater rated each classroom. The CLASS Pre-K instrument includes 10 items scored with a seven point rating scale (1-7) measuring the following concepts:

- Positive climate
- Negative climate
- Teacher sensitivity
- Regard for student perspectives
- Behavior management
- Productivity
- Instructional learning formats
- Concept development
- Quality of feedback
- Language modeling

The CLASS data are used to demonstrate the potential for the Bayesian bivariate probit ordinal missing data model to be applied to real life judge rated data. For these data, the

observation scores from the Bayesian model are compared to those obtained using the Rasch rating scale model, and MFRM.

1. Evaluation Criteria

The purpose of this thesis is to test a new method against well-known traditional approaches. To facilitate this comparison, I established a set of a priori evaluation criteria for comparing the results across models. In particular, the models are evaluated based on the magnitude of the standard errors of measurement, parameter recovery bias, the reliability of the person estimates, consistency of the rank ordering of the persons, consistency of the distribution of scores, consistency of the determination of statistical significance, and the coverage probabilities of the confidence intervals based on the standard errors of measurement. Each of these criteria are described in further detail here.

a. Magnitude of the Standard Errors of the Person Parameters

Any psychometric model that produces a parameter estimate (such as the examinee ability estimate) typically also estimates the standard error of measurement. The concept of the standard error of measurement was developed out of true score theory based on the idea that an observed score for an individual varies from their true score due to measurement error. Classical test theory approaches traditionally calculate a single standard error of measurement (SEM) based on the reliability estimate of the scores, which is applied to all individuals, regardless of their score. More modern psychometric approaches such as item response theory and Rasch measurement produce a conditional standard error of measurement for each examinee. These SEMs vary with the score of the individual, conditional on their score on the measurement instrument. Scores near the middle of the distribution typically have lower SEMs than scores near the tails of the distribution. The methods applied to the different data sets

will be compared on the size and accuracy of their standard errors of measurement. For the MFRM and Rasch rating scale model, the standard conditional SEM is calculated. The Bayesian bivariate probit ordinal missing data model produces posterior intervals which are used to calculate the standard error of measurement for each examinee ability. The estimate of the SEM for the Bayesian approach is the standard deviation of the posterior distribution for each of the examinee parameters. For the CTT/GT approach using the linear regression correction, the standard errors for the coefficients of the examinee terms are used as the SEMs.

b. Bias in Parameter Recovery

For the models which estimate examinee ability parameters in the latent trait metric (Rasch rating scale model, MFRM, and Bayesian bivariate model), it is important to look at the extent to which the model recovers the generating parameters. For the simulated data sets, the MFRM was used to generate ratings based on known "true" ability parameters. The estimates of these ability parameters from these three models are compared to the true abilities by calculating the root mean square error (RMSE). The root mean square error is simply the square root of the mean square error, which can be calculated for a single parameter θ as follows:

$$MSE = E\left[\left(\hat{\theta} - \theta\right)^2\right] = \text{var}\left(\hat{\theta}\right) + \text{bias}\left(\hat{\theta}, \theta\right)^2 \tag{84}$$

Thus, the mean square error (and RMSE) can be used to determine the extent to which the parameter estimates are close or far from the true values they are trying to represent.

c. Reliability of Parameter Estimates

Reliability can be defined as the extent to which scores are free from measurement error. In classical test theory, reliability was defined as the ratio of true score

variance to observed score variance – or, the proportion of variance in the scores that was attributable to real differences in the objects of measurement. The different modeling approaches are compared across data sets using a variety of reliability indices. Rasch modeling approaches use a separation reliability index (the ratio of measurement error adjusted variance to observed variance). Generalizability theory calculates a G-coefficient (described previously). Classical test theory typically uses Cronbach's alpha, which is calculated as follows:

$$\alpha = \left(\frac{I}{I-1}\right)\left(\frac{SD_X^2 - \sum SD_i^2}{SD_X^2}\right) \tag{85}$$

where I is the number of test items, SD_X^2 is the squared standard deviation (or variance) of the observed scores (total score), and SD_i^2 is the variance of ratings for item i (Thorndike, 2005).

To compare reliability estimates across models, I propose a common statistic based on the variance of the examinee abilities, and the average size of the standard errors of measurement. The reliability statistic is calculated as follows:

$$\text{reliability} = 1 - \frac{(\overline{SEM})^2}{SD^2} \tag{86}$$

where $\overline{SEM}$ is the average standard error of measurement of the ability estimates and *SD* is the standard deviation of the ability estimates. If the average SEM is as large as the standard deviation of the scores, reliability will be 0. If the SEMS are very small relative to the spread of the scores, the reliability will be high. Note: the average SEM could be larger than the SD; in this scenario, the reliability is set to 0.

d. Rank Ordering of Person Abilities

Pearson correlation coefficients were calculated to examine the extent to which the different approaches are producing similar or different ordering of the examinees. Models that produce ability estimates with more correspondence to the generating ability parameters will exhibit higher correlations, providing evidence of better model performance.

e. Changes to the Distribution of Person Abilities

Correlations can examine the consistency of the measures across the methods at a macro level, but it is also important to examine the consistency of the metrics at the individual examinee level. To accomplish this, the changes to the distribution of the parameter estimates are examined by placing individual examinees into groups based on their percentile rank in the distribution of examinees (given a particular model). Correspondence of the measures to the true parameters is examined by determining the percentage of examinees that stay at the same location within the distribution. For example, what percent of examinees in the bottom decile remain in the bottom decile under a different modeling approach?

f. Prevalence of Statistical Significance

Each of the methods applied to the data sets produces a standard error of measurement (see the discussion above). These standard errors are used to produce 95 percent confidence intervals for the ability estimates for all methods. That is, 95 percent confidence intervals for each ability estimate are created by bounding the estimate by plus or minus 1.96 SEMs. These confidence intervals are then used to look for systematic differences in the methods in the number or percent of examinees who are statistically different from 0 (i.e., the average score in the standardized metric for each model). This analysis is intended to determine if any

methods are over- or under-estimating the precision of the estimates and the extent to which scores truly are different from one another.

g. Confidence Interval Coverage Probability

The confidence intervals defined above can be used to examine the extent to which these intervals accurately represent the uncertainty in the estimates. This is accomplished by looking at the percentage of confidence intervals that contain the true generating parameter value (for the simulated data sets). To be considered accurate, the confidence intervals should contain the true value close to 95 percent of the time on average. True ability parameters, and all model-based estimates, are standardized to a z-score (standard normal) metric to facilitate cross-model comparisons.

2. Research Hypotheses

In developing the proposed model for analyzing judge ratings in the presence of non-ignorable missing data, I generated a set of hypotheses to test as part of the evaluation of the new approach. Five major hypotheses are detailed here and written results speak directly to the findings related to these hypotheses.

a. Hypothesis 1

The person parameter estimates produced by the new method are as close as or closer to the true values than those produced by the existing methods. It is hypothesized that for data scenarios that allow for the calculation of rater effects with traditional approaches (GT and MFRM), the new method will perform similarly well. Under scenarios where no rater effects can be calculated with traditional approaches, the new method will be able to estimate these effects, correct for them, and produce examinee parameter estimates which retain less bias than a

method which does not take these into account (true score theory or the Rasch rating scale model).

b. Hypothesis 2

When data are missing not at random, the new method will produce examinee estimates with less bias than existing methods, for scenarios where existing correction methods are applicable (i.e., double rater designs). Furthermore, for single rater designs, the new method will perform better (less biased estimates) than the Rasch rating scale model or true score theory approaches when missing ratings are MNAR.

c. Hypothesis 3

The standard errors of measurement will be larger (and more accurate estimates of uncertainty) with the new method than with existing methods that either cannot estimate rater variance, or inaccurately estimate it due to non-ignorable missing data. The reliability of the scores from the new method will be lower and will more accurately reflect the uncertainty due to rater variance.

d. Hypothesis 4

The new method will produce a more accurate rank ordering of the examinees than the traditional approaches when compared to the "true" ranking of the persons as determine by the data generating parameters.

e. Hypothesis 5

The new method will produce confidence intervals that have coverage probabilities that are aligned with the definition of the interval.

E. Conclusion

The methods outlined in this chapter are implemented to address two major research objectives: 1) to develop and analyze a method for estimating the variance due to raters when data do not support the use of traditional methods; and 2) to use that method to reduce the bias in the examinee scores produced from an analysis of such data. Two groups of analyses are conducted. The first analysis focuses on simulated data sets with the intent of showing how bias is reduced when data are analyzed to account for error due to raters (as compared to methods which cannot treat non-ignorable missing data). The second analysis demonstrates the proposed approach (a Bayesian bivariate probit ordinal missing data model) on a real data set where individuals were rated by a single rater. To establish the robustness of the various approaches, simulated data sets are considered where the number of raters, examinees, and items are varied. Furthermore, the methods are compared on their ability to handle both normal and non-normal (bimodal mixture normal) distributions of examinee abilities. Two rater designs are explored (double rater and single rater). Finally, missing data patterns for the double and single rater designs are generated as both MCAR and MNAR.

III. SIMULATION STUDY RESULTS

This chapter presents comparative results from the various simulated data sets discussed in the previous chapter. Each data set was analyzed using the approaches described in Chapter II (the polytomous Rasch model, many faceted Rasch model, ordinary least squares regression adjustment model, generalizability theory, and Bayesian bivariate probit ordinal missing data model). Results include general descriptive statistics regarding the data themselves, with comparisons made to determine the accuracy and appropriateness of the various analytic approaches. The results for the application of the methods to real world data (based on observations of classrooms in early childhood centers) are presented in the next chapter.

As described in Chapter II, a total of four master data sets were simulated. These data sets varied in the number of persons, raters, and items (and in one case, the distribution of the person abilities). The four master data sets are:

1. 50 persons (normally distributed abilities), 5 items, and 5 raters (p50i5r5)
2. 50 persons (normally distributed abilities), 20 items, and 5 raters (p50i20r5)
3. 200 persons (normally distributed abilities), 5 items, and 10 raters (p200i5r10)
4. 200 persons (bimodally distributed abilities), 5 items, and 10 raters (p200i5r10BM)

Each of the data sets was then sampled to produce four data subsets based on the number of raters each person was rated by, and the method of sampling the raters. Two single rater data subsets were created and two double rater data sets were created, with the sampling occurring either at random (MCAR) or not at random (MNAR) within each pair. These data collection scenarios (or data subsets) are referred to as 1MCAR, 1MNAR, 2MCAR, and 2MNAR. This

simulation and sampling process resulted in 16 total data sets to be analyzed with the various approaches described in Chapter II.

The MCAR data subsets were simulated by randomly selecting one or two judges for each examinee (for the 1MCAR and 2MCAR data sets, respectively). The MNAR data sets were simulated by selecting judges in a pattern where higher ratings were more likely to be missing – that is, examinees were more likely to receive a severe rater than a lenient rater. Table IV shows the relationship between the observed and unobserved ratings, and the missing data indicator. The table provides the coefficients from 16 different binomial regression models which predicted the binary missing data indicator based on the observed/unobserved rating, fixed effects for items and judges, and random person effects. For parsimony, only the coefficient for the ratings are included in the table. The values in Table IV indicate that there is no relationship between the ratings and missingness for any of the MCAR scenarios (although the relationship is marginally significant for p50i5r5 1MCAR). However, for all of the MNAR data sets, there is a statistically significant relationship between the ratings and the missing data indicator. These results demonstrate that the missing data for the MNAR data sets are non-ignorable (as missingness is dependent on the value of the unobserved ratings, even after conditioning on the items, judges, and examinees.

TABLE IV

BINOMIAL REGRESSION ESTIMATES OF THE RELATIONSHIP BETWEEN RATINGS (OBSERVED AND UNOBSERVED) AND MISSINGNESS

Data Set	Scenario	Coefficient	Std. Err	Sig.
p50i5r5	1MCAR	–0.154	0.091	0.090
	1MNAR	0.460	0.096	0.000
	2MCAR	0.082	0.075	0.271
	2MNAR	0.244	0.086	0.005
p50i20r5	1MCAR	0.025	0.045	0.580
	1MNAR	0.209	0.046	0.000
	2MCAR	–0.008	0.038	0.836
	2MNAR	0.257	0.043	0.000
p200i5r10	1MCAR	0.000	0.042	0.992
	1MNAR	0.596	0.046	0.000
	2MCAR	–0.016	0.032	0.607
	2MNAR	0.417	0.034	0.000
p200i5r10BM	1MCAR	–0.012	0.038	0.760
	1MNAR	0.363	0.039	0.000
	2MCAR	0.006	0.028	0.820
	2MNAR	0.284	0.030	0.000

Table V provides descriptive statistics for each of the 16 data sets, including the mean and standard deviation of the true abilities and observed scores (ordinal ratings) for the persons. Abilities were randomly sample to have mean 0 and standard deviation 1 (except for the bimodal distribution), and the table shows values which mirror that except for random sampling variance. The average observed scores for the MNAR data sets are lower than those for the MCAR data sets due to the method in which the MNAR data were sampled. Ratings from more lenient raters were more likely to be unobserved (i.e., there are more ratings from severe raters in the data), therefore, the average score is lower than in the random case. In addition, the standard deviation of the observed scores in lower for the MNAR data. The standard deviation of scores is also higher for the p200i5r10BM data set than for the other data sets. This difference is due to the p200i5r10BM data set having sampled true abilities from a bimodal mixture distribution (two standard normal distributions centered at –1 and +1).

TABLE V

SIMULATED DATA SET DESCRIPTIVE STATISTICS

		True Ability		Observed Score	
Data Set	Scenario	M	SD	M	SD
p50i5r5	1MCAR	0.1412	0.8383	2.6640	0.8680
p50i5r5	1MNAR	0.1412	0.8383	2.0400	0.4314
p50i5r5	2MCAR	0.1412	0.8383	2.4640	0.5446
p50i5r5	2MNAR	0.1412	0.8383	2.1060	0.4283
p50i20r5	1MCAR	0.1561	1.0258	2.7610	0.7481
p50i20r5	1MNAR	0.1561	1.0258	2.1580	0.2809
p50i20r5	2MCAR	0.1561	1.0258	2.6310	0.6336
p50i20r5	2MNAR	0.1561	1.0258	2.1100	0.3121
p200i5r10	1MCAR	–0.0526	0.9842	2.3920	0.8000
p200i5r10	1MNAR	–0.0526	0.9842	1.8600	0.3132
p200i5r10	2MCAR	–0.0526	0.9842	2.4650	0.6498
p200i5r10	2MNAR	–0.0526	0.9842	1.9105	0.3418
p200i5r10BM	1MCAR	0.0997	1.4484	2.6100	0.9244
p200i5r10BM	1MNAR	0.0997	1.4484	2.0180	0.3966
p200i5r10BM	2MCAR	0.0997	1.4484	2.5795	0.8046
p200i5r10BM	2MNAR	0.0997	1.4484	2.0320	0.4125

Two of the four data sets (p50i5r5 and p50i20r5) were created with five total raters, and two of the data sets (p200i5r10 and p200i5r10BM) had 10 total raters. Rater severity parameters for data generation were not sampled, but rather were purposively, uniformly distributed across a range of severities. For the five rater data sets, severity ranged from –2.0 to +2.0; for the 10 rater data sets, severity ranged from –2.25 to +2.25. Tables VI to XIII show the severity for each rater, and the average true ability and average observed score (in the ordinal rating metric) for the examinees that were “assigned” to that rater.

Table VI and Table VII show the rater severity for the "p50i5r5" single rater and double rater data sets, respectively. These data sets had 50 persons or examinees, five raters, and five items. For the MCAR data, one (1MCAR) or two (2MCAR) raters were selected at random and ratings for all other raters were deleted (represented as missing data for the MCMC approach). In the case of the MNAR data, one (1MNAR) or two (2MNAR) raters were selected not at random. For the MCAR data, true ability averages vary randomly across raters; but, for the MNAR data, the average ability is lowest for judge 1 and highest for judge 5. That is, the most lenient raters (with negative severity) have the lowest ability examinees and the most severe raters (with positive severity) have the highest ability examinees. For both the MCAR and MNAR data, the average observed score decreases as rater severity increases. This pattern should result in ability estimates for low ability examinees being inflated, and estimates for high ability examinees being depressed (if severity is unaccounted for).

TABLE VI

MEAN TRUE ABILITY AND OBSERVED SCORE: P50I5R5 – SINGLE RATER

		1MCAR		1MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2	0.1186	3.6000	–0.9666	2.6000
2	–1	0.0238	2.9600	–0.6888	2.5000
3	0	0.4968	2.8889	–0.2829	2.0600
4	1	0.0823	2.0727	0.2049	1.8000
5	2	0.0157	1.6889	0.8956	1.9176

TABLE VII

MEAN TRUE ABILITY AND OBSERVED SCORE: P50I5R5 – TWO RATERS

		2MCAR		2MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2	–0.0165	3.4125	–0.9666	2.6000
2	–1	0.0540	2.9636	–0.7999	2.5400
3	0	0.2412	2.5882	–0.4351	2.1125
4	1	0.1723	1.9905	0.3765	2.1800
5	2	0.2283	1.7000	0.5963	1.7933

Table VIII and Table IX show the rater severity for the “p50i20r5” single rater and double rater data sets, respectively. These data sets have 50 persons or examinees, five raters, and 20 items. The data sets were generated similarly to those in Table VI and Table VII, and show similar patterns in terms of average true ability and average observed score in relation to rater severity. For the MCAR case, true ability varies randomly around 0 across raters, but is correlated with judge severity for the MNAR case. For both data scenarios, observed score decreases with the severity of the judge, but more dramatically for the MCAR case.

TABLE VIII

MEAN TRUE ABILITY AND OBSERVED SCORE: P50I20R5 – SINGLE RATER

		1MCAR		1MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2	0.2113	3.4500	–1.6390	2.5625
2	–1	0.1135	3.0906	–1.0055	2.3917
3	0	–0.0949	2.4455	–0.3695	2.2200
4	1	0.0313	1.9375	0.2565	2.0692
5	2	1.1149	2.0625	1.2209	2.0118

TABLE IX

MEAN TRUE ABILITY AND OBSERVED SCORE: P50I20R5 – TWO RATERS

		2MCAR		2MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2	0.3182	3.5452	–1.6390	2.5625
2	–1	0.1541	3.0759	–1.2589	2.3400
3	0	0.2316	2.5921	–0.6080	2.0781
4	1	–0.1578	1.9350	0.5099	2.2275
5	2	0.0848	1.6087	0.8030	1.8333

Table X and Table XI show the rater severities, average true abilities by rater, and average observed score by rater for the "p200i5r10" data sets. These data sets have 200 examinees or persons, five items, and 10 raters. Data sets for the 1MCAR, 1MNAR, 2MCAR, and 2MNAR cases were generated as above. Again, a similar pattern in the relationship between true ability and observed score by rater severity is seen.

TABLE X

MEAN TRUE ABILITY AND OBSERVED SCORE: P200I5R10 – SINGLE RATER

		1MCAR		1MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2.2500	–0.1653	3.3750	–2.0940	2.2000
2	–1.7500	–0.1557	3.2800	–1.5772	2.2500
3	–1.2500	–0.3144	2.9529	–0.9440	2.2500
4	–0.7500	–0.4800	2.5913	–0.9513	2.0333
5	–0.2500	0.2782	2.8000	–0.5869	1.9500
6	0.2500	0.0659	2.5120	–0.5785	1.6800
7	0.7500	–0.0584	2.0333	–0.0032	1.7000
8	1.2500	0.3394	2.1111	0.1469	1.6231
9	1.7500	–0.2266	1.7000	0.6399	1.7294
10	2.2500	0.1515	1.4897	1.1590	1.9765

TABLE XI

MEAN TRUE ABILITY AND OBSERVED SCORE: P200I5R10 – TWO RATERS

		2MCAR		2MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2.2500	–0.1072	3.4054	–2.0940	2.2000
2	–1.7500	0.0694	3.4000	–1.8356	2.2625
3	–1.2500	–0.2424	3.0162	–1.1973	2.2500
4	–0.7500	0.0195	2.9189	–0.9476	2.2500
5	–0.2500	–0.1656	2.5026	–0.7235	1.9750
6	0.2500	–0.0372	2.3391	–0.5827	1.8550
7	0.7500	0.0589	2.1489	–0.2533	1.7522
8	1.2500	0.0437	1.9613	0.0719	1.6962
9	1.7500	–0.1408	1.6632	0.6913	1.9362
10	2.2500	–0.0484	1.5064	0.8994	1.8118

Table XII and Tabl XIII show the rater severities, average true abilities by rater, and average observed score by rater for the "p200i5r10BM" data sets. Like the previous data sets described in Tables X and XI, these data sets have 200 examinees, five items, and 10 raters. However, unlike those data sets, the abilities for the examinees were not sampled from a standard normal distribution. Instead, they were sampled from a bimodal distribution (a mixture of two normal distributions, both with standard deviation of 1.0, with differing means (–1.0 and 1.0). The average true abilities for the MCAR data sets vary randomly around zero across all raters. Similar to the other data sets, the average abilities for the MNAR data sets also increase with rater severity. Again, the observed score for the MCAR cases decreases with the severity of the rater; however, for the MNAR data sets, the observed score does not follow a similar pattern.

TABLE XII

MEAN TRUE ABILITY AND OBSERVED SCORE: P200I5R10BM – SINGLE RATER

		1MCAR		1MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2.2500	0.1902	3.4500	–2.6063	1.9500
2	–1.7500	–0.2472	3.1167	–2.0665	2.0250
3	–1.2500	0.0384	3.0400	–1.8357	1.8833
4	–0.7500	0.4611	3.1905	–1.4601	1.7333
5	–0.2500	–0.3692	2.3875	–0.9630	1.8700
6	0.2500	–0.1287	2.4133	–0.3438	1.9100
7	0.7500	0.5782	2.4000	0.1324	1.8538
8	1.2500	0.1468	2.1067	0.4489	1.8538
9	1.7500	–0.0660	1.6364	1.0530	1.9706
10	2.2500	0.3342	1.8182	2.1203	2.6294

TABLE XIII

MEAN TRUE ABILITY AND OBSERVED SCORE: P200I5R10 – TWO RATERS

		2MCAR		2MNAR	
Judge	Severity	True Ability	Observed Score	True Ability	Observed Score
1	–2.2500	0.1786	3.5130	–2.6063	1.9500
2	–1.7500	0.2125	3.3171	–2.3364	2.0250
3	–1.2500	0.0595	3.0810	–1.9281	1.9200
4	–0.7500	0.0268	2.8600	–1.6479	1.8083
5	–0.2500	–0.0789	2.5758	–1.1494	1.8875
6	0.2500	0.5801	2.6880	–0.6534	1.8450
7	0.7500	–0.3574	1.9500	–0.0746	1.8609
8	1.2500	0.0750	1.9895	0.2906	1.8808
9	1.7500	0.2205	1.8432	1.2719	2.2468
10	2.2500	–0.1133	1.4909	1.5866	2.2676

A. Simulation Results

In Chapter II, a number of criteria were established to compare and test the adequacy of the different modeling approaches (MCMC bivariate missing data model, many facet Rasch model, Rasch rating scale model, and linear regression adjustment) for analyzing judge-rated data in the presence of non-ignorable missing data. These criteria utilized (for the most part) the fact that the true generating parameters for the simulated data are known. Therefore, I can compare the parameter estimates generated by each of the modeling approaches with the known "true" generating parameters. The results from each of the criteria are presented in the tables which follow in this section. These criteria include, the correlation between parameter estimates and the true abilities, the root mean squared error of the parameter estimates, the mean standard error of measurement of the parameter estimates, the reliability of the scores, the extent to which

examinees position within the distribution matches the true distribution (based on quantile membership), the percent of examinee scores significantly different from zero, and the coverage probability of the confidence intervals.

1. Correlation between Estimates and True Scores

Table XIV shows the correlation between the point estimates from each of the modeling approaches and the true abilities (generating parameters for the simulated data sets). For this table and subsequent tables, the point estimates for the MCMC approach are the mean of the posterior distribution, excluding a burn-in number of cases. In general, MCMC and MFRM approaches tend to have the highest correlations with the true ability, though the linear regression approach tends to perform nearly as well in double rater cases. Winsteps (for the single rater cases) produces the lowest correlations with true ability, except for the cases of the single rater, MNAR data set where abilities were simulated from a bimodal distribution. The most important finding though is that for all single rater cases, where raters were not randomly assigned, the MCMC approach yields higher correlations than the Facets approach. For the p50i5r5 1MNAR case, the correlation for the MCMC is .379 compared to .328 for the MFRM. When more items are added (p50i20r5 1MNAR), the correlation advantaged for the MCMC approach over the MFRM gets considerably larger (.710 versus .351). This might suggest that the capability of the MCMC approach to account for the non-ignorable missing data increases with the addition of more rating instances per examinee. When there are 200 examinees (with a unimodal distribution of abilities), the MCMC estimates correlate .364 with the true abilities in the single rater case (as compared to only .198 for the MFRM). When true abilities are bimodal, the correlation for the MCMC estimates is .345 as compared to .153 for the MFRM. While many of these correlations are low, the differences do matter (see the discussion of the coverage probabilities later in this

section). Finally, in one case (p200i5r10BM 2MNAR), the MCMC approach performs very well in comparison to the MFRM and linear regression adjustment results. The correlation for the MCMC approach is .494 in comparison to negative correlations for the other models. This seems to indicate that in an extreme case of non-ignorable missingness, the MCMC approach is able to correct for much of the bias.

TABLE XIV

CORRELATION BETWEEN TRUE ABILITY AND MODEL ESTIMATES

Data Set	Scenario	MCMC	MFRM	Winsteps	Linear Regression
p50i5r5	1MCAR	0.694	0.714	0.493	
p50i5r5	1MNAR	0.379	0.328	–0.107	
p50i5r5	2MCAR	0.870	0.862		0.864
p50i5r5	2MNAR	0.624	0.733		0.751
p50i20r5	1MCAR	0.797	0.898	0.593	
p50i20r5	1MNAR	0.710	0.351	–0.256	
p50i20r5	2MCAR	0.957	0.962		0.949
p50i20r5	2MNAR	0.978	0.978		0.980
p200i5r10	1MCAR	0.765	0.789	0.463	
p200i5r10	1MNAR	0.364	0.198	–0.071	
p200i5r10	2MCAR	0.868	0.868		0.864
p200i5r10	2MNAR	0.494	–0.108		–0.152
p200i5r10BM	1MCAR	0.854	0.873	0.661	
p200i5r10BM	1MNAR	0.345	0.153	0.549	
p200i5r10BM	2MCAR	0.935	0.938		0.924
p200i5r10BM	2MNAR	0.454	0.590		0.633

2. Root Mean Squared Error

Mean squared error is a measure which combines the error in the estimates due to variance (or measurement error) and bias (squared). Table XV shows the root mean squared error (RMSE) for each of the approaches and data sets, which is just the square root of the mean squared error. Values closer to zero indicate the estimates were closer to the true parameters than values that are larger. Similar to the correlation results, the RMSE results show that MCMC and MFRM approaches oftentimes are quite similar in the precision of the estimates. However, as was seen above for the correlations, the MCMC approach has lower RMSE than the MFRM approach in all cases with a single non-randomly assigned rater.

TABLE XV

ROOT MEAN SQUARED ERROR BY MODEL

Data Set	Scenario	MCMC	MFRM	Winsteps	Linear Regression
p50i5r5	1MCAR	0.775	0.749	0.997	
p50i5r5	1MNAR	1.103	1.148	1.473	
p50i5r5	2MCAR	0.506	0.518		0.516
p50i5r5	2MNAR	0.858	0.724		0.699
p50i20r5	1MCAR	0.631	0.447	0.893	
p50i20r5	1MNAR	0.753	1.128	1.569	
p50i20r5	2MCAR	0.290	0.274		0.317
p50i20r5	2MNAR	0.207	0.206		0.200
p200i5r10	1MCAR	0.683	0.649	1.034	
p200i5r10	1MNAR	1.125	1.263	1.460	
p200i5r10	2MCAR	0.512	0.512		0.520
p200i5r10	2MNAR	1.004	1.485		1.514
p200i5r10BM	1MCAR	0.539	0.502	0.821	
p200i5r10BM	1MNAR	1.141	1.298	0.947	
p200i5r10BM	2MCAR	0.360	0.352		0.388
p200i5r10BM	2MNAR	1.043	0.903		0.854

3. Mean Standard Error of Measurement

Different estimation approaches produced different magnitudes of standard errors of measurement for the ability estimates. Table XVI presents the average standard error of measurement for each modeling approach. It is important to note that the standard errors of measurement for the MCMC, MFRM, and Winsteps approaches are conditional SEMs (i.e., extreme ability estimates will have larger SEMs than average ability estimates). The SEMs for the linear regression approach do not have a relationship between ability and precision (the consequences of which will become apparent when looking at coverage probability results).

Results in Table XVI indicate that in general, the average SEMs for MNAR data sets are larger than those for MCAR data sets (within a data collection scenario – e.g., p50i5r5). However, in one case, the SEMS for the MFRM approach are larger for the 2MCAR case than the 2MNAR case (p50i20r5). Furthermore, the MCMC approach yields the largest average SEM, followed by the MFRM approach, with Winsteps producing the smallest average SEMs. The linear regression SEMs are never the smallest, and they tend to be on the larger side among the different models.

TABLE XVI

MEAN STANDARD ERROR OF MEASUREMENT BY MODEL

Data Set	Scenario	MCMC	MFRM	Winsteps	Linear Regression
p50i5r5	1MCAR	0.731	0.574	0.357	
p50i5r5	1MNAR	1.517	0.884	0.709	
p50i5r5	2MCAR	0.558	0.425		0.572
p50i5r5	2MNAR	0.704	0.522		0.918
p50i20r5	1MCAR	0.383	0.296	0.202	
p50i20r5	1MNAR	0.888	0.687	0.547	
p50i20r5	2MCAR	0.289	0.232		0.337
p50i20r5	2MNAR	0.263	0.199		0.313
p200i5r10	1MCAR	0.759	0.547	0.365	
p200i5r10	1MNAR	2.410	1.104	0.879	
p200i5r10	2MCAR	0.509	0.408		0.499
p200i5r10	2MNAR	1.063	0.498		0.836
p200i5r10BM	1MCAR	0.531	0.459	0.346	
p200i5r10BM	1MNAR	2.001	1.033	0.757	
p200i5r10BM	2MCAR	0.378	0.328		0.385
p200i5r10BM	2MNAR	0.842	0.521		0.787

4. Reliability and Generalizability Coefficients

Reliability estimates were calculated using the mean SEMs presented in the previous section along with the standard deviation of the standardized scores. The reliabilities presented in Table XVII were calculated as one minus the squared ratio of the mean SEM to the standard deviation of scores. The reliabilities therefore represent the proportion of variance in the observed scores that can be attributed to real differences among the examinees. Reliabilities of 0 indicate that all observed score variance may be measurement error; reliabilities closer to 1 indicate most variance is due to real differences.

The reliabilities are then largely a function of the size of the standard errors. In all cases, the estimated reliability of the MCMC approach is lower than the reliability of the MFRM model. This indicates that the MCMC approach is reporting greater uncertainty that differences in observed scores represent true differences in examinee ability.

TABLE XVII

ESTIMATED RELIABILITY BY MODEL

Data Set	Scenario	MCMC	MFRM	Winsteps	Linear Regression
p50i5r5	1MCAR	.466	.671	.873	
p50i5r5	1MNAR	.000	.219	.497	
p50i5r5	2MCAR	.688	.819		.673
p50i5r5	2MNAR	.505	.728		.157
p50i20r5	1MCAR	.854	.911	.959	
p50i20r5	1MNAR	.212	.528	.701	
p50i20r5	2MCAR	.916	.946		.887
p50i20r5	2MNAR	.931	.960		.902
p200i5r10	1MCAR	.425	.701	.867	
p200i5r10	1MNAR	.000	.000	.227	
p200i5r10	2MCAR	.741	.833		.751
p200i5r10	2MNAR	.000	.752		.302
p200i5r10BM	1MCAR	.718	.790	.880	
p200i5r10BM	1MNAR	.000	.000	.426	
p200i5r10BM	2MCAR	.857	.892		.852
p200i5r10BM	2MNAR	.292	.728		.381

Table XVIII provides another look at reliability by showing the D-study generalizability coefficients estimated from the fully-crossed data sets and the MCMC approaches (calculated from the data sets with missing data). The D-study G-coefficients are calculated as described in Chapter II. For the fully-crossed data estimates, the full simulated data set (with no missing data) was used to calculate the variance components. For the MCMC estimates, the variance components from the MCMC runs were used to calculate the G-coefficient. These estimates were based on the data set with unobserved cases deleted. Therefore, these results indicate the extent to which non-fully-crossed data can be used to determine estimates of the reliability under different data scenarios. The table shows the D-study G-coefficient that takes into account the number of raters and number of items used in the data set with missing data.

The results in Table XVIII show that when raters are randomly assigned, the MCMC estimate (which uses incomplete data) comes quite close to the G-coefficient calculated with the fully-crossed data using traditional G-theory variance component estimation techniques. In general, the MCMC approach does better when there are more items and more examinees. When only a single rater rates each examinee, the MCMC approach does not do well if the rater is not assigned randomly. Furthermore, the MCMC approach sometimes looks reasonable with two non-randomly assigned raters, and is again better when there are more items and more examinees. Since GT assumes raters drawn randomly from a universe of potential raters, it is not surprising that numbers calculated based on non-randomly sampled judges cannot replicate the GT results.

TABLE XVIII

D-STUDY GENERALIZABILITY COEFFICIENT: FULLY-CROSSED DATA SET VERSUS MCMC ESTIMATE

Data	Scenario	Fully-Crossed Data	MCMC Estimate
p50i5r5	1MCAR	.527	.664
p50i5r5	1MNAR	.527	.371
p50i5r5	2MCAR	.690	.767
p50i5r5	2MNAR	.690	.691
p50i20r5	1MCAR	.855	.887
p50i20r5	1MNAR	.855	.634
p50i20r5	2MCAR	.922	.921
p50i20r5	2MNAR	.922	.925
p200i5r10	1MCAR	.615	.606
p200i5r10	1MNAR	.615	.215
p200i5r10	2MCAR	.762	.780
p200i5r10	2MNAR	.762	.470
p200i5r10BM	1MCAR	.778	.769
p200i5r10BM	1MNAR	.778	.250
p200i5r10BM	2MCAR	.875	.864
p200i5r10BM	2MNAR	.875	.588

5. Distributional Effects

To examine the extent to which the examinee abilities follow a similar distribution to the true abilities, I calculated the percent of cases where the ability estimate from each modeling approach placed the examinee into the same quintile of the scoring distribution. The results from this analysis are summarized in Table XIX. For single rater data sets, the distributional match was higher for the MCAR case than the MNAR case in all cases. In general, for single rater MCAR cases, the MCMC and MFRM approach were quite similar in result (which mirrors the correlation results described earlier). Winsteps was less effective in placing individuals into the correct quintile. These results (similar to the correlation results) seem to indicate that there is not

a clear advantage among models (between MCMC and the MFRM) in placing individuals into the right place of the distribution. Results can be far from the true abilities under scenarios where only a single rater is used or raters are assigned non-randomly. Again, the MCMC approach outperformed the MFRM in all cases where there was a single non-randomly assigned rater.

TABLE XIX

PERCENT OF EXAMINEES IN THE SAME QUINTILE AS TRUE ABILITY BY MODEL

Data Set	Scenario	MCMC	MFRM	Winsteps	Linear Regression
p50i5r5	1MCAR	46.0%	40.0%	32.0%	
p50i5r5	1MNAR	20.0%	18.0%	10.0%	
p50i5r5	2MCAR	52.0%	52.0%		58.0%
p50i5r5	2MNAR	38.0%	44.0%		48.0%
p50i20r5	1MCAR	58.0%	66.0%	34.0%	
p50i20r5	1MNAR	34.0%	26.0%	24.0%	
p50i20r5	2MCAR	70.0%	72.0%		66.0%
p50i20r5	2MNAR	88.0%	84.0%		84.0%
p200i5r10	1MCAR	42.5%	45.0%	26.0%	
p200i5r10	1MNAR	30.5%	27.0%	17.5%	
p200i5r10	2MCAR	49.5%	50.5%		50.0%
p200i5r10	2MNAR	32.5%	10.2%		10.5%
p200i5r10BM	1MCAR	55.5%	64.0%	33.5%	
p200i5r10BM	1MNAR	30.5%	21.5%	33.5%	
p200i5r10BM	2MCAR	69.0%	67.0%		65.5%
p200i5r10BM	2MNAR	28.5%	40.0%		41.5%

6. Statistical Significance of Difference from Average

Table XX shows the percent of examinee ability estimates that are significantly different from average (using model SEMs for the abilities). Models with higher percentages indicate a greater confidence (rightfully or wrongfully) in the precision of the estimates. These results highlight the policy implications of different model choices. Depending on model choice, individual examinees may be more likely to be deemed statistically far below or far above average, which could lead to various consequences or rewards. In general, models with larger mean SEMs will have lower percentages in the table. Across the board, the MCMC approach indicates that fewer examinees are significantly different from average than the MFRM. Furthermore, the percent of cases that are different from average is lower for MNAR data sets than MCAR data sets. All models are taking into account the greater uncertainty due to the non-random assignment of raters. The linear regression adjustment estimates have a much higher percentage that is different from zero. This is largely a function of this approach having smaller SEMs which are not conditional on the ability of the examinee; that is, extreme abilities do not have larger SEMs to account for greater uncertainty in the ability estimate at the high and low end of the ability continuum.

TABLE XX

PERCENT OF ABILITY ESTIMATES SIGNIFICANTLY DIFFERENT FROM AVERAGE (ZERO) BY MODEL

Data Set	Scenario	MCMC	MFRM	Winsteps	Linear Regression
p50i5r5	1MCAR	10.0%	32.0%	62.0%	
p50i5r5	1MNAR	2.0%	6.0%	6.0%	
p50i5r5	2MCAR	26.0%	46.0%		24.0%
p50i5r5	2MNAR	14.0%	28.0%		6.0%
p50i20r5	1MCAR	44.0%	62.0%	72.0%	
p50i20r5	1MNAR	6.0%	14.0%	26.0%	
p50i20r5	2MCAR	52.0%	66.0%		50.0%
p50i20r5	2MNAR	72.0%	88.0%		58.0%
p200i5r10	1MCAR	13.0%	28.0%	58.0%	
p200i5r10	1MNAR	0.0%	1.5%	1.0%	
p200i5r10	2MCAR	34.5%	40.5%		34.5%
p200i5r10	2MNAR	3.5%	31.0%		12.0%
p200i5r10BM	1MCAR	31.0%	38.0%	60.5%	
p200i5r10BM	1MNAR	0.0%	4.5%	12.0%	
p200i5r10BM	2MCAR	48.0%	54.0%		51.0%
p200i5r10BM	2MNAR	9.0%	26.0%		9.0%

7. Confidence Interval Coverage Probability

Table XXI displays the proportion of examinees whose 95 percent confidence interval covers their true ability estimate by model, data set, and scenario. Most strikingly, the intervals for the linear regression adjustment estimates almost never cover the true value. Again, this is due to the non-conditional nature of the SEMs from this model, along with the SEMs being too small generally. The Winsteps model also doesn't do well, due to not taking into account any uncertainty due to rater effects. The MFRM and MCMC approaches do the best. However, in all cases, the MFRM approach does not reach the optimal coverage probability (.95). This indicates that the MFRM approach is yielding SEMs which are too small given the

measurement error attributable to rater effects. In nearly all cases, the MCMC approach achieves or exceeds the coverage probability target of .95.

TABLE XXI

95 PERCENT CONFIDENCE INTERVAL COVERAGE PROBABILITY BY MODEL

Data Set	Scenario	MCMC	MFRM	Winsteps	Linear Regression
p50i5r5	1MCAR	.920	.800	.480	
p50i5r5	1MNAR	1.000	.860	.360	
p50i5r5	2MCAR	1.000	.900		.020
p50i5r5	2MNAR	.920	.840		.020
p50i20r5	1MCAR	.860	.800	.740	
p50i20r5	1MNAR	.980	.760	.600	
p50i20r5	2MCAR	.960	.900		.040
p50i20r5	2MNAR	1.000	.940		.020
p200i5r10	1MCAR	.980	.905	.520	
p200i5r10	1MNAR	1.000	.915	.220	
p200i5r10	2MCAR	.945	.895		.080
p200i5r10	2MNAR	.975	.475		.220
p200i5r10BM	1MCAR	.955	.935	.475	
p200i5r10BM	1MNAR	1.000	.875	.130	
p200i5r10BM	2MCAR	.940	.940		.060
p200i5r10BM	2MNAR	.900	.760		.115

B. Conclusion

This chapter presented the results from the analysis of the simulated data sets for the purpose of comparing the effectiveness of several different approaches to analyzing judge rated data. In general, the MFRM and MCMC approaches were similar in their ability to produce

ability point estimates that were correlated with the true abilities, and these correlations were lower when judges were not randomly assigned. While comparisons of point estimates did not yield a clear winner between the MCMC and MFRM approach, other indicators of the relative performance of the models tended to favor the MCMC approach under certain circumstances. The MCMC approach produces much larger SEMs than the other approaches, and this leads to lower estimates of reliability, fewer cases classified as significantly different from average, and higher 95 percent coverage probabilities. One interpretation of this would be to say that traditional approaches underestimate the uncertainty in the ability estimate. The MCMC approach, by directly modeling the missing (perhaps non-ignorable) missing data mechanism, accounts for this uncertainty through larger standard errors of measurement. These larger SEMs then result in less confidence (appropriately) in the reliability of the point estimates, along with the statistical significance of any differences in those estimate.

In addition to the comparison of the approaches in their ability to estimate generating ability parameters, this simulation study also examined the extent to which the MCMC approach could produce G-Theory variance components, even when traditional approaches could not estimate them. Results generally showed that when raters were randomly assigned, the MCMC approach produced variance component estimates which could be used in D-studies for calculating G-coefficients, even in cases with only a single rater. With only a single rater, the traditional G-Theory approach cannot disentangle the rater effect from the person effect.

IV. REAL WORLD DATA RESULTS

To test the applicability of the Bayesian bivariate probit ordinal missing data model to real world data, I analyzed data from the Class Pre-K observation system. These data include ratings of 255 early childhood centers in an urban environment. Each center was rated on 10 items (scored from 1 to 7), which were grouped into three construct scores: emotional support (ES), classroom organization (CO), and instructional support (IS). Scores for the emotional support construct were based on four items, while scores for classroom organization and instructional support were based on three items. As with the simulations, the scores were standardized to facilitate comparison across methods. Therefore, the mean score for the MCMC, MFRM, and Winsteps approaches is 0, and the standard deviation of scores is 1. Rater severity/leniency, item difficulty/easiness, and rating scale threshold parameters were all standardized using the mean and standard deviation of the examinee scores (specific to the particular analysis – i.e., MCMC, MFRM, or Winsteps).

A. MCMC Results

Separate models were fit for each of the three constructs. Tables XXII, XXIII, and XIV show the estimates of rater leniency for the emotional support, classroom organization, and instructional support constructs, respectively. As opposed to rater severity, the MCMC approach estimates rater leniency – that is, higher values of the parameter indicate higher scores for examinees on average (for examinees assigned that rater). For the emotional support construct, the standardized rater leniency parameters range from 1.988 to 5.180, with an average standard error of about 0.4, indicating that there are statistically significant differences in rater leniency for this construct. One rater (judge 2) gave particularly high scores for this construct, after conditioning on the ability of the examinees.

TABLE XXII

MCMC JUDGE LENIENCY: EMOTIONAL SUPPORT

Judge	Leniency Estimate	SEM
1	3.268	0.399
2	5.180	0.418
3	2.722	0.397
4	2.394	0.394
5	1.988	0.397
6	2.358	0.428
7	2.364	0.421
8	2.123	0.399

For the classroom organization construct (Table XXIII), rater leniency ranged from 0.795 to 3.473, with an average standard error of about 0.41. As with the previous construct, judge 2 is again the most lenient; however, this judge did not appear to be quite as much of an outlier as for emotional support. Judge 8 appears to give quite lower ratings than the other judges for classroom organization.

TABLE XXIII

MCMC JUDGE LENIENCY: CLASSROOM ORGANIZATION

Judge	Leniency Estimate	SEM
1	3.415	0.404
2	3.473	0.406
3	2.682	0.402
4	1.719	0.404
5	1.434	0.437
6	1.735	0.438
7	1.765	0.419
8	0.795	0.449

Rater leniency ranged from –0.159 to 2.127 for instructional support (Table XIV), with an average standard error of about 0.37. Contrary to the other two constructs, judge 2 was not the most lenient for this construct. Judge 8 was again the most severe. The lack of consistency in general of the leniency ordering across constructs might indicate an interaction between the raters, items, and examinees, which is not accounted for by this model or any of the other models.

TABLE XXIV

MCMC JUDGE LENIENCY: INSTRUCTIONAL SUPPORT

Judge	Leniency Estimate	SEM
1	1.613	0.345
2	1.671	0.345
3	2.127	0.346
4	1.089	0.368
5	1.551	0.407
6	0.816	0.404
7	0.808	0.366
8	–0.159	0.410

Table XXV presents the standardized item easiness parameters by construct. Items with higher easiness parameter estimates are indicative of items where examinees tended to receive higher scores (conditional on examinee ability and rater leniency).

TABLE XXV

ITEM EASINESS PARAMETER ESTIMATES BY CONSTRUCT

	Emotional Support		Classroom Organization		Instructional Support	
Item	Easiness	SEM	Easiness	SEM	Easiness	SEM
1	1.320	0.432	0.589	0.428	–0.091	0.382
2	0.966	0.419	0.374	0.408	1.323	0.361
3	1.040	0.397	1.006	0.425	–0.042	0.392
4	1.222	0.418				

Table XXVI shows the threshold parameters for the three constructs. These parameters represent the dividing points between the seven rating categories. Looking across the constructs, there is some evidence that the space between categories is not consistent across items. A partial credit model may be more appropriate for these data, but is beyond the scope of this thesis.

TABLE XXVI

MCMC RATING SCALE THRESHOLD PARAMETER ESTIMATES BY CONSTRUCT

Construct	1/2	2/3	3/4	4/5	5/6	6/7
Emotional Support	–0.053	0.801	1.600	2.831	3.876	5.196
Classroom Organization	–0.038	0.679	1.354	2.490	3.734	5.223
Instructional Support	–0.021	1.004	1.868	2.689	3.436	4.293

B. MFRM Results

Similar to the MCMC results, the results for the MFRM are presented separately by construct. Tables XXVII, XXVIII, and XXIX show the rater severity for the emotional support, classroom organization, and instructional support constructs, respectively. In contrast to the MCMC results above, the tables show judge severity (as opposed to leniency). For the MFRM and Winsteps psychometric packages, judge and item parameters are generally reported as judge severity and item difficulty. For the MCMC approach, the model parameters represent the opposite (leniency and easiness). The MFRM was implemented with an assumption of average

group ability equal to 0 for each disjoint subset of examinees – that is, the average ability for examines was anchored to 0 for each judge.

For the emotional support construct (Table XXVII), standardized rater severity ranged from –2.884 to –1.457 with an average standard error of about 0.08. Judges 2, 3, and 4 tended to be the most lenient, and judge 6 tended to be the most severe.

TABLE XXVII

MFRM JUDGE SEVERITY: EMOTIONAL SUPPORT

Judge	Severity Estimate	SEM
1	–2.678	0.062
2	–2.884	0.113
3	–2.884	0.113
4	–2.755	0.067
5	–1.751	0.088
6	–1.457	0.082
7	–1.720	0.062
8	–1.612	0.088

Rater severity for the classroom organization construct (Table XXVIII) ranged from –2.033 to –0.768, with an average standard error of about 0.08. Judges 1 and 4 were the most lenient, and judge 7 was the most severe.

TABLE XXVIII

MFRM JUDGE SEVERITY: CLASSROOM ORGANIZATION

Judge	Severity Estimate	SEM
1	–1.993	0.065
2	–1.592	0.100
3	–1.858	0.110
4	–2.033	0.070
5	–0.974	0.090
6	–0.808	0.080
7	–0.768	0.060
8	–1.100	0.095

Standardized rater severity estimates for the instructional support construct (Table XXIX) ranged from –0.933 to 0.653, with an average standard error of about 0.06. Judge 5 was the most severe rater, and judge 2 was the most lenient. As with the MCMC results, there is evidence that the severity of judges is not consistent across constructs, which may indicate an interaction between judges, items, and examinees which is not accounted for by this model.

TABLE XXIX

MFRM JUDGE SEVERITY: INSTRUCTIONAL SUPPORT

Judge	Severity Estimate	SEM
1	–0.653	0.040
2	–0.933	0.073
3	–0.216	0.063
4	–0.346	0.043
5	0.653	0.067
6	0.523	0.060
7	–0.836	0.047
8	0.543	0.067

Table XXX shows the item difficulty parameters for the three constructs when analyzed with the MFRM. In contrast to the MCMC runs, higher values of the parameter estimates indicate more difficult items and lower overall scores (conditional on examinee ability and rater severity).

TABLE XXX

MFRM ITEM DIFFICULTY PARAMETER ESTIMATES

	Emotional Support		Classroom Organization		Instructional Support	
Item	Difficulty	SEM	Difficulty	SEM	Difficulty	SEM
1	0.144	0.052	–0.176	0.050	0.167	0.033
2	–1.576	0.093	–0.196	0.050	0.127	0.033
3	0.582	0.052	0.366	0.045	–0.293	0.033
4	0.855	0.046				

The rating scale thresholds are shown in Table XXXI. Similar to the MCMC results, there is some inconsistency in the size of the categories across constructs. This indicates that the categories are functioning differently across constructs, and may indicate that a partial credit model may be more appropriate for these data (if thresholds vary across items, within construct). While MFRM software can easily accommodate a partial credit model, the rating scale model is used here to serve as a comparison for the MCMC results.

TABLE XXXI

MFRM RATING SCALE THRESHOLD PARAMETER ESTIMATES BY CONSTRUCT

Construct	1/2	2/3	3/4	4/5	5/6	6/7
Emotional Support	–2.668	–1.190	–0.597	0.659	1.401	2.395
Classroom Organization	–1.722	–0.964	–0.753	0.155	1.089	2.199
Instructional Support	–2.012	–1.049	–0.313	0.433	1.052	1.892

There is one issue of significance to note in the item parameter tables for the MFRM and MCMC approach. The ordering of the item difficulties appears to be different. This is perhaps due to the existence of missing rater/item interactions, which neither approach is modeling. The MCMC results indicate that there is a relationship between missing data and the missing rating itself. As the MCMC approach accounts for this different, it may not be surprising that the two sets of item parameters are inconsistent with one another.

C. WINSTEPS Results

The observation data were also analyzed with Winsteps. This modeling approach cannot account for raters, so no rater severity estimates are presented here. These results were calculated to enable comparisons among the MCMC and MFRM model, and a model which does not adjust for rater severity. Table XXXII shows the item difficulty parameter estimates for the three constructs. As with the MFRM results, higher values indicate items which had lower scores on average (i.e., were more difficult for examinees to achieve a high score).

TABLE XXXII

WINSTEPS ITEM DIFFICULTY PARAMETER ESTIMATES

	Emotional Support		Classroom Organization		Instructional Support	
Item	Difficulty	SEM	Difficulty	SEM	Difficulty	SEM
1	–1.802	0.045	–1.451	0.045	–0.121	0.030
2	–2.221	0.071	–1.469	0.045	–0.156	0.030
3	–1.418	0.045	–0.977	0.040	–0.523	0.030
4	–1.177	0.040				

Table XXXIII shows the rating scale threshold parameters from the Winsteps rating scale model run, presented separately by construct.

TABLE XXXIII

WINSTEPS RATING SCALE THRESHOLD PARAMETER ESTIMATES BY CONSTRUCT

Construct	1/2	2/3	3/4	4/5	5/6	6/7
Emotional Support	–3.688	–2.622	–2.310	–1.195	–0.508	0.393
Classroom Organization	–2.806	–2.139	–1.965	–1.165	–0.347	0.628
Instructional Support	–2.025	–1.181	-0.541	0.113	0.649	1.386

D. Comparisons

Unlike with the simulation analyses, there is no "true" benchmark to which I could compare the scores across the different methods. However, I could analyze the results to see the extent to which the different methods would yield similar inference in an applied setting. To accomplish that, I looked at the correlation among the scores across methods, the average standard errors of measurement, the reliabilities of the scores, and the percent of scores that were significantly different from average. The results from each of these analyses are presented here.

A key component of the MCMC approach is the model specification for the missing data mechanism. In particular, the missingness is allowed to be conditional on the value of the missing rating itself. Table XXXIV shows the coefficient estimate for the rating (observed or unobserved) on the likelihood that a rating is missing. Positive coefficients, with posterior intervals which do not overlap zero are indicative of higher ratings being more likely to be missing. For all three constructs, the missing data model indicated that missing data are non-ignorable. These results imply that any model which treats the missing data as ignorable may lead to poor statistical inference.

TABLE XXXIV

ESTIMATES OF THE RELATIONSHIP BETWEEN RATINGS (OBSERVED AND UNOBSERVED) AND MISSINGNESS

Construct	Estimate	Posterior SD
Emotional Support	0.159	0.016
Classroom Organization	0.161	0.012
Instructional Support	0.088	0.018

To examine the extent to which different methods produced similar scores (by construct), I calculated the correlation among the scores. Table XXXV shows these correlations. Across all three constructs, the scores from the MCMC and MFRM approaches showed the highest correlation. Correlations were lower between Winsteps and the other approaches; however, Winsteps scores tended to correlate higher with the MCMC scores than the MFRM scores. These correlations imply that the point estimates of the scores from the MCMC and MFRM approaches are quite similar for this observation data set.

TABLE XXXV

CORRELATION AMONG ABILITY ESTIMATES BY MODEL AND CONSTRUCT

Construct	Model	MCMC	MFRM	Winsteps
Emotional Support	MCMC	1.000	0.981	0.937
	MFRM		1.000	0.871
	Winsteps			1.000
Classroom Organization	MCMC	1.000	0.986	0.937
	MFRM		1.000	0.880
	Winsteps			1.000
Instructional Support	MCMC	1.000	0.997	0.901
	MFRM		1.000	0.875
	Winsteps			1.000

Table XXXVI shows the average standard errors of measurement across constructs and methods. As with the simulation results, the average standard errors of measurement for the MCMC approach were larger than those from MFRM, which were larger than those from Winsteps. The MCMC approach takes into account the non-ignorable missing data mechanism which leads to larger SEMs. Given the relationship between the unobserved ratings and missingness, these larger SEMs seem appropriate.

TABLE XXXVI

MEAN STANDARD ERROR OF MEASUREMENT BY MODEL AND CONSTRUCT

Construct	MCMC	MFRM	Winsteps
Emotional Support	0.580	0.491	0.428
Classroom Organization	0.626	0.493	0.437
Instructional Support	0.422	0.317	0.280

These larger standard errors contributed to lower estimates of reliability for the MCMC scores, with MFRM scores having the second lowest estimated reliabilities, and Winsteps having the largest estimated reliabilities (see Table XXXVII).

TABLE XXXVII

RELIABILITY BY MODEL AND CONSTRUCT

Construct	MCMC	MFRM	Winsteps
Emotional Support	0.663	0.759	0.817
Classroom Organization	0.609	0.757	0.809
Instructional Support	0.822	0.899	0.922

In addition to these reliabilities, I also used the variance component estimates from the MCMC runs to calculated D-study G-coefficients using the correct number of items, and a single rater for each construct. The G-coefficients for the three constructs are .709 for emotional support, .711 for classroom organization, and .867 for instructional support. These numbers slightly lower than the estimated reliabilities from Winsteps and the MFRM, and slightly higher than those from the MCMC approach. This result may imply that the MFRM and Winsteps approaches are overestimating the reliability whereas the MCMC approach is underestimating the reliability.

Table XXXVIII shows the percent of centers with scores that were significantly different from average (when taking into account the standard error of measurement). The MCMC

approach identified significantly fewer observations as statistically significantly lower than or higher than average. As with the simulations, the MCMC approach represents a more conservative approach to the classification of scores as different from average.

TABLE XXXVIII

PERCENT OF CENTERS SIGNIFICANTLY DIFFERENT FROM AVERAGE

Construct	MCMC	MFRM	Winsteps
Emotional Support	22.4%	31.4%	23.5%
Classroom Organization	19.6%	27.5%	28.6%
Instructional Support	43.9%	59.6%	60.8%

E. Conclusion

The application of the Bayesian bivariate probit missing data model to a real world data set of Pre-K classroom observation demonstrates the applicability of this model in a practical setting. Budgetary constraints may have limited observations to a single non-randomly assigned rater per classroom. The analysis of these data showed that the particular model implemented mattered for the inference drawn about the objects of measurement. In particular, the MCMC results indicate that there was a relationship between the observed and unobserved ratings and the pattern of missingness. This implies that the data are likely MNAR, with a single rater. Simulation results in Chapter III indicate that the MCMC approach outperforms traditional approaches such as the MFRM or rating scale model in this type of scenario. The percent of

centers identified as significantly above or below average varies by model. Therefore, model selection has a potential real world impact for the objects of measurement. The MCMC approach best takes into account the uncertainty associated with the non-random assignment of rater.

V. DISCUSSION

Scores and ratings that rely on the judgment of raters with different levels of severity are subject to increased measurement error and bias if steps are not taken to account for this source of error. Failure to study and control for such error can lead to incorrect conclusions, and real-life consequences for the subjects of measurement. Solutions exist to estimate the extent of measurement error due to raters and adjust your data collection plan to minimize such error; that is, conduct a G-study followed by a D-study using the methods available in generalizability theory. Other solutions exist for estimating rater severity when some or all examinees are rated by multiple judges. These methods (e.g., the many-facet Rasch model and linear regression adjustments) estimate a fair score for an examinee by adjusting for the severity of the judges they were assigned. Each of these approaches requires that at least some of the examinees are rated by more than one judge. When data are collected without multiple judges per examinee, a solution needs to be found for estimating and accounting for the error and bias due to unequal judge assignment to examinees. This paper proposes a method to address that scenario.

This thesis tested a method for applying MCMC and missing data analysis techniques to a sparse data set of judge rated examinees (i.e., sparse to the point of only a single judge rating each examinee). The proposed methods were applied with the intent of generating estimates of reliability and error similar to traditional GT and correcting for rater bias including instances when missing data are non-ignorable. The two main objectives of this thesis were as follows:

- **Research Objective 1:** To develop statistical methods that allow one to investigate the extent to which measurement error is partitioned among the facets of measurement (with

an emphasis on rater bias) for data collection scenarios where traditional approaches in the literature cannot be applied.

- **Research Objective 2:** To develop methods which produce more accurate latent trait measure estimates that account for rater bias and error due to the other facets of measurement, compared to existing methods which either cannot account for rater bias, or assume non-ignorable missing data.

To accomplish these objectives, the new MCMC approach was compared to several existing approaches for analyzing judge rated data. The high-level details of each approach are reiterated here.

Generalizability theory provides methods for estimating the variance attributable to different facets of measurement (e.g., items, judges). The most straightforward method, which will produce unconfounded estimates of the variance components, requires fully crossed data (all examinees rated on all items by all raters). Once variance components are estimated, these parameter estimates can then be used to determine a generalizability coefficient for a given study design. This method doesn't correct for rater bias, but rather helps researchers be proactive about selecting an appropriate number of judges to rate each individual to yield measures with an appropriately small standard error of measurement.

Linear regression adjustment fits a linear model to the individual ratings, including fixed effects for raters, persons, and items. The person fixed effect estimates from these models adjust for the severity of the specific raters that each examinee was assigned. Examinees with more severe raters will receive a boost in their score whereas examinees with more lenient raters will have their scores adjusted lower. This model is in line with classical test theory and generalizability theory models in that it does not include a link function to account for the ordinal

nature of the ratings. This approach also requires that sufficient examinees are rated by more than one judge, and that there exists a data-path that connects the ratings of one examinee to another via common judges.

The many-facet Rasch model is similar to the linear regression model in that it estimates parameters for persons, items, and judges, but does do using a logistic linking function to account for the ordinal nature of the observed ratings. Furthermore, like the linear regression model, this approach requires that at least some of the examinees are rated by multiple judges, and that there exists a data-path from one examinee to another. However, it is possible to implement the MFRM even when no double ratings have occurred. By assuming that the average ability of the examinees assigned to each rater is equivalent, one can estimate rater severity effects. Of course, these estimates are dependent on the assumption of this group-mean anchoring approach – which may be less appropriate when examinees are not randomly assigned to raters.

The Rasch rating scale model is similar to the MFRM, but does not account for rater effects. This model was applied to data collection scenarios where a single judge rated each examinee. These estimates serve as a check on how biased scores might look when the fact that judges were used in the scoring process is completely disregarded.

Finally, the Bayesian bivariate probit ordinal missing data model described in this thesis offers a possible solution to accounting for error and bias due to raters, while minimizing assumptions about the data. This approach treats the unobserved responses for judges not assigned to a particular examinee as missing data. Based on that, I have proposed a bivariate model which simultaneously models the missing data mechanism and the measurement model.

This approach is implemented via Markov Chain Monte Carlo methods using a hybrid Gibbs/Metropolis algorithm.

To compare the different modeling approaches under different data collection approaches, I simulated four master data sets. Each of these data sets were generated with ordinal data that fit the MFRM. In these master data sets, each judge rated each examinee on all items. Then, data subsets were generated which deleted out data (either randomly or non-randomly) to emulate data collection scenarios where examinees were either rated by a single judge or two judges. The four simulated data sets had the following properties:

1. P50i5r5: 50 examinees (with normally distributed abilities), rated by 5 judges, on 5 items.
2. P50i20r5: 50 examinees (with normally distributed abilities), rated by 5 judges, on 20 items.
3. P200i5r10: 200 examinees (with normally distributed abilities), rated by 10 judges, on 5 items.
4. P200i5r10BM: 200 examinees (with bimodally distributed abilities), rated by 10 judges, on 5 items.

For each of these master data sets, data subsets where generated as 1MCAR (a single, randomly assigned rater), 1MNAR (a single, non-randomly assigned rater), 2MCAR (two, randomly assigned raters), and 2MNAR (two, non-randomly assigned raters).

Each of the modeling approaches described above (generalizability theory, MCMC, MFRM, and Rasch rating scale model) were fit to each of the 16 simulated data subsets to allow for comparison across a number of criteria. For some data subsets, a particular approach may not have been fit due to incompatibility of the method to the particular scenario. The results for each

of the comparison criteria were presented in chapter 3, and the nature of these criteria are described briefly here.

- **Pearson correlation with true abilities:** The correlation between the standardized measures from each approach were correlated with the standardized true generating parameters used to create the simulated data. This criterion demonstrates the correspondence between scores from each method and the targeted true abilities.
- **Root mean squared error:** The square root of the average squared deviation between the ability estimates and the true abilities was calculated to determine the extent to which each method yielded the lowest bias and variance.
- **Average standard error of measurement:** The average SEMs were calculated to show the extent to which each method represented the uncertainty in the ability estimates (which could then be used to estimate reliabilities and confidence intervals).
- **Reliability:** Reliability for each set of scores was estimated (for consistency across methods) by using the standard error of measurement and standard deviation of the observed scores. In addition to these reliabilities, D-study G-coefficients were calculated based on the variance components from the MCMC approach (using incomplete data) and standard G-Theory techniques (using the fully-crossed data – i.e., the complete data set).
- **Percent of cases different from average:** The percent of cases statistically different from average helps tell the story of the practical impact on examinees of using each of the modeling approaches.
- **Confidence interval coverage probability:** The proportion of cases where the 95 percent confidence interval overlapped the true generating parameter was used to

estimate the extent to which each approach accurately estimated the amount of uncertainty contained in the score.

- **Rank-ordering of examinees into quantiles:** The score estimates from each approach were grouped by quantile and compared to the quantiles groupings of the true abilities. The percent of cases lying along the table of the cross-tabulation served as a further indicator of the consistency between the estimates and the true abilities.

Chapter III presented the results from the analysis of the four main simulated data sets, with the results from a real world data set presented in Chapter IV. For each of the data sets, all four modeling approaches (plus G-Theory) were used to determine the extent to which different approaches yielded more accurate results, or led to different practical inferences. The MCMC and MFRM approaches were similar in their capability to generate ability estimates that were closely aligned with the true abilities used for the data simulation when raters were randomly assigned, or two raters rated each examinee. Correlations favored the MCMC approach when a single non-randomly assigned rater was used. The correlations (and RMSE) showed greater alignment with true abilities in cases where judges were randomly assigned, and when two raters were used to rate each examinee (as opposed to a single rater). The MCMC approach yielded larger SEMs than the MFRM or other approaches, which led to higher confidence interval coverage probabilities, but lower reliability estimates and fewer examinees identified as significantly different from average. These findings were also consistent for the real world data set. The MCMC estimates of G-coefficients tended to closely mirror the G-coefficients estimated using classical G-Theory techniques on the fully-crossed complete data, but only when raters were randomly assigned. Furthermore, the estimates were closer to the classical estimates in scenarios with more items and more examinees (i.e., larger sample sizes).

The remainder of this chapter provides a discussion of these findings in the context of the two main research objectives, along with a plan for potential future research.

A. Research Objective 1

The first objective of this thesis was to develop statistical methods that allow one to investigate the extent to which measurement error is partitioned among the facets of measurement (with an emphasis on rater bias) for data collection scenarios where traditional approaches in the literature cannot be applied. Generally speaking, this objective aimed to determine if the proposed MCMC approach could replicate classical G-Theory approaches in situations where fully-crossed data was not collected.

This thesis does not argue that the approach described here should be used in place of a traditional, rigorous G-Theory approach. To determine the proper number of items and raters needed to achieve a certain level of reliability (as estimated by the G-coefficient), researchers should conduct a G-study using randomly assigned raters, and using a fully-crossed data collection approach. However, when no such study has been conducted, or is not possible given time or resource constraints, it seems imperative to still analyze the reliability of the data you are collecting, and make decisions about how many judges are needed to collect data which meets the needs of your study. To that end, the MCMC method proposed here seems like a promising method for estimating the variance attributable to the different facets of measurement in situations where data are collected in a manner that doesn't allow for traditional approaches. In particular, this thesis demonstrated that variance component estimates and G-coefficients produced from non-fully-crossed data are quite similar to estimates from fully-crossed data when only a single rater is used per examinee (as long as the raters are randomly assigned, which is a requirement for implementation of traditional GT in practice).

B. Research Objective 2

The second objective of this thesis was to develop methods which produce more accurate latent trait measure estimates that account for rater bias and error due to the other facets of measurement, compared to existing methods which either cannot account for rater bias, or assume non-ignorable missing data. The theory behind this objective was that a modeling approach which modeled the missing data model in addition to the measurement model would yield ability estimates that were closer to the true ability estimates than approaches which did not explicitly model the missing data mechanism.

Chapter II proposed five research hypotheses that relate to this research objective. Hypothesis one proposed that the new method would produce estimates which were closer to the true generating parameters than estimates from traditional approaches. Hypothesis two proposed that these estimates would exhibit less bias. In cases with a single non-randomly assigned rater, the new approach yielded estimates that were closer to the true abilities than traditional approaches. Hypothesis three posited that standard errors of measurement would be higher for the new approach and reliabilities would be correspondingly lower. Results indicated that this was the case regardless of measurement scenario. Hypothesis four proposed that the rank ordering of ability estimates would more closely match the ordering of the true abilities when using the new method. This was true for non-randomly assigned single-rater cases, but not generally. Finally, hypothesis five proposed that the coverage probabilities (defined as the percent of true abilities covered by the 95 percent confidence interval for the estimate) would be better with the new approach. This was generally true, with the new method having coverages close to 95 percent (or higher) in most cases, while traditional approaches were often well under 95 percent coverage.

The results indicated that under many scenarios the MFRM approach (with the assumption of equal ability group means by rater) could produce estimates that were as close to the true abilities as those from the proposed MCMC approach. However, there were cases where the MFRM approach failed dramatically. The MFRM group anchoring approach is only as good as the quality of the assumption. The MCMC approach did not yield ability estimates that were less biased than the MFRM approach in general. However, the benefit of the MCMC approach appears to come in the form of more appropriately larger standard errors of measurement for the scores. The MCMC approach achieved the target 95 percent coverage probability in nearly all cases. To that end, the MCMC approach appears to be taking better account of the uncertainty due to rater severity and the non-random assignment of raters to examinees. In scenarios were there are high stakes for examinees, the MCMC approach is more cautious in indicating that examinees have scores which are reliably different from one another. The MCMC approach also takes greater care in indicating if examinees are either far below or far above average.

C. Conclusion

The Bayesian bivariate probit ordinal missing data measurement model proposed here provides a potential method for dealing with judge rated data in the presence of non-ignorable missing ratings. By treating rater assignment as a missing data problem, this modeling approach more accurately models the uncertainty in examinee scores by taking into better account the error due to rater severity, and non-random assignment of raters. In particular, the MCMC approach yields confidence intervals with better coverage probabilities than traditional approaches, and this finding is consistent when raters are randomly or non-randomly assigned to examinees. In addition, the MCMC approach provides a method for estimating the variance attributable to different facets of measurement, even when a traditional G-study with fully-crossed data has not

be implemented. While these techniques require significant computing time, as computing power continues to increase, they should be less time-consuming to implement for such measurement problems.

Future research can build upon the initial work on this topic presented in this thesis. In particular, the Bayesian ordinal probit missing data measurement model presented here could be expanded to models where the same rating scale structure is not used for all items (as in a partial credit model). Furthermore, this approach could be implemented with adjacent categories ordinal models (such as are traditional in the Rasch family of measurement models), rather than the cumulative logit models considered here. More work can also be done to better understand the conditions under which the G-Theory variance components obtained via the MCMC approach are best matched to those that would have been obtained under a traditional implementation of G-Theory with fully-crossed data. Finally, the algorithm could be updated to improve the speed of calculation. This step would allow for the generation of results for many simulated datasets, which would help demonstrate the generalizability of these results.

CITED LITERATURE

Adams, R., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Englehard, & M. Wilson (Eds.), *Objective measurement: Theory into practice -- vol. III* (pp. 296-321). Norwood, NJ: Ablex.

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Beale, E., & Little, R. (1975). Missing data in multivariate analysis. *Journal of the Royal Statistical Society, Series B, 37*, 129-145.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22*, 41-52.

Braun, H. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1-18.

Breland, H., & Jones, R. (1984). Perceptions of writing ability. *Written Communicaton, 1*, 101-119.

Brennan, R. L. (2001). *Generalizability Theory.* New York: Springer.

Briggs, D., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44*(2), 131-155.

Cronbach, L., Rajaratnam, N., & Gleser, G. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137-163.

Dean, M. (1980). Presentation order effects in product taste tests. *Journal of Psychology, 105*, 107-110.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Devroye, L. (1986). *Non-uniform random variate generation.* New York: Springer-Verlag.

Embretson, S. (1996). Item response theory and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*, 201-212.

Embretson, S. (2006). The continued search for nonarbitrary metrics in psychology. *Applied Psychological Measurement, 61*, 50-55.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum Publishers.

Englehard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

Englehard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

Englehard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Englehard, G. (1996). Evaluation rater accuracy in performance assessments. *Journal of Educational Measurement, 33*, 56-70.

Englehard, G., & Myford, C. (2009). Comparison of single- and double-assessor scoring designs for the Assessment of Accomplished Teaching. *Journal of Applied Measurement, 10*, 52-69.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*, 328-338.

Freedman, S., & Calfee, R. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Rosenthal, L. Tamor, & S. A. Walmsley, *Research on writing: Principles and methods* (pp. 75-98). New York: Longman.

Gelfand, A., & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398-409.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2009). *Bayesian data analysis - 2nd edition.* New York: Chapman and Hall/CRC.

Graham, J., & Donaldson, S. (1993). Evaluating interventions with differential attrition: the importatnce of nonresponse mechanisms and use of follow-up data. *Journal of Appled Psychology, 78*, 119-128.

Gruijter, D. d. (1984). Two simple models for rater effects. *Applied Psychological Measurement, 8*, 213-218.

Guilford, J. (1954). *Psychological methods.* New York: McGraw-Hill.

Gyagenda, I., & Englehard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement, 10*, 225-246.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction - 2nd edition.* New York: Springer.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97-109.

Hocking, R. (1996). *Methods and applications of linear models: Regression and the analysis of variance.* New York: John Wiley and Sons.

Holmes, C., & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis, 1*, 145-168.

Holmes, C., & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis, 1*, 145-168.

Houston, W., Raymond, M., & Svec, J. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15*, 409-421.

Hoyt, W. (1999). Magnitude and moderators of bias of observer ratings: A meta analysis. *Psychological Methods, 4*, 403-424.

Hoyt, W. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64-86.

Huang, L., Chen, M., & Ibrahim, J. (2005). Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates. *Biometrics, 61*, 767-780.

Ibrahim, J., Chen, M., Lipsitz, S., & Herring, A. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association, 100*, 332-346.

Kane, M., & Brennan, R. (1977). The generalizability of class means. *Review of Educational Research, 47*, 267-292.

Kim, S., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement, 10*, 408-423.

King, L., Hunter, J., & Schmidt, F. (1980). Haol in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psycyhology, 65*, 507-516.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoreing system for sample essays. *Applied Measurement in Education, 2*, 195-205.

Lang, W., & Wilkerson, J. (2005). Comparison of single- and double-assessor scoring designs for the Assessment of Accomplished Teaching. *Journal of Applied Measurement, 6*, 57-70.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (1995). Generalizability theory and many-facet Rasch measurement. In M. Wilson, K. Draney, & G. Englehard (Eds.), *Objective measurement: Theory into practice* (pp. 85-98). Norwood, NJ: Abex.

Linacre, J. M., & Wright, B. (2004). Construction of measures from many-facet data. In E. Smith, & R. Smith, *Introduction to Rasch measurement* (pp. 296-321). Maple Grove, MN: JAM Press.

Linacre, J. M., & Wright, B. (2004). Construction of measures from many-facet data. In E. Smith, & R. Smith (Eds.), *Introduction to Rasch measurement* (pp. 296-321). Maple Grove, MN: JAM Press.

Linn, R., & Miller, M. (2005). *Measurement and assessment in teaching, ninth edition.* Upper Saddle River, NJ: Pearson.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data - 2nd edition.* New York: Wiley.

Looney, M. (2004). Evaluating judge performance in sport. *Journal of Applied Measurement, 5*, 31-47.

Luntz, M., Wright, B., & Linarcre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

MacMillan, P. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education, 68,* 167-190.

MacMillian, P. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education, 68,* 167-190.

Maxwell, S., & Delaney, H. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97*, 85-93.

McIntyre, R., Smith, D., & Hassett, C. (1984). Accuracy of performance rating as affected by rater training and perceive purpose of rating. *Journal of Applied Psychology, 69*, 147-156.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1092.

Muckle, T., & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement, 46*, 198-219.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59-71.

Murphy, K., & Anhalt, R. (1992). Is halo error a property of the raters, ratees, or the spcific behaviors observed? *Journal of Applied Psychology, 72*, 494-500.

Murphy, K., & Jako, R. (1990). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. *Journal of Applied Psychology, 75*, 500-505.

Myford, C., & Wolfe, E. (2003). Understanding Rasch measurement: Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C., & Wolfe, E. (2004). Understanding Rasch measurement: Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227.

Oehlert, G. (2000). *A first course in design and analysis in experiments.* New York: W.H. Freeman.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics, volume 2b: Bayesian inference, 2nd edition.* New York: Oxford University Press.

Patz, R., & Junker, B. (1999a). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R., & Junker, B. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behaviora Statistics, 24*, 342-366.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Raudenbush, S., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2007). *The reliability of group-level measures and the power of group-randomized studies.* Whilliam T. Grant Foundation.

Raymond, M., & Roberts, D. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47*, 13-26.

Raymond, M., & Viswesvaran, C. (1983). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30*, 253-268.

Raymond, M., Webb, L., & Houston, W. (1991). Correcting performance-rating errors in oral examinations. *Evaluation and the Health Professions*, 100-122.

Raymond, M., Webb, L., & Houston, W. (1991). Correcting performance-rating errors in oral examinations. *Evaluation and the Health Professions, 14*, 100-122.

Rubin, D. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Shavelson, R., & Webb, N. (2006). Generalizability theory. In J. Green, G. Camilli, & P. Elmore (Eds.), *Handbook of complementary methods in education resarch* (pp. 309-322). Washington, D.C.: American Educational Research Association.

Smith, E., & Kulikovich, J. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement, 64*, 617-639.

Sudweeks, R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and many-facet Rasch measurement in the analysis of college sophmore writing. *Assessing Writing, 9*, 239-261.

Swanlund, A. (2012). Developing examinations that use equal raw scores for cut scores. *Journal of Applied Measurement, 25*, 50-62.

Thorndike, R. (2005). *Measurement and evaluation in psychology and education: 7th edition.* Upper Saddle River, NJ: Pearson.

Webb, N., Shavelson, R., & Haertel, E. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 1-44.

Wilson, H. (1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement, 48*, 69-81.

Wolfe, E. (1997). Identifying rater effects using latent trait models. *Psychology Science, 4*, 83-106.

Wolfe, E. (2004). Identifying rater effects using latent trait models. *Psychology Science, 4*, 83-106.

Wolfe, E. (2009). Understanding Rasch measurement: Item and rater analysis of constructed response items via the multi-faceted Rasch Model. *Journal of Applied Measurement, 10*, 335-347.

Wolfe, E., & Smith, E. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II - validation activities. In E. Smith, & R. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243-290). Maple Grove, MN: JAM Press.

Wolfe, E., Kao, C., & Ramney, M. (1998). Cognitive differences in proficient and nonproficient essay scores. *Written Communication, 15*, 465-492.

Wolfe, E., Moulder, B., & Myford, C. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*, 256-280.

Wright, B., & Lincare, J. M. (1989). Observations are always ordinal; Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857-960.

Wright, B., & Masters, G. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B., & Stone, M. (1979). *Best test design.* Chicago: MESA Press.

APPENDIX

This appendix provides example code of the implementation of the Bayesian bivariate probit missing data model as implemented in Matlab

```
%% Import the data
[~, ~, raw] = xlsread(datasetname.xlsx','sheetname');
raw = raw(2:end,:);

%% Replace non-numeric cells with 0.0
R = cellfun(@(x) ~isnumeric(x) || isnan(x),raw); % Find non-numeric cells
raw(R) = {0.0}; % Replace non-numeric cells

%% Create output variable
data = cell2mat(raw);
%% Allocate imported array to column variable names
personid = data(:,1);
judgeid = data(:,2);
itemid = data(:,3);
ability = data(:,4);
ability1 = data(:,5);
difficulty = data(:,6);
severity = data(:,7);
rating = data(:,8);
rating1 = data(:,9);
txct = data(:,10);
ab.strata = data(:,11);
keep = data(:,13);

missing = keep;
missing(keep==1) = 0;
missing(keep==0) = 1;


%% Clear temporary variables
clearvars data raw R columnIndices;

%% Create the design matrix X
X = dummyvar([personid itemid judgeid]);


% Create the response vector, Z, and starting values for missing data
Y = rating;
m = missing;


[n, p] = size(X);
np = 50;
ni = 5;
nj = 5;
```

```
% need random starting values
randstart = round(unifrnd(1,4,n,1));
Y(m==1) = randstart(m==1);

Z = [Y X];

[n, pz] = size(Z);
Zprop = Z;

% Number of iterations
niter = 100000;
% Number of response options
Q = 4;


% set up parameter vectors
beta = zeros([p niter]);
psi = zeros([pz niter]);
gamma = zeros([Q+1 niter]);
gamma(1,:) = -Inf;
gamma(2,:) = 0;
gamma(3,1) = 1.5;
gamma(4,1) = 3;
gamma(Q+1,:) = Inf;

% Auxilliary variable vectors
Ystar = zeros([n 1]);
mstar = zeros([n 1]);
Ystarmiss = zeros([n 1]);

% establish the prior information for beta and psi
mu_beta = zeros([p 1]);
mu_psi = zeros([pz 1]);

var_beta = eye(p);
var_psi = eye(pz);

sigmaY = ones([n 1]);
sigmam = ones([n 1]);

mtrunclow = [-Inf 0];
mtrunchi = [0 Inf];

Vbetastar = inv(inv(var_beta)+X'*X);
invVbeta = inv(var_beta);
invVpsi = inv(var_psi);

% MCMC Loop
v_beta = ones([1 niter]);
v_psi = ones([1 niter]);
sigma2_p = zeros([1 niter]);
sigma2_i = zeros([1 niter]);
sigma2_j = zeros([1 niter]);
sigma2_pi = zeros([1 niter]);
sigma2_pj = zeros([1 niter]);
sigma2_ij = zeros([1 niter]);
sigma2_pij = zeros([1 niter]);
```

```
gcoef = zeros([1 niter]);

for k=2:niter

% Step 1: Draw proposed v_beta:

v_betaProp = exp(normrnd(log(v_beta(k-1)),1));
LL1 = sum(log(normpdf(beta(:,k-1),0,sqrt(v_betaProp)))) +
log(real(v_betaProp<1000));
LL0 = sum(log(normpdf(beta(:,k-1),0,sqrt(v_beta(k-1))))) + log(real(v_beta(k-
1)<1000));
if rand<exp(LL1-LL0); v_beta(k)=v_betaProp; else v_beta(k)=v_beta(k-1); end;

% Step 2: Draw proposed v_phi:

v_psiProp = exp(normrnd(log(v_psi(k-1)),1));
LL1 = sum(log(normpdf(psi(:,k-1),0,sqrt(v_psiProp)))) +
log(real(v_psiProp<1000));
LL0 = sum(log(normpdf(psi(:,k-1),0,sqrt(v_psi(k-1))))) + log(real(v_psi(k-
1)<1000));
if rand<exp(LL1-LL0); v_psi(k)=v_psiProp; else v_psi(k)=v_psi(k-1); end;


% Step 3: Sample Ystar

    Ystar = norminv(normcdf(gamma(Y,k-1),X*beta(:,k-
1),sigmaY)+(unifrnd(0,1,n,1).*(normcdf(gamma(Y+1,k-1),X*beta(:,k-1),sigmaY)-
normcdf(gamma(Y,k-1),X*beta(:,k-1),sigmaY))),X*beta(:,k-1),sigmaY);

% Step 4: Sample mstar

    mstar = norminv(normcdf(mtrunclow(m+1)',Z*psi(:,k-
1),sigmam)+(unifrnd(0,1,n,1).*(normcdf(mtrunchi(m+1)',Z*psi(:,k-1),sigmam)-
normcdf(mtrunclow(m+1)',Z*psi(:,k-1),sigmam))),Z*psi(:,k-1),sigmam);

% Step 5: Metropolis step for threshold parameters (1, 2, and 5 are
    % constant)

    gammaprop = gamma(:,k-1);
    gammadenom = gamma(:,k-1);
    for q=3:Q
        gammaprop(q) = normrnd(gamma(q,k-1),.05);
        if gammaprop(q) > gammaprop(q-1) & gammaprop(q) < gammaprop(q+1)
            if unifrnd(0,1) <=
min(1,exp(sum(log(normcdf(gammaprop(Y+1),X*beta(:,k-1),sigmaY)-
normcdf(gammaprop(Y),X*beta(:,k-1),sigmaY)))-
sum(log(normcdf(gammadenom(Y+1),X*beta(:,k-1),sigmaY)-
normcdf(gammadenom(Y),X*beta(:,k-1),sigmaY)))));
                gammadenom(q) = gammaprop(q);
            else gammaprop(q) = gamma(q,k-1);
            end
        else gammaprop(q) = gamma(q,k-1);
        end
    end
    gamma(:,k) = gammadenom;
```

```
% Step 6: Sample the betas

    beta(:,k) =
mvnrnd(inv(inv(v_beta(k)*eye(p))+X'*X)*(inv(v_beta(k)*eye(p))*mu_beta +
X'*Ystar),inv(inv(v_beta(k)*eye(p))+X'*X));

% Step 7: Sample the psis

    psi(:,k) =
mvnrnd(inv(inv(v_psi(k)*eye(pz))+Z'*Z)*(inv(v_psi(k)*eye(pz))*mu_psi +
Z'*mstar),inv(inv(v_psi(k)*eye(pz))+Z'*Z));

% Step 8: metropolis the ystar for the missing ratings, and generate
    % missing values

    Ystarmiss = Ystar;
    Ystarmiss(m==1) = normrnd(Ystar(m==1),1);
    Zprop = Z;
    Zprop(m==1 & Ystarmiss < gamma(2,k),1)=1;
    Zprop(m==1 & Ystarmiss < gamma(3,k) & Ystarmiss >= gamma(2,k),1)=2;
    Zprop(m==1 & Ystarmiss < gamma(4,k) & Ystarmiss >= gamma(3,k),1)=3;
    Zprop(m==1 & Ystarmiss >= gamma(4,k),1)=4;

    randdraw = unifrnd(0,1,n,1);
    proplikelihood =
(normpdf(Ystarmiss,X*beta(:,k),sigmaY).*normpdf(mstar,Zprop*psi(:,k),sigmam))
./(normpdf(Ystar,X*beta(:,k),sigmaY).*normpdf(mstar,Z*psi(:,k),sigmam));
    Z(m==1 & randdraw <= min(1,proplikelihood),1) = Zprop(m==1 & randdraw <=
min(1,proplikelihood),1);
    %G = tabulate(Z(m==1,1));
    %distofY(:,k) = G(:,3);
    Y = Z(:,1);
    % insert G theory here


pmean = zeros([np 1]);
imean = zeros([ni 1]);
jmean = zeros([nj 1]);
pimean = zeros([np ni]);
pjmean = zeros([np nj]);
ijmean = zeros([ni nj]);
for ip = 1:np
    pmean(ip) = mean(Y(personid==ip));
    for ii = 1:ni
        pimean(ip,ii) = mean(Y(personid==ip & itemid==ii));
    end
    for ij = 1:nj
        pjmean(ip,ij) = mean(Y(personid==ip & judgeid==ij));
    end
end
for ii = 1:ni
    imean(ii) = mean(Y(itemid==ii));
    for ij = 1:nj
        ijmean(ii,ij) = mean(Y(itemid==ii & judgeid==ij));
    end
end
for ij = 1:nj
```

```
    jmean(ij) = mean(Y(judgeid==ij));
end
gmean = mean(Y);

ss_p = ni*nj*sum(pmean.^2) - np*ni*nj*(gmean^2);
ss_i = np*nj*sum(imean.^2) - np*ni*nj*(gmean^2);
ss_j = np*ni*sum(jmean.^2) - np*ni*nj*(gmean^2);
ss_pi = nj*sum(sum(pimean.^2)) - ni*nj*sum(pmean.^2) - np*nj*sum(imean.^2) +
np*ni*nj*(gmean^2);
ss_pj = ni*sum(sum(pjmean.^2)) - ni*nj*sum(pmean.^2) - np*ni*sum(jmean.^2) +
np*ni*nj*(gmean^2);
ss_ij = np*sum(sum(ijmean.^2)) - np*nj*sum(imean.^2) - np*ni*sum(jmean.^2) +
np*ni*nj*(gmean^2);
ss_pir = sum(Y.^2) - nj*sum(sum(pimean.^2)) - ni*sum(sum(pjmean.^2)) -
np*sum(sum(ijmean.^2)) + ni*nj*sum(pmean.^2) + np*nj*sum(imean.^2) +
np*ni*sum(jmean.^2) - np*ni*nj*(gmean^2);

ms_p = ss_p/(np-1);
ms_i = ss_i/(ni-1);
ms_j = ss_j/(nj-1);
ms_pi = ss_pi/((np-1)*(ni-1));
ms_pj = ss_pj/((np-1)*(nj-1));
ms_ij = ss_ij/((ni-1)*(nj-1));
ms_pij = ss_pir/((np-1)*(ni-1)*(nj-1));

sigma2_p(k) = (ms_p-ms_pi-ms_pj+ms_pij)/(ni*nj);
sigma2_i(k) = (ms_i-ms_pi-ms_ij+ms_pij)/(np*nj);
sigma2_j(k) = (ms_j-ms_pj-ms_ij+ms_pij)/(np*ni);
sigma2_pi(k) = (ms_pi-ms_pij)/nj;
sigma2_pj(k) = (ms_pj-ms_pij)/ni;
sigma2_ij(k) = (ms_ij-ms_pij)/np;
sigma2_pij(k) = ms_pij;

gcoef(k) = sigma2_p/(sigma2_p +
((sigma2_pi/ni)+(sigma2_pj/nj)+(sigma2_pij/(ni*nj))));
end


abilests = mean(beta(1:50,10001:niter),2);
judges = mean(beta(56:60,10001:niter),2);
items = mean(beta(51:55,10001:niter),2);
psiests = mean(psi(:,10001:niter),2);
gamma = mean(gamma(:,10001:niter),2);

se_abil = std(beta(1:50,10001:niter),0,2);
se_judges = std(beta(56:60,10001:niter),0,2);
se_items = std(beta(51:55,10001:niter),0,2);

sigma2_p_mean = mean(sigma2_p(10001:niter));
sigma2_i_mean = mean(sigma2_i(10001:niter));
sigma2_j_mean = mean(sigma2_j(10001:niter));
sigma2_pi_mean = mean(sigma2_pi(10001:niter));
sigma2_pj_mean = mean(sigma2_pj(10001:niter));
sigma2_ij_mean = mean(sigma2_ij(10001:niter));
sigma2_pij_mean = mean(sigma2_pij(10001:niter));
gcoef_mean = mean(gcoef(10001:niter));
```

```
sigma2_p_se = std(sigma2_p(10001:niter));
sigma2_i_se = std(sigma2_i(10001:niter));
sigma2_j_se = std(sigma2_j(10001:niter));
sigma2_pi_se = std(sigma2_pi(10001:niter));
sigma2_pj_se = std(sigma2_pj(10001:niter));
sigma2_ij_se = std(sigma2_ij(10001:niter));
sigma2_pij_se = std(sigma2_pij(10001:niter));
gcoef_se = std(gcoef(10001:niter));

abilities = [abilests se_abil];
judgeseverities = [judges se_judges];
itemdifficulties = [items se_items];

gtheory = [sigma2_p_mean sigma2_p_se;
    sigma2_i_mean sigma2_i_se;
    sigma2_j_mean sigma2_j_se;
    sigma2_pi_mean sigma2_pi_se;
    sigma2_pj_mean sigma2_pj_se;
    sigma2_ij_mean sigma2_ij_se;
    sigma2_pij_mean sigma2_pij_se;
    gcoef_mean gcoef_se;]
```