Unidimensional Models Do Not Fit Unidimensional Mixed Format Data Better than Multidimensional Models

By

Melinda Montgomery

Submitted to the Department of Psychology and Research in Education and the
Faculty of the Graduate School of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Accepted by:
]William P. Skorupski, EdD, Chairperson
Brue Frey, PhD
Neal Kingston, PhD
Vicki Peyton, PhD
Susan Twombly, PhD

December 2014

Abstract

This dissertation examines the scaling of large scale assessments containing both dichotomous and polytomous items, mixed format assessments. Because large scale assessments are generally built to measure one construct, e.g. eighth grade mathematics, unidimensional data was generated to simulate a mixed format assessment. The test length, number of polytomous to dichotomous items per assessment and the discrimination level between dichotomous and polytomous items were varied in this study. There were five item combinations and two level of discrimination defined.

The goal of this dissertation was to compare the fit of the generated data to three different Item Response Theory models; one unidimensional and two multidimensional. The first model used to fit the data was the same model type used to generate the data; a 3PL IRT model in combination with the Generalized Partial Credit model. The second model was the Hierarchical MIRT Model. The final model was the bi-factor model. The research questions examined in this study were; (1) Which of the models achieves the best model fit across simulation conditions?, and (2) Do the variables of item combination or discrimination affect the model fit?

The study showed that the bi-factor model fit unidimensional data, in mixed format, better than either the unidimensional or the hierarchical MIRT models. The criterion used to make this determination was the Bayesian convergence criterions; BIC, DIC and AIC. Overall, the bi-factor model fit the unidimensional mixed format data better than the generating model fit the data. The hierarchical MIRT model did not fit the data very well, and in a few cases, did not converge. The more polytomous item

included on the assessment the better the bi-factor model improved overall fit over the unidimensional model.

This result suggests that noise in the data from mixed format assessments can cause the unidimensional models to fail to fail to fit the data. This study illustrates the format alone can create the appearance of dimensionality. However since the data was generated as unidimensional, this format dimensionality affect was an attribute of the data alone, not of items or examinees interactions with the items. Mixed format assessments create an artifact in the data that causes the data to factor into dimensions that are not actually present. It appears there is noise in the data of mixed format assessment that needs to accounted for when scaling.

Acknowledgements

I would first like to thank God for the many blessing that have made this dissertation completion possible. If it were not for my faith during these past few years, this dissertation would not have been possible. I would also like to thank my church family for their prayers and support.

I must also thank my dissertation chairperson and advisor, William Skorupski, whose teaching and inspiration was the catalyse for this project. He encouraged me to pursue this degree path and provided invaluable support and encouragement during my entire program and especially during the dissertation process. I would also like the thank the other members of my committee Dr. Bruce Frey, Dr. Neal Kingston, Dr. Vicki Peyton, and Dr. Susan Twombly. You have each inspired me in various ways through the classes you taught, the projects we worked on and through a variety of interaction where each of you have supported and encouraged me. Without that support and encouragement, this would not have been possible.

I would like to thank Minho Scholle and the Mathematics Stack Exchange for help with the proof that the sum of quotients was not equal to the quotient of the sums. This proof was part of the methodology in determining the amount of error in the parameter estimates.

Finally, I would like to thank Dr. Paul Johnson, from the Center for Research Methods and Data Analysis at the University of Kansas, whose assistance with the Rocks cluster of Linux computer nodes was invaluable to completion of this project.

Contents

List of Figures

List of Tables

List of Equations

Chapter 1

Introduction

Education assessment is an integral part of the national debate on educational reform. The heart of the issue is a common goal that all children receive the best education possible. But, how do we measure that? What kind of assessment is needed to determine whether or not students are meeting educational goals. Most assessments are designed to measure how much students know about a particular subject: e.g. mathematics or English. But, could we design assessments that measure that ability more fully or more accurately so that students and educators have a better measure of students strengths and weaknesses. Some stakeholders suggest that part of the solution is to create better, more "authentic" assessments. This is not a new initiative. Dating back to World War II, there have been several waves of assessment reform (Linn, 2000). The more recent wave of assessment reform involves the inclusion of performance based assessments or at the very least more performance based items (Linn, 2000). Madaus and O'Dwyer (1999) claim that standardized multiple choice assessments are 'out' in popular and profession literature and that more "authentic" performance assessments are "in." Exactly what is meant by "authentic" performance assessment varies according the stakeholder involved, but it is clear that the current wave of assessment reform includes a move away from assessments containing only multiple-choice items and toward assessments that include more innovative and/or constructed response items.

Some argue that the inclusion of innovative items provides more information about examinees' understanding and true ability than multiple choice items alone (Bennett, Morley, & Quardt, 2000;

A. L. Zenisky & Sireci, 2001). Innovative items can range from "drop and drag" to "formulating hypothesis" (A. Zenisky & Sireci, 2002). In fact, A. Zenisky and Sireci (2002) defined twenty-one different polytomously scored innovative item types used in various testing situations, although most commonly in licensure exams. Particularly in the area of licensure exams, the variety of innovative items is designed to create assessments that providing examinees a variety of different methods to demonstrate their understanding, knowledge, and ability on the construct being measured.

While there is overlap between multiple choice and constructed response items, in terms of the processes they assess, constructed response items can provide additional information about examinee cognitive processes which cannot be easily duplicated with multiple choice items (Bennett, Rock, & Wang, 1991). The decision about which type of items to use on assessment depends on the intended purpose of the assessment outcome, mastery or a more refined evaluation of the examinee abilities (Bennett et al., 1991). If the purpose is merely to determine mastery, then a multiple choice assessment may provide all of the information required (Bennett et al., 1991). However, the trend in the educational assessment today is to use assessments to evaluate examinee abilities on a continuum rather than to determine mastery.

Education Testing Services (ETS), the SMARTER Balanced Assessment Consortia and The Partnership for Assessment of Readiness for College and Careers (PARCC) Consortia, consider innovative item types an important part of the next generation of assessments. A quick look at the ETS website's research section will find Cognitively Based Assessment of, for, and as Learning (CBAL) in addition to several current research projects on innovative assessments (ETS). One large component driving the push to integrate innovate item into large scale assessments is driven by the Common Core State Standards (CCSS). CCSS requires examinees to be assessed on more critical thinking and analytical thinking skills in the two categories: college and career readiness standards, and K-12 standards (*CCSS Process*, 2009). As stated above, innovative or constructed response items may provide a better means to assess these desired critical thinking skills.

While discussion of innovative items dates back to the nineties, the ability to create and score

those item types had been cost prohibitive. Advancements in assessment technology has not only increased the ability to machine score items (Madaus & O'Dwyer, 1999), but is also making it possible to create a variety of innovative items more efficiently. Of the twenty-one different item types defined by A. Zenisky and Sireci (2002), the range of difficult as well as the possibility of computer scoring the items varies widely. Items such as drag-and-drop, inserting text, or sorting might be easily computer scored, whereas items such as generating examples and analyzing situations or writing essays are less conducive to computer scoring. A. L. Zenisky and Sireci (2001) indicated that the best choice of scoring routines used on innovative item types remains unanswered.

1.1 Background of the Study

How to scale the next generation of assessments in an important ongoing topic. It is likely that future large scale assessments will contain both dichotomously scored as well as polytomously scored items. The choice of an incorrect model can result in incorrect conclusions with respect to parameter estimation and person fit (Kang & Cohen, 2007; Kang, Cohen, & Sung, 2009). While issues of parameter estimation, person fit as well as topic of equating are well understood for most unidimensional IRT models, the added complexity of a mixed format assessment requires further study. The choice of model that will provide the most valid and generalizable results is a growing focus of educational researchers. In particular, several studies have considered the use of item response theory (IRT) models, bi-factor models, testlet models and hierarchical IRT models to scale mixed format assessments (Cai, Yang, & Hansen, 2011; DeMars, 2006; Reise, Morizot, & Hays, 2007; Rijmen, 2010; Whittaker, Chang, & Dodd, 2012).

Details about these research studies will be presenting in the literature review section. As an overview, Cai et al. (2011) proposed an extended item bi-factor analysis framework and conducted a study on how this framework could handle item responses from multiple groups, with dichotomous, ordinal, and nominal response formats. The extended bi-factor model allows some items to load onto the general factor without loading onto one of the specific factors. They found that the generalized bi-factor models reliably fit the data with little bias and reported no convergence issues.

Cai et al. (2011) suggest that the generalized framework can be used to study dimensionality in data as wells as bi-factor based linking and equating studies. Based on the (Cai et al., 2011) study, this study examined a dimensionality affect resulting from the scoring associated with multiple formats on an assessment by utilizing the bi-factor model.

The DeMars (2006) study favored the more parsimonious testlet-effects model over the bi-factor model. However, the authors also stated that the speed at which the bi-factor model can be calibrated, in comparison to the testlet-effects model, might be of benefit to practitioners. In a study that compared the hierarchical MIRT, testlet and bi-factor models to real data, Rijmen (2010) found that the proportionality restrictions imposed by the hierarchical MIRT model were too stringent. The better fit of the bi-factor model suggests that practitioners might reconsider the tendency to use the testlet model over the bi-factor (Rijmen, 2010).

The study conducted by Whittaker et al. (2012) considered the accuracy of six model selection methods ability to choose the correct IRT models for mixed format data. They found that the proportion of polytomously scored to dichotomously scored items had an effect the accuracy of model selection. In particular, the 2PL Item Response Theory (IRT) model combined with the Generalized Partial Credit Model was correctly selected more often when the assessment consisted of more polytomously scored items than dichotomously scored items. They also found that sample size played a role in the accuracy of model selection.

These finding supported the findings in a preliminary study conducted by Montgomery and Skorupski (2012) which also found that the proportion of polytomously scored items to dichotomously scored items as well as sample size played a role in the rate of convergence in mixed format data fit to combined unidimensional IRT models. This finding led the author to consider the possibility of either dimensionally or simply data noise resulting from item format. If there is in fact noise in the data resulting from mixed format alone, it could cause the models to fail to fit the generating model in favor of more complex models that account for that noise. The findings by Cai et al. (2011); DeMars (2006); Rijmen (2010) indicating that the bi-factor model showed promise in fitting a variety of mixed format assessments and might be useful identifying any dimensionally

affect inherent in assessments of this type.

It is important to understand the underlying assumptions of Item Response theory models and higher order models used to model mixed format assessments. The following sections will explain some of the assumptions and complications with the unidimensional and hierarchical models.

1.2 IRT Model

Item response theory (IRT) models utilize a nonlinear, logistic model, based on the item difficulty, item discrimination, a parameter for guessing and an optional constant used in scaling (De Ayala, 2009). One benefit of IRT is that the analysis is at the item level as opposed to Classical Test Theory which is at the test score level. IRT links the item with the examinees ability whereas Classical Test Theory places examinee ability on the total score metric. In IRT the examinees' ability is placed on a scale from negative infinity to positive infinity allowing for a more accurate view of examinee ability based on the probability of answer the item correctly given the item difficulty and discrimination. The result is an ability score that is generalizable, a very desirable outcome. Given the advantages of IRT models in large scale assessments it is reasonable that testing organizations would first look to these models when scaling mixed format assessments. However, there are some important assumptions that must be considered.

1.2.1 IRT Assumptions

Two important, and related, assumptions when considering unidimensional IRT models are that the assessments are unidimensional and the items are locally independent. Unidimensionality is defined as independence of item responses after controlling for the a single latent variable (Reise et al., 2007). The assumption of unidimensionality simply means that the assessment measures only one construct. It has been argued that a slight deviations from unidimensionality, provided that the assessment is designed to assess the same general construct, continues to establish essential unidimensionality (Strout, 1990).

Some have suggested that there might be enough dimensionality created by the difference in item types (dichotomous/polytomous) in a mixed format assessment that the unidimensionality assumption has been violated to the point that a multidimensional approach is a better choice for scaling the assessment(Ercikan et al., 2005). A multidimensional IRT model allows the assessment to consist of more than one dimension and provides a method for assigning an overall ability score. It is possible that the introduction of constructed response items might introduce a level of multidimensionality in the form of noise in the data from the different score values. How much of this noise can be tolerated without violating the unidimensionality assumption is unclear. Too much noise int he data could result from construct irrelevant variance related to the item format alone or from item complexity. If true, that variance needs to be accounted for and factored out before an overall ability score is assigned.

1.2.2 IRT Model Combinations

It is well known that the properties of IRT models provide useful and desirable features to large-scale assessments. We count on the invariant property of IRT models that afford sample independence of item and person parameters, provided the model fits the data. IRT models can be used in combination with other IRT models, as was done in this study. The combined models considered for this study were created by paring a 3PL dichotomously scored model with the Generalized Partial Credit model (GPC) (Muraki, 1992).

There have been several recent studies evaluating the model fit and parameter recovery of mixed format assessments using the GPC and one or more of the dichotomous models (Chon, Lee, & Dunbar, 2010; Whittaker et al., 2012). The Chon et al. (2010) article consider item fit statics and found that sample size and test length was related to the performance of item fit statistics. They also found that they model type affected fits statics. In particular, the 3PL/GPC model demonstrated slightly higher error rates. However, based on a previous study, the 3PL/GPC model combination was selected for this study because it has been found to maintain a high convergence rate and has also demonstrated low bias and low RMSE(Montgomery & Skorupski, 2012).

1.2.3 Format Effect

There may also be a format effect that results from the variety of score values attributed to each set of item types. This format affect may cause a combined IRT model to fail to fit the data. Hence, it may turn out that a multidimensional model such as a hierarchical MIRT (mathematically equivalent to a testlet model) or a bi-factor model would be a better choice in scaling assessments with this type of complexity. If there is a format affect, not only is it important to find a model that will accurately fit the data, but also a model that might be used to better explain the complexity. Modeling the data is such a way as to highlight the format differences could provide a more complete picture of examinee ability. Format effect will be discussed more completely in Section 2.3.1 of the literature review.

1.2.4 Dimensionality

There are a number of methods that can be used to determine whether or not an assessment is unidimensional: inspection of the ratio of the fist and second eigenvalues, inspection of the residual distribution after one factor has been extracted, examining scree plots, and a confirmatory factor analysis (Reise et al., 2007). Another method that can be used is the bi-factor model. This model deviated from a typical factor analysis by allowing each item to have a positive loading onto a general trait in addition to allowing each item to load onto a group factor. More information about how the bi-factor model can be used to establish dimensionality will be provided in the literature review section.

So the question remains, in data from a mixed format assessment, will a multidimensional or a unidimensional model provide the most reliable information about examinee's performance on each item and on the examinees' overall ability? Since this study utilized simulate data only, the issue of scoring differences based on item type was the criteria considered. This study sought to answer that question by looking at three specific types of models: combined unidimensional model (3PL/GPC), hierarchical MIRT, and the bi-factor model.

1.3 Hierarchical Model

The hierarchical IRT model is one option for fitting data from mixed format assessments. The hierarchical model is mathematically equivalent to the testlet model. Assuming a standard normal distribution for the latent variables, the hierarchical MIRT model can be thought of as a restricted bi-factor model in that the loading on the specific dimensions are proportional to the loadings on the general dimension (Rijmen, 2010). Loadings in this case refers to the Item Response Theory a-parameters rather than loadings a factor analysis.

The hierarchical model contains both a general dimension and a specific dimension, just like the testlet and bi-factor models, but the items do not depend directly on the general dimension (Rijmen, 2010). In this model, item depend directly on the specific dimension which in term dependent upon the general dimension. This model assumes that the specific dimensions are conditionally independent and all associations between the specific dimensions are accounted for by the general dimension.

1.4 Bi-factor Model

In the bi-factor model, each item is dependent upon both the specific dimension and the general dimension (Holzinger & Swineford, 1937). The general dimension stands for the latent variable of central interest such as polynomials in an algebra class. The K other dimensions take into account additional dependencies such as format effect. In this model there are J items in which individual items load onto the general dimension and J_k items, $k = 1, \ldots, K$, that load onto the K specific dimensions.

The major difference between the bi-factor model and the hierarchical model is that, in the hierarchical model, the specific dimension are explained by the general dimension. By comparison, in the bi-factor model, the specific dimensions are not explained by the general dimension but by the items clusters alone.

1.5 Statement of the Problem

While the most commonly selected method of scaling mixed format assessment in large scale testing programs most likely utilizes combined unidimensional IRTs such as a 2PL for the dichotomous items and a GPC for the polytomous items, there are issues with these combined models. Several studies have found that the ability to select the model that fits the data best, as well as the ability to choose the correct model, is influenced by the IRT model selected, test length, sample size, and the proportion of score points attributed to polytomous and dichotomous items (Chon et al., 2010; Whittaker et al., 2012).

In addition to the ability to choose the best model to fit the data, there is also the question of whether or not the format of the items, or just the scoring associated with the format, creates dimensionality that is large enough to require a more complex model than the combined unidimensional IRT models. However, is unidimensional data is generated and a more complex model fits the data better, this may instead indicate that there is enough noise in mixed format data that is unexplained.

1.6 Purpose of the Study

The purpose of this study is to compare the fit of three types of IRT models to unidimensional data generated in mixed format: a combined unidimensional model, a bi-factor model and a hierarchical IRT model. This study used Fortran to generate unidimensional data based on the distributions of the parameters defined in the methods section. A Bayesian approach was then used to fit the data to the three models.

The model that fits the data best, based on the fit indices defined in the methods section will be considered the better fitting model. This study also considered the issue of dimensionality through the use of a bi-factor model. This bi-factor model created specific dimensions based on item format. The loading structure of both the bi-factor model and hierarchical MIRT model was used to determine if there is important information that can be obtained about the data structure

based on the item format.

1.7 Hypothesis

A unidimensional IRT model assumes an underlying unidimensional structure. A bi-factor model allows for the possibility of multiple dimensions by allowing each items to load onto the specific factors and to the specific factor. The hierarchical model also allows for the possibility of multiple dimensions by allowing the items to load on the specific factors and for the specific factors to load onto the general factor.

Literature has shown that model selection can be impacted by sample size and the proportion of polytomously scored items to dichotomously score items (Whittaker et al., 2012). (Rijmen, 2010) found that the bi-factor model, in particular, fit the data from mixed format assessments and should be considered more frequently.

Based on the belief that dimensionality, or noise, is present in mixed format assessments even when the data is generated as unidimensional, this study sought to determine whether or not there was enough unexplained dimensionality or noise for the data to fail to fit its generating model. The following hypothesis serve as the underlying premise under which data was evaluated.

Hypothesis. The bi-factor model will fit the unidimensional mixed format data better than the unidimensional IRT model or the hierarchical model. Since the specific dimensions in the bi-factor model are not accounted for by the general dimension, any format effect dimensionality will be evident in the loading structure of the bi-factor model.

This study fit a unidimensional model, bi-factor model, and a hierarchical IRT model to several variations of mixed format assessments. The two research questions examined in this study are:

1. How well does the unidimensional model recover item and examinee parameters across the simulation conditions?

2. Which model fits the unidimensional data best: 3PL/GPC, bi-factor, or hierarchical MIRT?

3. Is the model fit affected by the proportion of dichotomous to polytomous items or by the level of discrimination?

1.8 Definitions of Variables

The definitions of terms used in this study are summarized as follows:

Item Response Theory

Item Response Theory (IRT) models the correspondence between the latent variables of examinee ability and item difficulty, discrimination, and guessability as predictors of observed responses.

- 3PL Model

  The 3PL model is an IRT model that uses the latent variables of examinee ability, difficulty, discrimination and guessability as predictors of observed responses. The 3PL models dichotomous responses only.

- Generalized Partial Credit Model (GPCM)

  The GPCM is an IRT model that uses the latent variables of examinee ability, difficulty and discrimination as predictors of observed responses. This model is similar to the 2PL with the exception that this model is designed to model the latent variables from polytomous predictor variables. It divided the range of possible item scores into categories and models the probability of one category over another as a function of examinee ability.

Hierarchical Model

The hierarchical model is a multidimensional IRT model where each cluster of items, testlet, represents a specific dimension. Each item depends on the specific domain but do not depend directly upon the general dimension. The specific dimensions depend upon the general

dimension. This means that all associations between the specific dimensions is accounted for by the general dimension (Rijmen, 2010)

Bi-factor Model

The bi-factor model consists of more than one dimension: a general dimension and K other specific dimensions. The general dimension represents the overall latent construct (e.g. Solving polynomials) where the specific dimensions represent clusters of attributes that make up the general dimension (e.g. factoring, quadratic formula)(Rijmen, 2010). In this model items depend on both the specific and general dimensions.

1.9 Summary and Significance

The motivation for this study is the increasing desire of test developers to build and accurately score assessments that contain both dichotomously and polytomously scored item. Many testing companies may opt to use a mixed unidimensional IRT model for the purpose of scoring mixed format assessments. There has been concern in recent years that assessments that consist of both dichotomous and polytomous items may be inherently multidimensional and thus require a multidimensional approach to scaling.

If dimensionality is created by mixing item format, then this must be account for when scaling the assessment. Fitting a unidimensional model to multi-dimensional data could result in examinee score bias. Particulary in state assessments or other high stakes assessments this bias could result in misclassification of examinees. The bi-factor and multidimensional models may provide a method for accounting for the format affect and provide less bias in examinee scores.

Chapter Two will review the existing studies including two specific studies that informed this research, a study by Rijmen (2010) which compared the Bi-factor, testlet and hierarchical multidimensional IRT models and a study by Cai et al. (2011) that compared the Bi-factor model with a mixed unidimensional IRT models. Additionally this section will provide background for mixed format assessment and motivation for the simulation study approach incorporated by this study.

Chapter Three defines model specifications for each of the models: 3PL/GPCM, Bi-factor, Hierarchical MIRT. Next, this chapter details the simulation including defining the independent variable and a discussion of the Bayesian estimation approach utilized in this project.

Chapter Four details the findings of the study including comparing the Bayesian model fit criterion, monte carlo error and loading structure differences across models.

Finally, Chapter five will recap the study methodology and results as well as a discussion of limitations and implications.

Chapter 2

Literature Review

This literature review will discuss research on the different types item response theory (IRT) models that are commonly used on educational assessments in general and on mixed format assessments specifically. Issues of dimensionality in mixed format assessments and its affect on the of IRT scaling models will be discussed. Several authors have noted the possibility of a format affect resulting from assessments that consist of items that are scored both as polytomously and dichotomously. Finally this literature review will discuss methods that can be used to check dimensionality. The purpose of this literature review is to establish this study in the existing literature and illustrate the gaps in the literature that this study seeks to fill.

2.1 IRT Models

There are a number of possible choices for scaling mixed format assessments including unidimensional and multidimensional models. The combination could include the dichotomous response models including the 1PL, 2PL or a 3PL model along with polytomous models such as the partial credit model (PCM)(Masters, 1982), the generalized partial credit model (GPCM) (Muraki, 1992) or the graded response model (Samejima, 1997). These partial credit models can be viewed as nested models (Chon et al., 2010). The 1PL/PCM is nested within the 2PL/GPCM which is in turn nested within the 3pl/GPCM. When the slope parameters are constrained across items, the GPCM reduces to the PCM (Chon et al., 2010). However, equal discrimination is unlikely partic-

ularly in mixed format assessments (Sykes & Yen, 2000). The Rasch model, generalized as the Partial Credit Model (Masters, 1982; Muraki, 1992), takes a cumulative approach to partial credit scoring in that the examinee must answer the first part correctly in order to have the possibility of receiving points on subsequent parts. In other words, if the examinee makes an error in the first part of the calculation there is no partial credit awarded and the item is counted as incorrect. This type of scoring is typically used in mastery assessments such as the registered nursing exams (Julian, Wendt, Way, & Zara, 2001; O'Neill, Marks, & Reynolds, 2005).

In addition to the GPC or PCM models for scaling polytomous items, the Graded Response Model (GRM) have also been studies for many years with several authors finding important differences between the GRM and the GPCM. van der Ark (2005)found that ordering of the expected latent trait was violated more often by the GRM than the GPCM. DeMars (2008) confirmed those results but found that this result did not lead to differences in theta values matched on raw scores. Kang et al. (2009)found that the GPCM fit data generated by the GRM better than the GRM, itself particularly in small sample sizes. This issue of fitting generated data to the other models will not be tested in this study, but may be considered in subsequent studies. What is clear is that several authors have found that the GRM and the GPCM perform differently under certain criteria.

A simulation study comparing these two models in mixed format assessments found that the GPC maintained a higher rate of convergence than the GRM (Montgomery & Skorupski, 2012). Based on this finding and the difference cited above by other authors, this study will use the Generalized Partial Credit model (Muraki, 1992) for the polytomously scored items. The GPCM considers the probability that an examinee selected a particular response over the previous one and treats the response space as dichotomous.

2.1.1 Model Comparisons

Gibbons et al. (2007) conducted a simulation study to examine the GRM in unidimensional and bi-factor form to multidimensional data. This study varied; test length, number of items, number of dimensions, primary loadings, and domain loadings. The outcome results included; the standard

deviation of the theta estimates, posterior standard deviations, log-likelihood, differences between estimated and actual theta and the percentage change between unidimensional and bi-factor models of these variables (Gibbons et al., 2007). The significant likelihood ratio test for the improvement of model fit found that the bi-factor model fit the data better then the unidimensional graded response model (Gibbons et al., 2007). Gibbons et al. (2007) concluded that the bi-factor model provides an alternative to the traditional unidimensional IRT models when conditional dependence is likely as is the case where tests consist of two or more methods of item presentation.

For the type of polytomous item considered in this study, the GPC scoring model seems to best represent the scoring procedures that might be implemented in practice. Furthermore, since model convergence has been an issue in previous studies and the GPC converges more consistently than the GRM, the GPC model was selected.

2.2 IRT Assumptions

There are several assumption important Item Response Theory. The first assumption is unidimensionality which holds that the observations on the manifest variables are a function of a single continuous latent variable (De Ayala, 2009). In other words, the items measure the same construct. In an ideal situation, the unidimensionality assumption can be thought of as analogous to the homogeneity of variance assumption in analysis variance (De Ayala, 2009). All the variance in item responses can be accounted for by the latent dimension.

However, in data from real assessments there is likely to be some degree of violation to the unidimensionality assumption. It is possible for a unidimensional model to sufficiently fit data generated from two latent variables De Ayala (2009). A typical middle school math assessment might contain as many as five subcategories ranging form data to geometry and yet it will usually be considered and often scaled as a unidimensional assessment.

The second assumption for unidimensional IRT models is conditional independence. Conditional independence simply means that the response to a question is determined solely by the examinee's location on the ability continuum and not by any other question on the assessment.

(De Ayala, 2009). The testlet model is an example of a multidimensional model where the response to one question is not expected to be independent of other responses in a particular testlet. Even in unidimensional assessments this assumption can be violated. English Language Arts tests often require examinees to answer a number of questions based on the same passage. It is also possible in science or mathematics to have a stem followed by several response questions. In these cases, the responses to one question may be closely related to responses to other questions in the same section. When there is a substantial violation to the conditional independence assumption; accuracy in item parameter estimation is affedted and the the total information may be overestimated (De Ayala, 2009).

The third assumption is functional form, meaning that the date follow the function specified by the model (De Ayala, 2009). In a unidimensional 1Pl model this is reflected in parallel item response functions. While this assumption may not be perfectly met but as long as the item response functions are parallel within sampling error, model-data fit is indicated (De Ayala, 2009).

On assessments that are composed of both dichotomous and polytomously scored items, mixed format assessments, there may exist a type of format affect that violates one of more the above IRT assumptions (Traub, 1993; Kim & Kolen, 2006). There are a several possible methods for determining if there is a dimensionality in data including; the inspection of the ratio of the first to the second eigenvalues, inspection of the distribution of the residuals after extracting one factor, inspection of scree plots, and confirmatory factor analysis. Fitting the data to a bi-factor model is another method that can be used to test for dimensionality (Reise et al., 2007).

2.3 Unidimensionality

One problem with the mixed format assessment is that is unlikely to meet the unidimensionality assumption. Lee (2010) found that the Iowa Test of Basic Standards in Mathematics and the Science assessment, both of which are mixed format in design, violated this unidimensionality assumption. The failure to fits a unidimensional model may result from a type of formatting effect. A formatting effect occurs then the multiple choice and constructed response items are said to

measure different abilities and cause the presence of multidimensionality in the total test score (Kim & Kolen, 2006). A number of authors have discussed the issue of dimensionality in mixed format assessments (Lee, 2010; Kamata & Bauer, 2008; Kim & Kolen, 2006; Kim & Lee, 2006; Kim, Walker, & McHale, 2010; Yao & Schwarz, 2006; Cao, 2008).

Cao (2008), in a study of a two construct assessment using the Graded Response Model as the constructed response model, found that the multidimensional test structure showed more significant and systematic effects on the performance of the calibration of the data than other factors in study. Kim et al. (2010) discussed the possibility of a multidimensionality effect on equating, but did not find any difference associated with dimensionality. Yao and Schwarz (2006) stated that the issue of dimensionality is important but that dimensionality based on format could not be concluded from the factor analysis used in the study. The study went on to conclude that the skills and knowledge assessed by the item contributes as much to the dimensionality effect as does format (Yao & Schwarz, 2006). The fact that a mixed format assessment has two different item types may not be enough to cause some examinees to perform differently on the two item types but the complexity of the one of the item types might cause those items to measure different skills and knowledge.

2.3.1 Format Effect

Traub (1993) found that there can be a format effect resulting from examinees processing items differently. For example, if the polytomously score item is multiple select or matching that adds a level of complexity over a multiple choice item. When a formatting effect occurs, the multiple choice and constructed response items may measure different abilities and cause the presence of multidimensionality in the test total score Kim and Kolen (2006). Neither of these studies, or the studies listed in the previous section provided a comprehensive analysis of the issue of dimensionality in mixed format assessment.

A. L. Zenisky and Sireci (2001) mentioned the possibility that innovative items may introduce construct irrelevant variance. In other words, the level of complexity of the innovative item, as well

as how the examinee interfaces with the item format, may cause that item to appear either easier, or more difficult, than a multiple choice item measuring the same construct. Bennett et al. (1991) was unable to concluded that multiple choice items substantially measure different constructs but stated that the differences is the process used by the examinee might not be apparent in the factor analytic process used in their study.

If there is not enough noise in the data to affect model fit, but the ability to accurately classify examinees into pass/fail or similar categories is compromised, then the assessment results may be called into question. Lee (2010) examined the performance of classification and accuracy indices on mixed format assessments using real data. He did not find a difference in the performance of the indices across the models. However, Lee (2010) stated that the results were not generalizable due to the specific test examined in the study and the limited population sample. In addition, while Lee (2010) noted that there was some level of dimensionality in the data, the impact of dimensionality on classification was not examined.

2.3.2 Method of Examining Unidimensionality

There are several traditional methods used to establish dimensionality in data. Among them are; the inspection of the ratio of the first to the second eigenvalues, inspection of the distribution of the residuals after extracting one factor, inspection of scree plots, and confirmatory factor analysis. Reise et al. (2007) argues that the bi-factor representation can complement those more traditional methods. First they argue that the bi-factor analysis allows for the evaluation of the distortion that may occur when unidimensional IRT models are it to multidimensional data. When the factor loading and the item discriminations are different this may indicate dimensionality (Reise et al., 2007). The second argument for the bi-factor model is that it allows researches to empirically examine the possibility of forming subscores. In the data studied by Reise et al. (2007), they found that once the variance due to the general construct was removed the items did not provide sufficient information to scale individuals on sub-dimensions. While they did not find significant sub-scales this method can be used to determine dimensionality.

Thirdly, (Reise et al., 2007) argue that the bi-factor model provides an alternative to the non-hierarchical multidimensional models for scaling individual differences. By conducting a bi-factor analysis and partitioning the item response variance into general and group components the researcher can make an informed decision between the two models. When dimensions are modestly correlated $(r = 1 \text{ to } .4)$, the items will tend to have small loadings on the general factor and larger loadings on the group factors indicating the use of a non-hierarchical MIRT model (Reise et al., 2007). The bi-factor model representation will be a viable alternative when the dimensions are moderately or highly correlate $(r = .4 \text{ and above})$.

2.4 Mixed Format Assessments

Several studies have considered the use of combined item response theory (IRT) models, bi-factor models, and second order IRT models to scale educational assessments (Cai et al., 2011; DeMars, 2006; Reise et al., 2007; Rijmen, 2010; Whittaker et al., 2012). These studies served, in part, as motivation for this study.

Cai et al. (2011) conducted a study that considered fitting several different item response theory models including an extended bi-factor. In this extended bi-factor model some items are allowed to load onto the general factor without loading onto one of the specific factors. The study conducted two simulations and the analysis of item responses resulting from the 2000 Program for International Student Assessment (PISA) data. The first simulation was conducted to check the accuracy of the proposed estimation methods. In this simulation, data from an extended bi-factor model in two groups was generated ($N_1 = N_2 = 1000$). Group 1 consisted of $n_1 = 16$ dichotomously scored items all fit to the graded response model for two categories. The specific factors were defined as 4-item clusters. The latent variable in group 1 were assumed to have zero means and unit standard deviations. In group 2 there were only 3-item clusters composed from $n_2 = 12$ observed items. The item parameters in group 2 were constrained to be equal across groups to allow for measurement of model invariance.

The second simulation study conducted by Cai et al. (2011) modeled a complex assessment

that consisted of multiple choice (MC), constructed response (CR) and complex multiple choice items (CMC). Complex multiple choice item occur in clusters that make up mini testlests within a larger test structure (Cai et al., 2011). In this simulation the hypothesized test consisted of 9 MC items, 1 CMC item containing 2 questions and 5 CR items. The first 5 items formed one cluster and the CR items forms a second cluster each loading on to a specific factor. The factors were assumed to me normally distributed with zero means and unit variances. Data was simulated from a bi-factor model with $N = 3000$. For each data set the authors fit two models; the first model imposed equality constrains on the slopes of the CR items whereas the second model does not impost the equality constraints.

Cai et al. (2011) ran 500 replications and did not have any issues with convergence. They imposed a normal stochastic constraint on the lower asymptote parameters with a mean equal to -1.1 and standard deviation equal to 50 to help stabilize estimation for the MC items. Due to the imposed constraints, the guessing estimates centered around the true values (Cai et al., 2011). They found slightly larger bias estimates in the slope and intercepts in the MC items than in the CMC or CR items. Cai et al. (2011) state that this finding is consistent with the fact that unidimensional IRT 3-parameter model requires significantly larger N than the other items response models to achieve stable estimations.

Since the extended bi-factor model fit the data well, the authors conclude that the generalized modeling framework outlined and analyzed in the study opens up opportunities for full-information bi-factor-based multidimensional differential item functioning analyses and bi-factor-based linking/equating studies(Cai et al., 2011). The authors also stated that the proposed extended bi-factor model can be applied to explore the dimensionality of psychological and educational measurement instruments (Cai et al., 2011).

DeMars (2006) conducted a study to compare the ability, reliability, item difficulty and item discrimination estimates for the bi-factor model, the testlet effects model, the testlets-as-polytomous-items model and the independent items model. The testlet-as-polytomous models refers to model that estimates a unidimensional model but treats items within a testlet as a single polytomous item

(DeMars, 2006). This study consisted of both a simulation study and an examination of the models using real data. The authors found that using a more complex model when a less complex model was sufficient led to slightly higher RMSE but not to higher bias. Using a less complex model than necessary also led to a higher RMSE and to negatively biased slopes (DeMars, 2006). In general this study favored the use of the more parsimonious testlet-effects model over the bi-factor model. However, the authors also stated that the speed at which the bi-factor model could be ran, in comparison to the testlet-effects model, might be of benefit to practitioners. The additional parameters of the bi-factor model did not decrease the accuracy of the primary trait or slope estimates (DeMars, 2006).

Rijmen (2010) compared the formal relations between the bi-factor, the testlet and a second-order multidimensional IRT models as well as the use of real data to fit and compare the models. He showed mathematically that the testlet model and the second-order model were formally equivalent and furthermore that they are restricted version of the bi-factor model. The conditional dependencies between items on the same testlet were taken into account through the testlet-specific dimensions (Rijmen, 2010). The real data was collected from an international English assessment ($N = 13,508$) consisting of 20 reading comprehension items organized into four testlets. Rijmen (2010) found that the proportionality restrictions imposed on the data by the second-order model were too stringent. The better fit of the bi-factor model indicated that the use of the testlet model without even considering the bi-factor model in educational testing may not be the best practice (Rijmen, 2010).

Whittaker et al. (2012) conducted a simulation study to examine the performance of six model selection criteria on mixed-format IRT models. This study found that model selection indices more accurately distinguished between correct and incorrect models that were less parameterized; PC, 1PL, or 1PL/PC models. The accuracy of model selection was not as accurate for the 2PL, 3PL, 2PL/GPC, or 3PL/GPC. The authors also found that the models fit indices selected the mixed format (2PL/GPC) more accurately when the assessment consisted of more polytomously scored items than dichotomously scored items in terms of score points. When the score points were

more equally distributed are more came from dichotomously scored items sample size also became important. For example, the larger the sample size the better the LRC (G^2) model selection method was at correctly selecting the 3PL/GPC over the 2PL/GPC.

2.5 Summary

The body of research on mixed format assessments is still being developed. Gibbons compared the graded response model to a bi-factor model by with Likert scaled data rather than an education assessment (Gibbons et al., 2007). While they found that the bi-factor model fit the data better, how the models compare in data generated from a mixed format assessment was not addressed. Cao (2008) did consider the graded response model within a mixed format construct but created the data to represent an assessment that was built to more than on construct. Furthermore, the ratio of multiple choice to constructed response items was restricted to 8:1 and the primary object of the study was to consider the affects of this kind of data model on equating. (Whittaker et al., 2012) looked at model selection procedures for mixed format assessments but did not consider the bi-factor or second order model.

In the Cao (2008) article the authors found that the extended bi-factor model fit the data well. But, this study did not compare different test configurations in fitting the 3PL or the extended bi-factor model. The primary focus of this study was to examine how the extended bi-factor model would be used in fitting data from assessments in which some of the items are clustered together (Cao, 2008). Similarly, DeMars (2006) looked at a specific type of mixed format assessment in which the items clustered together were treated as one polytomously score item. Rijmen (2010) compared the bi-factor, testlet and second order models both empirically and with real testlet-based data. While the empirical finding where useful in the foundations of this study, the use of testlet-based data will not be used in this study.

From the research of literature discussed here, it remains unclear whether or not data resulting from a unidimensional construct but presented in a mixed format assessment, is better modeled by a unidimensional or multidimensional model. In a mixed format assessment, where the constructed

response items are innovative, there may be complexity and construct irrelevant variance associated with the precess required to answer the item in addition to possible dimensionality in the data itself. This study will look at the complexity of the data resulting from mixed format assessments and will unitize the study from this section as a framework for the methodology.

Chapter 3

Methods

This chapter is organized into three sections. The first section will discuss the specific models used in this study including the parameters estimated. The second section will discuss the simulation study design with discussions of the variables included in the study. The final section will discuss the evaluation and simulation criteria used to determine model fit.

3.1 Models

3.1.1 Combination 3PL/GPC Model

One method of fitting mixed format assessment data to an item response theory model (IRT), would be to combine an dichotomous IRT model for the dichotomous items and polytomous IRT model for the polytomous items. Selection of the unidimensional IRT model combination has many possibilities. For the dichotomous item; the Rasch, 1PL, 2PL or 3PL models, could be utilized. For the polytomously scored items, there are also a number of possibilities including; the rating scale method, the partial credit model and its generalized counterpart or the graded response model. This study used the 3PL model for the dichotomous items and the generalized partial credit model for the polytomously scored items.

The choice of the 3PL model is based on the fact that the 3PL provides the largest amount of information and consequently should fit the data. Furthermore, the 3PL model allows the possi-

bility that an examinee, who is not proficient with the skills required to answer a particular item correctly based on their understanding of the concept, can answer the item correctly by guessing through the inclusion of a c-parameter that sets a lower asymptote for the model. Below is the 3PL model equation where c represents the guessing parameter, a_j represents the item discrimination and is allowed to vary across items, b_j represents the item difficult and θ represents the examinee's ability estimate.

$$P(y_i = 1|\theta, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 - e^{a_j(\theta - b_j)}} \tag{3.1.1}$$

The selection of the generalized partial credit model (GPCM) model is based on the findings of a previous study which found that the combination of a 3PL with a GPCM provided a higher rate of convergence and model fit with little parameter recovery bias (Montgomery & Skorupski, 2012). The GPCM is an extension of the Partial Credit Model (PCM). The PCM forces the slope parameters to remain the same across all item whereas the GPCM allows the slope parameter to vary across items (Muraki, 1992).

$$P(y_j|\theta) = \frac{e^{\sum_{h=0}^{y_j} a_j(\theta - b_{jh})}}{\sum_{k=0}^{M_j} e^{\sum_{h=0}^{k} a_j(\theta - b_{jh})}} \tag{3.1.2}$$

The GPCM models the probability that an examinee responded to a particular response category over the previous one. In other words, it provides the probability of scoring a 1 over the probability of a 0, or of scoring a 2 over a 1. This model dichotomizes the probability of answering the item correctly to a comparison between categories. In the model below, M is the number of response categories and $b_{jh} = b_j - \tau_h$, is the threshold component used to identify the categories.

3.1.2 Second Order IRT Model

The second order multidimensional IRT model is formally equivalent to the testlet model (Rijmen, 2010). This model contains a specific dimension for each testlet and a general dimension for the overall assessment. Unlike the testlet model, and the bi-factor model discussed below, items

do not depend directly on the general dimension (Rijmen, 2010). Rather, each item depends on a specific parameter which is dependent upon the general dimension.

Rijmen (2010) showed that given the second-order equation below

$$g(\pi_j) = \alpha_{jk}\alpha_{kg}\theta_g + \alpha_{jk}\xi_k + \beta_j \tag{3.1.3}$$

it follows that:

$$\begin{aligned} g(\pi_j) &= \alpha_{jk}\alpha_{kg}\theta_g + \alpha_{jk}\frac{\alpha_{kg}}{\alpha_{kg}}\xi_k + \beta_j \\ &= \alpha_{jk}\alpha_{kg}(\theta_g + \frac{\xi_k}{\alpha_{kg}}) + \beta_j \\ &= \alpha^*_{jg}(\theta_g + C^*_k\xi_k) + \beta_j \end{aligned} \tag{3.1.4}$$

Where $\alpha^*_{jg} = \alpha_{jk}\alpha_{kg}$ and $C^*_k\xi_k = \frac{1}{\alpha_{kg}}$. Thus equation 3.1.4 is equivalent to the testlet equation, without a guessing parameter, as written by (Rijmen, 2010).

Second order factors were first described by Thurstone (1947). Factors obtained from the test item correlations are called first-order factors (Thurstone, 1947). Second order factors then, are factors obtained from the correlation of the first order factors. In the case of the second order multidimensional IRT model that is exactly what is being described. In this model, the general ability is obtained from the correlation of the specific abilities. Or, as it is often stated, the general ability explains the variance in the specific abilities which in turn explain the variance of the item responses. The concept of a single general second-order factor was born from controversies surrounding the Spearman general intellective factor (Thurstone, 1947).

As an educational example, suppose there is a 10-item mathematics test assessing solving polynomials. This test is subdivided into three specific content domains (e.g. solving 2-degree polynomial, solving 3-degree polynomials, and graphing) and one primary domain (e.g. solving polynomials). The second order model states that answering an item correctly depends on the gen-

eral ability to solve polynomials as a whole which in turn explains the ability to respond correctly to the individual content domains. In this study, the individual domains consisted of item format. Dichotomous and polytomous items each were forced to load onto a specific ability construct representing the item format type then the specific ability constructs loaded onto a general ability construct.

In the second order model, the position of the a-parameters for the primary and specific dimensions can be illustrated as the loading of the items on the specific construct and the loading of the specific construct on the general construct. In the diagram below. y_1 and y_2 refer to the set of dichotomously scored and set of polytomously scored items respectively.

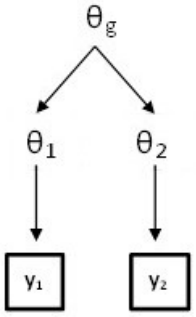

Figure 3.1: Diagram of Second-Order Model with 1-dichotomous (θ_1) and 1-polytomous specific domain (θ_2)

The model equations are as follows, let $\pi_j = P(y_{j(k)} = Y|\theta_g, \theta_k)$, then the linear function of the latent variables, in this model and the bi-factor model, can be written as the link function $g(\cdot)$, where the response probabilities (π_j) are linked to the predictor of latent variables with the logit link, $g(\pi_j) = \ln \frac{\pi_j}{(1-\pi_j)}$.

$$g(\pi_j) = a_{jk}\theta_k + \beta_j \tag{3.1.5}$$

$$\theta_k = a_{kg}\theta_g + \xi_k \tag{3.1.6}$$

Combining these equations yields

$$g(\pi_j) = a_{jk}a_{kg}\theta_g + a_{jk}\xi_k + \beta_j \tag{3.1.7}$$

Where a_{kg} indicates how much of the specific dimension θ_k is explained by the general dimension θ_g, and ξ_k is the unique contribution from θ_k. It can be assumed that all dependencies between the specific dimensions are accounted for in the general dimension and consequently that all ξ_k are statistically independent from each other and from the general dimension (Rijmen, 2010).

In this study, the θ_k dimensions represented item format. In this way, the assessment was assumed to be unidimensional on the general dimension, but the specific dimensions allowed the model to account for item fit differences between the dichotomously and polytomously scored items.

3.1.3 Bi-Factor Model

The final model considered is the bi-factor model. The bi-factor method was introduced in the (*Preliminary Reports on Spearman-Holzinger Unitary Trait Study*, 1930-1936). Holzinger and Swineford (1937) illustrated how the bi-factor method could be modified for analysis of variables that are more complex that originally considered in the preliminary report. The latter document is the first reference of this methodology identified as "bi-factor." The authors describe the theoretical framework as consisting of a general factor that runs through all variables with specific factors in each variable (Holzinger & Swineford, 1937). The general procedure for utilizing this analysis is to (1) resort the data so that items that are more highly correlated are clustered into small groups, (2) remove the general factor from each of the groups, and (3) examine the group factors. This process is repeated until the group factors show no greater complexity. At this point, the final factor pattern may be established (Holzinger & Swineford, 1937).

The bi-factor model is a theoretical framework where a general factor is assumed to run through all variables but in addition, a number of uncorrelated group factors is included in the model (Holzinger & Swineford, 1937). The lack of correlation between the group factors is the difference

between the second-order model and the bi-factor model. In the bi-factor model the group factors are not correlated with each other or with the general factor.

As an example, using the same hypothesized 10-item mathematics test assessing solving polynomials above, the test is subdivided into three specific content domains (e.g. solving 2-degree polynomial, solving 3-degree polynomials, and graphing) and one primary domain (e.g. polynomials). In the bi-factor model, each item loads onto the primary ability domain and onto the specific content domains. But, unlike the second-order structure, the specific and general domains are not directly related in the model. In this study, the group factors consisted of item format rather than content.

The directed acyclic graph of the bi-factor model is displayed in Figure 3.2 below.

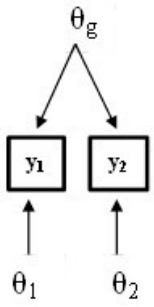

Figure 3.2: Diagram of Bi-Factor Model with 1-dichotomous (θ_1) and 1-polytomous specific domain (θ_2)

In the bi-factor model, the position of the a-parameters for the primary and specific dimensions can be illustrated in a simple structure matrix, S_b, as follows

$$\mathsf{S_b} = \begin{bmatrix} a_{1,1} & a_{1,2} & 0 \\ a_{2,1} & a_{2,2} & 0 \\ a_{3,1} & a_{3,2} & 0 \\ a_{4,1} & a_{4,2} & 0 \\ a_{5,1} & a_{5,2} & 0 \\ a_{6,1} & 0 & a_{6,3} \\ a_{7,1} & 0 & a_{7,3} \\ a_{8,1} & 0 & a_{8,3} \\ a_{9,1} & 0 & a_{9,3} \\ a_{10,1} & 0 & a_{10,3} \end{bmatrix} \tag{3.1.8}$$

In this structure matrix the primary domain items have a nonzero value on the discrimination parameter, $a_{j1} \neq 0$, and clusters of items that belong to the defined specific ability dimension have nonzero value on the item discrimination, e.g. $(a_{i2}, a_{i2}) \neq 0$. Generalizing this structure, a test with n items has clearly defined D - 1 orthogonal dimension of specific content domains and one dimension that represent the primary ability.

Bi-factor structure

$$Y = \lambda_y \Theta + \varepsilon$$

Where $Y = \begin{bmatrix} y_{ik} \end{bmatrix}$ $\lambda_y =$ loading structure matrix $\Theta = \begin{bmatrix} \theta_{g1} \\ \theta_{s1} \\ \theta_{s2} \end{bmatrix}$ $\varepsilon =$ item error

In the case of binary data, y_{jk} denotes the response on the j^{th} item, $j = 1, \ldots, J$, in the k^{th} testlet, $k = 1, \ldots, K$. There are J_k items within each testlet therefore, $\sum_{k=1}^{K} J_k = J$. The responses, conditional on the testlet specific latent variable θ_k and the general latent variable θ_g, are assumed

to be statistically independent (Rijmen, 2010),

$$P(y|\theta) = \prod_{j=1}^{J} (y_{j(k)}|\theta_g, \theta_k) \tag{3.1.9}$$

where $\theta = (\theta_g, \theta_1, \ldots \theta_k, \ldots \theta_K)$. Figure 3.2 illustrates the typical model where the latent variables are uncorrelated. The latent variables are also typically considered to be normally distributed. The linear logit function $g(\cdot)$ represents the relationship between the latent variables and the probability of a correct response, $\pi_j = P(y_{ik} = Y|\theta_g, \theta_k)$, as in the previous model, can be written

$$g(\pi_j) = a_{jg}\theta_g + a_{jk}\theta_k + \beta_j \tag{3.1.10}$$

where β_j is the intercept parameter and a_{jg} and a_{jk} are the slopes of item j on the general and specific latent variable (Rijmen, 2010). As in the case of the bi-factor model, this study considered the specific dimensions as representing item format and the general dimension as the overall attribute to be measured.

3.2 Scoring Procedure

3.2.1 Bayesian Estimation Method

In this study both dichotomously (0/1) and two levels of polytomously (0/1/2,0/1/2/3/4) scored items were examined. Let Y be the matrix of item responses such that the response pattern array for examinee i on item j is one row in the matrix $Y = [y_{ij}]_{Nxn}$, for N number of examinees and n number of items. Where the indicator function for the conditional category response probability $P(y_{ij(k)} = Y|\theta_g, \theta_k)$, is

$$y_{ij} = \begin{cases} 1 & \text{if } y^*_{ij} \geq \tau_j \\ 0 & \text{if } y^*_{ij} < \tau_j \end{cases} \tag{3.2.1}$$

Here, τ_j represents the threshold for item j. In the dichotomous case, $\tau_i = 1$. Note that the

latent response variable can be written as

$$y_{ij}^{*} = v_i + \lambda_i \xi + \varepsilon_i \tag{3.2.2}$$

where v_i is the intercept, λ_i is the factor loading, ξ is the latent factor score for a particular person, and the residual for item i is ε_i.

For all three models, assuming conditional local item independence on item and person parameters, the conditional joint probability function of item responses, Y, is

$$f(y_{ij}|\theta_g, \theta_k) = \prod_{k=0}^{k_j} P(y_{ij} = k|\theta_g, \theta_k)^{y_{ij}} \tag{3.2.3}$$

Note that if y_{ij} is a missing value then the indicator function will be zero. The general factor and the specific dimensions are assumed to be jointly normally distributed and mutually orthogonal (Cai et al., 2011). Theta is typically considered to have a multivariate normal distribution with a zero mean and identity covariance matrix. The orthogonality and normality of these latent variables reduces the density function to a product as illustrated in 3.2.4 (Cai et al., 2011).

$$f(\theta_g, \theta_k) = f(\theta_g) f(\theta_k), \text{for } k = 1, \ldots, K \tag{3.2.4}$$

The vector of observed item responses for respondent i is y_i (3.2.1). For the purposes of simplicity, we can consider all of the unknown parameters together and refer to them collectively as β. Thus the marginal likelihood of β, $L(\beta|y_i) = f_\beta(y_i)$, is defined as a function of unknown parameters β. Since the respondents are assumed independent, the marginal log-likelihood is a sum over respondents $\sum_{n=1}^{N} L(\beta|y_i)$. Then the conditional distribution of the observed responses in Equation 3.2.3 depends on β. The distribution of latent variable in Equation 3.2.4 does not depend on β because this is not a multi-group design where the means and variances are free.

3.2.2 Markov Chain Monte Carlo with Gibbs Sampling

Markov Chain Monte Carlo (MCMC) is an estimation method that constructs a set of random draws, for each parameter being estimated, from the posterior distribution. This process involves choosing a distribution that can be easily sampled from and then either accepting or rejecting the draws based on the likelihood that they represent the actual posterior distribution. Essentially, the prior distribution of draws multiplied by the likelihood function equals the posterior. Only the draws that make sense are kept. The benefit of this Bayesian approach is that examinees with identical response patterns will not get identical points which accounts for the fact that no two individuals are ever truly identical in their responses.

By defining a Markov chain $M_0, M_1, M_2, \ldots$ with states $M_k = (\theta^k, \xi^k)$, observations are simulated from the Markov chain. The distribution of $M_k = (\theta^k, \xi^k)$ will converge to the chain's stationary distribution $\pi(\theta, \xi)$. The Markov chain should be defined in such a way that $\pi(\theta, \xi)$ is the posterior of $p(\theta, \xi | U)$.

The MCMC transition kernel defines the probability of moving to a new draw given the current draw. This is used to determine whether or not to retain each new random draw from the posterior.

$$t[(\theta^0, \xi^0), (\theta^1, \xi^1)] = P[M_{k+1} = (\theta^1, \xi^1) | M_k = (\theta^0, \xi^0)] \tag{3.2.5}$$

Provided that the transition kernel is defined so that $\pi(\theta, \xi) = p(\theta, \xi | U)$, after throwing away the first K observations, the remaining observations are treated as draws from the posterior.

$$(\theta^1, \xi^1) = M_{K+1}, (\theta^2, \xi^2) = M_{K+2}, \ldots, (\theta^L, \xi^L) = M_{K+L}$$

This process is called the Gibbs sampling procedure. It is defined iteratively as:

- Draw $\theta^k \sim p(\theta | U, \xi^{k-1})$
- Draw $\xi^k \sim p(\xi | U, \theta^k)$
- repeat the process.

This study used OpenBugs (Thomas, O'Hara, Ligges, & Sturtz, 2006) to perform the MCMC with Gibbs sampling procedure.

3.3 Simulation Study Design

For this study data was generated as unidimensional in mixed format form consisting of both dichotomously and polytomously scored items. The programming language FORTRAN was used to generate the data based on the 3PL IRT and GPCM equations. A separate data set was generated for each of the five test types listed in Table 3.1. The polytomous items were generated at two levels (0/1/2) and (0/1/2/3/4). The generated data was fit to the 3PL/GPC, second-order, and bi-factor models using a Bayesian framework with OpenBUGS (Thomas et al., 2006). RMSE and Bias was used to examine parameter recovery for the unidimensional model.

Table 3.1: Simulation conditions for the 3PL/GPCM, Second Order Model, and Bi-factor model

N	Test Lengths				Score points
	Total Length	Dichotomous	Polytomous (2-pts)	Polytomous (4-pts)	
2000	40	36	2	2	48
2000	40	10	15	15	100
2000	40	20	10	10	80
2000	60	45	10	5	85
2000	75	69	3	3	87

Note: N: Number of examinees

3.3.1 Independent Variables

The independent variables included the number of dichotomously scored items to the number of polytomously score items in the form of 5 test types, five models combinations and two levels of discrimination. Varying the proportion of dichotomously scored to polytomously scored items was used to determine if there are some combinations of items from with a format affect creates the appearance of multidimensional results out of unidimensional data. The model combinations

are chosen to determine if varying the number of items and score point combinations of the polytomously items impacts the model fit.

3.3.1.1 Item Type and Test length

The dichotomous items were scored in the traditional way; 0 for incorrect and 1 for correct. Multiple choice questions with one correct answer and true/false questions are considered dichotomous or binary. The polytomously scored items were scored on two values; 2 points and 4 points. The mixed format assessments were constructed with the dichotomous items combined with the polytomously scored items scored as either 2 points and 4 points. The true item parameters, examinee parameter and item responses were generated from the combination of equation 3.1.1 and equation 3.1.2 for the unidimensional model as well as from equations 3.1.7 and 3.1.10 for the second order and bi-factor models, respectively.

There were two criteria selected to define the item combinations; 1) holding the number of items constant (40 items) and 2) allowing the number of items to vary to produces a variety of score points. The 40 items test range in score points from 48 points to 100 points with the former coming from mostly dichotomously scored items and the later from mostly polytomously score items. The other two tests contained 60 and 75 items with score points of 85 and 87, respectively.

The number of examinees was fixed at 2000. This sample size will be large enough for parameter estimation while small enough to run in a reasonable about of time given the use of a Monte Carlo estimation process for this simulation. Table 3.1 summarizes the test length and score point simulation conditions of the study. There are 5 test types(test length/score point) X 3 fitted models X 2 discrimination levels = 30 simulation conditions considered in this study. The 3 fitted models are: the 3PL/GPCM, a bi-factor with 2 specific constructs, and a 2nd order model with 2 specific constructs.

3.3.1.2 Item and Examinee Characteristics

The examinee θ vector was randomly drawn from a multivariate normal distribution with a zero mean vector and a variance-covariance matrix equal to the identity matrix, such that $\theta \sim MVN(0, I)$. The θ values were chosen between -3 to 3, on the ability spectrum of the general and specific domains. The α-parameters are bounded by zero and distributed as log-normal with a zero mean and uniform variance $U(0.2, 2.0)$.

For this study the discrimination parameter for the primary dimension ranged from (0.75, 1.25). Item discriminations on the specific domains were allowed to be equal to and greater than the primary domain for the second order and bi-factor models. One set of item combinations resulted from the a-parameter ranging from $(0.75, 1.25)$ for both item types. A second set of item combinations resulted from allowing the a-parameter on the polytomous items ranging from $(1.25, 1.75)$. Table 3.2 summarizes the discrimination parameters for the second order and bi-factor models in terms of their primary and secondary dimensions.

Table 3.2: Discrimination parameters

Primary dimension	Secondary Dimension
$a_{j1} \sim U\,(0.75, 1.25)$	$a_{j2} \sim U\,(0.75, 1.25)$
	$a_{j3} \sim U\,(1.25, 1.75)$

Examinee and item parameters were randomly drawn from the following distributions using FORTRAN. The ability parameter for all dimensions were randomly drawn from a MVN distribution with each examinee having a minimum of one latent ability for the combined unidimensional model and a maximum of three; one for the primary and as many as two for the specific content domains in the second order and bi-factor models. The 3PL/GPCM incorporated difficulty values drawn from $b \sim N(0, 1)$ and guessing parameters values drawn from $c \sim U\ (0.0, 0.3)$.

Threshold values were drawn from a $\sim N(0, 1)$ such that the sum of threshold values is zero. For example, in the case where the possible scores are (0/1/2), there are three threshold values $\beta_{ji}, \beta_{j2}, \beta_{j3}$. Each of these values established the transition between categories; 0 versus $\{1, 2\}$, $\{0, 1\}$ versus 2, and $\{0, 1, 2\}$ versus greater than 2, which has a probability of 1. The probability

of 1 is obtained by definition, when $\beta_{j1} = 0$ (Muraki, 1992). The other threshold values are randomly drawn and added to the b-values. FORTRAN was used to generate the threshold values and calculate the sum of b with each threshold for each item. The threshold values along with the a, b, and c values were recorded and retained.

3.4 Checking Model Convergence

This study focused on model fit and rate of convergence over parameter recovery. Based on the recommendation of Gelman and Shirley (2011), the $\hat{R}$ or the potential scale reduction factor was examined to determine if all parameters converged. This factor takes the mixture variance divided by the average within-chain variance, and computes the square root of the ratio. The rational behind this statistic is that at convergence the chains will have mixed. If the distribution between the within chains and between chains are identical then $\hat{R}$ should be equal to 1. As a rule of thumb, $\hat{R}$ values less than 1.2 or considered an indication the parameters have converged (Gelman, 1996; de la Torre & Hong, 2010).

After convergence was established for each item combination, based on the $\hat{R}$ values, the monte carlo error was calculated to verify that the chain extended long enough after convergence was established. The monte carlo error provided by OpenBUGS, when compared to the standard deviation of the parameter estimate, should be 0.05 or less. Smaller values indicate that the estimate contains less error.

3.4.0.3 Parameter Recovery

A comparison of the parameters generated from the unidimensional model was compared to the parameters estimated with OpenBugs. This was only done with the unidimensional model to verify that the parameters defined by the simulation were recovered with little error or bias. To evaluate each of the parameters; a, b, c, and θ, the Root Mean Square Error (RMSE) and bias was calculated to establish the amount of error present in parameter recovery. Below is the RMSE calculation for theta. By replacing theta with each of the other parameters, the following formula

can extended to the remaining parameters.

$$RMSE_\theta = \sqrt{\frac{\sum_{i=1}^{N}\sum_{i=1}^{R}(\widehat{\theta}_{ir} - \theta_i)^2}{NR}} \tag{3.4.1}$$

By replacing theta with each of the other parameters of interest, the bias formula below was altered to establish bias for each of the other parameters.

$$BIAS_\theta = \frac{\sum_{i=1}^{N}\sum_{i=1}^{R}(\widehat{\theta}_{ir} - \theta_i)}{NR} \tag{3.4.2}$$

The standard deviation of estimates over replications provided a standard error of the estimate. The following formula estimates the sampling error associated with theta across replications.

$$SD_\theta = \sqrt{\frac{\sum_{i=1}^{N}\sum_{i=1}^{R}(\widehat{\theta}_{ir} - \overline{\widehat{\theta}}_i)^2}{NR}} \tag{3.4.3}$$

The relationship between these parameters can be expressed, as the total error variance equal to the random error plus systematic error as follows:

$$RMSE^2 = SD^2 + BIAS^2 \tag{3.4.4}$$

3.4.1 Bayesian Fit

Model convergence was examined through three commonly used methods for establishing model fit and complexity; Akaike's Information Criterion (AIC),Bayesian information criterion (BIC) and the deviance information Criterion (DIC) (Akaike, 1974; Gelman, 2006; Gelman & Shirley, 2011; Rijmen, 2010; Spiegelhalter, Best, Carlin, & Van der Linde, 1998; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002; Schwarz et al., 1978). Comparisons among these indices can be useful in evaluating the relative effectiveness of model selection (Li, Cohen, Kim, & Cho,

2009).

The Akaike's Information Criterion (AIC), first identified by Akaike (1974), incorporates a penalty function on model complexity. In the equation below $\overline{D}$ is the posterior mean of the deviance in the MCMC estimation and $2PD$ serves as a penalty for overparameterization (Li et al., 2009). P_D is the number of parameters estimated calculate by OpenBUGS.

$$AIC = \overline{D} + 2P_D \tag{3.4.5}$$

The next recommended method for evaluation of the simulation model fit is the Bayesian informational criteria (BIC) (Schwarz et al., 1978). This model penalizes overparameterization with the use of the logarithm of the sample size multiplied by the number of parameters estimated. BIC tends to chose models that are simpler, have fewer parameters, than the AIC.

$$BIC = \overline{D} + P_D \log N \tag{3.4.6}$$

The final method that was used to compare model fit was the deviance information criterion (DIC) (Spiegelhalter et al., 1998).

$$DIC = D(\overline{\theta}) + 2P_D \tag{3.4.7}$$

In this equation, $D(\overline{\theta})$, is the posterior mean of the deviance. The DIC was designed to be a more generalized version of the AIC and suitable for hierarchical models (Li et al., 2009). Vales for all there criterion were calculated and compared for each item combination and model type.

Chapter 4

Results

In this chapter the results from the simulation study comparing data fit to three IRT models; 3PL/GPCM, bi-factor and hierarchical MIRT model, is discussed. Model fit, in terms of AIC, BIC, and DIC across models is compared. Parameter recovery from the 3PL/GPCM to the generated data will be discussed with respect to root mean square error (RMSE) and bias. Parameter recovery in this sense, will only be discussed for the 3PL/GPCM calibration because the the parameters estimated from the bi-factor and hierarchical MIRT models are on a different scale than than the unidimensional model making direct comparison inappropriate . Unidimensional parameter recovery RMSE and Bias is displayed in Table 4.1.

Once convergence for each model was established, $\hat{R}$ values were less than 1.2, Monte Carlo error estimates illustrating efficiency of the estimates after convergence were collected. Details of MCMC error by parameter within model is displayed in Appendix F, Tables F1, F2, and F3. Bayesian criterion for model fit, pD, DIC, AIC and BIC, by item combination and model type is compared in Table 4.2.

The loading structure of each item onto the general and specific dimensions for both the bi-factor and hierarchical MIRT models is displayed in apprentices G and H, respectively. In the hierarchical MIRT model, lambda represents the loading of the specific abilities onto the general ability. Consequently, there are only two lambda values for each hierarchical MIRT model.

4.1 Computing Procedures

4.1.1 FORTRAN

To begin the discussion, the computing procedures utilized in the study are detailed. Data sets were generated with FORTRAN 95 using the Integrated Development Environment; Plato. Plato is the Silverfrost text editor available bundled with FTN95 or available as a stand alone editor. Plato was used as a stand alone editor for this project. FORTRAN code was written and compiled in this text editor both to generate data and to analyze the results. For each item combination 50 data sets were generated in two versions; one with the same a-parameters for both dichotomous and polytomous items and one with higher a-parameters for the polytomous items. This resulted in 500 unidimensional data sets. In addition to the data sets, the FORTRAN code also provided output for the true parameter values for a, b, c, and $theta$ for the 3PL model and a, b, $threshhold$, and $theta$ for the general partial credit model (GPCM). The true parameter values were used to compute the RMSE and bias for these parameters.

FORTRAN code was then used to write the OpenBUGS script and batch files used to fit the models in OpenBUGS. Both OpenBUGS 3.2.1 and OpenBUGS 3.2.2 were used to fit the data depending on which version was loaded on the individual computers. Each OpenBUGS script was sent directly to the OpenBUGS executable program with a batch file. As defined by the script file, OpenBUGS outputs a log file and a set of coda files for each run. The coda files can then be read by R2OpenBUGS to check convergence (Sturtz, Gleman, & Ligges, 2005). Do to the memory requirements of the coda files, they were not retained after convergence was verified. The log files contained the parameter estimates, standard error, MCMC error as well as the deviance statistics required to compare the models.

Finally, FORTRAN code was used to calculate the BIAS and RMSE for each of the estimated parameters as compared to the true parameter values in the 3PL/GPC model. The final calculation was based on the average across the 50 data runs for each item combination within each model type. RMSE and BIAS were not calculated for the parameters estimated by the bi-factor or the

hierarchical MIRT model. Since the bi-factor and hierarchical MIRT model parameters are on a different scale, it is not appropriate to make direct comparisons of parameter recovery. In the bi-factor and the hierarchical MIRT model, the goal was to compare the model fit within this new scale rather than to recover original values.

4.1.2 Parallel Computing

The overall study examined a total of 30 simulation conditions. For each simulation condition, they were two parallel Markov chains observed. In the 3PL/GPC model, over all item conditions, the total number of MCMC iterations were 5,000 with the number of discarded burn-in iterations set at 1,000. Each run in the analysis of 50 replications took on average, approximately eight hours. These analysis were conducted on the the quantitative statistics WinStat machines at the University of Kansas. Since the WinStat machines are intended for short-term interactive sessions with statistical programs, and were are not intended for long running computations, another computer system was used for the bi-factor and the second-order models.

Because of the computing time requirement of the Bayesian MCMC calibration in OpenBUGS, as seen with the unidimensional models, the bi-factor and hierarchial MIRT models were fit on the HPC cluster maintained by the Center for Research Methods and Data Analysis at the University of Kansas. This is a Rocks Cluster of Linux compute nodes running on Dell Power Edge 2950. Servers that have 16GB RAM and dual quad-core Intel Xeon processors. Because of the added complexity of the higher dimensional models, each of the of the runs contained a total number of 33,000 MCMC iterations of which 15,000 were discarded during burnin. While this chain length is much larger than the required by the unidimensional models, these lengths are consisted with other studies using OpenBUGS to fit multidimensional models (Kang, 2006; Kang & Cohen, 2007; Md Desa, 2012). Item combinations two and three, containing 10 dichotomous/30 polytomous and 20 dichotomous/20 polytomous items respectively, took five to six days to compile on the cluster machines. Using the cluster computers allowed all 50 runs to finish in less time than it would have taken to complete one run on a PC machine. The time requirement proved to be the biggest

obstacle to study completion.

4.2 Model Convergence

A subset of runs from each model, within each item combinations, was evaluated to assure convergence. Once the length of the Monte Carlo chains required to ensure convergence was established, all data sets within that item combinations were calibrated to those chain lengths that ensured convergence. At least three thousand chains were calculated after the convergence criteria was met.

To verify the accuracy in the posterior estimates, the MCMC error was evaluated. The MCMC error is provided in the OpenBUGS log file. The MCMC error depends on the true variance of the posterior distribution, the number of MCMC iterations and autocorrelation in the MCMC sample. By dividing the MCMC error by the standard deviation, for each parameter estimate, the efficiency of the estimates after convergence were evaluated. Although the $\hat{R}$ values were less than 1.2, which is the rule of thumb for convergence, some models demonstrated slightly more error in the retained MCMC chains overall. As a rule of thumb, the simulation should run until the MCMC error is less than five percent of the sample standard deviation (*Bugs Tutorial*, 2012; *Columbia University*, 2012). In particular, the unidimensional models as a group displayed errors closer to 0.05 than desired. On closer examination, of each model, there were 3-4 runs that contained higher error values than other runs of the same model. Those runs were not removed and are included in the overall error rates reported in Appendix F.

The smaller the value of the error, the better the estimate. Appendix F contains table of MCMC error by parameter within model type and across item combinations. Order of calculation is important. The quotient of error to standard deviation for each model must be calculated first and then averaged across the replications. This value will not be the same as the quotient of the averages of the individual vales. (See Appendix E for proof of this fact.) It is important that the quotient be calculated at each run and then averaged over the runs so that the error is not underestimated. It would be better to overestimate the error and run the chains longer.

4.2.1 Parameter Recovery

Item combinations where there are equal numbers of dichotomous and polytomous items or where there are more polytomous items than dichotomous items had higher RMSE and BIAS in the $a-$ and $b-$ parameters, shown in Table 4.1. Item combination two containing 10 dichotomous and 30 polytomous items produced the highest level of bias and RMSE in the a- and b- parameters. These values were calculated using all of the runs for this model. But some of the run for this model had higher error level in the retained parameters. If those runs were removed the overall values would be closer to the other RMSE and bias for other item combinations. The theta parameters produced bias values close to zero across all item combinations. In general, item combinations with higher a-parameters for the polytomous items produced higher RMSE and bias when compared to the same item combinations with the equal a-parameter for both dichotomous and polytomous items.

Table 4.1: Unidimensional Parameter Recovery

Combination	**RMSE**				**BIAS**			
	A	B	C	THETA	A	B	C	THETA
36/2/2	0.3274	0.7265	0.0817	0.4309	0.0235	0.2163	0.0437	-0.0092
	0.5244	0.7765	0.0878	0.4342	0.0479	0.1376	0.0537	-0.0070
10/15/15	0.6792	1.4355	0.1194	0.8068	-0.5400	0.8377	0.0754	-0.0844
	1.0265	1.5557	0.1165	0.8422	-0.8742	0.9523	0.0744	-0.0947
20/10/10	0.6478	1.1557	0.1044	0.5708	-0.2157	0.4090	0.0764	-0.0505
	1.0092	1.2779	0.1058	0.5828	-0.3487	0.4593	0.0780	-0.0556
45/10/5	0.4795	1.0900	0.1023	0.4120	0.0483	0.3321	0.0713	-0.0412
	0.6463	1.1865	0.1037	0.4164	-0.0213	0.3404	0.0726	-0.0439
69/3/3	0.3313	0.0791	0.0859	0.0442	0.5567	0.3461	0.0759	-0.0106
	0.4294	0.6095	0.0788	0.3483	0.0658	0.0859	0.0439	-0.0140

Note: For each item combination, row one represents the same a-value for all items.
Row two represents item combinations with higher a-values for polytomous items.

4.3 Bayesian Criteria Comparison

From the table below, using the DIC criteria for model fit, the bi-factor model fits the data better than either the unidimensional model or the hierarchical MIRT model. This finding is consistent

with other researchers whose studies have also selected the bi-factor model over other models when fitting data in mixed format assessments. This pattern is consistent across all data sets.

The hierarchical MIRT model did not fit the data at all in model two or three where the there are more polytomous items than dichotomous and equal numbers of dichotomous and polytomous items respectively. The output by OpenBUGS found a negative pD value even when the length of the chain was increased to 40000. It should be noted that the initial 33000 chain length took 8 days for one run to complete and resulted in a negative pD. The longer chain may take several weeks and may or may not achieve convergence with a positive pD value. A negative pD value means that the deviance of posterior means is larger than the posterior mean of deviances. According to WinBUGS documentation, this can happen when when the posterior distribution for a parameter is non-normal or bimodal. In that case, the posterior mean is a very poor summary statistic and gives a very large deviance.

Table 4.2: Bayesian Fit

Combination		**Equal A**			**Higher A**		
		3PL/GPC	Hierarchical	bi-factor	3PL/GPC	Hierarchical	bi-factor
36/2/2	DIC	101164.4	100815	98133.2	101146	100792.0	99991.0
	AIC	102968.7	102379	101759.4	102936.4	102479.35	101532.9
	BIC	113102.2	111067	111023.9	112979.5	111930.74	110170.3
10/15/15	DIC	166031.6	*	164581	150610	*	148970
	AIC	167510.5	*	166139.4	152131.2	*	150513.8
	BIC	175764.5	*	174944.6	160653.7	*	159205.8
20/10/10	DIC	134158	*	132816.7	123044.9	*	121576
	AIC	135929.9	*	134598.7	124855.7	*	123358.3
	BIC	145887.9	*	144630.1	134973.3	*	133373.5
45/10/5	DIC	159334	158696.6	158020.8	149980	149357.9	148688.3
	AIC	161282.8	160487.1	159895.7	151942.4	151182.7	150610.8
	BIC	172173.2	170435.9	170431.9	162871	161378.1	161516.3
69/3/3	DIC	182428	182086.2	181206.7	179100	181135	178960
	AIC	184429.4	185727.5	183026.3	181113.7	184545.5	180842.2
	BIC	195651.9	19592.6	193162.8	192365	194134.4	191350.1

*-indicates combinations where the model did not fit the data. Missing values from combination five will be provided with updated table.

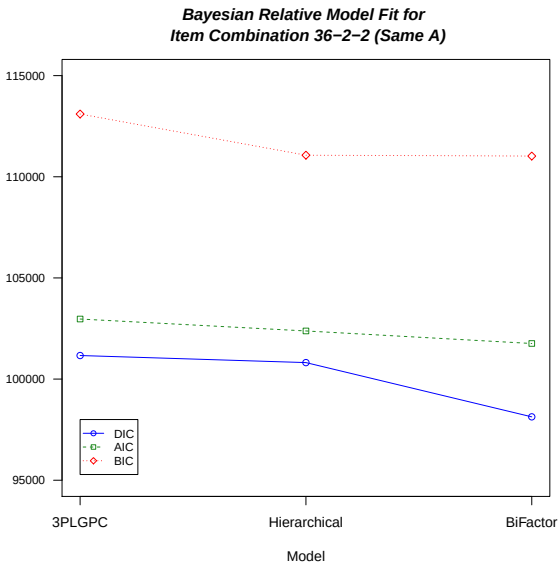

Figure 4.1: Same A

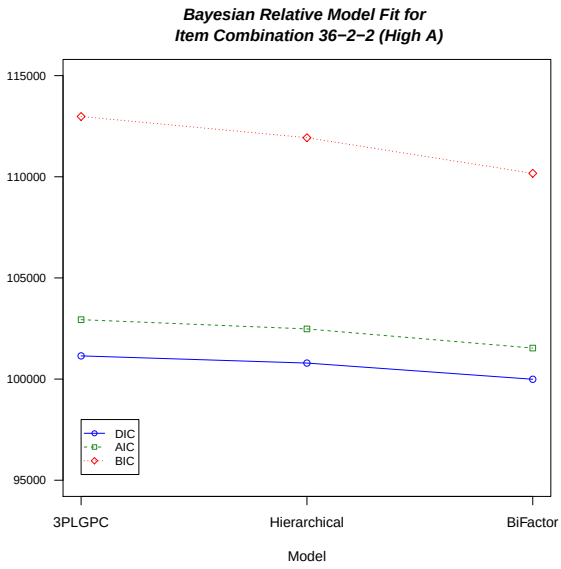

Figure 4.2: Higher A

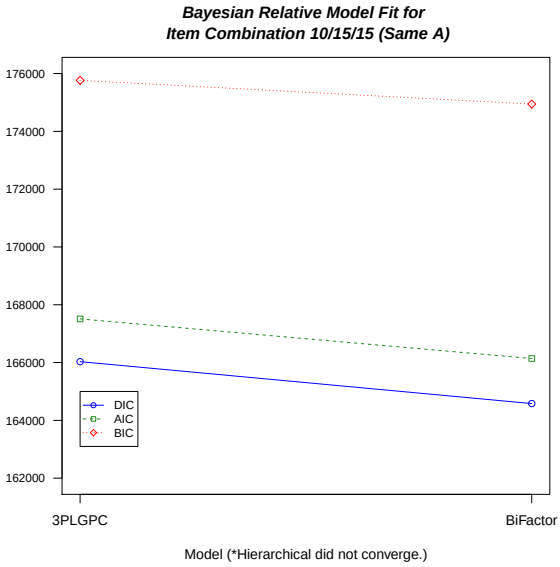

Figure 4.3: Same A

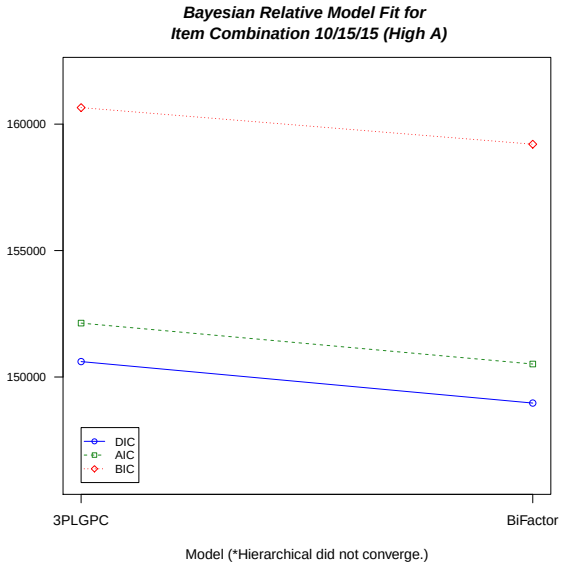

Figure 4.4: Higher A

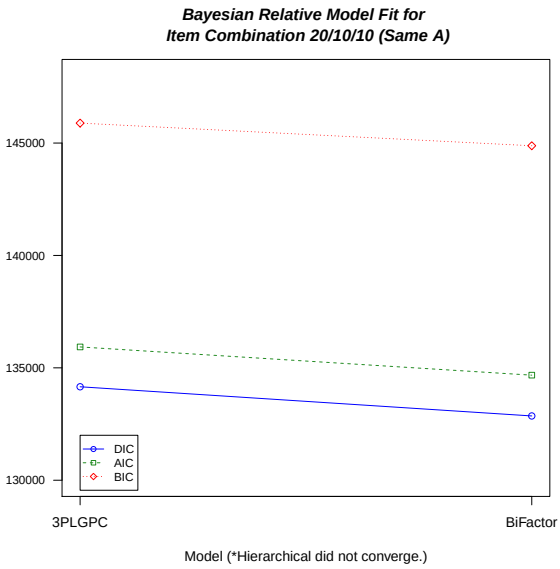

Figure 4.5: Same A

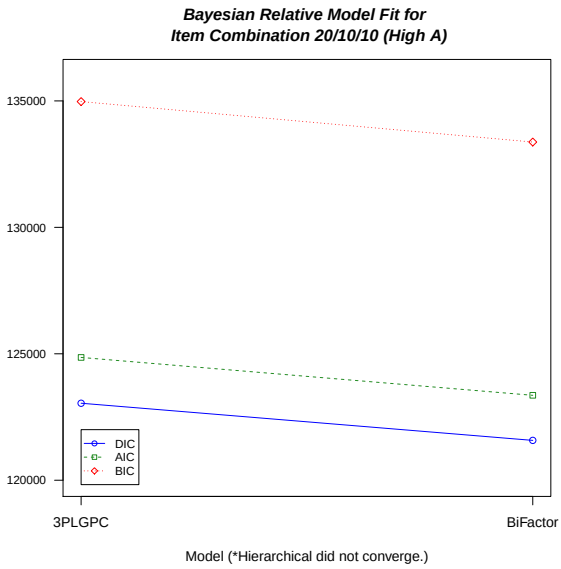

Figure 4.6: Higher A

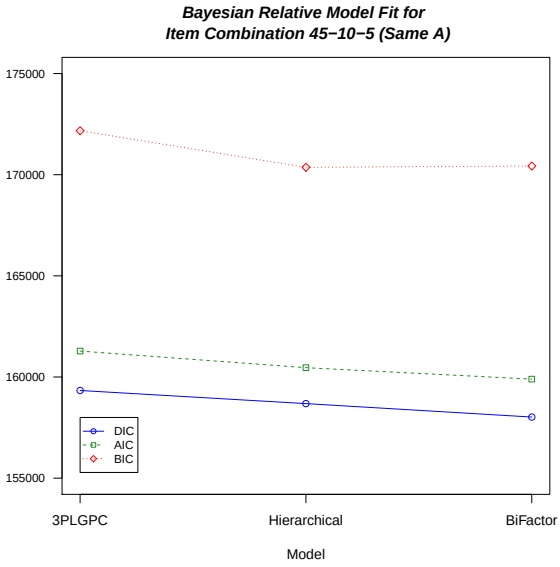

Figure 4.7: Same A

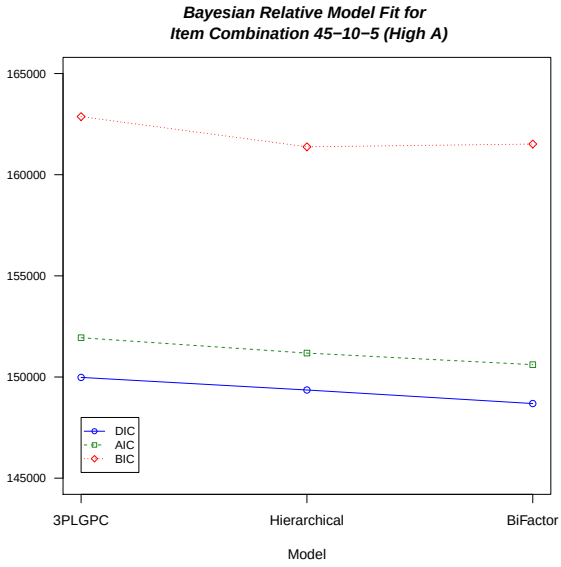

Figure 4.8: Higher A

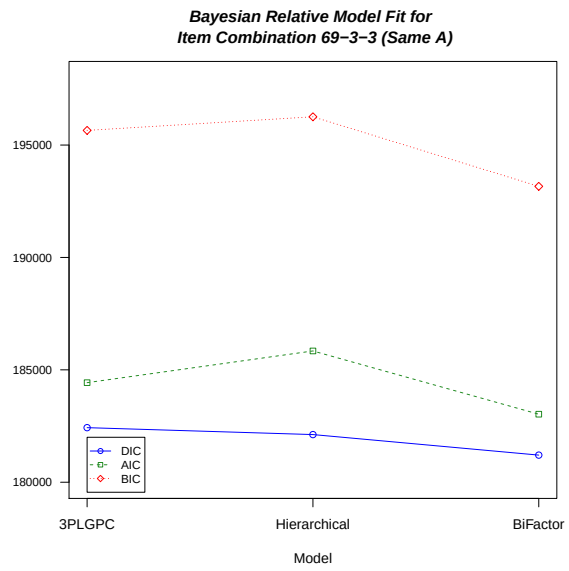

Figure 4.9: Same A

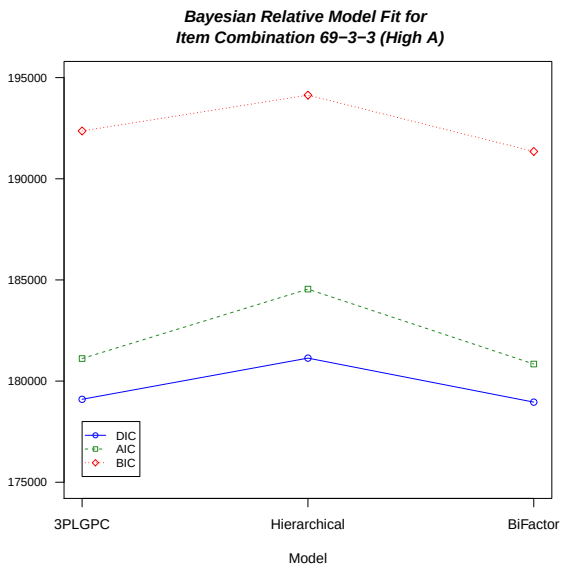

Figure 4.10: Higher A

Based on the AIC, BIC and DIC calculations of Bayesian convergence, it is clear that the Bi-factor model fits the unidimensional data better than the unidimensional model or the hierarchical IRT model for most of the mixed format data sets reviewed in this study. Answering the question of which model fit the data best satisfies one of the research questions but does not tell the whole story. In the next section the mean and standard deviation from the Bi-factor model will be discussed.

4.4 Bi-factor *A*-parameters

The table below illustrates the distributions for each of the *a*-parameters in the bi-factor models by items combination. Appendix G contains the full *a*-parameters structure for each of the item combinations. What is clear from the structure is that the specific factors representing the dichotomously scored and the polytomously scored items are larger than the *a*-parameters on the general dimension. This factoring of the unidimensional data, coupled with the evidence that the bi-factor model fit the data better, indicates that there is an element of mixed format data that causes a form of dimensionality or noise in the mixed format data.

The table below illustrates the distribution of the *a*-parameters on each of the general and

specific dimensions. In the polytomous items, in particular, the *a*-parameters on the general dimension is nearly zero in all cases except item combination four. Since these are distributions of the a-parameters, examination of the individual a-parameters should be examined. The complete table of *a*-parameters can be found in Appendix G. However, it is clear that unidimensional data in mixed format can be factored into specific dimensions based on the item format.

Table 4.3: Bi-factor Loading

		Equal A: Both Item types				**Higher A: Polytomous Items**			
		a-Gen Dich	a-Dichotomous	a-Gen Poly	a-Polytomous	a-Gen Dich	a-Dichotomous	a-Gen Poly	a-Polytomous
36/2/2	mean	0.6675	0.8718	0.0485	0.1768	0.6777	0.8896	0.0433	0.1760
	st. dev	0.0229	0.0274	0.0221	0.0230	0.1004	0.1437	0.0139	0.0357
10/15/15	mean	0.4221	1.0723	0.0555	0.0985	0.4179	1.0653	0.0649	0.1120
	st. dev	0.0570	0.1642	0.0220	0.0281	0.0521	0.2004	0.0284	0.0354
20/10/10	mean	0.5022	1.0260	0.0642	0.1234	0.5027	1.0385	0.0771	0.1454
	st. dev	0.0547	0.1711	0.0523	0.0589	0.0560	0.1718	0.0782	0.0853
45/10/5	mean	0.5771	0.9560	0.0666	0.1865	0.5831	0.9494	0.1005	0.2030
	st. dev	0.0721	0.1289	0.0478	0.0542	0.0797	0.1327	0.0860	0.1127
69/3/3	mean	0.6681	0.8274	0.0454	0.1610	0.6661	0.8302	0.0598	0.1787
	st. dev	0.0915	0.1191	0.0317	0.0663	0.0960	0.1198	0.0596	0.0809

4.5 Hierarchical MIRT *A*-parameters

The hierarchical MIRT models did not improve the fit of the data over the bi-factor models. By examining the loading structure, it is clear that these models did not explain the data (Appendix H). As noted above, model two and three resulted in a negative pD value. Since the negative pD value only occurred with model three and the same OpenBUGS script for the hierarchical model was used for all hierarchical models, this cannot be a result of an error in the script file. Additionally, the same data set was used to fit the unidimensional and the bi-factor models, so the issue is not with the data. It may be that these item combination take much longer to run than other hierarchical MIRT models in the study. Re-parameterizations of this model might also improve convergence.

In examining the loading structure of the model for this data set, the loading of the second specific dimension onto the general dimension was essentially zero. Although the polytomous items loaded onto the specific dimension that dimension did not load onto the general dimension. In a hierarchical MIRT model the general dimension accounts for the correlation between the specific dimensions. In this case, that correlation was very low therefore the model did not fit

the data. A similar issue occurred with model two where there are more polytomous items than dichotomous items. Although that model did not result in a negative pD value, from the loading structure (Appendix H) the lambda two which represents the polytomous items did not load onto the general dimension very well. In fact, that loading is nearly zero.

Additionally, notice that all of the *a*-parameters from the specific dimension onto the general dimension are nearly zero. The dichotomous items load onto the specific dimension and then onto the general dimension very strongly. In fact, the polytomous items load onto the polytomous specific dimension strongly but there does not seem to be correlation between the two specific dimensions as measured by the general dimension. However, it is important to note that two of the item combinations failed to converge consistently with this model.

One possible explanation for this phenomenon is the topology of the posterior distribution. Marin, Mengersen, and Robert (2005) stated that in a mixture model, instead of singling out one mode of the posterior the parameter space may include parts of several models resulting in a posterior mean that lay in a very low probability region. Even when the model specifies which region of the posterior the maximum is likely to be found, that does not guaranty better fit as the constraints may be at odds with the topology of the distribution (Marin et al., 2005). Since the Hierarchical MIRT seeks to identify a general ability which accounts for the correlation between the specific dimensions, rather than allowing the general and specific dimensions to be identified individually as in the bi-factor model, the topology of the space may account for the deviance in the poster mean of the general dimension. The Hierarchical model did not fit or explain the unidimensional mixed format data as well as the generating unidimensional model or the Bi-Factor model. Interpretation of the Hierarchical model results should be done with caution.

Chapter 5

Discussion

5.1 Bayesian Model Fit and Complexity

Unidimensional data was generated using FORTRAN in five mixed format item combinations. For each item combination there were two level of discrimination modeled. The model used to generate the data was the IRT 3PL model for all dichotomous items and the Generalized Partial Credit model for the polytomous items. The data was then fit to three models from a Bayesian approach using OpenBUGS. The three models consisted of; (1) the same unidimensional model used to generate the data - 3PL/GPC, (2) the bi-factor model, and (3) the hierarchical multidimensional IRT model (second order model). Fifty data sets for each of the ten model combinations were generated. In order to fit the fifty set per item combination, OpenBUGS was ran in batch mode using script files. OpenBUGS output the coda files used to determine convergence and log files for each run used to calculate deviance, error and in the unidimensional case RMSE and BIAS.

The log files output by OpenBUGS could not be read be FORTRAN because each line was missing a carriage return at the end of the line. While log files can also be generated from the coda files using R, those log files do not contain the deviance statistics. In addition, it took hours for R to read the coda files and output a new log files To correct this problem of reading the OpenBUGS log files an EXCEL macro was written to open and save each of the log files. In doing this the new log files contained the required carriage return and this process took only seconds to run.

Time turned out to be critical factor in this study. While the 3PL/GPC models generally took

less than 24 hours to compile, the bi-factor and hierarchical MIRT models took much longer. The items combinations containing more polytomous items than dichotomous items took the longest to converge. The longest running item combination was the 10/15/15 combination. That combination took a minimum of six day to run for both the bi-factor and the hierarchical MIRT model. The only way to accomplish this was to use the hpc computing system described in the results section This allowed all 50 replication to run consecutively on individual nodes. The 20/10/10 item combination also took about six days to run for both the bi-factor and the hierarchical MIRT model. This time concern is a barrier to future research using this methodology. Possible providing initial values rather than allowing OpenBUGS to generate initial values might shorten the time requirement. But, even when it appeared that initial values shortened the time required for burn-in the second phase of model updating often took multiple days for most models. This study could not have been completed in a timely manner without the HPC computer at the University of Kansas.

It was clear from the Bayesian convergence criteria of BIC, AIC and DIC, that the bi-factor model fit the unidimensional data better. This is consistent with the findings of (*Preliminary Reports on Spearman-Holzinger Unitary Trait Study*, 1930-1936). The bi-factor model better explains the data created from mixed format assessments. There does not seem to be a significant different in this affect across the item combinations. This indicated that even a few polytomous item added onto an assessment create a dimensionality effect that needs to be accounted for.

5.2 Parameter Structure

What is clear from the *a*-parameter structure of the bi-factor model is that there does appear to be level of dimensionality present in these unidimensional mixed format data sets. The bi-factor model in this study was designed to fit *a*-parameters onto the general ability dimension and then to fit *a*-parameters to the dichotomous items of one specific dimensions and *a*-parameter to the polytomous items of the second specific dimension. What we notice in the pattern, for all item combinations, is that the *a*-parameter value for the specific dimension representing the polytomous items was larger than the *a*-parameters of those same items onto the general construct. This seems

to indicate that the Bi-factor model is factoring the mixed format data away from from the general construct and into two distinct specific factors. Since this data is unidimensional, it cannot be dimensionality, but rather, is likely to be noise in the data set resulting from combining polytomous and dichotomous item the scaling process.

No matter the source, this noisse needs to be accounted for when scaling models and assigning ability score to examinees. This dimensionality effect may explain the convergence rates discovered by (Montgomery & Skorupski, 2012). In that study, data generated to specific model failed to fit then generated models when calibrated in PARSCALE (Montgomery & Skorupski, 2012). The current study provides an explanation for that phenomenon. When item combinations are more complex in terms of the number of polytomous to dichotomous item on the assessment, the data contains enough noise to create the appearance of dimensionality which the Bi-Factor model in the current study is trying to account for.

5.3 Hypothesis and Research Questions

This study began with the following hypothesis:

Hypothesis. The bi-factor model will fit the unidimensional mixed format data better than the unidimensional IRT model or the second order model. Since the specific dimensions in the bi-factor model are not accounted for by the general dimension, any format effect dimensionality will be evident in the loading structure of the bi-factor model.

The hypothesis was shown to be true in this study as in fact the bi-factor model did fit the unidimensional IRT data better than either the unidimensional model or the second order model (hierarchical MIRT). As strategy for verifying the hypothesis the following research questions were defined.

1. How well does the unidimensional model recover item and examinee parameters across the simulation conditions?

2. Which model fits the unidimensional data best; 3PL/GPC, bi-factor, or hierarchical MIRT?

3. Is the model fit affected by the proportion of dichotomous to polytomous items or by the level of discrimination?

In answer to question one, the parameter recovery in terms of Bias and RMSE in the unidimensional case showed that the parameter were recovered with little bias. Model two presented the most bias and that was found to be in part because several of the runs in that model had more error in the retained estimations. This model did not require as many monte carlo chains nor as much time as did the bi-factor and the hierarchical MIRT model.

As to questions two and three, the bi-factor was the clear winner in terms of the Bayesian convergence criterion BIC, AIC and DIC. This improvement in model fit did not seem to be affected by the item combination or the whether or not the level of discrimination was the same for both the polytomous and the dichotomous as opposed to a higher level of discrimination for the polytomous items. The issue model fit does not seem to be impacted by the number of polytomous or dichotomous items. However, item combination did have an impact on the fit of the hierarchical MIRT model. In data sets were there were equal numbers of polytomous and dichotomous items the MIRT model did not fit the data at all. Even when the MIRT models were able to converge, the fit of the models were worse than either the bi-factor or the 3PL/GPC models.

5.4 Limitations and Future Research

With any simulation study the main limitation is that the data was generated rather than collected from real examinees. The data used in this study was simulated. In this study the goal of the study was to compare a variety of mixed format assessments which were then fit to three different IRT models in order to answer the question which model fits the data best. Questions of this form are best answered initially with simulated data. To valuable follow up study would be to fit data from a mixed format assessment collected from real examinees to these three model and compare the outcome to those found in this study. One of the limitations to performing such a follow up exam is access to data from mixed format assessments. While there are many testing companies

piloting and moving toward large scale assessments that are mixed format, those data sets are not accessible to all researchers.

While several of the item combinations used in this simulation are less likely to currently exist in practice, that was by design. One of the research questions was to see if there model fit varied across the different item combinations. To answer that question required a use of item combinations that might not typically be utilized but are possible combination that test developers might use. This aspect of the study did not find a difference in model fit but it did highlight time required to fit a bi-factor and second order model using a fully Bayesian approach. Follow up studies could explore ways to reduce the time requirement.

Since this study showed the bi-factor model fit the unidimensional model the better than the other two model, more studies should be conducted to examine the information that could be obtained from using the bi-factor model in these types of mixed format assessments. In this study, the bi-factor model was used to model the data from dichotomous items versus polytomous items. Further study could be conducted to examine the bi-factor model to model the polytomous items in more than one specific dimension. For example, this study contain two score point levels for the dichotomous items. This study could be extend to model each of those score point levels as separate specific dimensions rather that considering them as one set of polytomous items as was the case with this study.

Additionally, the bi-factor model is a tough sell to stake holders. It is hard to explain why examinees now have multiple ability scores. It is all a challenge for testing companies when providing, interpreting and explaining the overall score from a bi-factor model. Given the difficulty of explaining and utilizing the bi-factor model in practice, even though it fits the data better, what this study should highlight is that the even thought it has been commonly believed that two unidimensional models could be combined that does not seem to be the case. There is the option of scaling the dichotomous and polytomous items separately but that is not a satisfactory answer either. There are unanswered questions that must be examined by future studies such as using a 2PL instead of the 3PL used in this study. Additional, using the GRM instead of GPC to determine

the the results are present with these different models.

Finally, if large scale assessments need to move toward a bi-factor or hierarchical MIRT model, or other complex model, better software is needed. OpenBUGS offers the ability to preform the model fitting with a monte carlo methodology in a purely Bayesian approach. But, the cost of this type of approach is time. While OpenBUGS allows the freedom to run the monte carlo chain as long as need to reach convergence there is no way of knowing from the outset how long it will take. In this study it was not uncommon for a multi-dimensional models to take five to tens to complete on run of 33000 chain lengths. This time could be shortened a little by providing initial values, more informative priors or simplifying the data sets or the models. Research should be conducted to examine ways to improve time including looking at other software programs.

5.5 Conclusion

Even though this was a simulations study and did not utilize real examinee scores, there are some important outcomes for this study that needs be considered in educational measurement practice as more and more large scale assessments include mixed format. This study also illustrates the possibility for follow up studies that could improve our understanding of mixed format assessments. First of all, this study showed that unidimensional data in mixed format assessments fit the bi-factor model better than the unidimensional model from which it was generated. Does this mean that large scale assessments in mixed format assessment form should be scaled using a bi-factor model? Probably not. Much more study is needed to understand why this phenomenon exists in mixed format data sets.

This study, and others like it, point out that we need a better way to scale mixed format assessment with IRT models. Since the bi-factor model fit the unidimensional model better than the unidimensional model, it is clear that there is much to learn about the latent space created by mixed format assessments. It may be that the mixed format assessment creates a bimodal distribution in the posterior or creates boundary conditions that cause difficulty in fitting the models. The bi-factor model may account for complexity of the space by not requiring correlations between items

of different item formats. But when measuring one unidimensional construct with a mixed format assessment, the bi-factor model does not make measurement sense. Research should continue into methodology that allows unidimensional data from mixed format assessment to be scaled together as one assessment measuring one construct.

This is a timely topic because many large assessment are adding polytomous items to their existing assessments. It is important that education measurement experts be able to accurately scale those assessments and place examinees accurately on the ability scale. Follow-up studies should include examination of the 2PL/GPC as well as the graded response model, as well as other unidimensional combinations, to see if those model types fit the bi-factor model better than the generating models.

Finally, more research into the complexity created by mixed format assessments including the topology of the posterior and the parameter space must continue. We do not know enough about the posterior or the parameter space created by mixed format assessments. The field had believed that we could combine two unidimensional models, one for the dichotomous and one polytomous items, and use that combined model to fit mixed assessments. But this study and the previous study by (Montgomery & Skorupski, 2012) illustrates that this does not work in practice. Research into the best way to scale mixed format assessments must continue.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Bennett, R., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement*, *24*, 294-309.

Bennett, R., Rock, D., & Wang, M. (1991). Equivalence of free response and multiple choice items. *Journal of Educational Measurement*, *28*(1), 77-92.

Bugs tutorial. (2012). `http://mathstat.helsinki.fi/openbugs/Manuals/Tutorial.html`. (Accessed: 2014-07-12)

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological methods*, *16*(3), 221.

Cao, Y. (2008). *Mixed-format test equating: effects of test dimensionality and common-item sets*. Proquest dissertations and thesis, University of Maryland, College Park, Retreived from; http://search.proquest.com/docview/288069188accountid=14556.(288069188).

Ccss process. (2009). The Common Core State Standards Initiative.

Chon, K. H., Lee, W.-C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed irt models. *Journal of Educational Measurement*, *47*(3), 318-338.

Columbia university. (2012). `http://www.columbia.edu/cjd11/charles-dimaggio/DIRE/styled-4/styl` (Accessed: 2014-07-12)

De Ayala, R. (2009). *The theory and practice of item response theory*. The Guilford Press.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order irt

model approach. *Applied Psychological Measurement*, *34*(4), 267–285.

DeMars, C. (2006). Application of the bifactor multidimensional item response theory model to testlet based tests. *Journal of Educational Measurement*, *43*(2), 145-168.

DeMars, C. (2008). Polytomous differential item functioning and violations of ordering of the expected latent trait by the raw score. *Educational and psychological measurement*, *68*(3), 379.

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (2005). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*(2), 137–154.

Gelman, A. (1996). Inference and monitoring convergence. In *Markov chain monte carlo in practice* (pp. 131–143). Springer.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, *1*(3), 515–534.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of Markov Chain Monte Carlo*, 163–174.

Gibbons, R., Bock, R., Hedeker, D., Weiss, D., Segawa, E., Bhaumik, D., ... Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4-19.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Julian, E., Wendt, A., Way, D., & Zara, A. (2001). Moving a national licensure examination to computer. *Nurse Educator*, *26*(6), 264.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136-153.

Kang, T. (2006). *Model selection methods for unidimensional and multidimensional irt models*. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Kang, T., & Cohen, A. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331-358.

Kang, T., Cohen, A., & Sung, H. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*(7), 499-518.

Kim, S., & Kolen, M. (2006). Robustness to format effects of irt linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357-381.

Kim, S., & Lee, W. (2006). An extension of four irt linking methods for mixed format tests. *Journal of Educational Measurement*, *43*(1), 53-76.

Kim, S., Walker, M., & McHale, F. (2010). Comparisons among designs for equating mixed format tests in large scale assessments. *Journal of Educational Measurement*, *47*(1), 36-53.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*(1), 1-17.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous irt models. *Applied Psychological Measurement*.

Linn, R. (2000). Assessments and accountability. *Educational researcher*, *29*(2), 4-16.

Madaus, G., & O'Dwyer, L. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan*, *80*, 688-695.

Marin, J.-M., Mengersen, K., & Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, *25*, 459–507.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Md Desa, D. Z. N. (2012). *Bi-factor multidimensional item response theory modeling for subscores estimation, reliability, and classification*. phdthesis, University of Kansas, Lawrence, KS.

Montgomery, M., & Skorupski, W. (2012). Investigation of irt parameter recovery and classification accuracy in mixed format. In *Paper presented at the annual meeting of the nation council of measurement in education.*

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, *16*(2), 159-176.

O'Neill, T., Marks, C., & Reynolds, M. (2005). Re-evaluating the nclex-rn passing standard. *Journal of Nursing Measurement*, *13*(2), 147-167.

Preliminary reports on spearman-holzinger unitary trait study (Tech. Rep. No. 1-8). (1930-1936). Statistical Labaratory, Department of Education, University of Chicago.

Reise, S., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19-31.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bifactor, the testlet, and a second order multidimensional irt model. *Journal of Educational Measurement*, *47*(3), 361-372.

Samejima, F. (1997). Graded response model. *Handbook of modern item response theory*, 85-100.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Spiegelhalter, D. J., Best, N., Carlin, B. P., & Van der Linde, A. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models* (Tech. Rep.). Citeseer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Strout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293-325.

Sturtz, S., Gleman, A., & Ligges, U. (2005). R2winbugs: A package for running winbugs from r. , *12*(3), 1-16.

Sykes, R., & Yen, W. (2000). The scaling of mixed item format tests with the one parameter and two parameter partial credit models. *Journal of Educational Measurement*, *37*(3), 221-244.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making bugs open. *R news*, *6*(1), 12–17.

Thurstone, L. (1947). *Multiple-factor analysis, a development and expansion of the vectors of mind*. Chicago, Illinois: The University of Chicago Press.

Traub, R. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. *Construction versus choice in cognitive measurement*, 29-44.

van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous irt models. *Psychometrika*, *70*(2), 283–304.

Whittaker, T., Chang, W., & Dodd, B. (2012). The performance of irt model selection methods with mixed-format tests. *Applied Psychological Measurement*, *36*(3), 159-180.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469-492.

Zenisky, A., & Sireci, S. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, *15*(4), 337-362.

Zenisky, A. L., & Sireci, S. G. (2001). Feasibility review of selected performance assessment item types of the computerized uniform cpa exam. *Laboratory of Psychometric and Evaluative Research Report*, *406*.

Appendix A

Script

modelCheck('PATH/3PLGPC.txt') !Defines the path to the OpenBUGS model file
modelData('PATH/DATA1.TXT ') !Defines the path to the data file
modelCompile(2)
modelGenInits() !Generates the initial values (could load values with ModelInits(path/filename))
modelUpdate(1000)
samplesSet(a)
samplesSet(b)
samplesSet(c)
samplesSet(theta)
summarySet(a)
summarySet(b)
summarySet(c)
summarySet(theta)
dicSet()
modelUpdate(5000)
samplesStats("*")
dicStats()
samplesCoda("*",'PATH/ 1') !Saves Coda files for each run (each run is numbered)
modelSaveLog('PATH/log1.TXT ') !Saves a text file with the the summary stats and the deviance statistics
modelQuit('yes')

Appendix B

3PLGPC Model

```
model{
#Models defined
      for (i in 1:N){
            for (j in 1:D) {                # dichotomous items
            r[i,j] ~ dbern(p[i,j])
            t[i,j] <- exp(-a[j]*(theta[i] - b[j]))
            p[i,j] <- c[j]+(1-c[j])/(1 + t[i,j])
            }
      for (j in (D+1):(D+P1)) {               #polytomous (1,2,3)
            r[i,j] ~ dcat(pd[i,j,1:nK1])
      for (k in 1:nK1){
            td[i,j,k] <- a[j] * (theta[i] - thresh[j,k])
            psum[i,j,k] <- sum(td[i,j,1:k])
            exp-psum[i,j,k] <- exp(psum[i,j,k])
            pd[i,j,k] <- exp-psum[i,j,k]/sum(exp-psum[i,j,1:nK1])
      } }
      for (j in (D+P1+1):(D+P1+P2)) {               #polytomous (1,2,3,4)
      r[i,j] ~ dcat(pd[i,j,1:nK2])
            for (k in 1:nK2)
            td[i,j,k] <- a[j] * (theta[i] - thresh[j,k])
            psum[i,j,k] <- sum(td[i,j,1:k])
            exp-psum[i,j,k] <- exp(psum[i,j,k])
            pd[i,j,k] <- exp-psum[i,j,k]/sum(exp-psum[i,j,1:nK2])
            } } }
# Priors
      for (i in 1:N){
      theta[i] ~ dnorm(0,1)
      }
      for (j in 1:D) {              #D == number of dichotomous items
      a[j] ~ dlnorm(0,1)
      b[j] ~ dnorm(0,1)
      c[j] ~ dbeta(5,17)
      }       for (j in D + 1:D + P1) {             #P1 == polytomous items modeled (1,2,3)
      thresh[j, 1] <- 0.0
      a[j] ~ dlnorm(0,1)
      for (k in 2: nK1) {             #nK1 == the threshold boundaries
      thresh [j, k] ~ dnorm(0, 1) }
      b[j] <- mean(thresh[j, 1:nK1])
      for (k in 1:nK1) {
      step[j, k] <- b[j] - thresh[j, k]
      } }
      for (j in (D + P1 + 1): (D + P1 + P2)){             #item modeled as (1,2,3,4)
      thresh[j, 1]<- 0.0
      a[j] ~ dlnorm(0,1)
            for (k in 2: nK2) {
            thresh [j, k] ~ dnorm(0, 1)
            }
      b[j] <- mean(thresh[j, 1:nK2])
            for (k in 1:nK2) {
            step[j, k] <- b[j] - thresh[j, k] }} } # MODEL END
```

Appendix C

Bifactor Model

```
model {
# Bifactor: 2d-3PL model calibration: MC items
      for (i in 1:N) {
            for (j in 1:D) {
            r[i,j]~ dbern(p1[i,j])
            t[i,j] ← exp(aG[j]*thetaG[i] + aS1[j]*thetaS1[i] + d[j])
            p1[i,j] ← c[j]+(1-c[j])/(1+t[i,j])
            }
            for (j in (D+1):(D+P1)) {        #Bifactor:polytomous (1,2,3)
            r[i,j] ~ dcat(pd[i,j,1:nK1])
            for (k in 1:nK1) {
            td[i,j,k] ← (aG[j]*thetaG[i] + aS2[j]*thetaS2[i] + d[j] + tt[j,k])
            psum[i,j,k] ← sum(td[i,j,1:k])
            exp-psum[i,j,k] ← exp(psum[i,j,k])
            pd[i,j,k] ← exp-psum[i,j,k]/sum(exp-psum[i,j,1:nK1])
            } }
            for (j in (D+P1+1):(D+P1+P2)) {        #Bifactor:polytomous (1,2,3,4)
            r[i,j] ~ dcat(pd[i,j,1:nK2])
            for (k in 1:nK2) {
            td[i,j,k] ← (aG[j]*thetaG[i] + aS2[j]*thetaS2[i] + d[j] + tt[j,k])
            psum[i,j,k] ← sum(td[i,j,1:k])
            exp-psum[i,j,k] ← exp(psum[i,j,k])
            pd[i,j,k] <- exp-psum[i,j,k]/sum(exp-psum[i,j,1:nK2])
            } } }

      #priors for dichotomous
#Priors Thetas
      for (i in 1:N) {
      thetaG[i] ~ dnorm(0,1)
      thetaS1[i] ~ dnorm(0,1)
      thetaS2[i] ~ dnorm(0,1)
      }
      for (j in 1:D) {
      aG[j]~ dnorm(0,1) I(0,)
      aS1[j]~ dlnorm(0,1)
      d[j]~ dnorm(0,.5)
      c[j]~ dbeta(5,17)
      }
      for (j in D+1:D +P1) {
      tt[j,1] ← 0
      aG[j]~ dnorm(0,1) I(0,)
      aS2[j]~ dlnorm(0,1)
      d[j]~ dnorm(0,.5)
      for (k in 2:nK1){
      tt[j,k]~ dnorm(0,1)
      } }
      for (j in D+P1+1:D +P1+P2) {
      tt[j,1] ← 0
      aG[j]~ dnorm(0,1) I(0,)
      aS2[j]~ dlnorm(0,1)
```

```
d[j]~ dnorm(0,.5)
for (k in 2:nK2){
tt[j,k]~ dnorm(0,1)
} } }
```

Appendix D

Second Order Model

```
model {
# 2nd Order: 2d-3PL model calibration: MC items
      for (i in 1:N) {
      thetaS[i,1] ← (lambda[1]*thetaG[i] + eta[1])
      thetaS[i,2] ← (lambda[2]*thetaG[i] + eta[2])
            for (j in 1:D) {
            r[i,j] dbern(p1[i,j])
            t[i,j]<- exp(aS[j]*thetaS[i,1] + d[j])
            p1[i,j]<- c[j]+(1-c[j])/(1+t[i,j])
      }
      for (j in (D+1):(D+P1)) {              #2nd order: polytomous (1,2,3)
            r[i,j] ∼ dcat(pd[i,j,1:nK1])
      for (k in 1:nK1) {
            td[i,j,k] ← (aS[j]*thetaS[i,2] + d[j] + tt[j,k])
            psum[i,j,k] ← sum(td[i,j,1:k])
            exp-psum[i,j,k] ← exp(psum[i,j,k])
            pd[i,j,k] ← exp-psum[i,j,k]/sum(exp-psum[i,j,1:nK1])
            } }
      for (j in (D+P1+1):(D+P1+P2)) {              #2nd Order: polytomous (1,2,3,4)
            r[i,j]∼ dcat(pd[i,j,1:nK2])
      for (k in 1:nK2) {
            td[i,j,k] ← (aS[j]*thetaS[i,2] + d[j] + tt[j,k])
            psum[i,j,k] ← sum(td[i,j,1:k])
            exp-psum[i,j,k] ← exp(psum[i,j,k])
            pd[i,j,k] ← exp-psum[i,j,k]/sum(exp-psum[i,j,1:nK2])
            } } }
#priors for specific dimensions
      for (k in 1:2) {
      lambda[k]∼ dnorm(0,1) I(0,)
      eta[k]∼ dnorm(0,1)
      } #Priors Thetas
      for (i in 1:N) {
      thetaG[i] ∼ dnorm(0,1)
      }
      for (j in 1:D) {
      aS[j] ∼ dlnorm(0,1)
      d[j] ∼ dnorm(0,1)
      c[j] ∼ dbeta(5,17)
      }
      for (j in D+1:D +P1) {
      tt[j,1] ← 0
      aS[j] ∼ dlnorm(0,1)
      d[j] ∼ dnorm(0,1)
      for (k in 2:nK1){
      tt[j,k] ∼ dnorm(0,1)
      } }
      for (j in D+P1+1:D +P1+P2) {
      tt[j,1] ← 0
      aS[j] ∼ dlnorm(0,1)
      d[j] ∼ dnorm(0,1)
```

```
for (k in 2:nK2){
tt[j,k] ~ dnorm(0,1)
} } }
```

Appendix E

Proof

$\sum_{i=1}^{n} \frac{a_i}{b_i} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$ does not hold for all a_i or $b_i \in \Re$.

Proof: If $a_i \geq 0$,$b_i > 0$ $\forall i$,

$\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \max_{1<i<n} \frac{a_i}{b_i}$ is true when $\frac{a_i}{b_i} = \ldots = \frac{a_n}{b_n}$

and

$\max_{1<i<n} \frac{a_i}{b_i} \leq \sum_{i=1}^{n} \frac{a_i}{b_i}$ is true only when at most one $a_i \neq 0$.

It follows that $\sum_{i=1}^{n} \frac{a_i}{b_i} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$ only when all $a_i = 0$

Therefore, $\forall i$ $a_i \geq 0$,$b_i > 0$ and when all $a_i \neq 0$

$$\sum_{i=1}^{n} \frac{a_i}{b_i} > \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

Appendix F

Monte Carol Estimate Error Tables

Table F.1: 3PL/GPC MCMC Average Error by Parameter

	All Items:Equal A				**Polytomous Items:Higher A**			
	a	b	c	Theta	a	b	c	Theta
36/2/2	0.0571	0.0690	0.0683	0.0242	0.0554	0.0670	0.0644	0.0243
10/15/15	0.0531	0.0601	0.0352	0.0128	0.0538	0.0658	0.0460	0.0127
20/10/10	0.0533	0.0584	0.0677	0.0242	0.0561	0.0604	0.0680	0.0242
45/10/5	0.0551	0.0654	0.0655	0.0249	0.0563	0.0660	0.0656	0.0250
69/3/3	0.0561	0.0688	0.0648	0.0246	0.0566	0.0691	0.0691	0.0246

Table F.2: Bifactor MCMC Average Error by Parameter

	All Items:Equal A					
	aG	aS1	aS2	ThetaG	ThetaS1	ThetaS2
36/2/2	0.0483	0.0451	0.0505	0.0250	0.0143	0.0348
10/15/15	0.0299	0.0467	0.0213	0.0157	0.0135	0.0121
20/10/10	0.0422	0.0470	0.0244	0.0202	0.0178	0.0128
45/10/5	0.0422	0.0468	0.0245	0.0247	0.0235	0.0119
69/3/3	0.0469	0.0491	0.0350	0.0288	0.0285	0.0123
	Polytomous Items:Higher A					
	aG	aS1	aS2	ThetaG	ThetaS1	ThetaS2
36/2/2	0.0487	0.0507	0.0436	0.0261	0.0256	0.0137
10/15/15	0.0300	0.0462	0.0216	0.0158	0.0135	0.0121
20/10/10	0.0349	0.0453	0.0225	0.0192	0.0119	0.0119
45/10/5	0.0424	0.0473	00244	0.0249	0.0238	0.0118
69/3/3	0.0474	0.0495	0.0348	0.02955	0.0292	0.0123

Table F.3: 2^{nd} Order MCMC Average Error by Parameter

	All Items:Equal A				
	aS1	aS2	ThetaG	ThetaS1	ThetaS2
36/2/2	0.0515	0.0505	0.0130	0.0324	0.0324
45/10/5	0.0482	0.0485	0.0119	0.0437	0.0437
69/3/3	0.0458	0.053	0.0120	0.0611	0.0610
	Polytomous Items:Higher A				
	aS1	aS2	ThetaG	ThetaS1	ThetaS2
36/2/2	0.0519	0.0544	0.0125	0.0370	0.0364
45/10/5	0.0490	0.0571	0.0119	0.0463	0.0463
69/3/3	0.0463	0.0539	0.0120	0.0610	0.0610

Appendix G

Bifactor Loadings

Table G.1: Bifactor Loading Structure Matrix

Combination 36/2/2

	Equal A: Both Item types				**Higher A: Polytomous Items**			
Item	Intercept	aG	aS1	aS2	Intercept	aG	aS1	aS2
1	0.1679	0.6667	0.8672		0.4340	0.6858	0.8582	
2	0.2468	0.6598	0.8458		0.9893	0.6576	0.9037	
3	-0.0783	0.6626	0.8744		-2.2447	0.4883	0.5857	
4	0.1273	0.6841	0.8487		-1.1563	0.8009	1.0742	
5	0.2015	0.6399	0.8882		-0.3854	0.6429	0.8130	
6	-0.1042	0.6409	0.8161		-1.5932	0.7687	1.0105	
7	-0.0069	0.6686	0.8583		1.3373	0.6672	0.8434	
8	0.3405	0.6892	0.9129		-0.7809	0.7160	0.9294	
9	0.2466	0.7006	0.9201		2.0841	0.6514	0.9093	
10	0.1565	0.6836	0.8983		0.7809	0.5625	0.7298	
11	0.1824	0.6460	0.8548		-0.2298	0.7939	1.0472	
12	0.1844	0.6711	0.8547		1.0352	0.6580	0.9114	
13	0.0496	0.6573	0.8544		1.2265	0.6187	0.8410	
14	0.1495	0.6590	0.9016		1.2496	0.7084	0.8901	
15	-0.0214	0.6789	0.8688		-0.6425	0.7113	0.9439	
16	-0.0503	0.6573	0.8310		0.6872	0.8203	1.0827	
17	0.2500	0.7146	0.8936		1.7917	0.8638	1.1353	
18	0.4126	0.6749	0.8654		0.7338	0.6736	0.8518	
19	0.0261	0.6530	0.8365		0.4418	0.7102	0.9242	
20	0.1704	0.6489	0.8924		-1.8364	0.5466	0.6482	
21	0.2267	0.6758	0.8900		0.5070	0.7843	1.0383	
22	0.1583	0.6784	0.9186		-0.1562	0.8436	1.1013	
23	0.0027	0.6799	0.8930		0.7515	0.6649	0.8655	
24	0.1573	0.6928	0.8866		1.1507	0.6843	0.9275	
25	0.0390	0.6644	0.8732		0.8714	0.8286	1.1072	
26	-0.0101	0.6339	0.8324		0.6766	0.6984	0.9128	
27	0.3084	0.6157	0.9016		1.8387	0.5930	0.7237	
28	0.0955	0.6544	0.8809		2.8920	0.4324	0.5361	
29	0.2649	0.6634	0.8707		-1.2287	0.7137	0.9948	
30	0.0925	0.7230	0.8885		-0.1637	0.7034	0.9391	
31	-0.1481	0.6237	0.8315		1.3175	0.6662	0.9116	
32	0.0341	0.6698	0.9042		-0.1231	0.5495	0.7471	
33	0.2531	0.6962	0.8830		1.1108	0.7445	0.9631	
34	0.1762	0.6742	0.8304		1.5572	0.6151	0.8343	
35	0.0352	0.6700	0.8465		0.5871	0.5349	0.7136	
36	0.0446	0.6569	0.8703		2.6944	0.5944	0.7754	
37	-0.3637	0.0697		0.1915	0.6576	0.0606		0.2256
38	-0.3766	0.0653		0.1946	0.5365	0.0484		0.1721
39	-0.4433	0.0312		0.1768	-0.8185	0.0333		0.1652
40	-0.3941	0.0276		0.1442	-0.7605	0.0307		0.1410

Table G.2: Bifactor Loading Structure Matrix

Combination 10/15/15

	Equal A: Both Item types				**Higher A: Polytomous Items**			
Item	Intercept	aG	aS1	aS2	Intercept	aG	aS1	aS2
1	0.1929	0.4275	1.1388		0.5910	0.4464	1.1907	
2	0.2447	0.3779	1.0112		1.0698	0.3720	1.1319	
3	0.2494	0.3609	0.7201		-2.2458	0.3500	0.7044	
4	0.2572	0.4977	1.3300		-1.1216	0.4300	1.4070	
5	0.2503	0.4216	1.0001		-0.3207	0.3891	0.9612	
6	0.2523	0.5375	1.2112		-1.5687	0.4704	1.1825	
7	0.1781	0.3956	1.0806		1.4052	0.4246	1.0706	
8	0.2662	0.4017	1.2413		-0.7119	0.4072	1.1906	
9	0.1767	0.4354	1.0762		2.0356	0.5207	0.9524	
10	0.2312	0.3652	0.9136		0.7467	0.3687	0.8618	
11	0.5994	0.0716		0.1321	-1.6794	0.0873		0.1691
12	0.9393	0.0955		0.1434	-1.5127	0.1060		0.1609
13	-2.2578	0.0772		0.1280	-1.8427	0.1219		0.1620
14	-1.0874	0.0610		0.1092	0.1826	0.0693		0.1209
15	-0.3233	0.0593		0.1178	-0.9047	0.0690		0.1309
16	-1.5971	0.0850		0.1421	-1.6552	0.1132		0.1709
17	1.3813	0.1003		0.1399	-1.7441	0.1153		0.1811
18	-0.7627	0.0847		0.1200	-0.5308	0.0735		0.1218
19	2.1038	0.0657		0.1183	-0.2998	0.0584		0.1205
20	0.8459	0.0706		0.1012	0.6094	0.0749		0.1172
21	-1.0694	0.0639		0.1180	-0.6646	0.0706		0.1281
22	-1.1613	0.0627		0.1071	-0.0465	0.0707		0.1167
23	-1.1765	0.0639		0.1087	0.3509	0.0591		0.1163
24	0.0878	0.0797		0.1442	-1.8674	0.0924		0.1518
25	-0.6380	0.0861		0.1323	-1.6690	0.1287		0.1682
26	-1.0362	0.0333		0.0708	-0.3891	0.0428		0.0756
27	-1.0467	0.0324		0.0763	-0.5153	0.0419		0.0826
28	-0.2037	0.0360		0.0722	-0.5306	0.0369		0.0817
29	-0.2514	0.0322		0.0725	-0.5552	0.0395		0.0775
30	0.4008	0.0390		0.0685	-0.4844	0.0476		0.0786
31	-0.3962	0.0410		0.0780	-0.5582	0.0506		0.0795
32	-0.1629	0.0339		0.0772	-0.5742	0.0448		0.0871
33	0.1984	0.0379		0.0751	-0.5079	0.0481		0.0814
34	-1.4481	0.0395		0.0776	-0.6120	0.0467		0.0841
35	-1.1428	0.0377		0.0653	-0.2969	0.0375		0.0757
36	-0.2852	0.0398		0.0733	-0.4694	0.0437		0.0851
37	-0.3960	0.0364		0.0674	-0.4469	0.0406		0.0767
38	-0.3908	0.0276		0.0642	-0.6234	0.0336		0.0788
39	-0.4558	0.0348		0.0742	-0.3085	0.0370		0.0902
40	-0.3054	0.0369		0.0807	-0.5923	0.0439		0.0891

Table G.3: Bifactor Loading Structure Matrix

Combination 20/10/10

	Equal A: Both Item types				**Higher A: Polytomous Items**			
Item	Intercept	aG	aS1	aS2	Intercept	aG	aS1	aS2
1	0.4839	0.5112	1.0234		0.9275	0.5177	1.0251	
2	0.8035	0.4583	0.9571		1.8586	0.4552	1.0082	
3	-2.1884	0.3801	0.6872		-1.8067	0.3889	0.6608	
4	-1.1155	0.5586	1.2584		-0.7149	0.5669	1.2543	
5	-0.3796	0.4413	1.0052		0.1925	0.4609	0.9830	
6	-1.4891	0.5320	1.1500		-1.1576	0.5531	1.1697	
7	1.2055	0.4446	0.9515		2.0937	0.5055	0.9751	
8	-0.6919	0.5336	1.0670		-0.2926	0.5011	1.0665	
9	1.8791	0.5130	0.8872		3.1006	0.5668	0.9432	
10	0.7107	0.4429	0.8254		1.5948	0.4380	0.8327	
11	-0.1573	0.5389	1.2652		0.2114	0.5439	1.2756	
12	1.0475	0.5000	1.0115		1.6448	0.5215	1.0272	
13	1.2551	0.4951	0.9627		1.9157	0.4686	1.0016	
14	1.1948	0.5464	1.0199		1.8816	0.5329	1.0137	
15	-0.5652	0.5255	1.1039		-0.1892	0.4864	1.1039	
16	0.7670	0.5788	1.2804		1.2153	0.5630	1.2978	
17	1.7318	0.5977	1.2789		2.4881	0.5955	1.2994	
18	0.7644	0.4772	1.0030		1.3436	0.4667	0.9854	
19	0.4085	0.5283	1.0738		1.0128	0.5153	1.1057	
20	-1.7900	0.4397	0.7078		-1.4199	0.4056	0.7407	
21	-0.3591	0.0605		0.1295	0.2744	0.0648		0.1335
22	-0.1254	0.0504		0.1150	0.8470	0.0601		0.1284
23	0.0984	0.0596		0.1176	1.2268	0.0579		0.1248
24	-1.1681	0.0662		0.1549	-0.7946	0.0864		0.1779
25	-0.8241	0.0748		0.1519	-0.5612	0.0931		0.1907
26	-0.2432	0.0613		0.1189	0.5284	0.0675		0.1344
27	-0.7123	0.0690		0.1617	-0.1073	0.0734		0.1530
28	-2.8897	0.2742		0.3389	-2.8842	0.3964		0.4777
29	1.1036	0.0691		0.1490	2.8232	0.0936		0.1892
30	1.3151	0.0966		0.1652	3.2007	0.1139		0.1880
31	-0.3300	0.0391		0.0915	0.1169	0.0487		0.1073
32	-0.3930	0.0372		0.0829	0.2106	0.0445		0.0981
33	-0.4079	0.0429		0.0874	0.1287	0.0456		0.1038
34	-0.2781	0.0386		0.0805	0.0503	0.0416		0.0921
35	-0.4099	0.0392		0.0880	0.0877	0.0449		0.1007
36	-0.4084	0.0389		0.0885	0.0913	0.0452		0.1021
37	-0.4236	0.0396		0.0873	0.0760	0.0489		0.1000
38	-0.3940	0.0333		0.0875	0.0810	0.0395		0.1079
39	-0.1814	0.0293		0.0852	0.2497	0.0355		0.1002
40	-0.2499	0.0636		0.0876	0.1360	0.0398		0.0974

Table G.4: Bifactor Loading Structure Matrix

Combination 45/10/5

	Equal A: Both Item types				**Higher A: Polytomous Items**			
Item	Intercept	aG	aS1	aS2	Intercept	aG	aS1	aS2
1	0.1897	0.5284	0.8865		0.1671	0.4595	0.9013	
2	1.0397	0.5881	0.9283		1.0908	0.6100	0.9684	
3	-3.3192	0.6055	0.9076		-3.2514	0.6558	0.7836	
4	-0.8278	0.4705	0.7900		-0.8151	0.4546	0.8010	
5	-0.1360	0.4549	0.7957		-0.1716	0.4555	0.7523	
6	-1.4434	0.6337	1.1522		-1.4763	0.6394	1.1132	
7	1.1784	0.4627	0.8317		1.0122	0.4698	0.7653	
8	-0.5591	0.6122	0.9628		-0.5087	0.5537	1.0085	
9	2.4518	0.7004	1.1291		2.4861	0.6696	1.2184	
10	0.7751	0.4415	0.7949		0.7241	0.4131	0.7984	
11	-0.1273	0.5384	0.9370		-0.1484	0.4940	0.9636	
12	1.1208	0.6701	1.0940		1.0995	0.7011	1.1465	
13	1.0543	0.4735	0.7895		1.0441	0.4382	0.7773	
14	0.9776	0.5316	0.8380		0.9605	0.5020	0.8424	
15	-0.6586	0.6348	1.0773		-0.6838	0.6360	1.0463	
16	0.4767	0.6404	1.0833		0.3656	0.6326	1.0097	
17	1.7681	0.6764	1.1612		1.7111	0.5425	1.1672	
18	0.7086	0.5771	0.9534		0.6627	0.5335	0.9297	
19	0.4966	0.6815	1.1381		0.5067	0.6415	1.2056	
20	-2.0964	0.5469	0.8983		-2.0626	0.5377	0.8851	
21	0.5343	0.7082	1.1119		0.5333	0.6966	1.1194	
22	0.1143	0.5273	0.9344		0.0833	0.6156	0.8673	
23	0.6062	0.5443	0.9166		0.6537	0.5777	0.8971	
24	0.9269	0.5867	1.0373		1.0071	0.6261	1.0902	
25	0.7120	0.6268	1.0119		0.6707	0.6063	0.9498	
26	0.8357	0.6551	1.1294		0.9614	0.7741	1.1759	
27	2.1217	0.5782	0.8937		2.1154	0.6357	0.8549	
28	3.1318	0.5416	0.7967		3.1405	0.5730	0.8058	
29	-0.5274	0.5383	0.8820		-0.5543	0.3794	0.9538	
30	-0.1500	0.5683	0.9219		-0.1767	0.5599	0.9332	
31	1.3328	0.5517	0.8400		1.3399	0.5575	0.8496	
32	-0.1257	0.6150	1.1057		-0.1367	0.6717	1.0364	
33	1.1757	0.6356	1.1036		1.1380	0.6281	1.0270	
34	1.6733	0.5789	0.9522		1.6696	0.6071	0.9407	
35	0.7787	0.6377	0.9938		0.7905	0.6678	0.9798	
36	2.2324	0.4547	0.6779		2.3810	0.5426	0.8157	
37	-0.8579	0.6146	1.0349		-0.8089	0.5904	1.0989	
38	0.2551	0.5077	0.9049		0.2219	0.5587	0.8335	
39	0.1533	0.4876	0.7268		0.1526	0.4303	0.7379	
40	0.8369	0.5560	0.9070		0.8986	0.5890	0.8893	
41	-0.0059	0.6418	1.0011		-0.0028	0.5949	1.0207	
42	1.7138	0.6512	1.1198		1.7040	0.6406	1.1562	
43	0.5166	0.6621	1.1441		0.5569	0.7463	1.1673	
44	-0.9516	0.4961	0.8102		-0.9346	0.5934	0.7125	
45	-0.5405	0.5374	0.9124		-0.5438	0.5750	0.8951	
46	-2.9329	0.1954		0.2160	-3.7593	0.2792		0.4451
47	-1.3852	0.0749		0.1589	-2.3190	0.1431		0.2825
48	-2.2233	0.1319		0.2049	-3.2044	0.2757		0.3489
49	0.0942	0.0450		0.1171	0.0258	0.0505		0.1061
50	1.7385	0.0859		0.1645	2.6890	0.1859		0.2915
51	-1.2752	0.0801		0.2233	-2.0801	0.1665		0.2772
52	0.1601	0.0456		0.2314	0.2494	0.0423		0.1299
53	-0.6112	0.0552		0.2342	-0.9860	0.0643		0.1502
54	0.3761	0.0485		0.2020	0.7454	0.0502		0.1751
55	-1.9641	0.1026		0.2426	-2.8651	0.1590		0.3056
56	-0.5523	0.0284		0.1363	-0.6289	0.0269		0.1092
57	-0.2403	0.0261		0.2119	-0.2647	0.0266		0.1216
58	-0.2784	0.0267		0.0801	-0.5476	0.0262		0.0937
59	-0.3824	0.0282		0.1153	-0.5452	0.0391		0.0934
60	-0.4195	0.0250		0.2595	-0.6248	0.0233		0.0984

Table G.5: Bifactor Loading Structure Matrix

Combination 69/3/3

	Equal A: Both Item types				Higher A: Polytomous Items			
Item	Intercept	aG	aS1	aS2	Intercept	aG	aS1	aS2
1	0.2960	0.6201	0.7659		0.3181	0.5656	0.8486	
2	1.3017	0.7723	0.9905		1.3022	0.7687	0.9771	
3	-2.9997	0.6156	0.7782		-2.9922	0.6372	0.7514	
4	-0.8973	0.6445	0.7846		-0.8859	0.6616	0.7822	
5	-0.6178	0.6666	0.8732		-0.6403	0.6643	0.8667	
6	-1.2174	0.5387	0.6653		-1.2041	0.5171	0.6822	
7	1.4620	0.7756	0.9330		1.4110	0.7239	0.9329	
8	-0.3498	0.6325	0.8383		-0.3503	0.5925	0.8523	
9	1.5737	0.5772	0.6540		1.5769	0.6305	0.6189	
10	1.0924	0.6929	0.8454		1.0858	0.6443	0.8657	
11	-0.2898	0.7496	0.9026		-0.2666	0.7233	0.9656	
12	1.0656	0.7906	0.9779		1.0715	0.8308	0.9721	
13	1.5080	0.8226	0.9897		1.5183	0.8137	0.9942	
14	0.9721	0.6031	0.7857		0.9789	0.6605	0.7306	
15	-0.5973	0.6731	0.7776		-0.6013	0.6419	0.7976	
16	0.4228	0.6016	0.6926		0.4305	0.5979	0.7089	
17	1.2433	0.6131	0.7734		1.2517	0.6339	0.7651	
18	0.7671	0.6230	0.7908		0.8082	0.6015	0.8378	
19	0.2733	0.6122	0.7944		0.2472	0.5732	0.7980	
20	-2.5244	0.6722	0.8502		-2.5544	0.7250	0.8170	
21	0.3007	0.5870	0.7254		0.3440	0.6143	0.7230	
22	0.0328	0.6745	0.8714		0.0240	0.6982	0.8484	
23	0.5622	0.6557	0.7330		0.5850	0.6006	0.8027	
24	0.8559	0.6729	0.8092		0.8227	0.6486	0.8027	
25	0.5734	0.6900	0.8808		0.5660	0.6491	0.9178	
26	0.5639	0.6797	0.8191		0.5597	0.6715	0.8412	
27	2.2078	0.6534	0.8740		2.2313	0.7036	0.8477	
28	2.7122	0.4713	0.5662		2.6967	0.4861	0.5440	
29	-1.0543	0.6216	0.7903		-1.0645	0.6117	0.7892	
30	-0.4556	0.7265	0.8736		-0.4650	0.6818	0.9010	
31	1.6186	0.8199	1.0031		1.6601	0.7988	1.0334	
32	-0.1551	0.6228	0.7449		-0.1453	0.6366	0.7120	
33	0.6416	0.5374	0.6429		0.6160	0.5026	0.6629	
34	1.8660	0.8081	0.9688		1.8149	0.7613	0.9758	
35	1.1124	0.9864	1.1581		1.1177	0.9872	1.2071	
36	2.6100	0.6590	0.7540		2.6150	0.6319	0.7688	
37	-0.6961	0.7477	0.8843		-0.7050	0.6974	0.9346	
38	-0.0489	0.6562	0.8053		-0.0280	0.6621	0.7989	
39	0.1196	0.5621	0.6393		0.1034	0.5109	0.6830	
40	0.5114	0.6757	0.8053		0.5096	0.6446	0.8314	
41	0.0183	0.6762	0.8454		0.0222	0.6883	0.8437	
42	1.2508	0.5983	0.7249		1.3036	0.5646	0.7814	
43	0.3944	0.5992	0.7464		0.4074	0.5762	0.7603	
44	-0.7976	0.6291	0.7636		-0.8105	0.6590	0.7519	
45	-0.9326	0.6523	0.8411		-0.9530	0.6637	0.8312	
46	2.4242	0.8430	1.0916		2.3651	0.8074	1.0781	
47	1.4905	0.5750	0.7023		1.5269	0.6005	0.6938	
48	1.9444	0.6066	0.7679		1.9380	0.6377	0.7427	
49	1.6094	0.6444	0.7708		1.5916	0.6305	0.7803	
50	-1.1300	0.6024	0.7209		-1.1184	0.6169	0.7299	
51	1.1408	0.6022	0.7384		1.1324	0.6082	0.7344	
52	0.6014	0.7880	0.9985		0.5945	0.8059	0.9824	
53	-0.5993	0.7107	0.8433		-0.6089	0.7095	0.7952	
54	-0.7935	0.7322	0.9875		-0.7859	0.7378	0.9838	
55	0.4973	0.7976	0.9943		0.4763	0.8207	0.9497	
56	-0.4238	0.5932	0.7500		-0.4116	0.5961	0.7463	
57	0.3700	0.6846	0.8837		0.3517	0.7379	0.8353	
58	0.2655	0.5712	0.7045		0.2536	0.5797	0.7005	
59	-0.4900	0.6960	0.8025		-0.4725	0.6709	0.8218	
60	0.8728	0.8151	1.0620		0.9098	0.8889	1.0491	
61	0.2584	0.8098	1.0235		0.2541	0.8199	0.9930	
62	-0.8350	0.7185	0.9869		-0.8489	0.7876	0.9281	
63	2.1456	0.7439	0.9661		2.1284	0.7664	0.9322	
64	0.3075	0.5409	0.6951		0.3026	0.5300	0.6929	
65	-2.7075	0.5508	0.7071		-2.6751	0.5220	0.7232	
66	0.9210	0.6374	0.7492		0.9201	0.5975	0.7864	
67	-0.6855	0.6204	0.8385		-0.6707	0.6208	0.8604	
68	-3.0694	0.5380	0.7075		-3.0807	0.6104	0.6613	
69	2.6531	0.7182	0.8647		2.7266	0.7006	0.9247	
70	-0.5058	0.0436		0.1671	-0.5283	0.0428		0.1744
71	0.2521	0.0413		0.1605	0.4004	0.0440		0.1840
72	2.1123	0.1079		0.2874	2.8413	0.1808		0.3337
73	-0.4118	0.0274		0.1144	-0.4553	0.0308		0.1208
74	-0.3901	0.0241		0.1098	-0.4713	0.0289		0.1147
75	-0.4157	0.0277		0.1269	-0.4574	0.0316		0.1445

Appendix H

Hierarchical MIRT Loading

Table H.1: Hierarchical Loading Structure Matrix

Combination 36/2/2

	Equal A: Both Item types				**Higher A: Polytomous Items**			
Item	Intercept	aS1	aS2	Lambda	Intercept	aS1	aS2	Lambda
1	0.0227	1.0022			-0.0351	1.0284		
2	-0.1069	0.9290			0.5007	1.0513		
3	-0.0215	1.0059			-2.5219	0.6889		
4	-0.1129	1.0005			-1.6892	1.2384		
5	-0.1603	0.9631			-0.8188	0.9761		
6	-0.5005	0.9906			-2.0841	1.1648		
7	-0.1706	0.9606			0.8369	0.9702		
8	0.0156	1.0356			-1.2668	1.0828		
9	-0.0974	1.0259			1.6004	1.0542		
10	-0.2846	1.0769			0.3719	0.8621		
11	0.0249	0.9915			-0.7908	1.2184		
12	-0.3039	1.0317			0.5461	1.0293		
13	-0.2171	0.9916			0.7880	0.9699		
14	-0.0617	1.0128			0.7628	1.0736		
15	0.0064	0.9734			-1.1377	1.0870		
16	-0.1890	0.9647			0.1155	1.2810		
17	0.1522	1.0220			1.1944	1.3656		
18	0.2842	0.9905			0.2497	0.9799		
19	-0.1116	0.9965			-0.0555	1.0889		
20	-0.1662	0.9900			-2.1472	0.7698		
21	-0.0755	1.0127			-0.0312	1.2327		
22	-0.1259	1.0052			-0.7315	1.2818		
23	-0.1884	1.0134			0.2977	1.0236		
24	-0.0773	1.0319			0.6523	1.0780		
25	-0.1592	1.0154			0.2697	1.2730		
26	-0.2776	0.9872			0.1860	1.0756		
27	0.0431	0.9816			1.4330	0.8814		
28	-0.2327	1.0150			2.5402	0.6278		
29	0.1019	1.0387			-1.7068	1.1166		
30	-0.2660	1.0854			-0.6373	1.0927		
31	-0.2561	0.9422			0.8184	1.0524		
32	-0.3081	0.9810			-0.5284	0.8664		
33	-0.1081	0.9936			0.6055	1.1480		
34	0.0611	0.9728			1.0895	0.9681		
35	-0.2909	0.9609			0.2276	0.8298		
36	-0.2008	0.9492			2.2420	0.9151		
37	-0.4216		1.1285		0.2447		1.0721	
38	-0.1211		1.2872		0.8152		0.9842	
39	-0.3944		0.7403		-0.8210		0.8541	
40	-0.3044		0.8121		-0.6213		0.8303	
Dichotomous				1.1224				1.1037
Polytomous				0.0184				0.0150

Table H.2: Hierarchical Loading Structure Matrix

Combination 40/10/5

	Equal A: Both Item types				**Higher A: Polytomous Items**			
Item	Intercept	aS1	aS2	Lambda	Intercept	aS1	aS2	Lambda
1	-0.1960	0.9351			-0.1895	0.9100		
2	0.5737	0.9926			0.6278	0.9940		
3	-3.5489	0.9610			-3.6080	0.9540		
4	-1.1519	0.8497			-1.1430	0.8284		
5	-0.4983	0.8386			-0.4648	0.8113		
6	-1.8882	1.2179			-1.8950	1.1830		
7	0.8033	0.8857			0.7832	0.8616		
8	-0.9801	1.0524			-0.9576	1.0290		
9	1.9014	1.2285			1.9575	1.2195		
10	0.3769	0.8130			0.4170	0.8169		
11	-0.5333	0.9911			-0.5440	0.9647		
12	0.6166	1.1899			0.6174	1.1582		
13	0.6492	0.8301			0.6938	0.8012		
14	0.5226	0.8995			0.5322	0.8532		
15	-1.1212	1.1552			-1.1184	1.1296		
16	-0.0150	1.1263			-0.0141	1.0780		
17	1.2067	1.2232			1.2595	1.2045		
18	0.2613	1.0164			0.2661	0.9713		
19	-0.0177	1.2385			0.0177	1.2166		
20	-2.4125	0.9574			-2.4225	0.9291		
21	0.0224	1.2177			0.0579	1.1801		
22	-0.2970	0.9977			-0.2799	0.9661		
23	0.1958	0.9927			0.2276	0.9826		
24	0.4567	1.1326			0.5042	1.1136		
25	0.2236	1.0934			0.2778	1.0654		
26	0.3614	1.2301			0.4086	1.2170		
27	1.6490	0.9717			1.6747	0.9615		
28	2.5892	0.8481			2.7146	0.8690		
29	-0.9111	0.9415			-0.9044	0.9138		
30	-0.5481	0.9927			-0.5306	0.9720		
31	0.8999	0.9073			0.9475	0.9147		
32	-0.6154	1.1499			-0.5875	1.1231		
33	0.6524	1.1672			0.6834	1.1350		
34	1.1512	0.9749			1.2023	0.9755		
35	0.2929	1.0924			0.3240	1.0619		
36	1.6792	0.6633			1.8466	0.7558		
37	-1.2751	1.1038			-1.2664	1.0925		
38	-0.1390	0.9510			-0.1452	0.9144		
39	-0.1811	0.7986			-0.1624	0.7839		
40	0.4323	0.9920			0.4900	0.9812		
41	-0.4565	1.1210			-0.4358	1.0701		
42	1.1958	1.2018			1.2416	1.2023		
43	0.0059	1.2337			0.0306	1.2200		
44	-1.2596	0.8639			-1.2497	0.8603		
45	-0.9377	0.9889			-0.9241	0.9600		
46	-1.1764		3.8226		-1.8005		4.7634	
47	-0.7543		1.4946		-1.1174		2.2583	
48	-1.0487		2.3774		-1.7928		3.3772	
49	0.3970		1.0633		0.7203		0.9605	
50	1.5317		0.8295		2.7882		0.7870	
51	-0.3713		1.8412		-1.0185		2.1142	
52	0.3752		1.1121		0.5680		1.0095	
53	0.0022		1.3131		-0.3436		1.2837	
54	0.6769		0.9327		1.1896		0.9194	
55	-1.0208		1.9244		-1.6755		2.5484	
56	0.0440		1.2929		-0.0682		1.4405	
57	-0.0706		1.0076		-0.3524		0.8503	
58	-0.0525		1.0607		-0.1201		1.0800	
59	-0.0104		1.0068		-0.1381		1.1454	
60	-0.0732		0.9609		-0.2461		0.9842	
Dichotomous				1.0971				1.1260
Polytomous				0.0059				0.0072

Table H.3: Hierarchical Loading Structure Matrix

Combination 69/3/3

	Equal A: Both Item types				**Higher A: Polytomous Items**			
Item	Intercept	aS1	aS2	Lambda	Intercept	aS1	aS2	Lambda
1	-0.0619	0.9578			-0.0321	0.9849		
2	0.8353	1.2228			0.8567	1.2033		
3	-3.2781	0.9348			-3.2617	0.9374		
4	-1.2527	0.9757			-1.2294	0.9890		
5	-1.0067	1.0660			-1.0075	1.0527		
6	-1.5195	0.8168			-1.4983	0.8200		
7	0.9956	1.1766			0.9920	1.1576		
8	-0.7066	1.0122			-0.6974	0.9977		
9	1.1765	0.8111			1.2246	0.8319		
10	0.6787	1.0480			0.6982	1.0435		
11	-0.6976	1.1458			-0.6729	1.1712		
12	0.6015	1.2224			0.6187	1.2362		
13	1.0356	1.2533			1.0697	1.2575		
14	0.6002	0.9451			0.5995	0.9391		
15	-0.9627	0.9906			-0.9496	0.9858		
16	0.0829	0.8785			0.0980	0.8965		
17	0.8497	0.9487			0.8687	0.9491		
18	0.3961	0.9771			0.4428	0.9935		
19	-0.0933	0.9680			-0.1031	0.9449		
20	-2.8389	1.0082			-2.8663	1.0351		
21	-0.0453	0.8958			-0.0190	0.8922		
22	-0.3659	1.0628			-0.3563	1.0672		
23	0.1895	0.9478			0.2311	0.9726		
24	0.4711	1.0223			0.4375	0.9838		
25	0.1676	1.0907			0.1678	1.0844		
26	0.1578	1.0136			0.1760	1.0427		
27	1.7908	1.0615			1.8287	1.0754		
28	2.3628	0.6764			2.3824	0.6788		
29	-1.4047	0.9733			-1.3936	0.9582		
30	-0.8600	1.0830			-0.8513	1.0937		
31	1.1192	1.2541			1.1923	1.2806		
32	-0.5020	0.9342			-0.4812	0.9172		
33	0.3334	0.8087			0.3108	0.7986		
34	1.4150	1.2507			1.3795	1.2159		
35	0.5610	1.4737			0.5799	1.5128		
36	2.2285	0.9773			2.2501	0.9758		
37	-1.1122	1.1090			-1.0990	1.1201		
38	-0.4260	0.9937			-0.3855	1.0076		
39	-0.2139	0.7916			-0.1964	0.8043		
40	0.1382	1.0267			0.1518	1.0192		
41	-0.3771	1.0518			-0.3543	1.0579		
42	0.8785	0.9038			0.9421	0.9276		
43	0.0271	0.9244			0.0725	0.9167		
44	-1.1488	0.9511			-1.1443	0.9645		
45	-1.3159	1.0228			-1.3141	1.0220		
46	1.9059	1.3576			1.8929	1.3289		
47	1.1039	0.8556			1.1458	0.8592		
48	1.5764	0.9510			1.5479	0.9340		
49	1.2320	0.9828			1.2352	0.9754		
50	-1.4471	0.9056			-1.4339	0.9240		
51	0.7543	0.9125			0.7748	0.9171		
52	0.1353	1.2266			0.1508	1.2376		
53	-0.9847	1.0573			-0.9667	1.0293		
54	-1.2234	1.1896			-1.1870	1.1845		
55	0.0327	1.2233			0.0403	1.2161		
56	-0.7641	0.9244			-0.7281	0.9298		
57	-0.0343	1.0858			-0.0377	1.0818		
58	-0.0620	0.8806			-0.0680	0.8770		
59	-0.8612	1.0089			-0.8302	1.0210		
60	0.3849	1.3037			0.4217	1.3323		
61	-0.2070	1.2766			-0.1929	1.2509		
62	-1.2567	1.1642			-1.2566	1.1668		
63	1.6892	1.2028			1.7016	1.1893		
64	-0.0175	0.8510			-0.0072	0.8426		
65	-2.9711	0.8569			-2.9280	0.8492		
66	0.5457	0.9453			0.5639	0.9486		
67	-1.0418	0.9900			-1.0246	1.0191		
68	-3.3005	0.8323			-3.3143	0.8476		
69	2.2182	1.1097			2.2891	1.1370		
70	-0.3050	0.9995			-0.3917	1.0614		
71	0.3219		1.1105		0.4242		1.0982	
72	1.8821		1.4797		2.1005		1.5148	
73	-0.1727		1.0644		-0.3008		1.0171	
74	-0.2476		0.9251		-0.1750		0.9382	
75	-0.4213		0.8670		-0.3850		0.8775	
Dichotomous				1.0478				1.0474
Polytomous				0.0102				0.0117