Within-Item Interactions in Bifactor Models for Ordered-Categorical Item Responses

Meghan L. Fager

Submitted to the graduate degree program in the Department of Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

David Hansen, Ph.D., Co-Chairperson

Jonathan Templin, Ph.D., Co-Chairperson

Committee members

Lesa Hoffman, Ph.D.

Vicki Peyton, Ph.D.

Kandace Flemming, Ph.D.

Paul Johnson, Ph.D., Outside Member

Date defended: May 14th, 2019

The Dissertation Committee for Meghan L. Fager certifies
that this is the approved version of the following dissertation :

Within-Item Interactions in Bifactor Models for Ordered-Categorical Item Responses

David Hansen, Ph.D., Co-Chairperson

Date approved: May 14th, 2019

Abstract

Recent research in multidimensional item response theory has introduced within-item interaction effects between latent dimensions in the prediction of item responses. The objective of this study was to extend this research to bifactor models to include an interaction effect between the general and specific latent variables measured by an item. Specifically, this research investigates model building approaches to be used when estimating these effects in empirical data and the potential adverse impact of ignoring interaction effects when present in items modeled with the bifactor model. Two simulation studies were conducted with data generated to follow a bifactor 2-parameter normal ogive model and a bifactor graded response model without interaction effects and with varying numbers of items with interaction effects. Model parameters were then estimated from a bifactor model without interactions, with all possible interactions, and with interactions estimated to match the data-generated interactions. The data-generating model was generally favored in relative model comparisons, indexed by deviance information criteria (DIC). Item and respondent parameters were recovered best when the generating model matched the estimated model across all data-generating conditions. Item interaction parameters had small bias, absolute bias, and root mean squared errors decreased with a larger sample size. Regarding model refinement strategies, the highest density intervals and credible intervals correctly identified noninteracting items as not having an interaction at a higher rate compared to interacting items that were generated to have an interaction. A bifactor model with all, none, and reduced interactions was estimated in two empirical data sets with applications in educational measurement and psychological assessment. Results were evaluated in light of the poor performance of the parameter refinement and model comparison strategies investigated in the simulation studies. Implications of this research and future directions of study are discussed.

Keywords: Bifactor model, interactions, moderation, multidimensional item response theory, latent variable modeling, model misspecification

Acknowledgements

I would first like to thank my undergraduate mentor, Dr. Elizabeth Spievak, who inspired me to pursue this degree and for introducing me to research. I would not be where I am today without your profound belief in my abilities. I have been comforted to know that I can always rely on you for valuable advice and encouragement through this process.

I also want to express my immense gratitude to my graduate mentors, Dr. Jonathan Templin and Dr. Lesa Hoffman, for providing tremendous support throughout my graduate career. Words cannot express how much you both have done for me since we first met. You both have influenced me in so many ways to think outside the box and use my imagination when approaching research. Your passion for the field and dedication to teaching your students is truly an inspiration.

I would also like to thank my co-chair, Dr. David Hansen, for his helpful guidance and applied perspective on this research. I am especially grateful for your feedback and questions that have helped shaped this dissertation into what it is now. My other committee members, Drs. Vicki Peyton and Kandace Fleming, also gave invaluable insight and unique perspectives that have helped me become a better researcher and communicator. You offered suggestions and challenged me to think carefully about connecting this research to practice.

I am extremely grateful to Dr. Paul Johnson, my outside committee member who joined my committee even after I had proposed my dissertation. As my supervisor while I was a graduate research assistant at the Center for Research Methods and Data Analysis (CRMDA), you taught me about tools that I continue to use and love today, including Git, LaTeX, cluster computing, and Linux, that have been invaluable to the completion of this dissertation. My time at the CRMDA has shaped my career and aspirations more than I can express, and I would not have made it to where I am today without your guidance and friendship (and humor!).

Estimation of the models in this project was completed on the University of Kansas Community

Cluster. I would like to thank the folks at the Center for Research Computing for offering technical support to complete this project. Sponsorship for my account was provided by the CRMDA and from Dr. Jonathan Templin.

My success would not be possible without the support and love from my family and friends. Thank you so much to my parents, Mike and Jenn, for encouraging me to do my best. Yang Zhao, you have always been there for me and I am so grateful to have met you. I doubt I would have survived this journey without our ventures downtown and around Lawrence, and our fun conversations about life, love, and surviving grad school. And lastly but most importantly, my husband Travis. You have provided me with unrelenting love and encouragement and have proven over and over that you can be leaned on in the most difficult times. I am so lucky to have you in my life, and I could not have done this without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently, psychometric methods involving multidimensional latent variable models have become more popular in psychological and educational measurement research. The popularity of multidimensional models is likely driven by their focus on obtaining more fine-grained information about latent constructs that is not readily available with unidimensional models. One such multidimensional model that has received increased attention is the bifactor model (Reise, 2012; Gibbons & Hedeker, 1992; Holzinger & Harman, 1941), also called the direct hierarchical model (Gignac, 2008) or the nested-factors model (Chen et al., 2006; Gustafsson & Blake, 1993). Originally introduced in the 1930s (Holzinger & Swineford, 1937), the bifactor model has primarily been used in intelligence research (Gustafsson & Blake, 1993), but also frequently in the measurement of psychological constructs, such as well-being (Longo et al., 2016), quality of life (Chen et al., 2006), and psychiatric distress (Thomas, 2012). These constructs are characterized by several interrelated factors, making the bifactor model ideal for application.

The bifactor model simultaneously models multiple hypothesized specific dimensions (i.e., latent variables) and a single general dimension that is independent of the specific dimensions. It is useful in situations where construct heterogeneity exists, but specific dimensions are related by an underlying dimension. By taking this modeling approach, researchers are afforded with conceptually narrow inference about construct subcomponents alongside a more abstract and general view that is often sought, such as when validating the use of subscales (Reise et al., 2010). The bifactor model can estimate the reliability of subscores and determine the extent to which subscores can provide reliable information after controlling for the unique influence of the general factor (Reise et al., 2013). Furthermore, with the bifactor model, researchers interested in general inferences about a construct can statistically control for multifaceted data that correlate for reasons other than the underlying latent variable. These "nuisance" factors, such as method factors (common item stems) or testlets (DeMars, 2012), introduce systematic covariability in responses that when left unmodeled can bias model parameters and overestimate parameter precision (Sireci et al., 1991; Li et al., 2006). Vice versa, researchers interested in a narrower scope can statistically control for the general dimension and test the unique contribution of construct subcomponents. Other applications of the bifactor model in psychometric research include computerized adaptive testing (Seo & Weiss, 2015; Weiss & Gibbons, 2007), vertical scaling (Li & Lissitz, 2012), and differential item functioning in testlet-based tests (Jeon et al., 2012). Given these uses of the bifactor model, it is not surprising that there has been increasing empirical research using these models (Reise, 2012).

The bifactor model is also useful in cases where multifaceted constructs exist that may relate differently to an outcome (Chen et al., 2006). If the multidimensional structure is ignored but exists in the population, a composite score or unidimensional model may mask the differential contributions of the construct facets, causing incorrect characterizations of the relationships with other variables (Carver, 1989). Interpretations based on the definition of a broader construct may be confounded if only a subset of its unmodeled components is responsible for strong associations with dependent variables. Comparatively, considering only the construct facets by modeling the specific dimensions and ignoring the general dimension neglects to parcel out the variance the facets share. In this case, the true pattern of associations between the construct facets and an external criterion may be masked. To remedy this, the bifactor model targets the isolated effects of the general dimension and the specific dimensions to drive theory refinement and improve the predictive validity of the construct as a whole.

1.1 Extending the Bifactor Model

At present, bifactor models are applied with items having at most one parameter per dimension, often called a discrimination parameter in item response theory or a factor loading in factor analysis literature. For the purpose of this study, this parameter is called a main effect once an interaction is included. Including only main effects assumes that the effect of the specific dimension on the predicted item response does not depend on the general dimension (or vice versa). This research extends the bifactor model to include interaction effects in the prediction of item responses between the general and specific dimensions. This allows predicted relationships between the item response and the general (or specific) dimension to *change* as a function of the specific (or general) dimension, a process that is often called moderation in applied linear models. Thus, researchers can test the effect that high scores on one dimension have on the observed item response, conditional on the scores on the other dimension the item also measures.

As an example, consider a test of math ability with a variety of item types. Often, math ability is assessed with word problems where students use their reading comprehension to solve a problem (Vilenius-Tuohimaa et al., 2008). The response options may include an equation to reach the solution or the solution itself, and the item is scored correct or incorrect. However, if a student struggles with word problems that require strong reading comprehension skills, they may do poorly on the item, regardless of their overall math ability. In this context, a bifactor model is specified with a reading comprehension dimension for the subset of items that involve word problems and a general math ability dimension. With an interaction, the item's relationship with one measured dimension would depend on the values of the other. Thus, for example, the rate of increase in the probability of a correct response with more math ability can become augmented with more reading comprehension. Comparatively, having a low reading ability may also diminish the rate of increase in probability with increasing math ability. Theoretically, patterns of interactions, coupled with patterns for the conditional main effects, may reflect testing situations where respondents need sufficient lower-level skills before they are able to complete more difficult items or exams.

1.2 Goals of this Research

The goal of adding an interaction parameter is to ensure accurate inferences and parameter estimates that are unharmed by omitted variable biases in underspecified models that ignore interaction effects. Similar to cases where interaction effects are present but unmodeled in a univariate regression or analysis-of-variance model, the inferences made about the relationship between the observed and latent variables will be erroneous and there may be significant bias in other parameter estimates. Thus, this research sought to extend the bifactor model to include interactions in order to ensure the soundness of inferences made in subsequent analyses, such as in tests of differential item functioning and the effects of explanatory variables.

Using generated and empirical data, this dissertation addresses the following three research questions:

1. How well does the proposed bifactor model with interaction effects recover item and person parameters in realistic data generating conditions?

2. What is the impact of ignoring interaction effects on model recovery of item and person parameters when fitting a bifactor model with only main effects?

3. How well do model selection and parameter refinement strategies perform in identifying underlying interactions in a bifactor model?

The first research question asks whether the proposed model can be estimated accurately and efficiently in tests with varying sample sizes and proportions of items in an assessment with interactions. The second question expands on the first to investigate the impact on parameter estimates when a misspecified model is fit to the data. The objective is to determine whether extending the bifactor model to include interactions is even necessary and whether the true data-generating model can be accurately recovered. The final question inquires whether confirmatory model building procedures that rely on alternate model comparisons and parameter refinement strategies can accurately identify interactions for use in empirical data.

1.3 Chapter Overview

This dissertation is organized as follows. The first chapter introduces the multidimensional item response theory models surveyed in this manuscript, with sections that 1) describe previous research of latent variable models with and without interactions, and 2) present the traditional bifactor model and the proposed extension to have interactions for items having dichotomous and polytomous item response options. In Chapter 3, the design of two simulation studies is described, which is followed with results in Chapter 4. Informed by the results of the simulation studies, Chapter 5 applies the knowledge learned to demonstrate two empirical analyses with applications in education and psychology. These illustrations are presented with the goal of concretizing the relevance of the proposed model in real-world constructs and to show how the model can be specified and refined with model building techniques. The final chapter concludes this dissertation with a discussion of research findings and relevance to educational and psychological research.

Chapter 2

Modeling Framework

Latent variable models are used to infer, from manifest variables that can be directly observed, underlying hypothetical constructs that are hidden. For example, intelligence is one multifaceted construct that can be inferred from observable responses to tasks or answers to questions of memory, verbal ability, and reasoning (Gustafsson & Blake, 1993). The latent variable, intelligence, is a random variable with an assumed distribution that is formed by combining multiple observed measurements about individual units (e.g., people). The goal is to determine the common cause or causes (i.e., latent constructs) that account for the variation and covariation among a set of manifest variables in terms of a smaller set of hypothetical variables. After taking into account the influence of the latent variable(s) on the responses to manifest variables, the observed responses are thought to be independent, a condition called conditional independence (Lord & Novick, 1968). For the remainder of this manuscript, observable, manifest variables will be referred to as *items* measuring the construct and units of analysis as *respondents* to those items.

In this chapter, the theoretical framework for bifactor models with interaction effects is discussed based on a latent variable framework called item response theory (IRT; e.g., Hambleton et al., 1991; Embretson & Reise, 2000). The collection of item response models presented are intended for ordered categorical item responses and continuous latent variables. The subsequent sections offer a review of IRT model extensions to include multiple dimensions, bifactor structures, and interaction effects.

2.1 Multidimensional Item Response Theory

Item response theory is a family of latent variable models for categorical item types that are related by a continuous latent variable, often symbolized by θ and termed *latent trait*. IRT models are designed to rank order individuals on one or more latent variables that are assumed to be normally distributed and traditionally unidimensional and broad in scope (Lord, 1980; Lord & Novick, 1968). Multidimensional IRT (MIRT; Reckase, 2009) is a field that generalizes unidimensional IRT models to include more than one latent trait, such as the bifactor model.

There are a variety of IRT and MIRT models having different assumptions for data involving a varying number of categories. Specifically, there are models designed for dichotomous and polytomously ordered item responses (Samejima, 1969), unordered or nominal item responses (Bock, 1972), and partially ordered item responses (Muraki, 1992; Masters, 1982). For the purpose of this dissertation, the focus was on models for ordered dichotomous and polytomous item responses that do not model guessing (Birnbaum, 1968) and do not allow partial orderings. These models are the multidimensional 2-parameter normal ogive model (M2PNO; Samejima, 1969; Lord & Novick, 1968, pp. 365-384) and the multidimensional graded response model (MGRM; Samejima, 1969), as described next.

2.1.1 Multidimensional Two-Parameter IRT Models

The multidimensional 2-parameter normal ogive model is a model for dichotomous item responses, such as those scored correct or incorrect, that relates the probability of an endorsed response to a given latent variable through the use of a Bernoulli distribution. The M2PNO uses a probit link function to transform the conditional mean to be predicted onto the probability metric, which is bounded between 0 and 1. To illustrate the M2PNO, the following equation yields the probability of a correct or endorsed response to a binary item, i, taken by a respondent, r:

$$P(y_{ri} = 1|\theta) = \Phi(a_i(\theta_r - b_i)) = \int_{-\infty}^{(a_i(\theta_r - b_i))} \phi(z)dz. \tag{2.1}$$

The difficulty parameter of the item, b_i, is on the same metric as the latent variable, θ_r, and represents the location where $P(y_{ri} = 1)$ = 0.5. Put another way, when $\theta_r = b_i$, the respondent is equally likely to answer the item with a 0 or 1 response. The difficulty parameter reflects how much of the latent variable is needed to have a 50% chance of a correct or endorsed response. The item discrimination, a_i, represents how well the item can differentiate respondents with low and high latent dimension levels. An assumption of IRT and MIRT models is that the probabilities of endorsed responses are monotonically increasing, such that the sign of a_i that multiples the latent variable governs the direction of the monotonic function (e.g., Hambleton et al., 1991).

Because latent variables are unobserved, the scale must be assigned depending on the assumed distribution of the latent variable. In the IRT framework, the latent variables are normally distributed with a mean and variance, $\theta_r \sim N(\mu, \sigma^2)$. For model identification, either the mean of the latent variable can be fixed to 0 to allow estimation of the item difficulties, or the latent variable mean can be estimated by fixing the mean across item difficulties to zero or one item difficulty per latent variable to 0. Similarly, the variance of the latent variable can be fixed to 1 to estimate all item discriminations, or the discrimination parameter of a "marker" item can be fixed to 1 to allow estimation of the latent variance. IRT models typically have standardized latent variables where the latent variable means are fixed to 0 and variances are fixed to 1. The majority of latent variable estimates will then fall between -4 and 4, analogous to z-scores or standard-deviation units (though the actual range is from $-\infty$ to ∞). Correlations between latent variables can be estimated if desired.

When items measure one latent variable from a test with multiple dimensions, they have a factorially simple loading structure and are referred to as having between-item multidimensionality (Adams et al., 1997). Comparatively, if items have a factorially complex loading structure because they measure a combination of latent variables, they are said to have within-item multidimensionality (Adams et al., 1997). In these cases, computation becomes easier when the response function is written in slope-intercept form. The item discrimination parameter is multiplied through to result in $(a_i\theta_i - a_ib_i)$. The term $-a_ib_i$ is replaced with β_{i0}, an intercept for the item. The item discriminations become slopes or main effects, $a_i = \beta_i$. Main effects are estimated per dimension for complex items and are indexed with subscripts, β_{id}. With standard normally distributed latent variables, the item intercept becomes the probability of an endorsed response at the mean of all measured latent variables. This transformation of item difficulties into intercepts and item discriminations into main effects makes the response function more easily translatable to parameterizations of regressions and factor analytic models for ordered categorical data (Cai et al., 2011). For more detail on the relations between IRT and factor analysis models for categorical items, see Takane and Leeuw (1987).

Traditionally in MIRT literature, the M2PNO is specified as a variation of the multidimensional 2-parameter logistic model (M2PL; Birnbaum, 1968). Item response theory was originally developed based on the normal ogive model building from probit regression (Samejima, 1969; Lord & Novick, 1968). The primary difference in these models is that the M2PL uses a logit link rather than a probit link. The logistic version of the M2PNO written in slope-intercept form for an item response y_{ri} is specified as follows:

$$P(y_{ri} = 1|\theta) = \Psi[D(\beta_{i0} + \sum_d \beta_{id}\theta_{rd})] = \frac{exp(D(\beta_{i0} + \sum_d \beta_{id}\theta_{rd}))}{1 + exp(D(\beta_{i0} + \sum_d \beta_{id}\theta_{rd}))} \tag{2.2}$$

where Ψ is the cumulative logistic function. The M2PL and M2PNO are nearly indistinguishable when a constant scaling factor D = 1.702 is included in the exponent of the logistic response function to scale results from the logistic model onto the normal ogive model (Haley, 1952; Camilli, 1994). Given the close resemblance of the two models and their use in item response theory literature, both models will be referenced interchangeably throughout this manuscript.

2.1.2 Multidimensional Graded Response Model

A generalization of the M2PL and M2PNO is the multidimensional graded response model (MGRM; Samejima, 1969) for items with ordered polytomous response categories, such as Likert-type items (e.g., categories ranging from strongly disagree, disagree, neutral, agree, to strongly agree; Likert, 1932). The MGRM can be thought of as an extension of the M2PNO and M2PL that generalizes to more than two categories and uses a cumulative probit link function instead of a probit link function (or a cumulative logit link function if using a logistic distribution). The MGRM assumes that responses to an item require a number of ordered steps, such that previous steps must be accomplished before reaching the next steps. The categories k of an item begin at 1 up to the number of response options, M $(k = 1, 2, 3, ..., M)$. The probability of responding to a specific category is conceptualized as the probability of responding in or above the category, minus the probability of responding in or above the next category. Specifically, the probability of responding to category k is monotonically increasing with the measured latent variable, θ_r, and is conceptualized as the difference between two cumulative probabilities:

$$P(y_{ri} = k|\theta) = P(y_{ri} \geq k|\theta) - P(y_{ri} \geq k+1|\theta) \tag{2.3}$$

For an example item with three categories (coded 1, 2, 3), the probabilities of answering among the different categories are given as "sub-items":

$$P(y_{ri} = 1|\theta) = 1 - P(y_{ri} \geq 2|\theta) \tag{2.4}$$

$$P(y_{ri} = 2|\theta) = P(y_{ri} \geq 2|\theta) - P(y_{ri} \geq 3|\theta) \tag{2.5}$$

$$P(y_{ri} = 3|\theta) = P(y_{ri} \geq 3|\theta) - 0 \tag{2.6}$$

The first option, $P(y_{ri} = 1|\theta)$, is the probability of responding to categories 1-3, versus the probability of responding to categories 2 through 3. The probability of responding to category 2 is the difference in the probability of responding to categories 2 through 3 versus category 3. The probability of responding to the highest category is the probability of responding to category 3 minus 0 because it is the highest category. The MGRM models each of these directly as "sub-items," as defined next.

The MGRM directly models the probability of responding in category k or greater. With the normal ogive link, the following equation yields the MGRM in slope-intercept form:

$$P(y_{ri} = k|\theta) = \Phi(\beta_{i(k)} + \sum_d \beta_{id}\theta_{rd}) - \Phi(\beta_{i(k+1)} + \sum_d \beta_{id}\theta_{rd}). \tag{2.7}$$

The item main effect, β_{id}, for a latent variable θ_{rd}, has a similar interpretation as the M2PNO, but there are now $M-1$ ordered category-intercept parameters, $\beta_{i(k)}$. The higher categories require higher levels of the latent variable to have a higher endorsed response probability. Because there is one main effect parameter per latent variable, the relationship between the sub-items are parallel in the probit function. As with dichotomous items, the logistic version of the MGRM is specified by substituting the cumulative normal density function, Φ, with the cumulative logistic density function, Ψ. The joint distribution of the latent variables is multivariate normal and standardized for model identification. As before, inter-dimension correlations may be estimated.

2.2 Compensatory and Partially Compensatory MIRT

Compared to the multidimensional models defined previously which are compensatory in nature, such that an increase in any latent variable measured by the item will increase the probability of an endorsed response, some complex items are best modeled with “partially compensatory” or “noncompensatory” models. These models suppose that high levels of one latent variable cannot compensate for low levels of another dimension that is also measured by the item (Embretson & Reise, 2000; Ackerman, 1989; Whitley, 1980). This reasoning is analogous to a conjunctive relationship between latent variables that generates the latent response rather than a disjunctive relationship because both abilities are necessary for a correct response (Maris, 1999).

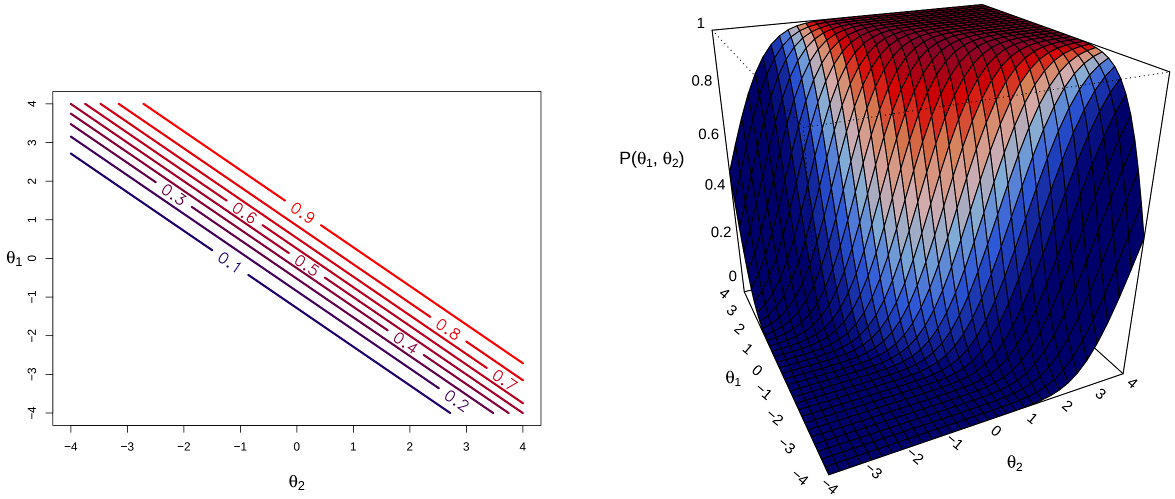
Partially compensatory models were originally introduced to model item responses as a product of probabilities of the individual response curves from separate unidimensional models (Sympson, 1977; Whitley, 1980). Conceivably, this modeling strategy is congruous with the interpretation that the probability of an item response should be limited by the lowest ability and therefore cannot exceed the lowest predicted probability of the unidimensional models (Chalmers & Flora, 2014). Adapted from Sympson's (1977) model, a partially compensatory M2PL product model (PC-M2PL) for dichotomous ordered-categorical item responses with a logit link function and notation consistent with Equation 2.2, is given as:

$$P(y_{ri} = 1|\theta) = \prod_{d=1}^{D} \frac{exp(\beta_{0id} + \beta_{id}\theta_{rd})}{1 + exp(\beta_{0id} + \beta_{id}\theta_{rd})} \tag{2.8}$$

where d is the dimension corresponding to the d element of the θ vector, ($\theta = [\theta_1, \theta_2, ..., \theta_D]$). The item now has an intercept parameter, β_{0id}, as in Equation 2.2, estimated for each dimension d of item i. Thus, for an item measuring two dimensions, there will be two item intercept parameters instead of one (for ordered polytomous items, there would be $D * (k-1)$ category-intercepts) and two main effects (as in a compensatory model). As before, the PC-M2PL can be adapted to the normal ogive model with a probit link function in place of the logit link function and is abbreviated as PC-M2PNO.

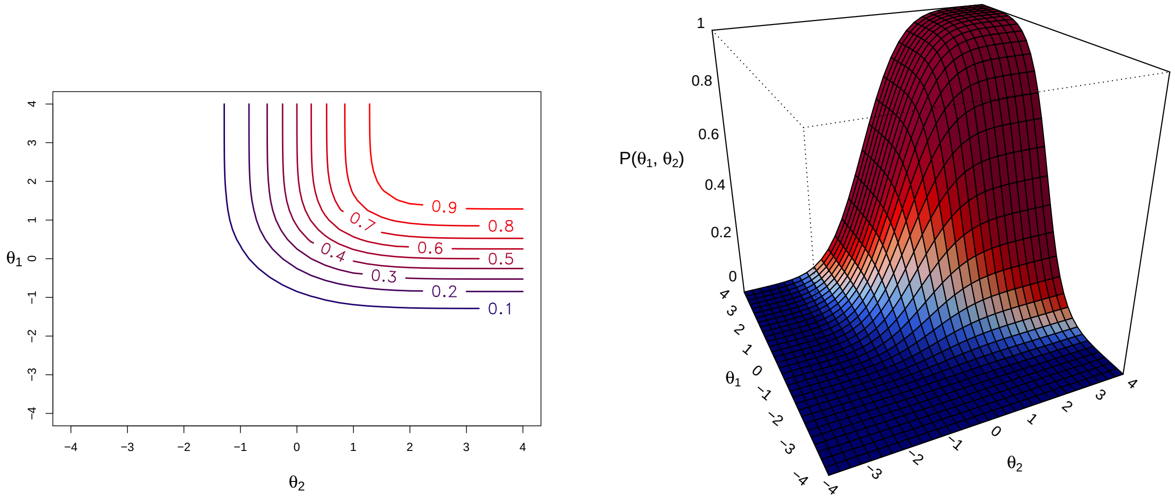
As an illustration, consider the contour plots and surface plots in Figures 2.1 and 2.2 that show the response surfaces of a dichotomously scored item that measures two latent dimensions equally ($\beta_{i1} = \beta_{i2} = 1$) when modeled using the compensatory M2PNO and the partially compensatory product PC-M2PNO. Each contour corresponds to the probability of correctly responding to the item for a respondent with a given level of θ_1 and θ_2. For compensatory models, the contours are equally spaced and parallel. As the slopes become steeper, the contours become closer. The rate of change of the probabilities corresponds to the size of the main effects and is constant for a given point on the coordinate axes in the multidimensional θ-space. Comparatively, the PC-M2PNO contours have curvilinear relationships between the item and the latent variables, which indicates that the rate of change of the probabilities can vary depending on the location of the latent variables.

Figure 2.1: Contour plot and surface plot of the probability of a one response for an item modeled using a compensatory M2PNO



Item Parameters: Intercept = 0, Main Effects = 1

Figure 2.2: Contour plot and surface plot of the probability of a one response for an item modeled using a partially compensatory M2PNO



Item Parameters: Intercepts = 0, Main Effects = 1

Variations of the PC-M2PL involving different combinations of estimated or fixed intercepts, slopes, or asymptotes (i.e., “guessing” parameters) also exist, such as the independent components model (Whitley, 1980) when there are no discrimination (i.e., slope) parameters, and the multicomponent latent trait model (MLTM; Embretson, 1984) when there are no discrimination parameters and an estimated upper asymptote. A variant of the MLTM is the conjunctive Rasch model (Maris, 1995) that does not use a two-step procedure to model complex items, though this model has greater difficulty in implementation. Additionally, the generalized MIRT model (GMIRT; Simpson, 2005) is a hybrid of the compensatory M2PL and the PC-M2PL that allows the level of compensation to vary by using an item-level variable to proportionally weight the two models.

Partially compensatory models for dichotomous item responses specified as products of probabilities impose a strict set of assumptions about how the latent variables combine to produce an item response. Further, these models have shown greater estimation difficulty and require more data for accurate parameter estimates compared to compensatory models. Bolt & Lall (2003) used an exploratory approach (i.e., all items measure all dimensions) to compare the MLTM and the M2PL model with two dimensions using Markov Chain Monte Carlo methods (see Patz & Junker, 1999b; Albert & Chib, 1993, for applications in item response theory). They found that the MLTM had much larger standard errors in an empirical dataset compared to the compensatory M2PL. In their simulation study, Bolt & Lall (2003) found that compared to the M2PL, the MLTM intercepts had larger root mean squared errors (RMSEs) and required more examinees ($N = 3,000$) and items ($I \leq 50$) with low to moderately correlated latent dimensions ($\rho \leq 0.30$) to accurately recover them. With highly correlated dimensions ($\rho \geq 0.60$), recovery was inadequate. To resolve the metric indeterminacy problem of exploratory models, the MLTM was identified by fixing the difficulty parameters of the first items per dimension to zero to estimate parameters, which were then equated onto the correct metric using the data-generating parameters (Bolt & Lall, 2003). Because true parameters are unknown in real data, this exploratory approach is impossible in empirical research.

With the addition of slope parameters for each dimension, Babcock (2011) used a confirmatory approach where items measured a subset of dimensions to investigate the parameter recovery of the PC-M2PL compared to the M2PL. He found that the PC-M2PL required a minimum of $N = 4,000$ respondents for accurate estimation and six unidimensional items per dimension to position the axes of the latent variables for model identification. Similar to Bolt & Lall (2003), the PC-M2PL functioned poorly when the correlation between dimensions was high. Another study by DeMars (2016) found that with highly correlated dimensions, the M2PL fit data generated with the PC-M2PL better than the PC-M2PL. Chalmers & Flora (2014) found decreased root mean-squared deviation (RMSD) of item parameter estimates of the PC-M2PL with more simple structure items (i.e., measuring one dimension) and larger sample size, but increasing the inter-dimension correlations increased it. These authors note that the poor estimation of data with correlated dimensions is because there are too few examinees in the response regions where one dimension is high and the other is low. Thus, it appears that a large amount of both items and respondents and dissimilar dimensions are necessary for accurate estimation of the PC-M2PL.

2.2.1 MIRT Interaction Effects

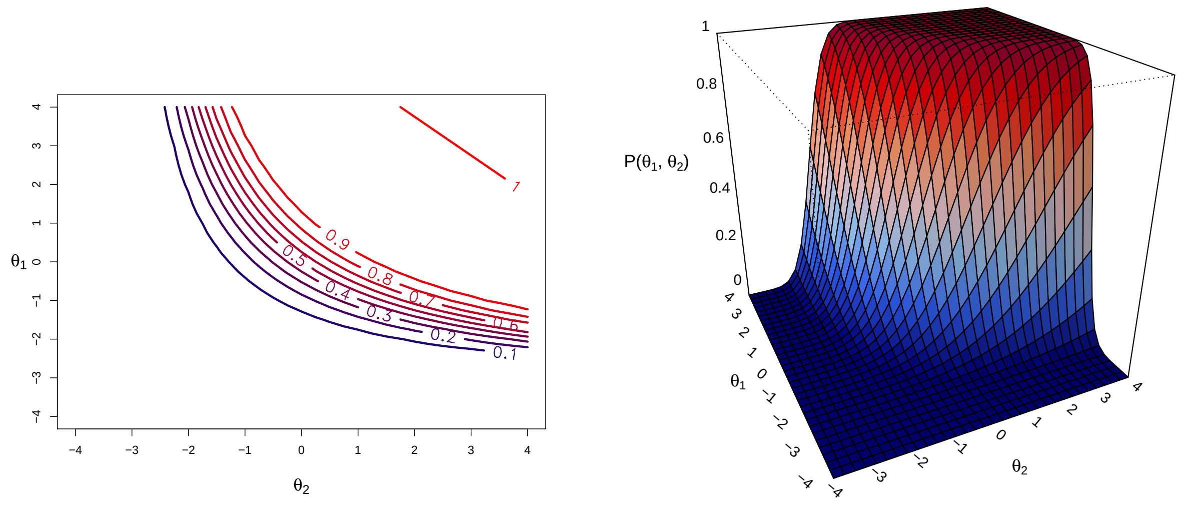
As an alternative, researchers have introduced interaction effects in multidimensional IRT models to model items having a partially compensatory relationship (DeMars, 2016; Chalmers & Flora, 2014; Rizopoulos & Moustaki, 2008). Compared to the PC-M2PL and the MLTM, the interactive model estimates an interaction effect to model partially compensatory relationships or determine whether the model can be simplified to a compensatory model with only main effects. The relationship between latent dimensions in the prediction of an item response is accounted for because the interaction term is inside of the item response function, whereas multiplying the individual item response probabilities together assumes that they are independent. Specifically, the M2PNO interaction model (I-M2PNO) with a probit link function for a binary item response, y_{ri}, for a respondent, r, to an item, i, that measures two dimensions indexed as 1 and 2, is given as:

$$P(y_{ri} = 1|\theta) = \Phi(\beta_{0i} + \beta_{i1}\theta_{r1} + \beta_{i2}\theta_{r2} + \beta_{i(1*2)}\theta_{r1}\theta_{r2}). \tag{2.9}$$

The probability of a correct response from respondent r to the item i, given the θ vector of two latent variables, $\theta_{\cdot 1}$ and $\theta_{\cdot 2}$, is a function of two conditional slopes or "simple main effects," β_{i1} and β_{i2}, for each of the measured latent dimensions, and their interaction effect, $\beta_{i(1*2)}$. The intercept,

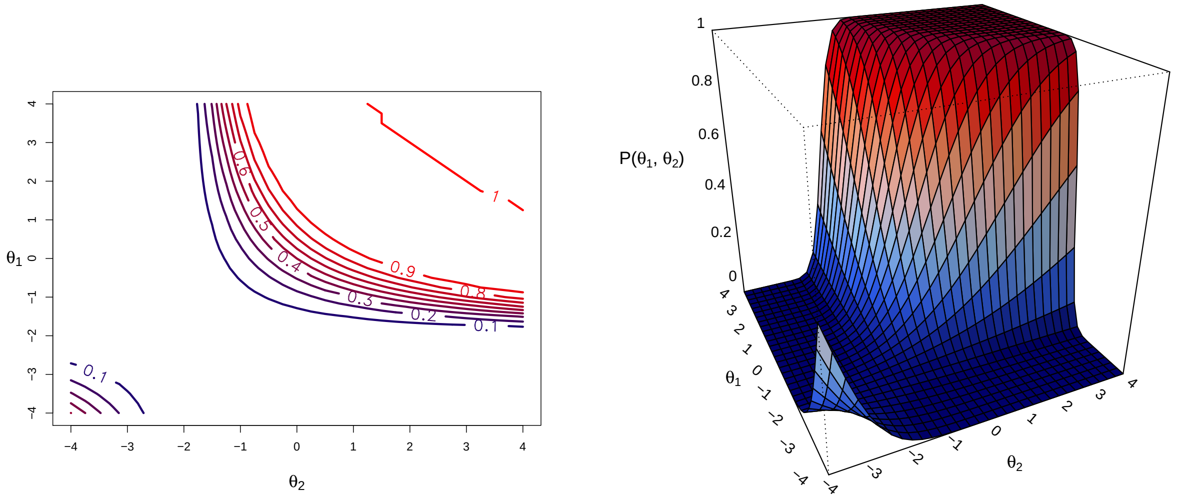
β_{0i}, is interpreted similarly to the M2PNO as the expected probability of a correct response for a respondent with a mean value of the two standardized multivariate normally distributed latent variables, $\theta_D \sim MVN(0,\Sigma)$. The first subscript of the interaction term, i, represents the item with which the parameter is associated, and the second subscript, $(1*2)$, represents which interacting dimensions are included in the effect.

Figure 2.3: Contour plot and surface plot of the probability of a one response for an item modeled using an M2PNO with an interaction of $\beta_{i(1*2)} = 0.3$



Item Parameters: Intercept = 0, Main Effects = 1, Interaction = 0.3

Figure 2.4: Contour plot and surface plot of the probability of a one response for an item modeled using an M2PNO with an interaction of $\beta_{i(1*2)} = 0.5$



Item Parameters: Intercept = 0, Main Effects = 1, Interaction = 0.5

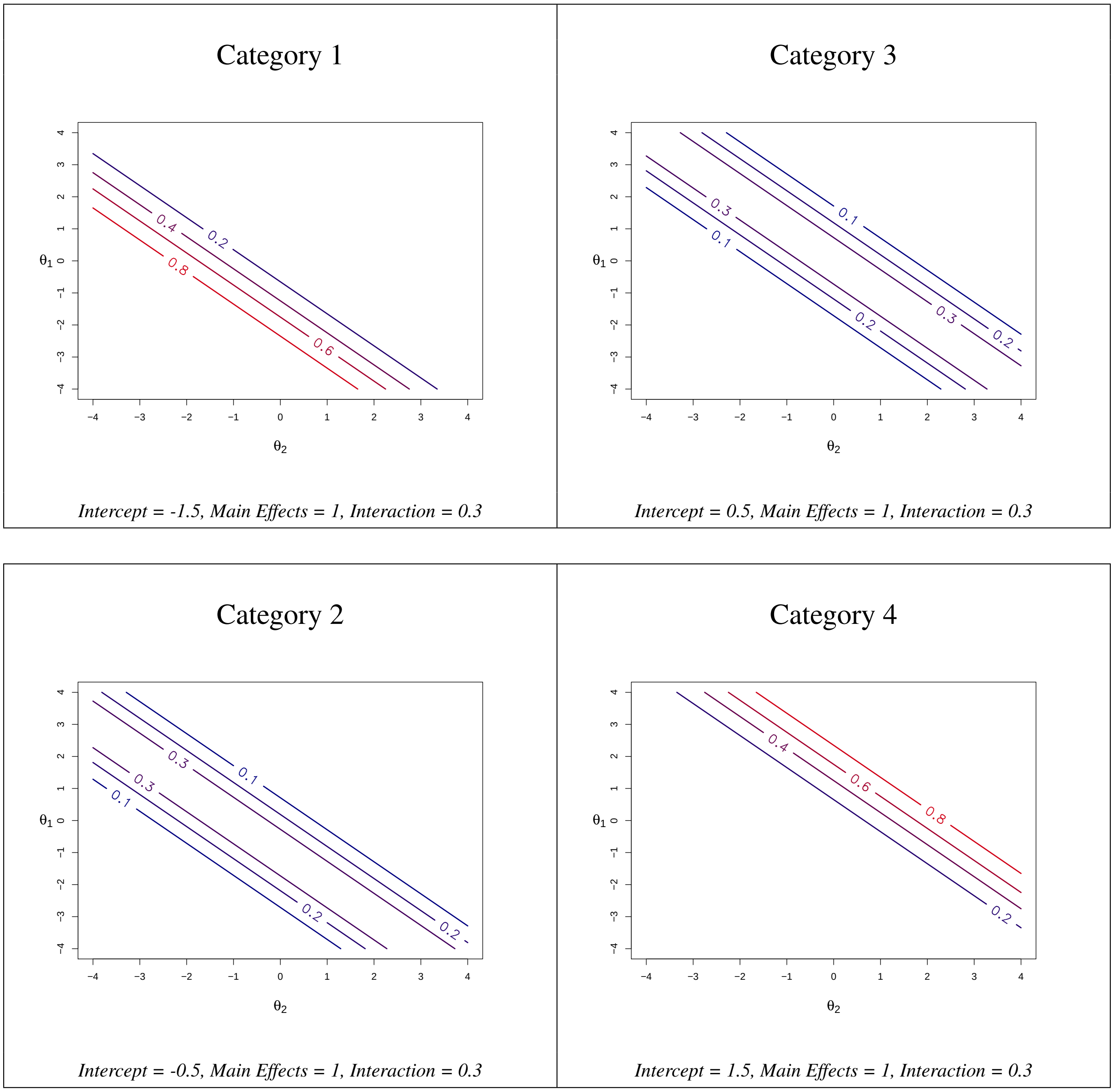
Figures 2.3 and 2.4 show contour and surface plots of a binary item that measures two dimensions equally with positive interaction effects of $\beta_{i(1*2)} = 0.3$ and $\beta_{i(1*2)} = 0.5$, respectively. With positive main effects ($\beta_{i1} = \beta_{i2} = 1$), a strong positive interaction of the I-M2PL would indicate that the item is partially compensatory and requires high values on both latent variables to endorse the item. As either dimension increases, the probability of an endorsed response also increases, but increases in both dimensions simultaneously can have an additional boost. Comparatively, if the interaction coefficient is small and approaching zero, high values of either latent variable can increase the probability of an endorsed response on the item, but increases together do not augment the rate of increase. This would indicate that the item response process could be simplified to a compensatory model where the contours are linear and parallel. Negative interaction coefficients are also possible. Positive main effects can become less positive with negative interactions, corresponding to diminishing gains in the probability of responding to higher categories with each unit increase of the interacting latent variable.

As shown in Figure 2.4, a large interaction may cause response probabilities to start increasing as θ values decrease after a certain point (Chalmers & Flora, 2014), which may not correspond to cognitive tests that assume that high ability levels relate to increasing probabilities of correct responses to test items. It may be that this shift in direction from decreasing probabilites to increasing probabilities occurs in the response regions where θ values are less frequently observed. However, this change in direction is conditional on the size and direction of the main effects (Buchholz, 2014, as cited in DeMars, 2016); with larger main effects of $\beta_{i1} = \beta_{i2} = 2$ and an interaction of $\beta_{i(1*2)} = 0.5$, the inflection point (θ_1, θ_2) occurs at $(-\beta_{i1}/\beta_{i(1*2)}, -\beta_{i2}/\beta_{i(1*2)}) = (-4, -4)$. Comparatively, smaller main effects of $\beta_{i1} = \beta_{i2} = 1$ and an interaction of $\beta_{i(1*2)} = 0.5$, the direction of the function changes at an inflection point of $(-2, -2)$, as observed in the contour plot and surface plot in Figure 2.4. Researchers have suggested placing boundaries on the main effects (Babcock, 2009), though this strategy has not been investigated.

Though not studied at the time of this research, the I-M2PL can be generalized to multiple-category item response models like the MGRM (labeled as I-MGRM with interactions). The difference between the I-M2PL and the I-MGRM is that the main effects and interaction effect model the probability of responding to higher categories of the item, and the model intercept corresponds to multiple ordered category-intercepts as previously described for the MGRM. Each “sub-item” that gives the probability of responding to a particular category, based on differences between two cumulative probabilities as in Equation 2.3, uses the same interaction effect and two main effects that are estimated, but with different category-intercepts. As an illustration, Figure 2.5 and 2.6 display the contour surfaces for an item having four response categories with category-intercepts of $\{-1.5, -0.5, 0.5, 1.5\}$. The contours of each sub-item are linear for the compensatory MGRM as in Figure 2.1 and nonlinear for the interactive MGRM as in Figure 2.3.

Past studies have yielded some important insights into the performance of the I-M2PL in simulated and empirical data. Rizopoulos & Moustaki (2008) fit the I-M2PL using an exploratory approach to simulated data with two dimensions and to example data from a section of a survey measuring workplace relations and employment practices. In their simulation, they manipulated the sample sizes of respondents (N = 500, 1,000) and the number of items (I = 10, 30). For each condition, 1,000 datasets were generated. Results showed a small bias of item parameter esti-

Figure 2.5: Contour plots of the item response surface of a MGRM item with 4 ordered categories



Category 1

Intercept = -1.5, Main Effects = 1, Interaction = 0.3

Category 3

Intercept = 0.5, Main Effects = 1, Interaction = 0.3

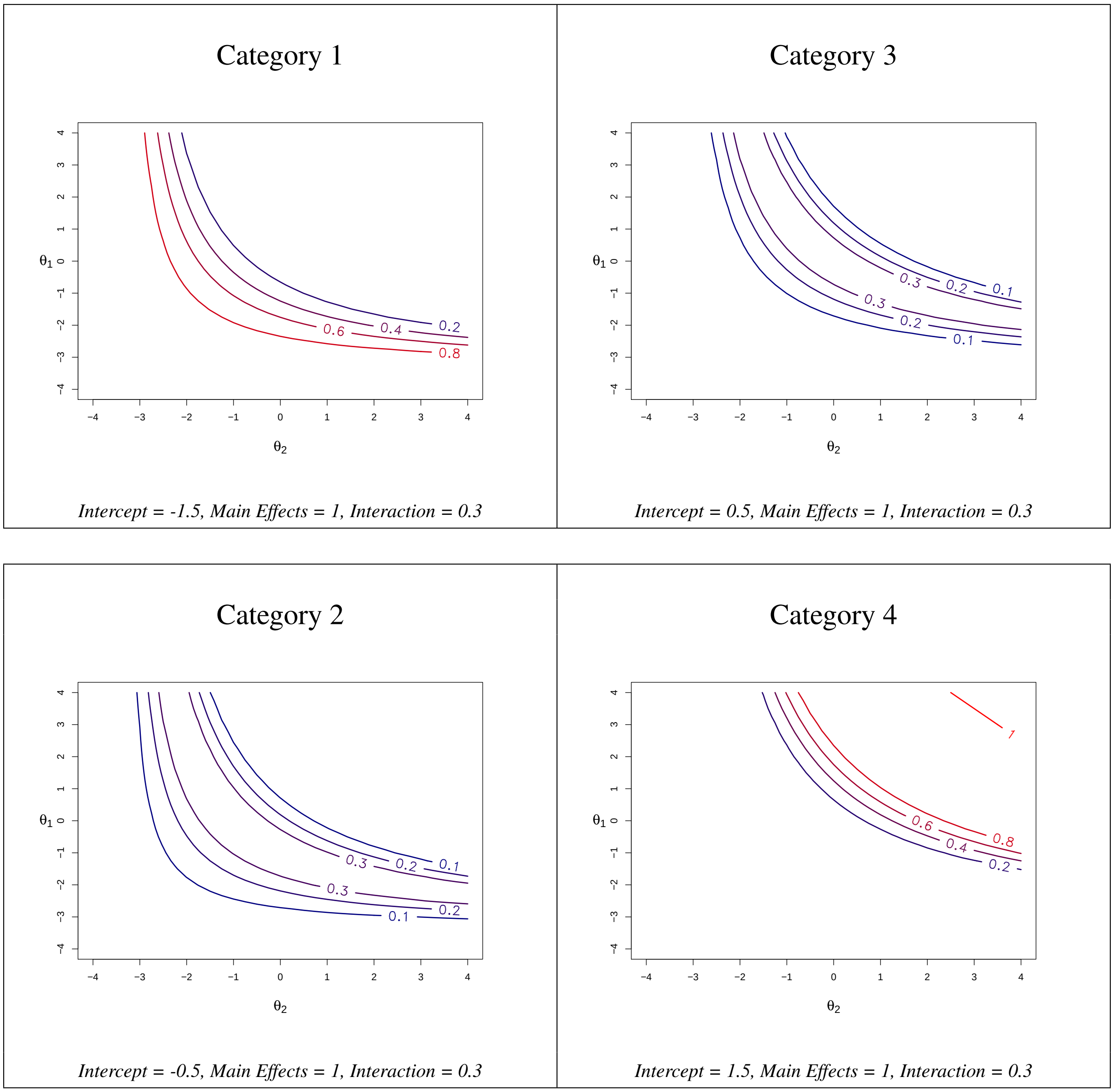
Category 2

Intercept = -0.5, Main Effects = 1, Interaction = 0.3

Category 4

Intercept = 1.5, Main Effects = 1, Interaction = 0.3

Figure 2.6: Contour plots of the item response surface of a MGRM interaction model item with 4 ordered categories



Intercept = -1.5, Main Effects = 1, Interaction = 0.3

Intercept = 0.5, Main Effects = 1, Interaction = 0.3

Intercept = -0.5, Main Effects = 1, Interaction = 0.3

Intercept = 1.5, Main Effects = 1, Interaction = 0.3 mates, which decreased with larger sample sizes and test lengths. Of note, the main effects in this study were simulated to be positive and negative, which may not reflect testing scenarios where endorsed item responses are monotonically increasing with the latent variable. Using empirical data, Rizopoulos & Moustaki (2008) also tested an interaction term between two latent dimensions at the structural model level (Cudeck & Harring, 2009; McDonald, 1962), given the short test length of 6 items. The interaction term was significant according to its Wald-based p-value and the model was favored in model comparisons. Though this interaction effect does not directly resemble the item-level interaction effects discussed in this manuscript, the example offers a conceptually interpretable application of interactions in latent variable models.

Another study by Chalmers & Flora (2014) used a confirmatory approach to compare the PC-M2PL and the I-M2PL. They fit the I-M2PL to data simulated from the product model and found faster convergence, although the log-likelihood was higher and G^2 goodness-of-fit statistic was lower for the product model, indicating its better fit to the data. The recovery of latent dimensions was greater for the product model, however this study did not compare the recovery of the I-M2PL when data are generated from the I-M2PL.

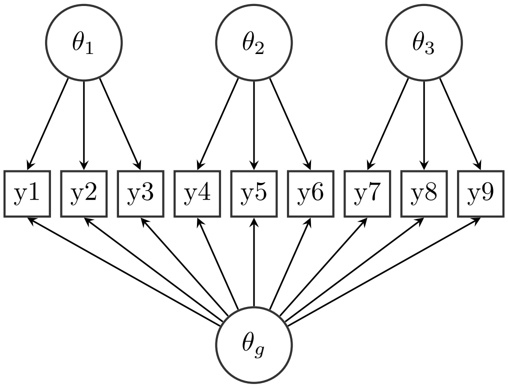
Demars (2016) is the only published paper at the time of this research that generated simulated data from the I-M2PL to investigate parameter recovery of item interaction effects in MIRT models. In her study, Demars (2016) compared the I-M2PL to the compensatory M2PL and PC-M2PL with and without lower asymptote parameters (an extension of the 2PL variants to include $[c_i + (1 - c_i)]$ in the item response function). Manipulated data-generating conditions were two numbers of dimensions (2- and 4-dimensions), three inter-dimension correlations ($\rho = 0, 0.7, 0.9$), and two generating models (interaction and product model). The number of items per dimension was fixed, with complex items measuring up to all possible dimensions. To address the issue of local maxima in the multidimensional θ-space, Demars (2016) generated interaction coefficients at 0.3 for all interacting items, with half of the item discriminations equal to 0.9 for both dimensions and the other half varying but satisfying the condition, $\sqrt{\beta_{i1}^2 + \beta_{i2}^2} = 1.3$. These values were chosen to restrict predicted probabilities from increasing as θ values decrease to regions where there are few respondents, an effect of large interactions as previously described (Chalmers & Flora, 2014). This study generated 100 replications of the data generating conditions and fit each dataset with a compensatory model, product model, and interaction model. Results showed that model fit indices consistently favored the correct model (generated equals estimated) with no correlations between dimensions, though the M2PL was selected more frequently for data generated with the PC-M2PL model with increasing inter-dimension correlations. Recovery of item response functions, according to their root mean squared errors (RMSE), was the lowest for the correct model across all data generating conditions. The inter-dimension correlations were adequately recovered, though using the product model on data generated as the interaction model overestimated the correlation(s) in all conditions, whereas using the interaction model on product model data underestimated them in all but the 4-dimension, $\rho = 0.9$ condition. Parameter recovery of thetas was similar for the different models, suggesting that a compensatory model may be sufficient if one is primarily interested in θ estimates. However, bias and RMSE for item parameters of the product model and interaction model increased with more correlated dimensions. Specifically, with two uncorrelated dimensions, the discrimination parameters of interacting items with varying values had small bias; however, when $\rho = 0.9$, the larger discrimination parameter was negatively biased and the smaller one was positively biased, indicating that higher correlations between dimensions tend to estimate the discrimination parameters with similar magnitude. Similarly, for the product model, the intercepts and discrimination parameters of items with different parameters per dimension had larger RMSEs and bias when the correlations between dimensions were high.

In summary, each of the aforementioned studies of partially compensatory models has claimed that 1) the interaction MIRT models are favored compared to the alternative product models, and 2) increasing inter-dimension correlations contribute to poor recovery of item parameters. With highly correlated dimensions, it may be that a unidimensional model is preferred because the dimensions are homogeneous (Drasgow & Parsons, 1983). Alternatively, a bifactor model that parcels out common variance among items with a general dimension and simultaneously models the unique influence of specific dimensions can offer an interpretive benefit. As such, this dissertation investigates interaction effects in bifactor models that do not have correlations estimated between dimensions, which in turn results in a computational advantage compared to traditional MIRT models. The next section describes bifactor models in depth and how they may be applied with interaction effects.

2.3 Bifactor Models

The bifactor model is widely used in the social sciences, stemming from the work of Holzinger and Swineford (1937) and Holzinger and Harman (1941), with traditions in factor analysis (Thurstone, 1947; Spearman, 1904). The bifactor model is a measurement model for exploratory and confirmatory latent variable models, such as item response theory, that specifies that the covariance among a set of items is accounted for by specific or grouping latent dimensions alongside a general latent dimension. Though frequently applied to continuous item responses, the bifactor model has been applied to data with dichotomous (Gibbons & Hedeker, 1992), polytomous (Gibbons et al., 2007), and nominal responses (Cai et al., 2011) using item response theory methods.

Figure 2.7: Bifactor model with three specific dimensions



To demonstrate the relationships between items and latent variables, a path diagram in Figure 2.7 depicts a bifactor structure. As shown, all items $(y_1 ... y_9)$ measure the general latent dimension, θ_g, and at most one of multiple specific dimensions, θ_1, θ_2, or θ_3. The pattern of the bifactor model for the items displayed in Figure 2.7 can be described as

$$\beta = \begin{bmatrix} \beta_{1g} & \beta_{11} & 0 & 0 \\ \beta_{2g} & \beta_{21} & 0 & 0 \\ \beta_{3g} & \beta_{31} & 0 & 0 \\ \beta_{4g} & 0 & \beta_{42} & 0 \\ \beta_{5g} & 0 & \beta_{52} & 0 \\ \beta_{6g} & 0 & \beta_{62} & 0 \\ \beta_{7g} & 0 & 0 & \beta_{73} \\ \beta_{8g} & 0 & 0 & \beta_{83} \\ \beta_{9g} & 0 & 0 & \beta_{93} \end{bmatrix}. \tag{2.10}$$

The first column is the general factor, θ_g, and the subsequent columns are the three grouping or specific dimensions in the pattern matrix. Typically, items will measure at most one specific dimension alongside the general dimension, though recent research has extended the bifactor model to include multiple general dimensions (Cai, 2010) or multiple sets of specific dimensions in confirmatory factor analysis (Jeon et al., 2018).

Following the MGRM model for a polytomous item responses as in Equation 2.7 and the bifactor model depicted in Figure 2.7, the bifactor model for an item measuring the general dimension and at most one of the specific dimensions, completed by respondent r, is defined as follows:

$$P(y_{ri} \geq k|\theta) = \Phi(\beta_{i(k)} + \beta_{ig}\theta_{rg} + \beta_{is}\theta_{rs}). \tag{2.11}$$

The probability of an endorsed item response equal to k or greater becomes a function of the general latent dimension, θ_{rg}, and the specific dimension, θ_{rs}, which are contained in vector θ ($\theta = [\theta_g, \theta_{s1}, ..., \theta_S]$) with one element for the general dimension and up to S elements for specific dimensions. The category intercept, $\beta_{i(k)}$, for category k is interpreted as the expected probability of a k response for a respondent with average levels of the latent variables (where $\theta_g = \theta_s = 0$)

measured by the item. There are two main effect parameters estimated per item i to indicate the strength of association between the items and the corresponding dimension: β_{ig} for the θ_{rg} general dimension and β_{is} for the θ_{rs} specific dimension. When interpreting scores on these dimensions, they must be conditioned at a point on the latent metric of the other dimension, such as at 0 (the mean). For model identification, the general and specific dimensions are jointly standard normally distributed and mutually orthogonal (i.e., uncorrelated) to each other (Gibbons & Hedeker, 1992).

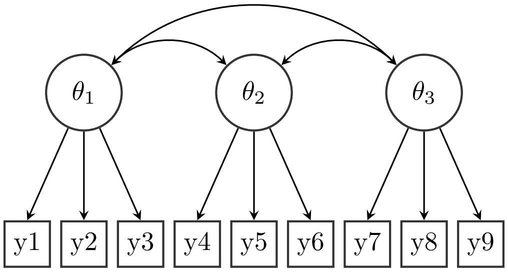
The specific dimensions of the bifactor model measure subsets of items to allow modeling of conditional dependence due to the unique influence of construct-relevant groupings or caused from a “nuisance” factor, such as with method factors (similar item phrasing) or testlets (common item stems) that threatens the validity of the psychometric model (Reise et al., 2010; Li et al., 2006; DeMars, 2012). The general dimension is presumed to directly influence all facets underlying the hypothetical construct by accounting for the relationships between all items measuring the construct. This is in contrast to traditional multidimensional models that can only capture variance associated with specific dimensions or unidimensional models that have a general dimension but are distorted by conditional dependence among item subsets. With the bifactor model, dimensionality assessment is more interpretable because the specific dimension structure can be discovered when the influence of the global construct and the unique contributions of the specific dimensions are parceled out (Reise et al., 2007; Chen et al., 2006). The analysis becomes more easily interpretable compared to other commonly applied models, as described next.

2.3.1 Frequently Compared Models

The bifactor model is advantageous for researchers interested in modeling specific constructs and a single “breadth” dimension that is not hierarchical in nature. To further illustrate the bifactor model and its applications, this explanation can be compared to competing models with similar latent variable interpretations that are also used to explain covariation among sets of item subsets, such as the correlated factor model (also called the oblique factor model, see Figure 2.8; DeMars, 2013; Jennrich & Bentler, 2012) and the second-order or higher-order factor model (see Figure

2.9). These models estimate correlations between specific latent dimensions, which can imply the presence of a general dimension when correlations are high. Though not the focus of this dissertation, the correlated factor model and second-order model are frequently contrasted with bifactor models. Thus, this section provides a brief comparison of these alternate models that have comparable conceptual interpretations.

Figure 2.8: Correlated factor model with three dimensions

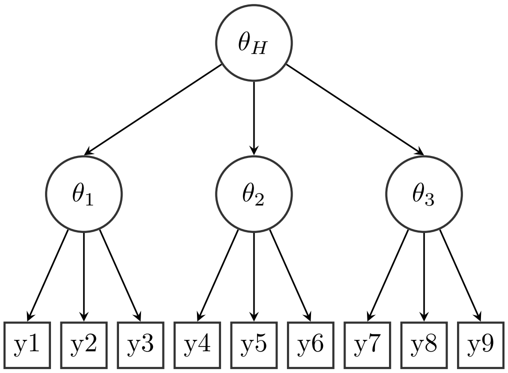


The compensatory and partially compensatory MIRT models previously described are correlated factor models when correlations between primary factors are estimated. The interpretation of these models is straightforward, where correlations among dimensions are what represent the general dimension. However, compared to the bifactor model, the correlated factors model is more computationally intensive (Cai, 2010; Cai et al., 2011; Gibbons et al., 2007; Gibbons & Hedeker, 1992). For a more in-depth explanation of the differences in estimation methods between the correlated factor model and the bifactor model, see DeMars (2013).

In the second-order model, a hierarchy represents the interrelationships between observed items and latent variables. The general, higher-order factor accounts for covariation between the lower-order (specific) dimensions, which account for the relationships between observed items. In this model, the higher-order dimension is only related to items through its indirect relationships with lower-order factors, such that the second-order factor is fully mediated by the first order-factors (Yung et al., 1999). The second-order model and the bifactor model are mathematically equivalent in confirmatory models when the ratio of general factor main effects to lower-order main effects in the bifactor model are equal within group factors (Yung et al., 1999; Gignac, 2008; Rijmen, 2010), often referred to as a proportionality condition (Gignac, 2016).

In exploratory bifactor modeling, the Schmid-Leiman orthogonalization procedure (Schmid & Leiman, 1957) allows researchers to partition variance common among all item subsets in a general factor from unique, systematic variance attributable to specific dimensions. However, this method relies on the proportionality condition which may not always result in an accurate solution because it relies on items having strictly simple structure (i.e., items measure at most one dimension) (Brunner et al., 2012; Jennrich & Bentler, 2011, 2012; Yung et al., 1999; Mansolf & Reise, 2016). The Schmid-Leiman orthogonalization works by starting with a correlated factor model obtained from a factor analysis with an oblique rotation that involves only specific dimensions, which is then converted into a second-order model with the addition of a general higher-order dimension. The result is then orthogonalized into an exploratory bifactor solution with proportionality constraints. As the ratio of factor loadings diverge from the proportionality constraints, the bifactor model tends to fit better and is preferred in likelihood ratio tests of nested models because it is less restrictive (Yung et al., 1999). The bifactor model is favored when additional complexities exist, such as cross-loadings and correlated errors (Murray & Johnson, 2013), though fit indices tend to be biased in favor of the bifactor model compared to the second-order model (Mansolf & Reise, 2017).

Figure 2.9: Second-order factor model with three lower-order dimensions



2.4 Bifactor Models with Interactions

The objective of this research was to extend the bifactor model to include interaction effects between the general and specific dimensions measured by an item. As described earlier, past research of MIRT models with interaction effects have shown difficulties in estimation and poor recovery of the item and person parameter estimates with increasing correlations between latent dimensions (Chalmers & Flora, 2014; DeMars, 2016; Rizopoulos & Moustaki, 2008). This study offers a bifactor model as a solution because correlations between specific dimensions are orthogonal based on the assumption that any relationships between specific dimensions are due to their shared variance with the general dimension (Reise, 2012; Reise et al., 2007). In other words, the correlations among specific dimensions are attributed to their common link among items, which is accounted for by the general dimension.

Similar to the I-M2PL and I-MGRM, bifactor models with interaction effects are useful in situations where the effect that one dimension has on the observed item response is dependent on the values of the other dimension the item measures. For example, consider a bifactor model of the M2PL with dichotomous item responses and continuous latent dimensions. Positive main effects and a positive interaction between the general and specific dimensions could correspond to an over-additive effect (or an under-additive effect if the interaction coefficient is negative) on the probability of endorsing the item when there are high scores on both the general dimension and specific dimensions. However, this implies that the predicted probability can only be bolstered if scores on both dimensions are high. If scores on the specific dimensions are low, high scores on the general dimension may not greatly increase the probability of endorsing the item or a higher category.

The proposed bifactor model with interactions for ordered-categorical responses, based on the I-M2PNO model from Equation 2.9 and the bifactor MGRM model given in Equation 2.11, for an item i completed by respondent r, is given as:

$$P(y_{ri} \geq k|\theta) = \Phi(\beta_{i(k)} + \beta_{ig}\theta_{rg} + \beta_{is}\theta_{rs} + \beta_{i(g*s)}\theta_{rg}\theta_{rs}). \tag{2.12}$$

The two main effects for the general and specific dimension, β_{i1} and β_{i2}, are now simple main effects as they are conditional on the interaction term, $\beta_{i(g*s)}$. Interpreting the relative contribution of one dimension is now dependent on the level of the other dimension and the value of the interaction parameter. For example, a unit increase in a respondent's score on the specific dimension increases the probability of a correct response, depending on the level of their general dimension. The model could lend itself to a conjunctive interpretation, depending on the values of the conditional main effects and the interaction. Specifically, with positive simple main effects and a positive interaction parameter, higher values of one dimension but not the other may not greatly increase the probability of a k item response. Comparatively, if the interaction is close to zero, one dimension can make up for low values of the other dimension as in a compensatory model, depending on the size of the main effects.

Consider the illustrations in Figure 2.10, 2.11, and 2.12 for three binary items that both measure the general and specific dimensions equally with main effects of 1 and an intercept of 1. The item in Figure 2.10 does not have an interaction effect, whereas the items in Figures 2.11 and 2.12 have interactions of positive and negative 0.3, respectively. The main effects in Figure 2.10 are marginal whereas those in Figures 2.11 and 2.12 are conditional on their interaction effect. As either dimension increases, the probability of an endorsed response also increases, but increases in both dimensions simultaneously can have an added boost or a diminishing return, depending on the value and direction of the interaction. Comparatively, if the interaction coefficient is small and approaching zero, high values of either latent variable will have approximately the same rate of increase in the probability of an endorsed response on the item, similar to the contours in a compensatory model.

Figure 2.10: Probability of a one response for an item modeled with a bifactor model having no interaction, $\beta_{i(g*s)} = 0$

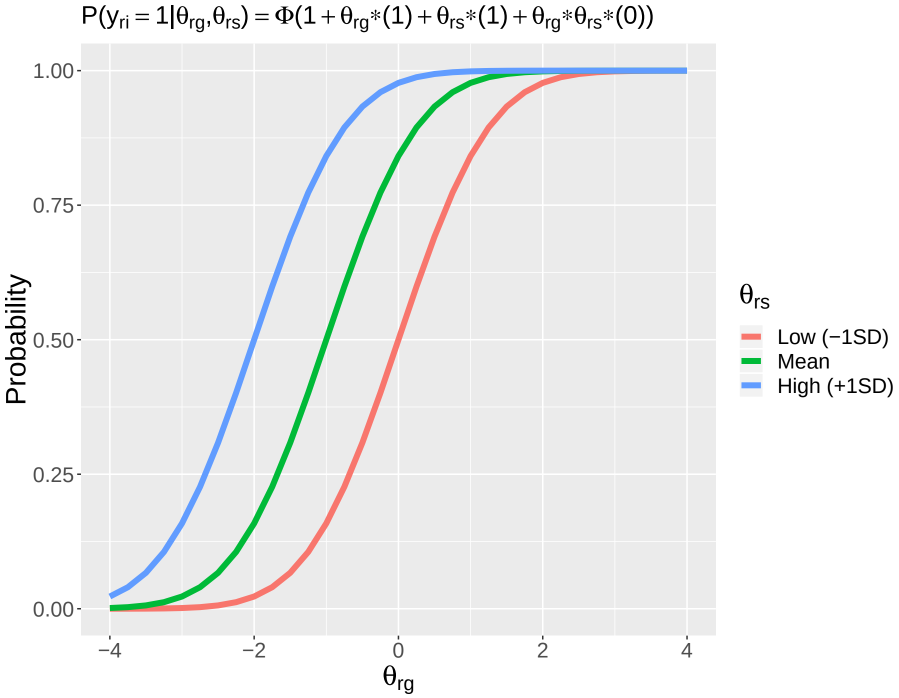


Figure 2.11: Probability of a one response for an item modeled with a bifactor model having a positive interaction, $\beta_{i(g*s)} = 0.3$

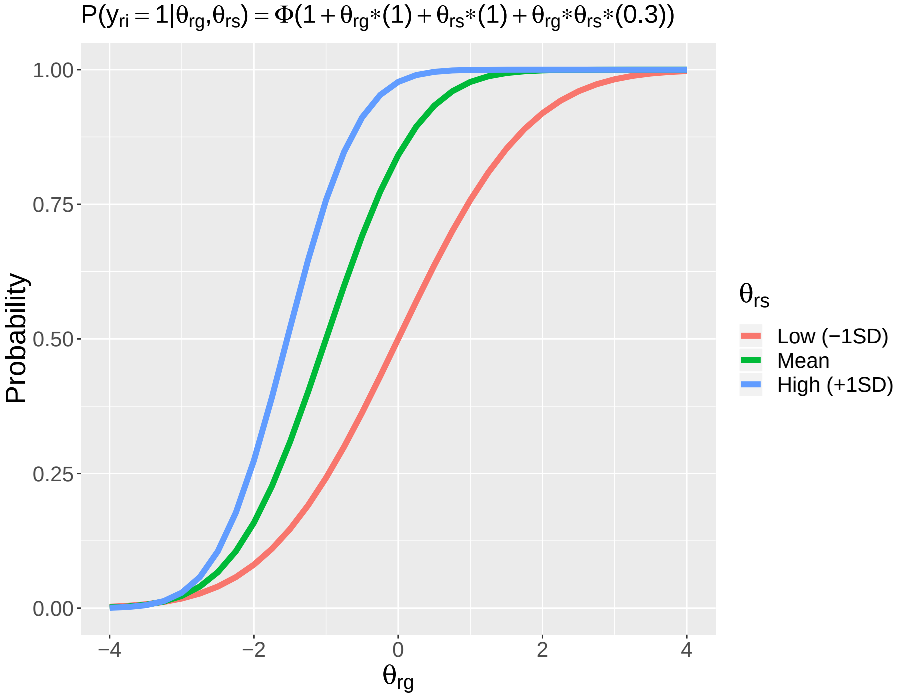
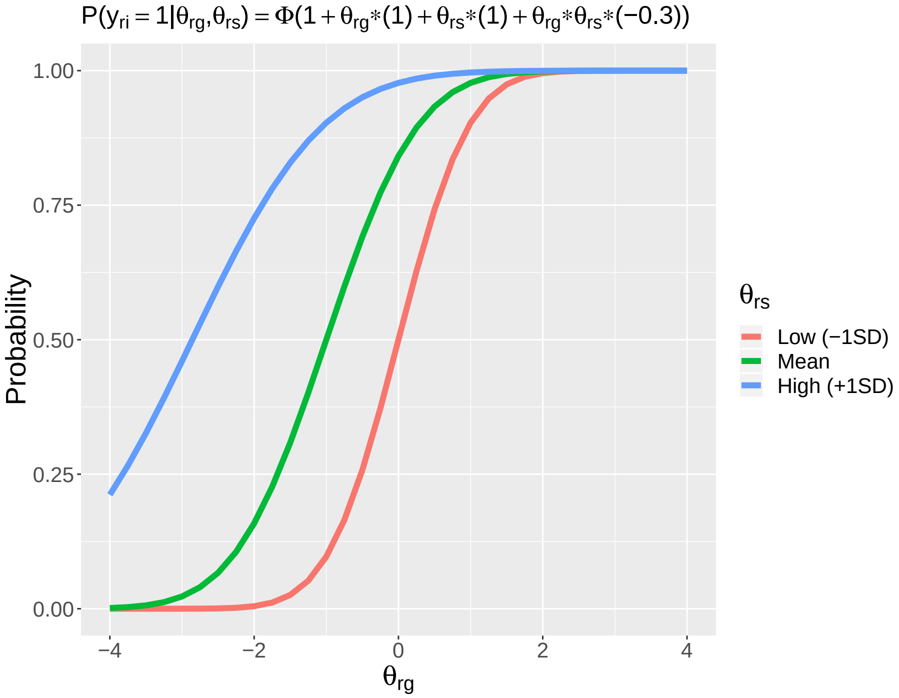


Figure 2.12: Probability of a one response for an item modeled with a bifactor model having a negative interaction, $\beta_{i(g*s)}$ = -0.3



2.5 The Current Study

Previous research of interaction effects in item response models have focused on models that have correlated factors and lack a general dimension to account for relationships between items. The aims of this dissertation were to extend this area of investigation by assessing bifactor models with interaction effects that specify a general dimension in addition to multiple specific dimensions to model latent constructs. In the current study, simulated and empirical data were used to evaluate three primary research questions. The first question investigates how well the proposed bifactor model with interactions performs under a range of data generating conditions and whether it can be estimated accurately and efficiently. The second question investigated the potential consequences of ignoring interaction effects as in a traditional bifactor model. Finally, the third question aimed to determine whether interactions can be uncovered when they are not known *a priori* and whether commonly used model selection and parameter refinement strategies can identify the correct data-generating model. Using generated and empirical data, this dissertation addressed the following three research questions:

1. How well does the proposed bifactor model with interaction effects recover item and person parameters in realistic data generating conditions?

2. What is the impact of ignoring interaction effects on model recovery of item and person parameters when fitting a bifactor model with only main effects?

3. How well do model selection and parameter refinement strategies perform in identifying underlying interactions in a bifactor model?

To address research questions 1 and 2, two Monte Carlo simulation studies outlined in Chapter 3 demonstrate model recovery for the bifactor model with and without interactions present and evaluate adverse effects that may arise when fitting a misspecified model. Informed by the results of these studies (Chapter 4), analyses of empirical data with applications in psychology and education are demonstrated to further investigate research question 3 in Chapter 5.

Chapter 3

Methods

Before investigating interaction effects in bifactor models with empirical data and demonstrating its usefulness as an approach to modeling latent constructs, the current study evaluated the bifactor model with interactions using Monte Carlo simulation. Monte Carlo simulation is a method that allows researchers to create a controlled environment to test statistical models because population parameter values are set by the researcher. With a simulation study, estimated model parameters from different models, either correctly specified or misspecified, can be compared against true values that were used to generate the data. As interaction effects in bifactor models have not been studied in previous research, little is known about the sampling distribution of interaction effects and how including them may affect model convergence in a range of data-generating conditions. Furthermore, recovery of bifactor model parameters and the potential consequences that can arise when fitting a misspecified model that ignores interaction effects has not been studied. Thus, Monte Carlo simulation is an appropriate choice for the research questions posed in this dissertation.

The following chapter provides readers with a description of the design of two Monte Carlo simulation studies: one for 2-category ordinal data that is typical of educational testing environments and one for 6-category ordinal data traditionally applied in psychological assessment research. Although the models and data-generating conditions for the simulation studies are different, the research questions are the same. Thus, this chapter begins with a description of the different study conditions that were chosen to investigate the theoretical properties of the bifactor model with interactions present in fitting ordered-categorical data. Then, for each research question, a description of the approach taken and various evaluation criteria used is given. Chapter 4 presents the results of the simulation studies, which is followed by empirical applications in

Chapter 5.

3.1 Simulation Study Design

The design of the Monte Carlo simulation studies was selected to address the three primary research questions. The first question focuses on the practicality of estimating the bifactor models with interactions and the quality of parameter estimates under a range of data-generating conditions. The second question compares the bifactor model with alternative specifications that do not match the data-generating model to examine the extent to which model misspecification due to omitted variables and model overfitting may impact the accuracy of the parameter estimates and respondent scores. The third question shifts focus from model quality and accuracy to implementation in empirical data. The performance of methods commonly used in the process of confirmatory model building and parameter refinement are evaluated in terms of their ability to identify interaction effects at the item level and select the true data-generating model.

3.1.1 Study Conditions

The variables manipulated in the simulation studies are presented in Table 3.1. The data-generating conditions of the simulation studies were the sample sizes of respondents and the percentage or number of items with interaction effects present. These variables were manipulated because it was expected that the accuracy of the estimated parameters of the bifactor model with interactions would depend on the number of interactions estimated and the sample size of respondents. The sample sizes were chosen to be representative of common sample sizes found in psychology research and large scale educational testing environments. The varying number of items with interaction effects present was chosen to investigate the effect that increased model complexity has on the ability of model building methods to detect interactions and the impact of model misspecifications resulting from over- and under-specified models. Furthermore, given the complexity of the model, it may be impractical to estimate and a more parsimonious alternative may be sufficient.

Table 3.1: Simulation study conditions

Label	Condition	2-Category Levels	6-Category Levels
N	Sample Size	500, 3,000	500, 1,000
I_{gen}	Items with Interaction Effects	0, 16 (25%), 32 (50%)	0, 8 (25%), 16 (50%)
I_{est}	Item Interactions Modeled*	Match, All, None	Match, All, None

* Indicates a nested condition

Thus, this condition also tests whether an increasingly complex underlying data structure affects parameter recovery accuracy and convergence. Accordingly, there are two sample size conditions and three interaction effect conditions, resulting in six data-generating conditions per simulation.

For each replication, the six datasets were estimated with up to three variations of a bifactor model (I_{est} = Match, All, None). The aim of estimating the data sets differently according to which interactions are assumed was to ensure the recovery of the estimated parameters when the model estimated matches the data-generating model (research question 1), weigh the potential consequences of fitting a misspecified model that omits interaction effects (research question 2), and to investigate model reduction strategies for use in empirical research when all possible item interactions were estimated (research question 3). The first model estimated was the correctly specified bifactor model that matches the data-generating model. The second fitted model had all item interaction effects estimated between the general and specific dimensions (no items measure more than one specific dimension). The final fitted model was a bifactor model that omits interactions (i.e., main effects only bifactor model). The resulting design factors are fully crossed as displayed in Table 3.1, with one exception: the models generated without interactions cannot be estimated with matching interactions. Thus, each simulation has 2x3x2(3) = 16 conditions examined.

3.1.2 Data Generation

Data sets were generated according to the bifactor MGRM with item interactions as in Equation 2.9 using the R software package, version 3.5.1 (R Core Team, 2018). Test length was fixed; the 2-category simulation had 64 items and the 6-category simulation had 32 items. The 2-category simulation was designed based on large-scale testing environments that typically involve large numbers of items for an accurate estimation of θ. Comparatively, psychological assessments typically involve shorter assessments and several dimensions to measure a latent construct.

The items measured at most one of four normally distributed and orthogonal specific dimensions and the general dimension. Thus, each specific dimension had an equal number of items measuring each (calculated as the total number of items divided by the number of specific dimensions). When data were generated with interactions, the number of items having interactions per specific dimension was also equal. Though items could theoretically measure multiple general or specific dimensions with or without an interaction (Jeon et al., 2018; Cai, 2010), this study only considers items that measure one specific dimension and one general dimension.

Data-generating parameters for the two simulations are shown in Table 3.2 and described in terms of the parameters of Equation 2.12. General and specific dimensions were simulated from a multivariate standard normal distribution ($\theta \sim MVN_D(\mathbf{0}, \Sigma)$) with mean $\mathbf{0}$ and an identity covariance matrix, Σ, to correspond to bifactor model identification. Item parameters for the two simulations were simulated for each replication from distributions to generalize results to empirical data where items have varying item parameters. Item intercepts were obtained by randomly drawing values using $\beta_{i0} \sim uniform(-1, 1)$ for the 2-category simulation to simulate tests with conditional item means centered around the latent variable means. For the item intercepts of the 6-category simulation study, the random draws were $\beta_{i(k)} \sim uniform(-3, 3)$ and sorted to reflect the cumulative ordering of the response options. The main effects, β_{ig} and β_{is}, were drawn per item and per dimension using $\sim uniform(0, 2)$ to reflect the assumption that items do not necessarily relate to the general and specific dimensions equally.

Given the relatively few studies on parameter recovery of partially compensatory MIRT models with interaction effects, this study used parameter-generating values for interaction effects informed by DeMars (2016). In the present study, it is believed that item interaction effects can be positive, negative, or zero. Thus, item interactions were drawn from a uniform distribution, $\beta_{i(g*s)} \sim uniform(-0.5, 0.5)$, with an interval that contains positive and negative values, centered at zero. Interaction values for an item were not conditional on their main effects.

Table 3.2: Data-generating parameters

Parameter	Label	Distribution
2-Response Item Intercepts	β_{i0}	$\sim Unif(-1,1)$
6-Response Item Intercepts	$\beta_{i(k)}$	$\sim Unif(-3,3)$
Main Effects	β_{ig}, β_{is}	$\sim Unif(0,2)$
Interactions	$\beta_{i(g*s)}$	$\sim Unif(-0.5,0.5)$
Latent Variables	θ_{rs}, θ_{rg}	$\sim MVN_D(\mathbf{0},\Sigma)$

3.2 Model Estimation

This study used Bayesian estimation with Markov chain Monte Carlo (MCMC) simulation techniques. This method was chosen given its greater flexibility in the estimation of complex models with high-dimensional parameter spaces and large numbers of item parameters that would otherwise be problematic using traditional frequentist methods such as maximum likelihood. A number of studies estimating IRT models have used MCMC methods and found comparable results to other common estimation methods such as Marginal Maximum Likelihood and the Expectation-Maximization (EM) Algorithm (Patz & Junker, 1999a,b; Kim & Bolt, 2007; Wollack et al., 2002; Kieftenbeld & Natesan, 2012; Gelfand et al., 1990).

All models were estimated using JAGS (version 4.3.0), abbreviated from “Just another Gibbs Sampler,” a Bayesian software program (Plummer, 2017). JAGS uses a variety of sampling methods to estimate models, depending on the distribution of parameters (Plummer, 2017). In this study, slice sampling (Neal, 2003) was used to sample item parameters coming from continuous scalar prior distributions and an adaptive random walk Metropolis algorithm (Metropolis et al., 1953; Hastings, 1970; Chib & Greenberg, 1995) for respondent parameters having a multivariate normal prior distribution. JAGS model code for ordered 2-category data and K-category data is presented in Appendix A and B, respectively. Because the models considered here were confirmatory, the model code includes two $I \times D$ indicator matrices (“designMain” and “designInt”) with ones and zeros to specify whether to estimate main effects and interaction effects.

Informed by preliminary testing, MCMC estimation was completed with the following specifications: 7,000 iterations per 2 chains with 3,500 discarded burn-in iterations and a thinning interval of 3. Priors to estimate item parameters were $\beta_{i\cdot} \sim N(0,1)$, with the exception of main effects, which were drawn from a standard normal distribution truncated at 0 for positive main effects. These informative priors were chosen to restrict the range of the intercepts and interactions to values that tend to lie between -4 and 4 and the main effects (conditional and marginal) to tend to lie between 0 and 4. Latent variable priors for the general and specific dimensions were drawn from a multivariate normal distribution, $\theta \sim MVN_D(\mathbf{0}, \Sigma)$, with $\mathbf{0}$ mean vector and identity covariance matrix, Σ, with 1's on the diagonal and 0's on the off-diagonals. Estimates of parameter values were obtained from the mean of the retained iterations.

For both simulations, each replication had unique random seeds and random number generators per MCMC chain. Starting values were set to their prior means of zero for all parameters, with the exception of the category-intercepts in the 6-category simulation: starting values for category-intercepts were sorted draws from normal distributions with means, $\mu = [-3, -1, -0, 1, 3]$, and standard deviations of 1. These starting values were used for all models estimated within a replication. Random seeds for generating starting values were set to their replication seeds, per chain.

3.3 Evaluation Criteria

Output analysis was completed with R (R Core Team, 2018) using built-in functions and functions within the coda package (Plummer et al., 2006). All models were checked for convergence using traditional methods, including the Gelman-Rubin estimated potential scale reduction factor (PSRF; Gelman & Rubin, 1992; Brooks, S. & Gelman, 1998) and Geweke's Z-score diagnostic (Geweke, 1992). Geweke's Z-score assesses convergence by comparing the mean and variance at the beginning and end of a single chain. If the z-scores are comparable at the two endpoints, it can provide evidence of convergence. The PSRF uses multiple chains to compare how similar they are according to their between-chain variance and within-chain variance. A PSRF threshold value of 1.2 was used to indicate convergence. A replication was included in subsequent analyses if at least 80% of all parameters converged according to both Geweke's Z score and the estimated PSRF.

Model parameter recovery was evaluated by looking at the posterior mean of the parameters compared to their true generated values. Since item parameters (intercepts, main effects, and interactions) and respondent scores on the latent variables (general and specific dimensions) were generated as random variables within a replication, recovery for parameter type was aggregated at the test level. For each estimated model parameter within each simulated replication of a condition, this simulation study used the following statistics to evaluate parameter recovery: average Pearson correlations between true and estimated respondent parameters, average parameter bias, average absolute parameter bias (ABS), and average root mean squared errors (RMSE). These statistics are calculated as follows:

$$Average\,Bias(\delta) = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{K}\sum_{k=1}^{K}(\hat{\delta}_k - \delta_k)\right] \tag{3.1}$$

$$Average\,Absolute\,Bias(\delta) = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{K}\sum_{k=1}^{K}|(\hat{\delta}_k - \delta_k)|\right] \tag{3.2}$$

$$Average\,RMSE(\delta) = \frac{1}{R}\sum_{r=1}^{R}\left[\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\delta}_k - \delta_k)^2}\right] \tag{3.3}$$

where δ_k is the true value of a parameter of type k compared against its estimated parameter, $\hat{\delta}_k$, averaged over K parameters for a condition within a replication, r, which is then averaged over R replications.

To compare the model-data fit of the different models estimated and evaluate the performance of model selection strategies in identifying the true underlying model, this study used deviance information criterion (DIC; Spiegelhalter et al., 2002), a Bayesian measure of the relative quality of a model. The DIC is a Bayesian analog of the commonly used Akaike's information criterion (AIC; Akaike, 1974) that combines the goodness-of-fit of a model and a penalty factor for model complexity. As with AIC, the DIC is an estimate of the expected predictive error, and therefore the model with smaller values is preferred when comparing alternative models.

The DIC of a model is calculated as

$$DIC = \bar{D} + pD \tag{3.4}$$

where $\overline{D}$ is the posterior mean of the deviance ($\bar{D} = E_{\theta|\mathbf{Y}}[D]$). The deviance D is calculated at each iteration as the sum of the deviance of all the observed stochastic nodes (Plummer, 2017). The penalty factor, *pD*, is an estimate of the effective number of parameters of the model (Spiegelhalter et al., 2002). The value of *pD* used in this study is described by Plummer (2002) and is based on the Kullback-Leibler information divergence between predictive distributions at two different values of θ. Defined in mathematical form, the Kullback-Leibler information divergence is:

$$I(\theta^0, \theta^1) = E_{Y_i|\theta^0}\left[log\left\{\frac{p(Y_i^0|\theta^0)}{p(Y_i^0|\theta^1)}\right\}\right]. \tag{3.5}$$

In the context of MCMC sampling, *pD* for an iteration *i* is the sample mean of the Kullback-Leibler information divergence between the distributions of θ for two chains (indexed by superscripts 0 and 1). The sample mean of *pD* was used to calculate the DIC of the model.

When estimating interactions, the rate in which items with and without data-generated interactions were correctly identified was evaluated by calculating the average proportion of items across replications within a data-generating condition that contained their true parameter value, either $\beta_{i(g*s)}$ or zero, within their 95% highest posterior density intervals (HDI) and credible intervals (CI). The HDI gives the endpoints of the highest part of the density, which are created from the empirical cumulative distribution function (CDF) of the posterior simulations that contain 95% of the posterior probability. The CI endpoints are the upper and lower quantiles around the mean of the posterior. The lower and upper limits of the HDI and CI used in this study were 0.025 and 0.975 quantiles of sampled draws, post burn-in. For non-interacting items, false positives would be items identified as having interactions when they are not present in the underlying data structure (they do not contain zero within the estimated interaction HDI or CI). Comparatively, the proportion of misidentified interacting items is the rate in which generating values for true interacting items were not contained within their HDI or CI.

3.3.1 Pilot Study

An initial pilot study was completed to evaluate the between-simulation sampling variability of the aforementioned Monte Carlo estimators using Monte Carlo error (MCE; Koehler et al., 2009) to ensure that there were enough replications of each simulation condition. The MCE is a measure of uncertainty that estimates the extent to which differences across simulation replications depends on the differences between simulation runs (i.e., the structure and distributional assumptions of the simulation) and the number of replications (Koehler et al., 2009). Given a condition, the MCE for a φ quantity of interest (e.g., average RMSE of intercept parameters) becomes:

$$MCE(\varphi) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\varphi}_r - \bar{\varphi})^2}{R-1}} \tag{3.6}$$

where $\hat{\varphi}_r$ denotes the estimate of φ for a replication, r, and $\bar{\varphi}$ is the mean of φ over R replications. Because parameter estimates were generated from a range of possible values instead of fixed for all items for each replication and the model itself is complex, there may be high MCE variability that demands more replications. Thus, an initial run of 100 replications per data-generating condition was done to estimate MCE and determine whether more replications were needed.

Chapter 4

Results

In this chapter, results are presented from the simulation studies dealing with the estimation and parameter accuracy of the bifactor model with interactions, the potential consequences of model misspecifications, and the effectiveness of model selection and parameter refinement strategies in recovering the true data-generating model. Specifically, the primary purpose of the simulation studies was to evaluate the convergence of models and recovery of parameters for the bifactor model when interactions are present using simulated data with known generating models and parameters. Additionally, this study investigated the effects on parameter recovery when interactions were omitted and when interactions were estimated for all items. Finally, the third objective was to determine strategies for use in empirical data that are most effective in identifying underlying interactions for individual parameters and choosing the correct model when making model comparisons.

4.1 Simulation Run Time

The cluster computing facility at the University of Kansas was used to complete the data generation and estimation for the simulation studies. Given the complexity of the model combined with Bayesian methods for estimation, a replication required a large amount of time to run. JAGS was initiated via the command line interface on a single computing core. For the 2-category simulation, time and RAM utilized to estimate a single model took on average 0.50 days with 623 MB of RAM when $N = 500$ and 7.49 days with 3.41 GB of RAM when $N = 3,000$ across all other study conditions. Comparably, an estimated model in the 6-category simulation with $N = 500$ required approximately 2.29 days and 560 MB of RAM, whereas data with $N = 1,000$ required 4.79 days and 1 GB of RAM. The amount of time to complete 100 replications (with three models estimated per replication) took several months for each simulation. In light of this, practitioners should allocate enough time for model testing and estimation when using these models.

4.2 2-Category Simulation

4.2.1 Convergence

Of the 1600 estimations (16 conditions x 100 replications), the proportion of replications that had over 80% of all parameters converge according to both Geweke's Z-score diagnostic and the estimated PSRF for each condition was 100%, except for data generated with a sample size of 500 and all interactions estimated. For the three levels of interactions generated, I_{gen} = 0, 16, and 32, overall convergence rates were 0.94, 0.95, and 0.93, respectively. From these results, it appears that the degree of estimation complexity, defined by the increase in estimated model parameters, increases the rate of nonconvergence regardless of data complexity. However, this effect was not observed in the $N = 3,000$ sample size conditions where all model estimations converged. This finding suggests that larger sample sizes may overcome convergence issues, though looking closer at nonconvergence according to Geweke's Z-score diagnostic for each of the different types of parameters as shown in Table 4.1 can help illuminate this finding.

Overall, the differences in convergence rates did not vary substantially among the different simulation conditions, and differences within a parameter type tended to be in the hundredth decimal place. The intercept parameters had the lowest rates of convergence among all the parameters, regardless of the data-generating condition or model estimated. The general and specific latent dimensions tended to have slightly higher rates of convergence in the larger sample size condition, which could explain the higher overall rate of convergence when $N = 3,000$. Considering main effects, the general dimension main effects had higher rates of convergence compared to specific main effects for all conditions. Interactions converged at a higher rate when they were estimated to match the data-generating model compared to when all interactions were estimated, and the overall rates of convergence tended to be higher when the sample size was $N = 3,000$.

When there were no interactions generated, estimating all interactions tended to have lower convergence rates for all model parameters compared to estimating no interactions (matching data-generating model) in the $N = 500$ conditions, but not in the $N = 3,000$ conditions. With a larger sample size and no generated interactions, convergence rates were equal for the different estimated parameters with the exception of the intercept, which had a minimal increase in the rate of convergence by 0.02 when all interactions were estimated compared to none. A similar finding was not observed when the data had either 16 or 32 interactions generated. For an item parameter type, the convergence rates per data-generating condition when interactions were generated did not appear consistently higher or lower when all interactions were estimated compared to the none or match estimated interactions conditions. The 18 non-converged estimations were excluded in subsequent analyses.

Table 4.1: Binary response average proportion of converged parameters according to Geweke's Z-score diagnostic under different conditions

			Estimated Parameter					
N	I_{gen}	I_{est}	β_{i0}	β_{ig}	β_{is}	$\beta_{i(g*s)}$	θ_{rg}	θ_{rs}
500	0	All	0.80	0.88	0.92	0.91	0.92	0.95
		None	0.84	0.90	0.95	*	0.94	0.96
	16	All	0.82	0.88	0.92	0.91	0.92	0.95
		Match	0.84	0.88	0.95	0.95	0.94	0.96
		None	0.82	0.90	0.94	*	0.94	0.96
	32	All	0.82	0.87	0.91	0.90	0.91	0.94
		Match	0.84	0.91	0.93	0.95	0.94	0.96
		None	0.81	0.90	0.94	*	0.93	0.96
3000	0	All	0.85	0.89	0.95	0.94	0.96	0.97
		None	0.83	0.89	0.95	*	0.96	0.97
	16	All	0.87	0.89	0.95	0.94	0.96	0.97
		Match	0.82	0.89	0.94	0.95	0.96	0.97
		None	0.81	0.90	0.95	*	0.96	0.97
	32	All	0.80	0.89	0.93	0.95	0.96	0.97
		Match	0.83	0.90	0.93	0.95	0.97	0.97
		None	0.86	0.89	0.95	*	0.97	0.97

* Non-estimated parameter

4.2.2 Parameter Recovery

The recovery of item parameters and respondent parameters according to their bias, absolute bias, and RMSE can be seen in Figures 4.1-4.3 with corresponding Monte carlo errors (MCE) in Tables 4.2 and 4.3. The pilot study showed small MCEs for item and respondent parameters, so additional replications were not completed. For the different data-generating conditions, MCE was largest for bias parameters in the smaller sample size condition. Recall that the data generated with the main effects only bifactor model has a nested estimation condition, such that the no interactions estimated condition is the same as the match interactions estimated condition when there are no interactions generated.

RMSE for the general and specific main effects shown in Figure 4.1 did not vary substantially among the study conditions, and was only slightly larger when the main effects only bifactor model was fitted to data with underlying interactions. Comparatively, bias and absolute bias for the general and specific main effects differed among the study conditions. Firstly considering the main effects, bias increased for all estimated models when data were generated with more items having interactions, though absolute bias was comparable for the different levels of generated interactions. For the baseline model with no interactions generated, the bias was negative when estimated to match the data-generating model, and positive when all interactions were estimated. Absolute bias of the general main effects within this condition was only slightly larger when all interactions were estimated versus none. With a larger sample size, bias and absolute bias were close to zero when no interactions were generated. In the $N = 500$ condition when either 16 or 32 items had interactions generated, general main effect bias was more negative and absolute bias was larger when the model estimated omitted interactions compared to models estimated with all or matching interactions, which reversed when $N = 3,000$. The negative direction of bias indicated that the general main effects tended to be more underestimated with more omitted underlying interactions, but when $N = 3,000$, bias was largest when interactions were modeled to match the data-generating model. In fact, the bias of general main effects when matched interactions were estimated was similar to bias when no interactions or all interactions were modeled. Interestingly, bias of general main effects was smaller when estimated with all interactions than with matching interactions for both sample sizes, however the absolute bias was approximately the same for the different estimated models in the larger sample size condition.

Compared to the general main effects, recovery was best for specific main effects also shown in Figure 4.1 when the data-generating model matched the estimation model for $N = 500$, and when all interactions were estimated when $N = 3,000$. However, in the larger sample size, bias and absolute bias when interactions were present and modeled with either all or matched interactions was nearly the same. As in the general main effects, specific main effects were positively biased when the sample size was small and all interactions were estimated, substantially more so compared to the correctly specified model. Omitting interactions had the largest bias, absolute bias, and RMSE when interactions were generated for all data-generating conditions, though overall main effect parameters were recovered well for the different study conditions.

Table 4.2: Binary response Monte Carlo error (MCE) of bias, absolute bias, and root mean square error (RMSE) of respondent parameters under different conditions

			Bias MCE		Absolute Bias MCE		RMSE MCE	
N	I_{gen}	I_{est}	θ_{rg}	θ_{rs}	θ_{rg}	θ_{rs}	θ_{rg}	θ_{rs}
500	0	All	0.036	0.020	0.022	0.013	0.019	0.024
		None	0.038	0.021	0.024	0.013	0.018	0.024
	16	All	0.036	0.020	0.021	0.011	0.020	0.020
		Match	0.035	0.019	0.020	0.011	0.018	0.020
		None	0.035	0.019	0.020	0.011	0.018	0.021
	32	All	0.041	0.020	0.025	0.012	0.020	0.019
		Match	0.040	0.019	0.024	0.012	0.019	0.019
		None	0.041	0.020	0.025	0.012	0.020	0.021
3000	0	All	0.018	0.008	0.011	0.005	0.016	0.016
		None	0.019	0.008	0.012	0.005	0.015	0.016
	16	All	0.018	0.010	0.010	0.007	0.014	0.016
		Match	0.019	0.010	0.010	0.007	0.013	0.016
		None	0.018	0.010	0.010	0.007	0.014	0.017
	32	All	0.020	0.009	0.013	0.005	0.017	0.017
		Match	0.020	0.009	0.013	0.005	0.017	0.017
		None	0.020	0.010	0.013	0.006	0.019	0.019

Table 4.3: Binary response Monte Carlo error (MCE) of bias, absolute bias, and root mean square error (RMSE) of item parameters under different conditions

			Bias MCE				Absolute Bias MCE				RMSE MCE			
N	I_{gen}	I_{est}	β_{i0}	β_{ig}	β_{is}	$\beta_{i(g*s)}$	β_{i0}	β_{ig}	β_{is}	$\beta_{i(g*s)}$	β_{i0}	β_{ig}	β_{is}	$\beta_{i(g*s)}$
500	0	All	0.036	0.037	0.023	0.033	0.023	0.027	0.018	0.019	0.016	0.018	0.019	0.02
		None	0.039	0.035	0.023	*	0.024	0.025	0.015	*	0.018	0.017	0.016	*
	16	All	0.041	0.037	0.024	0.041	0.025	0.026	0.017	0.026	0.015	0.019	0.018	0.026
		Match	0.039	0.036	0.024	0.051	0.023	0.034	0.015	0.035	0.014	0.022	0.016	0.039
		None	0.040	0.037	0.024	*	0.023	0.035	0.019	*	0.014	0.023	0.016	*
	32	All	0.044	0.031	0.027	0.041	0.027	0.028	0.020	0.026	0.018	0.022	0.021	0.025
		Match	0.043	0.030	0.027	0.038	0.027	0.030	0.016	0.021	0.018	0.023	0.019	0.032
		None	0.042	0.033	0.026	*	0.026	0.033	0.023	*	0.017	0.024	0.022	*
3000	0	All	0.022	0.017	0.010	0.014	0.012	0.011	0.007	0.008	0.009	0.008	0.007	0.008
		None	0.021	0.017	0.009	*	0.012	0.010	0.006	*	0.009	0.008	0.007	*
	16	All	0.021	0.020	0.014	0.02	0.013	0.020	0.008	0.012	0.009	0.017	0.011	0.01
		Match	0.022	0.020	0.014	0.024	0.012	0.020	0.010	0.014	0.009	0.017	0.011	0.018
		None	0.021	0.021	0.013	*	0.013	0.021	0.012	*	0.010	0.019	0.013	*
	32	All	0.022	0.021	0.014	0.025	0.014	0.021	0.009	0.015	0.008	0.021	0.011	0.017
		Match	0.023	0.022	0.014	0.019	0.014	0.022	0.009	0.013	0.008	0.022	0.011	0.027
		None	0.025	0.022	0.014	*	0.015	0.022	0.014	*	0.011	0.021	0.016	*

* Non-estimated parameter

Figure 4.2 shows bias, absolute bias, and RMSE for intercepts and interactions. The intercepts were recovered well for all estimated models, particularly in the larger sample size condition. Intercept bias and absolute bias was largest for data with a larger proportion of items and a small sample size that was fit with the main effects only bifactor model. Intercept RMSE was comparable between estimated models within a sample size condition, though fitting only main effects when interactions are present resulted in a larger RMSE compared to estimating either all or matched interactions. Interaction parameters had adequate parameter recovery. Bias and absolute bias was small, though RMSE was highest compared to other item parameters when $N = 500$ for all estimation conditions. This finding supports the notion that more data is required for increased accuracy of the interaction effects.

Correlations between true and estimated respondent scores on the latent variables were consistent between conditions, having values of either 0.94 or 0.95 for the general dimension and 0.88 or 0.89 for specific dimensions. As shown in Figure 4.3, RMSE of respondent parameters was recovered similarly for the data-generating and model estimation conditions, with average RMSE equal to 0.309 for the general dimension and 0.466 for the specific dimensions. Bias and absolute bias also shown in Figure 4.3 for the respondent parameters was small and decreased with a larger sample size for all other study conditions. With $N = 500$, the general dimension bias for this condition became slightly less negative when estimating all interaction effects when there were none. The decrease in bias for the general dimension was not observed in other conditions, where bias was generally the same for the different estimated models within a data-generating condition. For the specific dimension with $N = 500$, the bias and absolute bias was the same when no interactions were generated and models were fit with either no or all interactions.

Figure 4.1: Binary response bias, absolute bias, and root mean squared error (RMSE) for general and specific main effect parameters

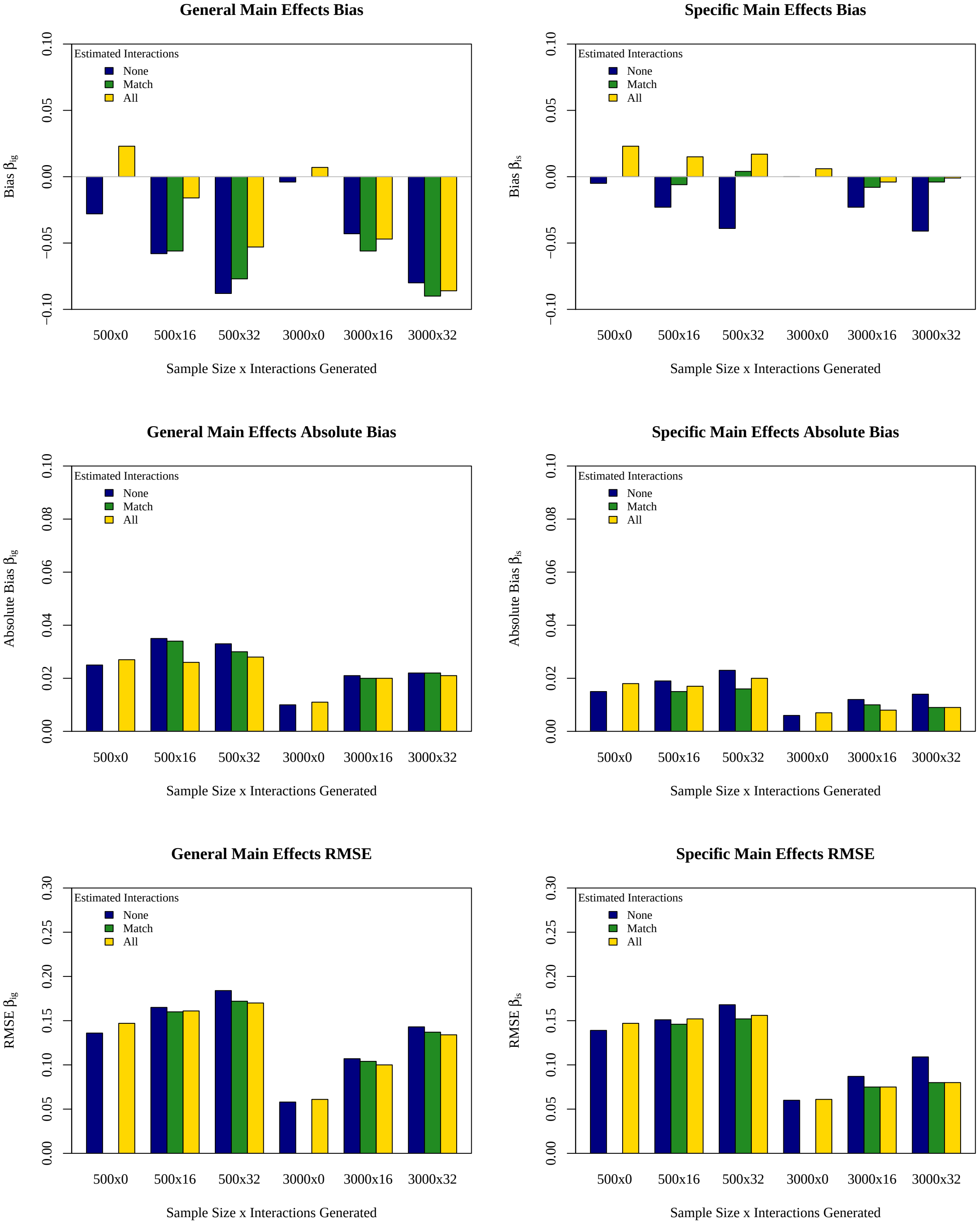


Figure 4.2: Binary response bias, absolute bias, and root mean squared error (RMSE) for intercept and interaction parameters

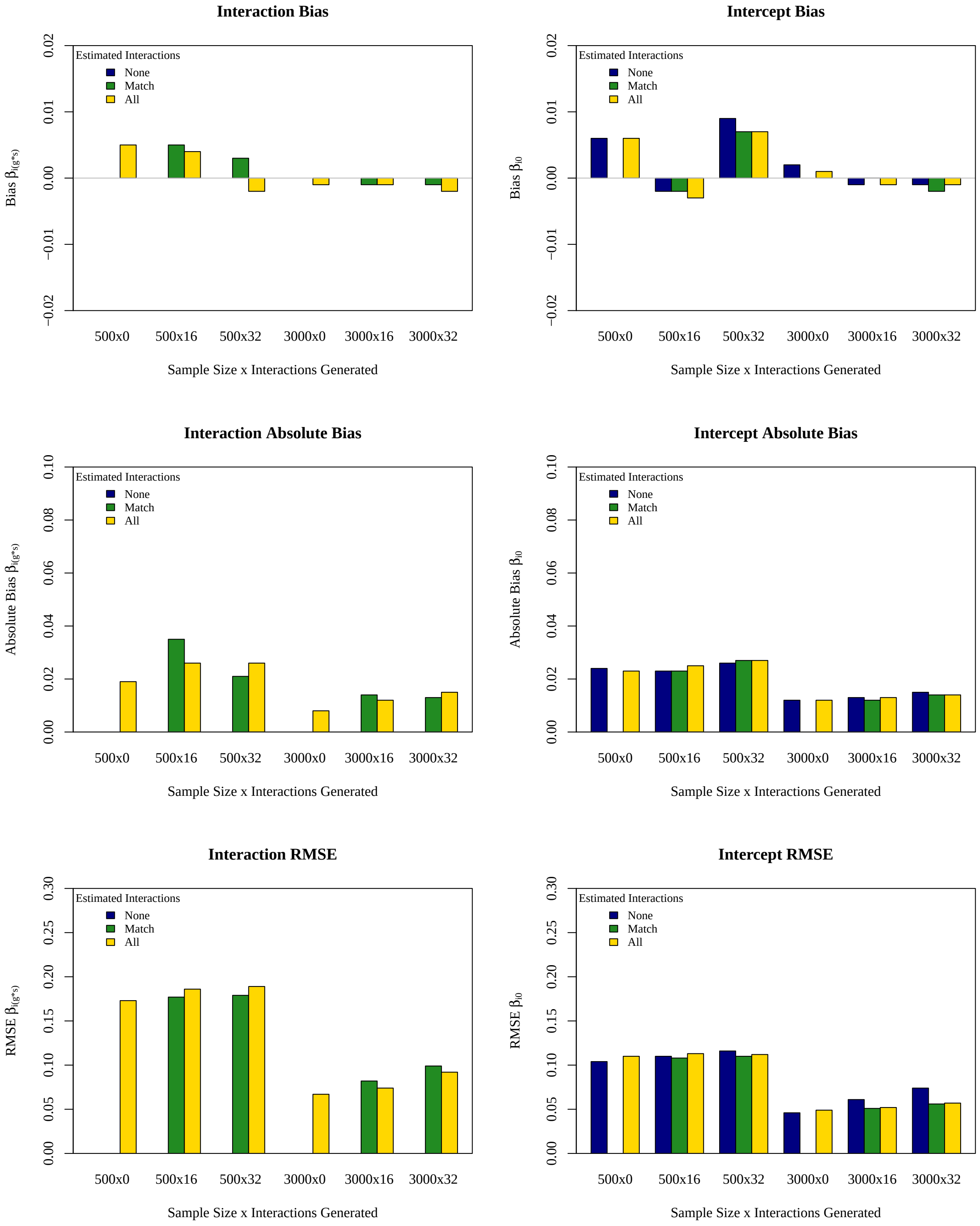
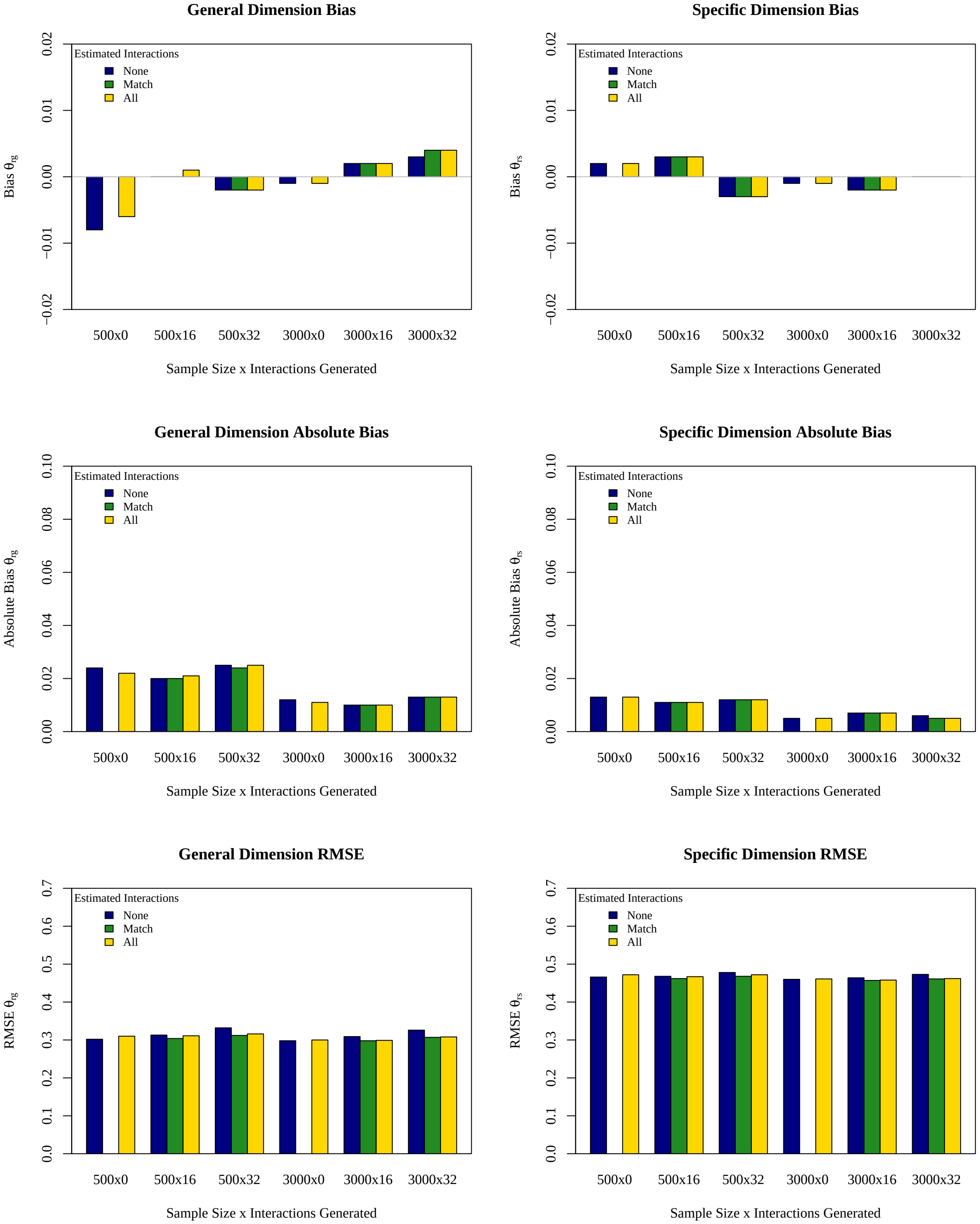


Figure 4.3: Binary response bias, absolute bias, and root mean squared error (RMSE) for general and specific dimension respondent parameters



4.2.3 Model Selection

Model selection frequencies are displayed in Table 4.4. Model comparisons were only made for a replication within a data-generating condition if it had all variations of interactions estimated (none, all, and matching) available to compare. Thus, the totals displayed in Table 4.4 will not sum to 100 for some conditions because there were 18 analyses removed due to nonconvergence. Generally, the true data-generating model was favored in relative fit model comparisons, regardless of sample size or the proportion of items having underlying interactions. The DIC preferred the misspecified main effects only bifactor model in only 9 out of 282 estimations (3.2%) for data with sample sizes of $N = 500$ but otherwise selected the correct model. Comparatively, the main effects only bifactor model was not favored over the data-generating model in larger sample sizes, though the DIC favored all interactions in 11 estimations out of 300 (3.7%) compared to the correct model. These results may be due to the penalty parameter, *pD*, which penalizes more complex models.

Table 4.4: Binary response model selection frequencies

		Estimated Interactions, I_{est}			
N	I_{gen}	None	Match	All	Total
500	0	*	94*	0	94
	16	8	87	0	95
	32	1	92	0	93
3000	0	*	100*	0	100
	16	0	100	0	100
	32	0	89	11	100
Total		9	562	11	582

* Indicates a nested condition

4.2.4 Model Reduction

To evaluate the effectiveness of highest density intervals (HDI) and confidence intervals (CI) in identifying underlying interactions, the average proportion of items that contained their true data-generating interaction value within the corresponding interval were evaluated for items estimated and generated with interactions and items estimated but not generated to have interactions. Firstly, the average proportion of interacting items containing their generating value within each interval was evaluated for the model estimated that matched the data-generating model to serve as a baseline and is presented in Table 4.5. The proportion of items that did not contain their data-generating value within their HDI and CI was higher in the larger sample size condition, regardless of whether there were 16 or 32 items having interactions. Specifically, when there are 32 interacting items in the larger sample size of 3,000, 19.3% of items on average did not contain their true interaction value within their highest density interval, which amounts to approximately 6 items. Results for the CI and HDI were nearly the same across the different data-generating conditions.

Table 4.5: Binary response average proportions (and Monte Carlo error) of true parameters outside of highest density intervals and credible intervals in models estimated with matched interactions

N	I_{gen}	HDI	CI
500	16	0.084 (0.076)	0.089 (0.079)
	32	0.078 (0.055)	0.089 (0.060)
3000	16	0.161 (0.095)	0.159 (0.094)
	32	0.193 (0.100)	0.194 (0.102)

Figure 4.4: Binary response proportion of items with true parameters outside of highest density intervals in models estimated with all possible interactions

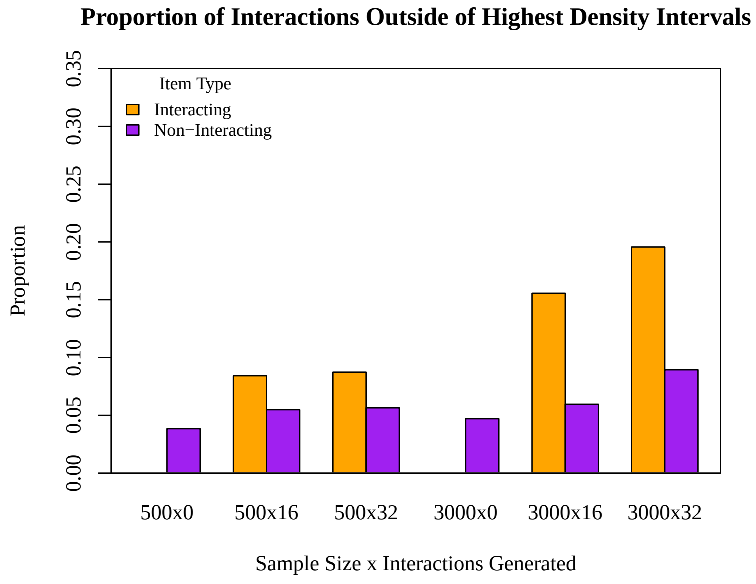
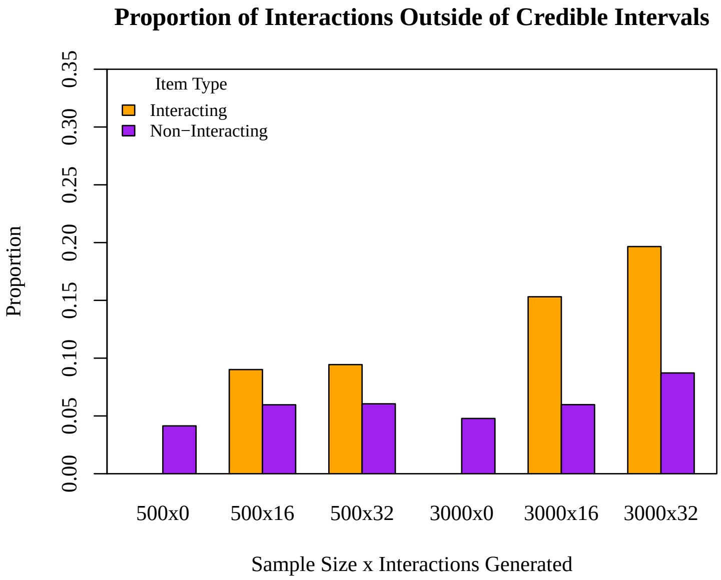


Figure 4.5: Binary response proportion of items with true parameters outside of credible intervals in models estimated with all possible interactions



For the models estimated with all possible interactions, Figures 4.4 and 4.5 show the proportion of misidentified interactions for items without interactions generated and interaction-generated items according to their highest density intervals (HDI) and confidence intervals (CI). As in the model estimated with interactions to match the data-generating model, the proportion of misidentified underlying interactions when $N = 3,000$ increased with more underlying interactions, but not when the sample size was $N = 500$. A similar effect occurred with items that were not generated to have interactions, though it was less pronounced; the proportion of non-interacting items that were falsely identified as having interaction increased in the larger sample size condition and in conditions with more items within a test were generated to have interactions.

4.3 6-Category Simulation

4.3.1 Convergence

Convergence rates in the 6-category simulation were comparable to the 2-category simulation. For the 100 replications, only 18 of the 1600 estimations had below 80% of all parameters converged as indicated by both the estimated PSRF and Geweke's Z-score diagnostic. The nonconverged estimations occurred only when all possible interactions were estimated. With a sample of $N = 500$, overall convergence rates were 0.92, 0.98, and 0.95 for the three levels of interactions generated, I_{gen} = 0, 16, and 32, respectively. In the $N = 1{,}000$ sample size condition with all interactions estimated, overall convergence was 0.99 for all levels of interactions generated.

The proportion of converged parameters according to Geweke's Z-score diagnostic by parameter type is presented in Table 4.6. As in the 2-category simulation, intercept parameters (β_{i1} through β_{i5}) had the lowest convergence rates for all data-generating and model estimation conditions, particularly for the highest and lowest categories. When no interactions were generated, intercepts converged at a higher rate when all interactions were estimated compared to none in the larger sample size of $N = 1{,}000$, but otherwise estimating the matching data-generating model with no interactions had a higher convergence rate for all parameters. In conditions when there were either 8 or 16 interactions generated, convergence rates for intercepts tended to be higher when the model estimated matched the model generated, except in the smaller sample size of $N = 500$ and all interactions were estimated. However, these comparisons were in the hundredth decimal place and likely not significantly different.

The rates of convergence for general and specific respondent parameters were comparable in the different conditions, and the general dimension parameters tended to have a higher rate of convergence than the specific dimension parameters. Interaction parameters had a higher rate of convergence when the model estimated matched the data-generating model, particularly for the sample size of $N = 1{,}000$. The 18 non-converged solutions were excluded in the following analyses.

Table 4.6: Graded response average proportion of converged parameters according to Geweke's Z-score diagnostic under different conditions

			Estimated Parameter									
N	I_{gen}	I_{est}	β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}	β_{ig}	β_{is}	$\beta_{i(g*s)}$	θ_{rg}	θ_{rs}
500	0	All	0.73	0.74	0.75	0.74	0.73	0.87	0.92	0.90	0.90	0.94
		None	0.74	0.75	0.76	0.76	0.75	0.90	0.94	*	0.93	0.96
	8	All	0.78	0.78	0.79	0.80	0.78	0.90	0.93	0.92	0.92	0.95
		Match	0.79	0.81	0.81	0.81	0.80	0.90	0.94	0.94	0.94	0.96
		None	0.77	0.77	0.78	0.76	0.76	0.92	0.93	*	0.94	0.96
	16	All	0.77	0.78	0.78	0.78	0.76	0.91	0.92	0.91	0.91	0.95
		Match	0.76	0.77	0.78	0.78	0.77	0.88	0.92	0.93	0.93	0.96
		None	0.76	0.77	0.78	0.77	0.76	0.91	0.94	*	0.94	0.96
1000	0	All	0.71	0.73	0.73	0.72	0.70	0.89	0.93	0.93	0.93	0.96
		None	0.69	0.70	0.71	0.70	0.69	0.91	0.94	*	0.94	0.96
	8	All	0.71	0.72	0.73	0.73	0.72	0.90	0.92	0.92	0.93	0.96
		Match	0.75	0.76	0.77	0.76	0.75	0.90	0.94	0.94	0.94	0.96
		None	0.69	0.71	0.71	0.70	0.69	0.89	0.94	*	0.94	0.96
	16	All	0.73	0.75	0.75	0.75	0.73	0.91	0.91	0.93	0.94	0.96
		Match	0.74	0.76	0.77	0.76	0.73	0.90	0.92	0.94	0.94	0.96
		None	0.73	0.74	0.75	0.74	0.73	0.92	0.94	*	0.94	0.97

* Non-estimated parameter

4.3.2 Parameter Recovery

Parameter recovery of the 6-category simulation is presented in Figures 4.6 through 4.11 with Monte Carlo error in Tables 4.7 and 4.8 for item and respondent parameter bias, absolute bias, and RMSE. MCEs of bias, absolute bias, and RMSE were small, and largest for the general main effects and intercept parameters. As in the 2-category simulation, models estimated with matched interactions are nested in the no estimated interactions condition when data are generated with only main effects.

Figure 4.6 shows the bias, absolute bias, and RMSE of the general and specific main effects. The pattern of results was similar to the 2-category simulation. Considering first the general main effects, bias and absolute bias was smallest in data generated with no interactions, a larger sample size of $N = 1,000$, and the model estimated had no interactions to match the data-generating model. The recovery of this baseline bifactor model without interactions generated or estimated had only a small negative bias when $N = 500$. When these data were fit with a bifactor model having all possible interactions, the bias was positive and only slightly larger in magnitude. However, the absolute bias and RMSE when only main effects were generated for a sample size of $N = 1,000$ was about the same for the models estimated with matched (no) interactions and all interactions estimated. There was minor fluctuation of RMSE for the different estimated models within a data-generating condition. RMSE was largest when interactions were generated but not estimated, and smallest when all interactions were estimated. Bias and absolute bias was smallest when all interactions were estimated when a proportion of items had generated interactions. A similar finding occurred with the 2-category simulation where the data-generating model with interactions had a larger bias and absolute bias than the model with all possible interactions. However, the differences in parameter recovery between the models estimated for the different data-generating conditions were minor, and overall general main effect parameters were recovered well.

Table 4.7: Graded response Monte Carlo error (MCE) of bias, absolute bias, and root mean square error (RMSE) of respondent parameters under different conditions

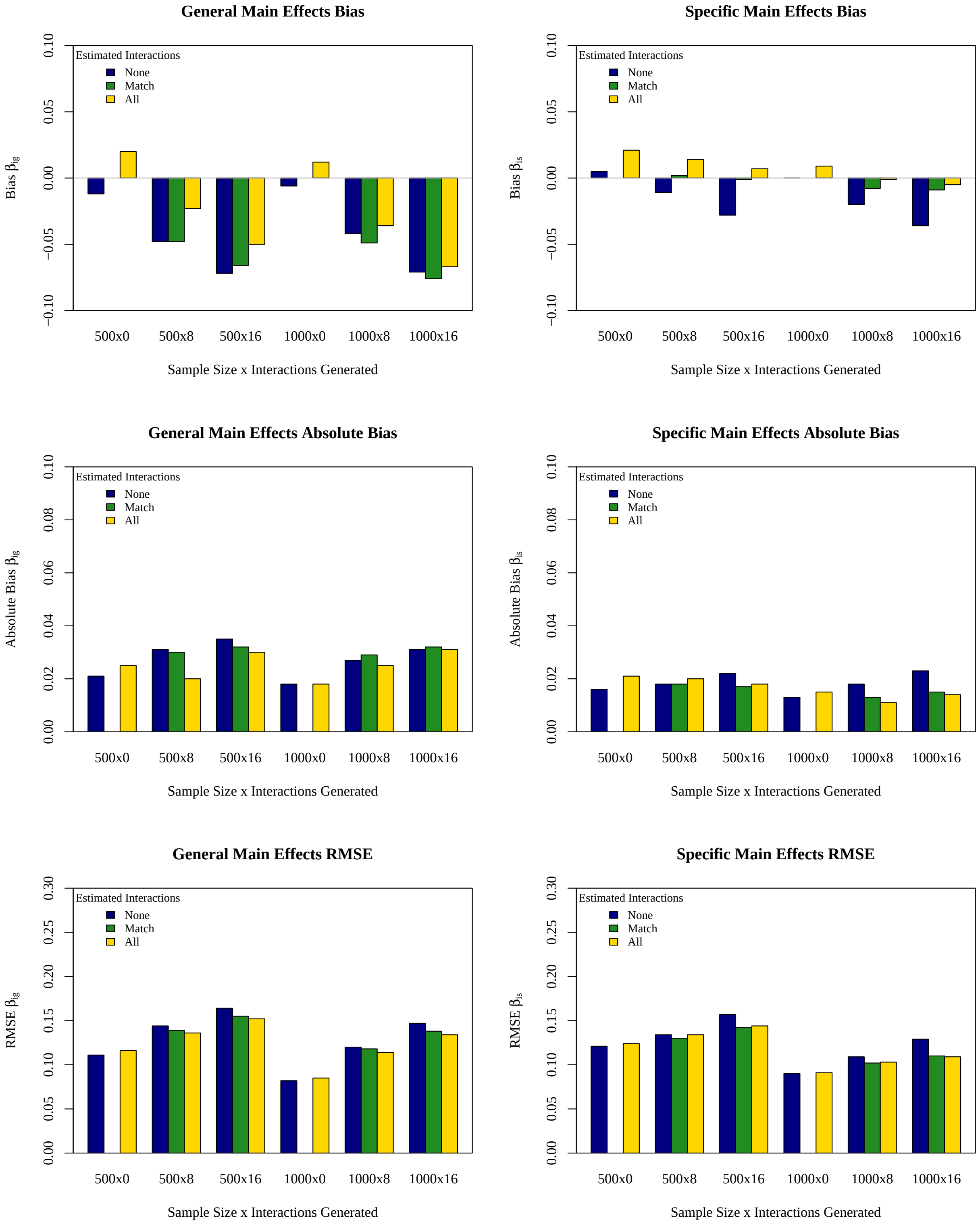
			Bias MCE		Absolute Bias MCE		RMSE MCE	
N	I_{gen}	I_{est}	θ_{rg}	θ_{rs}	θ_{rg}	θ_{rs}	θ_{rg}	θ_{rs}
500	0	All	0.032	0.023	0.019	0.013	0.029	0.030
		None	0.033	0.023	0.021	0.013	0.028	0.030
	8	All	0.035	0.022	0.021	0.013	0.030	0.030
		Match	0.033	0.023	0.020	0.013	0.030	0.030
		None	0.033	0.023	0.020	0.013	0.031	0.030
	16	All	0.035	0.022	0.021	0.013	0.033	0.030
		Match	0.035	0.022	0.021	0.013	0.035	0.029
		None	0.033	0.023	0.020	0.013	0.036	0.031
1000	0	All	0.030	0.015	0.018	0.009	0.023	0.027
		None	0.030	0.015	0.017	0.009	0.022	0.027
	8	All	0.029	0.015	0.017	0.009	0.022	0.026
		Match	0.030	0.015	0.018	0.009	0.022	0.026
		None	0.029	0.015	0.017	0.009	0.022	0.027
	16	All	0.030	0.015	0.018	0.009	0.024	0.026
		Match	0.030	0.015	0.017	0.009	0.024	0.027
		None	0.029	0.015	0.017	0.009	0.026	0.028

Table 4.8: Graded response Monte Carlo error (MCE) of bias, absolute bias, and root mean square error (RMSE) of item parameters under different conditions

			Bias MCE				Absolute Bias MCE				RMSE MCE			
N	I_{gen}	I_{est}	β_{ik}	β_{ig}	β_{is}	$\beta_{i(g*s)}$	β_{ik}	β_{ig}	β_{is}	$\beta_{i(g*s)}$	β_{ik}	β_{ig}	β_{is}	$\beta_{i(g*s)}$
500	0	All	0.038	0.037	0.026	0.038	0.025	0.025	0.021	0.024	0.016	0.021	0.022	0.022
		None	0.037	0.036	0.025	*	0.024	0.021	0.016	*	0.016	0.020	0.021	*
	8	All	0.037	0.034	0.028	0.043	0.023	0.020	0.020	0.026	0.015	0.023	0.020	0.03
		Match	0.036	0.035	0.028	0.057	0.022	0.030	0.018	0.039	0.014	0.024	0.020	0.056
		None	0.037	0.037	0.027	*	0.022	0.031	0.018	*	0.014	0.025	0.021	*
	16	All	0.036	0.036	0.030	0.045	0.023	0.030	0.018	0.027	0.014	0.027	0.022	0.032
		Match	0.035	0.035	0.029	0.046	0.022	0.032	0.017	0.029	0.014	0.027	0.023	0.041
		None	0.036	0.038	0.030	*	0.023	0.035	0.022	*	0.015	0.029	0.027	*
1000	0	All	0.035	0.028	0.021	0.023	0.021	0.018	0.015	0.015	0.015	0.016	0.017	0.015
		None	0.034	0.028	0.020	*	0.021	0.018	0.013	*	0.014	0.014	0.016	*
	8	All	0.034	0.031	0.021	0.033	0.019	0.025	0.011	0.021	0.012	0.024	0.020	0.023
		Match	0.034	0.032	0.021	0.038	0.020	0.029	0.013	0.024	0.013	0.026	0.021	0.033
		None	0.034	0.031	0.023	*	0.019	0.027	0.018	*	0.012	0.028	0.024	*
	16	All	0.030	0.032	0.023	0.039	0.018	0.031	0.014	0.024	0.012	0.025	0.021	0.024
		Match	0.032	0.032	0.023	0.035	0.019	0.032	0.015	0.022	0.012	0.026	0.021	0.031
		None	0.033	0.032	0.025	*	0.019	0.031	0.023	*	0.012	0.028	0.026	*

* Non-estimated parameter

Figure 4.6: Graded response bias, absolute bias, and root mean squared error (RMSE) for general and specific main effects



Specific main effects were recovered well, with less overall bias and absolute bias compared to the general main effects as displayed in Figure 4.6. Bias and absolute bias was largest when a bifactor model having only main effects was fit to data with generated interactions. In the larger sample size of $N = 1,000$, the model estimated with all interactions when interactions were generated had the least amount of bias and absolute bias. Comparatively, in the smaller sample size condition, bias and absolute bias of the model fit with all interactions were larger when data were generated with interactions compared to the model with matched interactions. The direction of bias when all interactions were estimated was positive in the $N = 500$ condition, indicating that estimated values were overestimated, compared to underestimated as observed for the general main effects. However, the overall bias was small and absolute bias was less than or approximately equal to 0.02 for the specific main effects. RMSE decreased in the larger sample size data and was largest when the model estimated only included main effects and omitted generated interactions. The models estimated with matched interactions and all interactions when interactions were simulated had approximately the same RMSE in the different sample size conditions.

Recovery of interaction parameters according to bias, absolute bias, and RMSE is presented in Figure 4.7. Bias was generally negative, with the exception of data generated with 16 items (50%) having interactions and $N = 500$, where bias was positive when data were estimated with matched interactions. Bias and absolute bias tended to be smaller when the model was estimated with all interactions compared to the data-generated matched interactions, though RMSE for these models estimated were similar within a data-generating condition. Thus, in data with more respondents and interactions generated, the data-generating model performed best at recovering the interactions but otherwise fitting all interactions had the least bias and absolute bias. Compared to general and specific main effects, interactions had less bias, but similar absolute bias and RMSE.

The bias and absolute bias of the 5-category intercepts are displayed in Figures 4.8 and 4.9 with RMSE in Figure 4.10. Overall, bias and absolute bias was small, and was smallest in the middle category 3 intercept, and largest in the extreme category intercepts, 1 and 5. When there were no interactions generated, the lower categories had a negative bias which became positive with increasing categories up to category 5. The opposite pattern occurred when there were 16 items with simulated interactions; bias was positive for category 1 and became less positive to negative with increasing categories. However, the bias and absolute bias of the intercepts were small and differences between the study conditions were minor. RMSE for the category intercepts was small overall and decreased with a larger sample size.

To understand the recovery of general and specific latent dimensions for respondents, bias, absolute bias, and RMSE are displayed in Figure 4.11. Bias and absolute bias was minimal for all study conditions for both the general and specific dimensions, and there was little fluctuation among estimated models within a data-generating condition. Bias was lowest for both general and specific dimensions in the larger sample size conditions, though the decrease was minimal. When the model estimated omitted interactions when generated in the data, RMSE was only slightly worse. RMSE was largest for the specific dimension across all study conditions (Mean = 0.496) compared to the general dimension (Mean = 0.328). This was expected given the larger number of items measuring the general dimension compared to the specific dimensions. Correlations between true generating parameters and estimated parameters were between 0.94-0.95 for the general dimension and 0.86-0.87 for the specific dimension for all study conditions.

Figure 4.7: Graded response bias, absolute bias, and root mean squared error (RMSE) for interaction effects

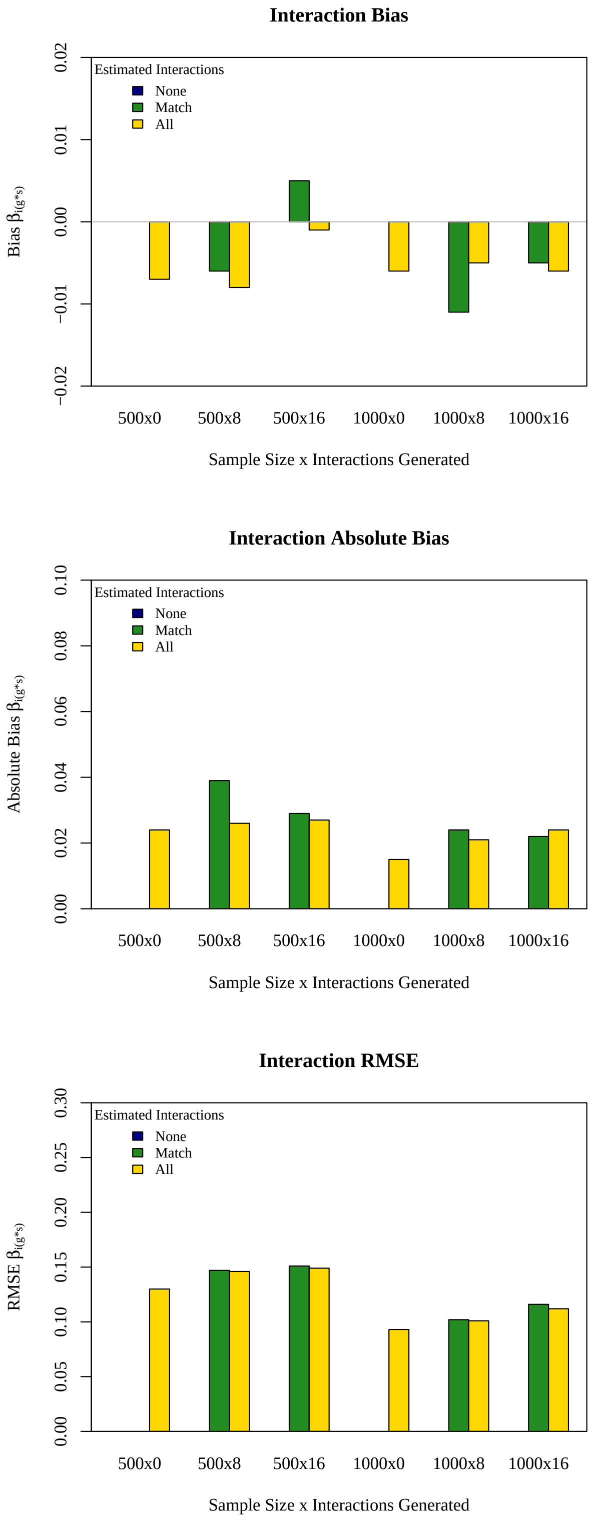
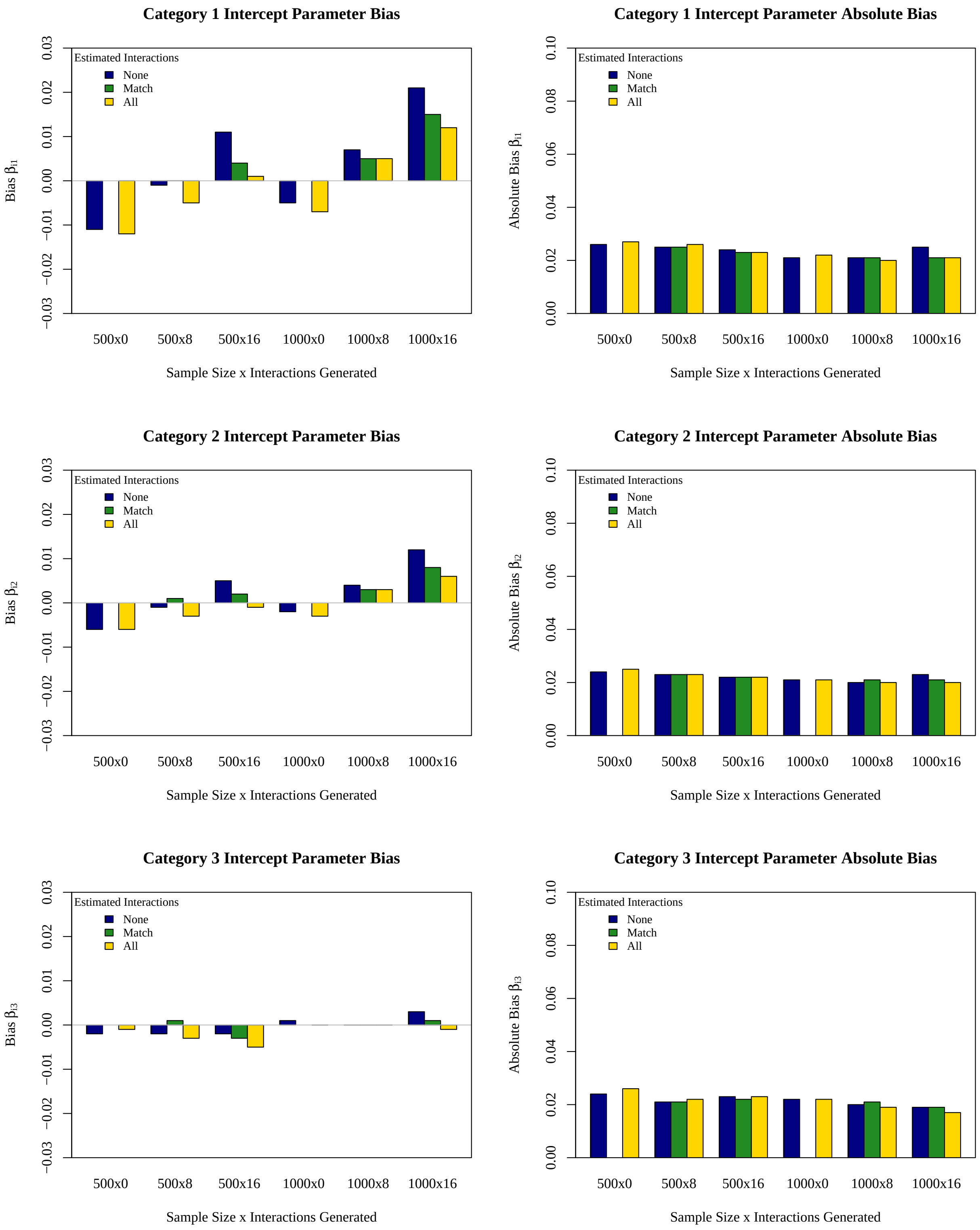


Figure 4.8: Graded response bias and absolute bias for categories 1 through 3 intercept parameters



Category 1 Intercept Parameter Bias

Category 1 Intercept Parameter Absolute Bias

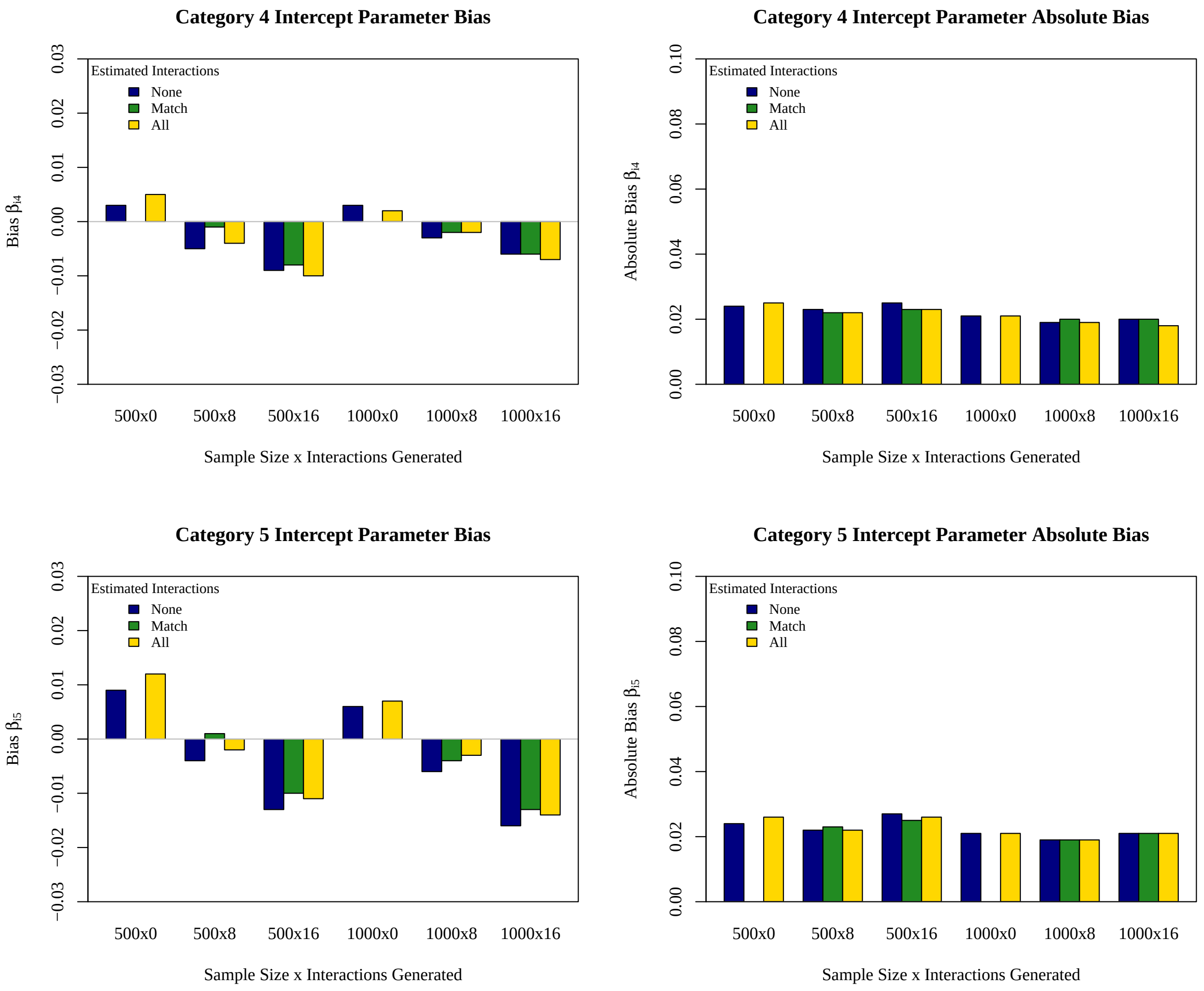
Category 2 Intercept Parameter Bias

Category 2 Intercept Parameter Absolute Bias

Category 3 Intercept Parameter Bias

Category 3 Intercept Parameter Absolute Bias

Figure 4.9: Graded response bias and absolute bias for categories 4 and 5 intercept parameters



Category 4 Intercept Parameter Bias

Category 4 Intercept Parameter Absolute Bias

Category 5 Intercept Parameter Bias

Category 5 Intercept Parameter Absolute Bias

Figure 4.10: Graded response root mean squared errors (RMSE) for item category intercept parameters

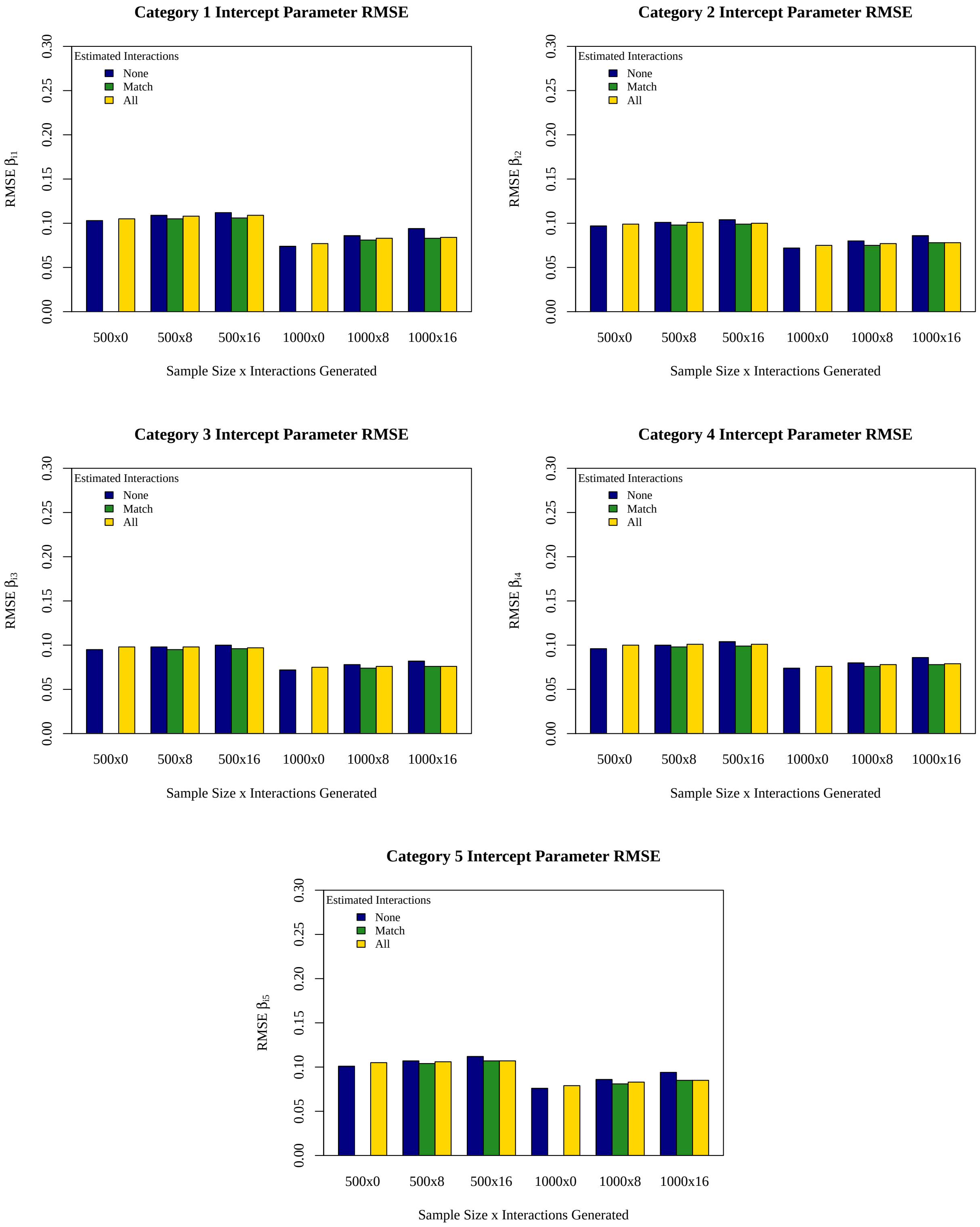
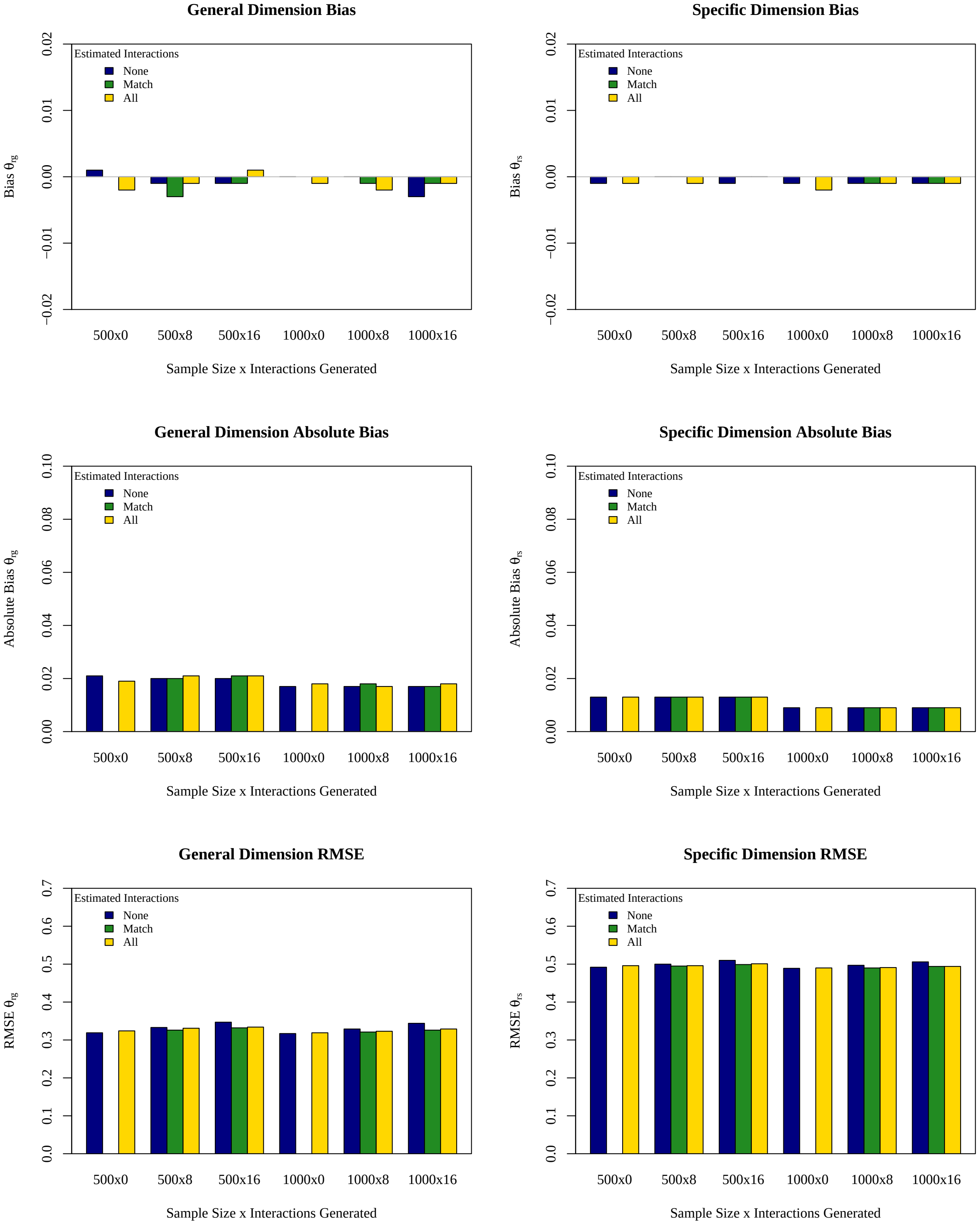


Figure 4.11: Graded response bias, absolute bias, and root mean squared error (RMSE) for general and specific dimension respondent parameters


General Dimension Bias

Specific Dimension Bias

General Dimension Absolute Bias

Specific Dimension Absolute Bias

General Dimension RMSE

Specific Dimension RMSE

4.3.3 Model Selection

For some or all iterations of the Markov chains for estimations in the 6-category simulation, the penalty parameter, *pD*, was estimated to be infinity. The majority of estimations had at least one occurrence of an infinite *pD* for an iteration (across both chains), and 42 of the 1,582 estimations (after 18 were removed due to nonconvergence) had an infinite *pD* for all iterations of the two chains. An infinite *pD* can occur because of how it is calculated as shown in Equation 3.5 and again defined as

$$log\left\{\frac{p(Y_i^0|\theta^0)}{p(Y_i^0|\theta^1)}\right\} \tag{4.1}$$

for an iteration, *i*. In the context of MCMC sampling defined by Plummer (2002), the Kullback-Leibler (KL) information divergence is between the two parallel chains (superscripted as 0 and 1). The KL can be estimated as infinity if the denominator of the function, $p(Y_i^0|\theta^1)$, is zero and the numerator, $p(Y_i^0|\theta^0)$, is greater than zero. Given this, another variant of *pD* was also compared and is calculated as the posterior mean deviance minus the deviance of the posterior means (Spiegelhalter et al., 2002):

$$pD = E_{\theta|\mathbf{Y}}[D] - D(E_{\theta|\mathbf{Y}}[\theta]) = \bar{D} - D(\bar{\theta}). \tag{4.2}$$

KL-based *pD* was calculated as the mean *pD* for an estimation with infinite iterations removed, and the 42 estimations with infinite *pD* were excluded in model comparisons. If a replication within a data-generating condition had missing estimations due to nonconvergence, they were also excluded. Thus, totals displayed in the following tables will not sum to tne number of estimations performed in the corresponding condition.

Table 4.9: Graded response model selection frequencies according to KL-based penalty parameter, *pD*

		Estimated Interactions, I_{est}			
N	I_{gen}	None	Match	All	Total
500	0	*	88*	2	90
	8	11	84	2	97
	16	3	78	9	90
1000	0	*	81*	8	89
	8	2	79	12	93
	16	0	65	27	92
Total		16	475	60	551

* Indicates a nested condition

Table 4.10: Graded response model selection frequencies according to variant penalty parameter, *pD*

		Estimated Interactions, I_{est}			
N	I_{gen}	None	Match	All	Total
500	0	*	16*	76	92
	8	0	25	73	98
	16	0	22	73	95
1000	0	*	18*	81	99
	8	0	25	74	99
	16	0	27	71	98
Total		0	133	448	581

* Indicates a nested condition

Model selection frequencies according to these different estimates of *pD* are displayed in Tables 4.9 and 4.10. As shown, the KL-based *pD* tended to perform better than the variant *pD* in selecting the true data-generating model, with 475 of the 551 estimations (86%) selected as the correct data-generating model. The variant *pD* tended to favor the more complex model, suggesting that the penalty for the effective number of parameters was not large enough to identify the true model that has fewer parameters. The inaccuracy of the KL-based *pD* results compared to the 2-category simulation are likely due to the small number of finite *pD* used to estimate the sample average of *pD*. In fact, some estimations had at most one iteration with a finite *pD* estimated.

Though not ideal, DIC comparisons based on the KL-based *pD* may at the very least identify the correct data-generating model with accuracy ranging between 70.65% to 97.78%, depending on the sample size and underlying interactions. As in the 2-category simulation, DIC misidentifications of the true model in the larger sample size condition of $N = 1000$ tended toward the model estimated with all interactions compared to no interactions, particularly when more underlying interactions were present. In the larger sample with interactions generated, the main effects only bifactor model was rarely selected over the data-generating model or the model with all estimated interactions. However, I emphasize that these results should be interpreted with caution given the unreliability in estimating pD as noted previously.

4.3.4 Model Reduction

The rate of misidentified interactions according to the highest density intervals (HDI) and credible intervals (CI), as defined by the proportion of estimated interaction parameters that did not contain their true data-generating parameter within the interval, are presented in Table 4.11 for models estimated with interactions that matched the data-generating model. This baseline model had higher misidentification of interactions in the larger sample size of $N = 1,000$ compared to $N = 500$. However, like the 2-category simulation, MCE of the average proportion over simulated replications was high, though it is unlikely that increasing the accuracy of these results with more replications will provide a more favorable inference. The rate of misidentified interactions was large, even when the model estimated matched the data-generating model, suggesting that an alternate strategy of model parameter refinement should be used instead.

Table 4.11: Graded response average proportions (and Monte Carlo error) of true parameters outside of highest density intervals and credible intervals in models estimated with matched interactions

N	I_{gen}	HDI	CI
500	8	0.118 (0.135)	0.114 (0.143)
	16	0.108 (0.099)	0.111 (0.103)
1000	8	0.121 (0.125)	0.121 (0.127)
	16	0.144 (0.101)	0.144 (0.102)

Figures 4.12 and 4.13 display the inaccuracy of the HDI and CI in correctly identifying the true interaction parameter when models are estimated with interactions for all items. The differences

in HDI and CI are nearly indistinguishable. In the smaller sample size condition of $N = 500$, there does not appear to be a large difference in the proportion of correctly identified interaction items when the data is generated with either 8 or 16 items having interactions, similar to the 2-category simulation. Comparatively, the proportion of misidentified interacting items in data with $N = 1,000$ respondents increases with more underlying interaction effects. The accuracy in identifying non-interacting items as not having an interaction was better for all the data-generating conditions at a rate of around 5% of items not containing their true value within their HDI, with a slight increase with more items having interactions generated in the data.

Figure 4.12: Graded response proportion of items with true parameters outside of highest density intervals in models estimated with all possible interactions

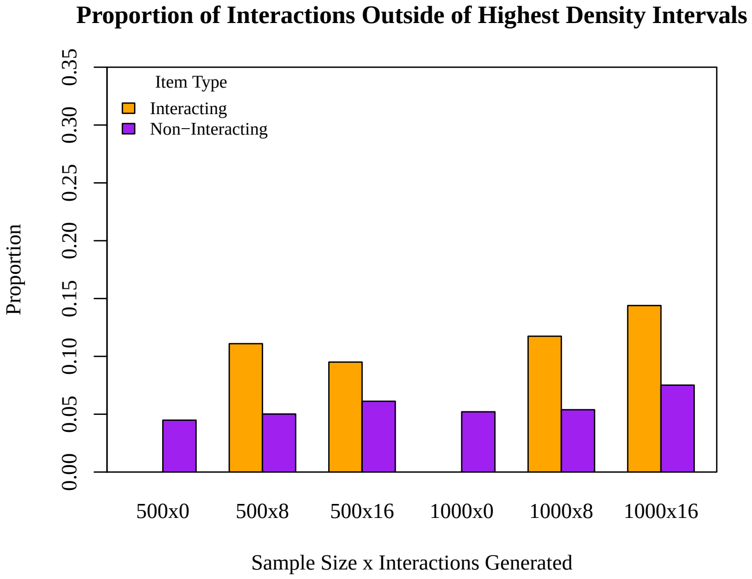
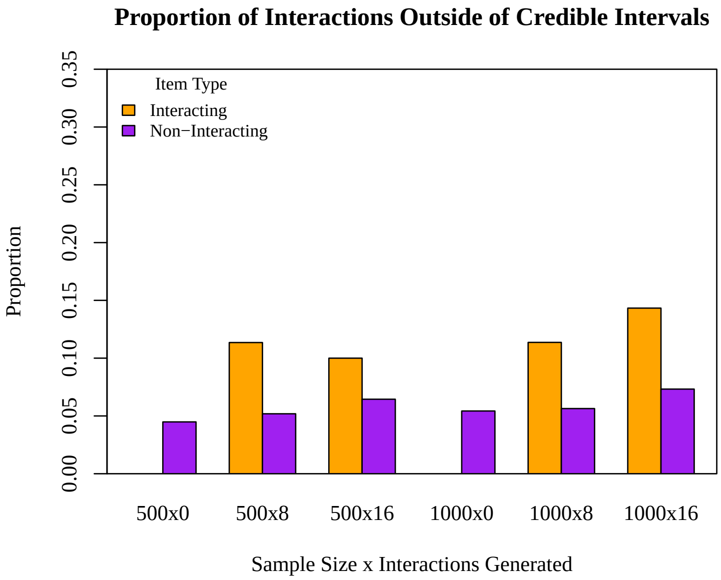


Figure 4.13: Graded response proportion of items with true parameters outside of credible intervals in models estimated with all possible interactions



Chapter 5

Empirical Applications

In simulation studies, data are generated to match the model under study and true parameters are known. Comparatively, empirical data will generally not fit as well because the true underlying model is not known. Thus, this chapter presents an example of the bifactor model using two empirical data sets: one for dichotomous data in educational testing, and one for graded response data in psychology. The aim was to evaluate the rate in which different model parameters converge, assess the prevalence of interactions, show how model parameters compare among different estimated models, and demonstrate how interpretations of interactions can be made in real-world constructs. The bifactor model with interactions and the competing main effects only bifactor model were used to test the underlying structure of the measures and illustrate confirmatory model building procedures and parameter refinement techniques. Model refinement and comparison methods that were assessed in the simulation study were used to identify interactions when the true parameters are not known. Specifically, the bifactor model with interactions and traditional main effects only bifactor model were compared to determine whether building a model with all interaction effects and refining individual item parameters using highest density intervals results in the selection of a reduced model compared to a simpler model that omit interaction effects. To this end, each data set was fitted with three variations of the bifactor model:

1. A main effects only bifactor model that omits interaction effects,
2. A bifactor model with all interactions estimated, and
3. A reduced bifactor model with interactions estimated for items that did not contain zero within their highest density intervals from the model estimated with all interactions.

5.1 Application 1: Mathematics Ability

The 2011 cycle of the Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2012) math test for 8th grade students was selected to illustrate the bifactor model with interactions in dichotomously scored data. The data selected for analysis was taken from students in the top 10 highest-scoring countries completing Booklet 8, resulting in a sample size of 3,720 students. Sample sizes by country are presented in Table 5.1. Students having unique identifiers appearing in multiple countries were removed (N = 26). For a complete description of the sample characteristics and data collection process, see the original report published by TIMSS in Mullis et al. (2012).

Table 5.1: TIMSS sample sizes by country

Rank	Country	N
1	Republic of Korea	369
2	Singapore	421
3	Chinese Taipei	353
4	Hong Kong SAR	280
5	Japan	299
6	Russian Federation	357
7	Israel	324
8	Finland	305
9	United States	740
10	England	272

The TIMSS math assessment has 34 items that measure one of four content domains: Number ($I = 11$), Data and Chance ($I = 5$), Algebra ($I = 13$), and Geometry ($I = 5$). There were 15 multiple choice items that were scored correct and incorrect and 19 constructed response items that were scored with up to 2 points; for the purpose of this study, responses to constructed response items were recoded from partial credit scoring to either correct (score equal to 1 or 2 becomes a score of 1) or incorrect (score equals 0) so that a dichotomous item response bifactor model could be used. Though the constructed response items could be modeled with a partial credit bifactor model having interactions, this illustration is restricted to only dichotomous ordered item responses.

5.1.1 Model Estimation

The same MCMC specifications and prior values of the simulation study were used to estimate the model. Estimation was completed with 7,000 iterations per 2 chains with 3,500 burn-in iterations and a thinning interval of 3. Priors were standard normal ($\sim N(0,1)$) for item parameters, which was truncated at zero for main effects. Latent dimensions had multivariate standard normal priors, $\theta_{r.} \sim MVN(\mathbf{0}, \Sigma)$, with identity covariance matrix. The JAGS model code for dichtomously scored items as specified in Appendix A was used.

Convergence assessment was completed using the estimated potential scale reduction factor (PSRF; Gelman & Rubin, 1992; Brooks, S. & Gelman, 1998), and Geweke's Z-score diagnostic (Geweke, 1992). The proportion of converged parameters for the three estimated models per each criterion is presented by parameter type and overall in Table 5.2. Similar to the convergence rates by parameter type observed in the 2-category simulation study, the TIMSS bifactor models had poor convergence of intercept parameters according to Geweke's Z-score diagnostic. Overall, the bifactor model estimated with all interactions had the lowest convergence rates for both the estimated PSRF and Geweke's Z-score diagnostic. To address the poor convergence according to Geweke's Z-score, a more diffuse prior could be specified, especially because true parameters are not known. However, a more diffuse prior would require more iterations of the Markov chains to reach convergence, thereby requiring more time to estimate the model. Thus, practitioners should carefully consider the choice of prior distribution and number of MCMC iterations in preliminary testing of their specified model.

Table 5.2: Proportion of converged parameters for TIMSS data analyzed with different estimated interactions

	Estimated Interactions, I_{est}					
	None		All		Reduced	
Parameter	PSRF	Geweke	PSRF	Geweke	PSRF	Geweke
Intercepts	0.941	0.647	0.882	0.235	0.971	0.912
Main Effects						
General	0.941	0.765	0.912	0.765	1.000	0.882
Specific	0.941	0.794	0.765	0.676	0.971	0.706
Interaction Effects	*	*	0.735	0.647	1.000	0.857
Latent Variables						
General	1.000	0.966	0.994	0.878	1.000	0.967
Specific	0.999	0.902	0.854	0.710	1.000	0.936
Total	0.999	0.914	0.881	0.742	0.999	0.941

*Non-estimated parameter

5.1.2 Model Selection

Differences in DIC between the bifactor models that have estimated interactions and the traditional bifactor model with only main effects estimated were compared to see if adding interaction effects and reducing to only those interactions that do not contain zero within their highest density intervals (HDI) results in a significant improvement in model fit. The bifactor model with all interactions between the general and specific dimensions estimated was favored for having a smaller DIC value of 116,607.06 ($\bar{D}$ = 105,827.2, pD = 10,779.82) compared to the DIC of the main effects only model, 116,836.37 ($\bar{D}$ = 106,639.7, pD = 10,196.68). As displayed in Table 5.4, 58% of the interactions estimated contained zero within their HDI. These parameters were then removed and the dataset was fit with reduced interactions. The DIC for this reduced model was 116,265.9 ($\bar{D}$ = 105,455.9, pD = 10,810.03), which is smaller than the DIC of the models estimated with all interactions or no interactions, thereby making it the preferred model.

5.1.3 Parameter Estimates

Because true parameters are not known in empirical data, parameter estimates from the different models estimated were compared to each other rather than to true parameters. Posterior means of the main effect and interaction estimates and their corresponding standard deviations for the three estimated models, no interactions, all possible interactions, and reduced interactions, are presented in Tables 5.3, 5.4, and 5.5, respectively. Intercept mean estimates for all models are presented in Table 5.6. Generally across the different bifactor models estimated, general and specific main effects and intercept parameter estimates were comparable with some fluctuation. All main effects for all the bifactor models estimated did not contain zero within their highest density intervals, supporting the hypothesized bifactor pattern matrix that defines the relationships between the items and the latent dimensions.

The general dimension main effects in the bifactor models estimated with all interactions and no interactions were similar, though parameter estimates tended to be closer to zero in the main effects only bifactor model for all but 8 of the 34 items. Considering the results of the simulation study, the main effects only bifactor model main effect parameters tended to be downwardly biased when more underlying interactions were present but unmodeled in the data. Interaction effects in the reduced bifactor model were predominantly for items measuring the number dimension (4/14) and algebra dimension (9/14), with one item of the data and chance dimension having an interaction. The direction of the main effects were positive for 8 interactions and negative for 6 of the 14 interactions. Of the reduced interactions, the majority of items were constructed response (10/14) rather than multiple choice.

Two example illustrations of the probability of correct responses for items 5 and 16 of the TIMSS assessment are displayed in Figures 5.1 and 5.2. Item 5 was estimated to have a negative interaction, such that the rate of increase in probability of answering the item correctly for a respondent with average general math ability is lower with high (+1 standard deviation) levels of the Number dimension. However, with more general math ability, the difference in the rate of increase in the probability is less noticeable because the probability with slightly above average ability has approached the ceiling of a one probability. The second example item on the other hand has a much different interpretation. The three predicted curves of the specific dimensions at plus and minus one standard deviation around the mean cross at below average levels of general math ability. The rate of change in the probability of a correct response becomes less positive with lower Data and Chance ability. However, at lower general math ability levels, the probability of getting the item correct is actually higher for a respondent with low versus high data and chance ability. This conflicts with inferences about cognitive ability that is typically modeled such that the probability of correctly responding is monotonically increasing with the latent ability.

Figure 5.1: Probability of a one response for item 5 of the TIMSS assessment modeled with a bifactor model having reduced interactions

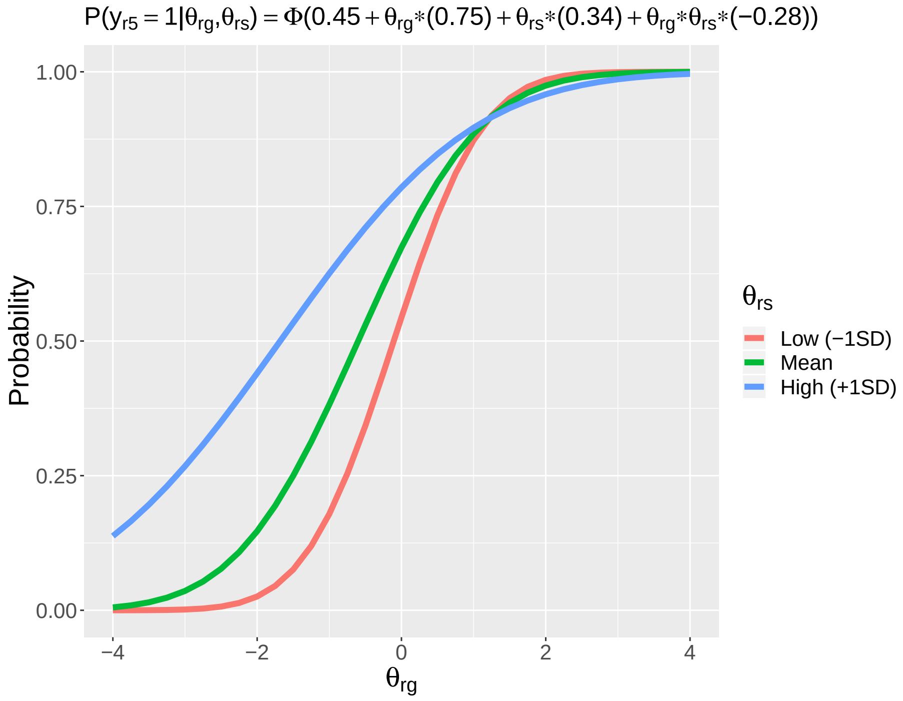
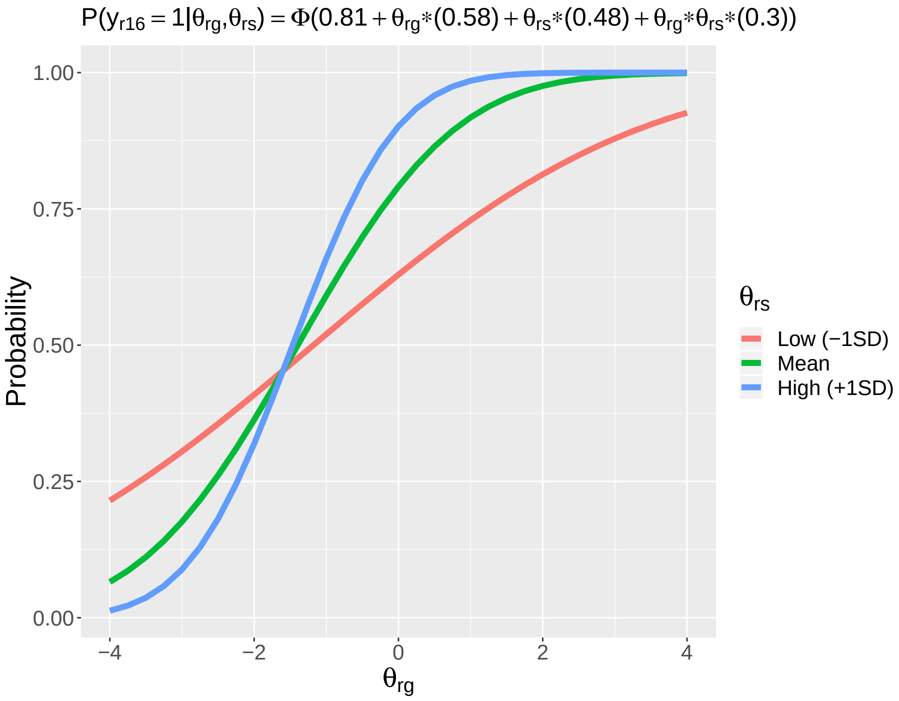


Figure 5.2: Probability of a one response for item 16 of the TIMSS assessment modeled with a bifactor model having reduced interactions



$P(y_{r16} = 1|\theta_{rg},\theta_{rs}) = \Phi(0.81 + \theta_{rg}*(0.58) + \theta_{rs}*(0.48) + \theta_{rg}*\theta_{rs}*(0.3))$

Table 5.3: TIMSS main effect parameter means (and standard deviations) estimated from a bifactor model with no estimated interactions

Item	Reference	General	Number	Algebra	Geometry	Data & Chance
1	M052413	0.69 (0.03)*	0.02 (0.02)*			
2	M052134	0.81 (0.04)*	0.47 (0.06)*			
3	M052078	0.56 (0.03)*	0.34 (0.05)*			
4	M052034	0.70 (0.03)*	0.27 (0.05)*			
5	M052174A	0.78 (0.04)*	0.59 (0.09)*			
6	M052174B	0.82 (0.04)*	0.49 (0.08)*			
7	M052130	0.71 (0.03)*		0.24 (0.04)*		
8	M052073	1.01 (0.04)*		0.20 (0.04)*		
9	M052110	1.33 (0.05)*		0.06 (0.03)*		
10	M052105	0.95 (0.04)*		0.02 (0.01)*		
11	M052407	0.90 (0.14)*			0.73 (0.32)*	
12	M052036	0.72 (0.03)*			0.14 (0.07)*	
13	M052502	0.90 (0.04)*				0.25 (0.08)*
14	M052117	0.40 (0.03)*				0.07 (0.05)*
15	M052426	0.67 (0.04)*				0.33 (0.09)*
16	M042183	0.46 (0.03)*	0.19 (0.05)*			
17	M042060	0.89 (0.04)*	0.45 (0.06)*			
18	M042019	0.50 (0.03)*	0.35 (0.05)*			
19	M042023	1.25 (0.04)*	0.07 (0.04)*			
20	M042197	0.83 (0.03)*	0.05 (0.03)*			
21	M042234	0.92 (0.04)*		0.15 (0.04)*		
22	M042066	0.66 (0.03)*		0.03 (0.02)*		
23	M042243	1.22 (0.05)*		0.10 (0.04)*		
24	M042248	1.33 (0.05)*		0.11 (0.04)*		
25	M042229A	7.42 (0.45)*		5.85 (0.35)*		
26	M042229B	1.88 (0.08)*		0.83 (0.05)*		
27	M042229Z	7.45 (0.45)*		5.88 (0.35)*		
28	M042080A	1.03 (0.04)*		0.18 (0.04)*		
29	M042080B	1.39 (0.06)*		0.25 (0.04)*		
30	M042120	0.90 (0.04)*			0.04 (0.03)*	
31	M042203	1.17 (0.28)*			0.67 (0.46)*	
32	M042264	0.79 (0.03)*			0.03 (0.03)*	
33	M042255	0.58 (0.03)*				0.42 (0.10)*
34	M042224	0.81 (0.05)*				0.51 (0.12)*

*Does not contain zero within its Highest Density Interval

Table 5.4: TIMSS main effect and interaction parameter means (and standard deviations) estimated from a bifactor model with all possible estimated interactions

Item	General	Number	Algebra	Geometry	Data & Chance	Interaction
1	0.71 (0.04)*	0.06 (0.04)*				0.12 (0.06)*
2	0.81 (0.04)*	0.39 (0.06)*				-0.09 (0.08)
3	0.59 (0.03)*	0.42 (0.05)*				-0.12 (0.07)
4	0.72 (0.03)*	0.31 (0.05)*				0.02 (0.06)
5	0.76 (0.03)*	0.33 (0.05)*				-0.26 (0.08)*
6	0.91 (0.05)*	0.28 (0.06)*				-0.42 (0.08)*
7	0.70 (0.03)*		0.25 (0.04)*			0.05 (0.07)
8	1.01 (0.04)*		0.23 (0.04)*			-0.05 (0.07)
9	1.37 (0.06)*		0.09 (0.04)*			-0.20 (0.08)*
10	0.95 (0.04)*		0.02 (0.02)*			0.06 (0.06)
11	0.80 (0.05)*			0.30 (0.21)*		0.11 (0.11)
12	0.77 (0.05)*			0.29 (0.15)*		0.17 (0.10)
13	1.06 (0.16)*				0.23 (0.20)*	0.15 (0.52)
14	0.40 (0.03)*				0.05 (0.04)*	-0.01 (0.09)
15	0.66 (0.05)*				0.08 (0.07)*	0.02 (0.38)
16	0.61 (0.05)*	0.50 (0.07)*				0.35 (0.08)*
17	0.91 (0.04)*	0.46 (0.06)*				-0.04 (0.07)
18	0.55 (0.03)*	0.44 (0.05)*				0.11 (0.06)
19	1.26 (0.05)*	0.04 (0.03)*				-0.01 (0.07)
20	0.83 (0.03)*	0.07 (0.04)*				0.00 (0.06)
21	0.96 (0.05)*		0.13 (0.04)*			-0.25 (0.08)*
22	0.68 (0.03)*		0.03 (0.02)*			0.13 (0.05)*
23	1.28 (0.05)*		0.09 (0.04)*			-0.27 (0.08)*
24	1.37 (0.06)*		0.16 (0.04)*			-0.20 (0.09)*
25	7.32 (0.44)*		6.02 (0.37)*			0.85 (0.29)*
26	1.86 (0.08)*		0.85 (0.06)*			0.07 (0.09)
27	7.31 (0.46)*		6.01 (0.37)*			0.85 (0.28)*
28	1.19 (0.07)*		0.11 (0.04)*			0.48 (0.08)*
29	1.44 (0.07)*		0.12 (0.06)*			0.42 (0.10)*
30	0.94 (0.04)*			0.11 (0.07)*		0.03 (0.16)
31	1.20 (0.10)*			0.30 (0.23)*		-0.10 (0.68)
32	0.82 (0.04)*			0.03 (0.02)*		-0.13 (0.12)
33	0.57 (0.03)*				0.18 (0.15)*	0.11 (0.17)
34	0.93 (0.11)*				0.49 (0.36)*	0.36 (0.12)*

*Does not contain zero within its Highest Density Interval

Table 5.5: TIMSS main effect and interaction parameter means (and standard deviations) estimated from a bifactor model with reduced estimated interactions

Item	General	Number	Algebra	Geometry	Data & Chance	Interaction
1	0.72 (0.04)*	0.06 (0.04)*				0.14 (0.06)*
2	0.82 (0.04)*	0.44 (0.06)*				
3	0.58 (0.03)*	0.40 (0.05)*				
4	0.71 (0.03)*	0.32 (0.04)*				
5	0.75 (0.04)*	0.34 (0.06)*				-0.28 (0.09)*
6	0.92 (0.05)*	0.30 (0.05)*				-0.44 (0.08)*
7	0.70 (0.03)*		0.25 (0.04)*			
8	1.00 (0.04)*		0.23 (0.04)*			
9	1.37 (0.05)*		0.08 (0.04)*			-0.2 (0.08)*
10	0.94 (0.04)*		0.02 (0.02)*			
11	0.97 (0.15)*			0.91 (0.30)*		
12	0.72 (0.03)*			0.16 (0.05)*		
13	0.90 (0.04)*				0.20 (0.06)*	
14	0.40 (0.03)*				0.05 (0.04)*	
15	0.66 (0.04)*				0.19 (0.08)*	
16	0.58 (0.05)*	0.48 (0.07)*				0.30 (0.07)*
17	0.91 (0.04)*	0.48 (0.06)*				
18	0.52 (0.03)*	0.41 (0.05)*				
19	1.25 (0.04)*	0.04 (0.03)*				
20	0.83 (0.03)*	0.07 (0.04)*				
21	0.95 (0.04)*		0.13 (0.04)*			-0.24 (0.08)*
22	0.67 (0.03)*		0.03 (0.02)*			0.13 (0.05)*
23	1.27 (0.05)*		0.09 (0.04)*			-0.27 (0.07)*
24	1.37 (0.06)*		0.15 (0.04)*			-0.19 (0.08)*
25	7.33 (0.44)*		5.99 (0.37)*			0.78 (0.26)*
26	1.86 (0.08)*		0.85 (0.05)*			
27	7.32 (0.47)*		5.98 (0.38)*			0.78 (0.27)*
28	1.20 (0.07)*		0.10 (0.04)*			0.48 (0.09)*
29	1.46 (0.08)*		0.10 (0.06)*			0.45 (0.10)*
30	0.90 (0.04)*			0.04 (0.03)*		
31	1.03 (0.06)*			0.43 (0.11)*		
32	0.79 (0.03)*			0.03 (0.02)*		
33	0.57 (0.03)*				0.30 (0.07)*	
34	1.17 (0.18)*				1.14 (0.31)*	0.31 (0.10)*

*Does not contain zero within its Highest Density Interval

Table 5.6: TIMSS item intercept parameter means (and standard deviations) estimated from a bifactor model with different estimated interactions

	Estimated Interactions, I_{est}		
Item	None	All	Reduced
1	1.00 (0.03)	1.01 (0.03)	1.02 (0.03)
2	1.18 (0.04)	1.14 (0.04)	1.17 (0.04)
3	0.02 (0.02)	0.00 (0.03)	0.02 (0.03)
4	0.50 (0.03)	0.51 (0.03)	0.51 (0.03)
5	0.52 (0.03)	0.45 (0.03)	0.45 (0.03)
6	-0.42 (0.03)	-0.46 (0.04)	-0.47 (0.04)
7	0.05 (0.03)	0.05 (0.03)	0.05 (0.03)
8	0.51 (0.03)	0.51 (0.03)	0.51 (0.03)
9	0.06 (0.03)	0.06 (0.04)	0.06 (0.03)
10	-0.82 (0.03)	-0.82 (0.03)	-0.82 (0.03)
11	1.12 (0.17)	0.96 (0.06)	1.21 (0.19)
12	0.12 (0.03)	0.14 (0.03)	0.12 (0.03)
13	1.06 (0.04)	1.18 (0.14)	1.05 (0.04)
14	-1.02 (0.03)	-1.02 (0.03)	-1.02 (0.03)
15	1.40 (0.05)	1.39 (0.04)	1.36 (0.04)
16	0.68 (0.03)	0.83 (0.05)	0.81 (0.05)
17	1.00 (0.04)	1.00 (0.04)	1.01 (0.04)
18	0.19 (0.02)	0.21 (0.03)	0.19 (0.03)
19	0.20 (0.03)	0.20 (0.03)	0.20 (0.03)
20	-0.26 (0.03)	-0.25 (0.03)	-0.26 (0.03)
21	0.80 (0.03)	0.81 (0.03)	0.81 (0.03)
22	0.15 (0.03)	0.15 (0.03)	0.15 (0.02)
23	0.76 (0.04)	0.78 (0.04)	0.78 (0.04)
24	-0.16 (0.03)	-0.16 (0.04)	-0.16 (0.04)
25	-0.74 (0.20)	-0.38 (0.22)	-0.41 (0.22)
26	-0.40 (0.05)	-0.37 (0.05)	-0.39 (0.05)
27	-0.75 (0.2)	-0.39 (0.22)	-0.41 (0.22)
28	0.13 (0.03)	0.13 (0.03)	0.13 (0.03)
29	-1.03 (0.05)	-1.04 (0.05)	-1.05 (0.05)
30	0.86 (0.03)	0.88 (0.04)	0.86 (0.03)
31	1.06 (0.26)	0.93 (0.05)	0.93 (0.05)
32	-0.67 (0.03)	-0.68 (0.03)	-0.67 (0.03)
33	0.75 (0.04)	0.73 (0.03)	0.73 (0.03)
34	0.63 (0.04)	0.73 (0.12)	0.98 (0.17)

5.2 Application 2: Exercise Dependence

The Exercise Dependence Scale-Revised (EDS; Downs et al., 2004; Hausenblas & Downs, 2002) is a self-report measure of exercise dependence that is characterized by excessive exercise engagement as well as physiological, psychosocial, and cognitive symptoms. The EDS is a 21 item measure with seven subscales (3 items each) that are based on the seven criteria of the DSM-IV criteria for substance dependence: withdrawal, intention effects, lack of control, time, reduction in other activities, continuance, and tolerance (Hausenblas & Symons Downs, 2002; American Psychiatric Association, 1994). The EDS instructs participants to answer questions in reference to their current exercise beliefs and behaviors that have occurred in the past three months on a 6-point Likert scale (1 = Never, 6 = Always). The items and corresponding subscales are presented in Table 5.7. The general bifactor dimension may indicate the general severity of exercise dependence disorder, while the subscales indicate the different criteria that can indicate the various dimensions of exercise dependence.

Participants were 406 undergraduate students from a northeastern university in the United States, between the ages of 17 and 53 (*M* = 20.23, *SD* = 3.65). There were 273 females and 130 males in the study (*N* = 3 unspecified). The sample was mostly Caucasian (*N* = 329), but also included African American (*N* = 20), Latin-American (*N* = 20), Asian-American (*N* = 18) and others (*N* = 15). The students were recruited through the psychology subject pool and were mostly from introductory psychology, but some professors of other courses grant extra credit in return for their participation. After reviewing a consent letter that informed participants of their rights, participants responded to a paper and pencil survey in groups of one to ten. IRB approval was obtained for this study and participants were treated in accordance with APA guidelines.

5.2.1 Model Estimation

The models were estimated using the same MCMC specifications as were used in the simulation study: 2 chains estimated with 7,000 total iterations, 3,500 burn-in iterations, and a thinning interval of 3. JAGS model code for *K*-category data used for this analysis is presented in Appendix B.

Table 5.7: Exercise Dependence Scale-Revised items and subscales

Item	Subscale	
1	Withdrawal	I exercise to avoid feeling irritable.
2	Withdrawal	I exercise to avoid feeling anxious.
3	Withdrawal	I exercise to avoid feeling tense.
4	Continuance	I exercise despite recurring physical problems.
5	Continuance	I exercise when injured.
6	Continuance	I exercise despite persistent physical problems.
7	Tolerance	I continually increase my exercise intensity to achieve the desired effects/benefits.
8	Tolerance	I continually increase my exercise frequency to achieve the desired effects/benefits.
9	Tolerance	I continually increase my exercise duration to achieve the desired effects/benefits.
10	Lack of Control	I am unable to reduce how long I exercise.
11	Lack of Control	I am unable to reduce how often I exercise.
12	Lack of Control	I am unable to reduce how intensely I exercise.
13	Reduction in Other Activities	I would rather exercise than spend time with family/friends.
14	Reduction in Other Activities	I think about exercise when I should be concentrating on school/-work.
15	Reduction in Other Activities	I choose to exercise so I can get out of spending time with family/friends.
16	Time	I spend a lot of time exercising.
17	Time	I spend most of my free time exercising.
18	Time	A great deal of my time is spent exercising.
19	Intention Effects	I exercise longer than I intend.
20	Intention Effects	I exercise longer than I expect.
21	Intention Effects	I exercise longer than I plan.

Table 5.8: Proportion of converged parameters for exercise dependence data analyzed with different estimated interactions

	Estimated Interactions, I_{est}					
	None		All		Reduced	
Parameter	PSRF	Geweke	PSRF	Geweke	PSRF	Geweke
Intercepts	1.000	0.829	0.952	0.914	0.942	0.581
Main Effects						
General	0.857	0.905	0.857	0.762	0.714	0.571
Specific	0.952	0.857	0.952	1.000	0.904	0.952
Interaction Effects	*	*	0.714	0.810	1.00	1.00
Latent Variables						
General	1.000	0.929	1.000	0.921	1.000	0.857
Specific	1.000	0.956	1.000	0.954	1.000	0.954
Total	0.999	0.948	0.996	0.947	0.996	0.929

*Non-estimated parameter

5.2.2 Model Selection

Model selection according to DIC could not be completed using *pD* because it was estimated to be infinity for all iterations in the chain, as discussed in the 6-category simulation study results. Though the variant *pD* (calculated as the posterior mean deviance minus the deviance of the posterior means) could be used instead, it did not provide reliable identification of the true underlying model in the simulation. Thus, model selection in this application relies on a priori specification of the pattern matrix, the substantive interpretation of the parameters, the convergence of the parameters, and the highest density intervals for individual parameters.

5.2.3 Parameter Estimates

Posterior means of the intercepts, main effects, and interaction effects and corresponding standard deviations for the three estimated models (no interactions, all interactions, and reduced interactions) are presented in Tables 5.9 through 5.14. Main effects were strong in the different models estimated and did not contain zero within their HDI, supporting the confirmatory structure of the bifactor pattern. As in the TIMSS data, parameter estimates were comparable among the different estimated models with minor fluctuation.

As indicated by the highest density intervals, 85.7% of the interaction effects estimated in the model with all interactions contained zero within the interval. Thus, these interactions were removed and the model was re-fit with a reduced model having 3 interactions out of the 21 possible. These interactions were estimated for the following three items (and subscale):

1. I am unable to reduce how often I exercise. (Lack of Control)

2. I am unable to reduce how intensely I exercise. (Lack of Control)

3. I exercise longer than I intend. (Intention Effects)

The first two items are for the lack of control subscale and had positive interactions of 0.48 ($SD =$ 0.15) and 0.86 ($SD = 0.14$), respectively. The third item belongs to the intention effects subscale and had a negative interaction of -0.21 ($SD = 0.15$) that when re-estimated with reduced interactions did contain zero within its HDI. Given this, it is likely that items' relationships with the general exercise dependence dimension and the specific lack of control dimension is a conditional one, but not for items measuring the intention effects dimension.

Table 5.9: Exercise dependence scale intercept parameter means (and standard deviations) estimated from a bifactor model with no estimated interactions

Item	β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}
1	-1.80 (0.14)	-0.77 (0.11)	-0.07 (0.11)	0.97 (0.12)	1.89 (0.15)
2	-1.22 (0.16)	-0.46 (0.14)	0.41 (0.13)	1.56 (0.17)	2.68 (0.25)
3	-1.84 (0.19)	-0.68 (0.15)	0.09 (0.14)	1.22 (0.17)	2.60 (0.25)
4	-1.11 (0.15)	-0.15 (0.13)	0.49 (0.13)	1.73 (0.18)	2.61 (0.24)
5	-0.38 (0.12)	0.54 (0.12)	1.33 (0.14)	2.18 (0.18)	3.22 (0.27)
6	-0.31 (0.14)	0.53 (0.13)	1.32 (0.14)	2.25 (0.17)	3.74 (0.29)
7	-2.81 (0.19)	-1.84 (0.15)	-0.86 (0.13)	0.35 (0.13)	1.83 (0.17)
8	-2.60 (0.22)	-1.37 (0.18)	-0.29 (0.15)	0.96 (0.16)	2.67 (0.24)
9	-1.85 (0.18)	-1.12 (0.16)	-0.01 (0.14)	1.24 (0.15)	3.01 (0.24)
10	-0.74 (0.12)	0.48 (0.12)	1.32 (0.15)	2.39 (0.22)	3.08 (0.29)
11	-0.45 (0.14)	0.85 (0.14)	1.92 (0.17)	2.64 (0.20)	3.38 (0.27)
12	-0.44 (0.13)	0.76 (0.13)	1.60 (0.14)	2.50 (0.18)	3.05 (0.22)
13	-0.14 (0.11)	1.04 (0.14)	1.83 (0.17)	2.60 (0.23)	3.35 (0.32)
14	-0.35 (0.11)	0.66 (0.11)	1.20 (0.12)	1.84 (0.14)	2.97 (0.25)
15	0.58 (0.11)	1.67 (0.14)	2.30 (0.17)	2.79 (0.21)	3.53 (0.35)
16	-2.19 (0.20)	-0.80 (0.14)	0.48 (0.13)	1.59 (0.16)	2.72 (0.23)
17	-0.96 (0.15)	0.37 (0.14)	1.41 (0.16)	2.41 (0.20)	3.58 (0.29)
18	-1.09 (0.17)	0.32 (0.14)	1.43 (0.15)	2.54 (0.18)	4.09 (0.31)
19	-1.40 (0.17)	-0.03 (0.14)	1.06 (0.14)	2.12 (0.17)	3.53 (0.27)
20	-1.20 (0.19)	0.06 (0.15)	1.42 (0.15)	2.77 (0.20)	4.18 (0.32)
21	-0.79 (0.14)	0.20 (0.14)	1.52 (0.15)	2.59 (0.19)	4.05 (0.34)

Table 5.10: Exercise dependence scale main effect parameter means (and standard deviations) estimated from a bifactor model with no estimated interactions

Item	General	Withdrawal	Continuance	Tolerance	Lack of Control	Reduction in Other Activities	Time	Intention Effects
1	1.28 (0.12)*	0.92 (0.11)*						
2	1.65 (0.18)*	1.37 (0.17)*						
3	1.88 (0.17)*	1.40 (0.17)*						
4	1.56 (0.18)*		1.24 (0.16)*					
5	1.62 (0.14)*		0.89 (0.11)*					
6	1.94 (0.17)*		1.17 (0.13)*					
7	1.72 (0.14)*			1.18 (0.12)*				
8	2.18 (0.19)*			1.51 (0.16)*				
9	2.08 (0.17)*			1.44 (0.13)*				
10	1.40 (0.14)*				0.95 (0.15)*			
11	2.04 (0.17)*				1.06 (0.15)*			
12	2.00 (0.16)*				0.88 (0.13)*			
13	1.53 (0.16)*					0.75 (0.16)*		
14	1.67 (0.14)*					0.23 (0.11)*		
15	1.55 (0.16)*					0.64 (0.13)*		
16	1.89 (0.16)*						0.73 (0.14)*	
17	2.36 (0.20)*						0.65 (0.13)*	
18	2.42 (0.19)*						0.73 (0.12)*	
19	2.20 (0.19)*							1.05 (0.12)*
20	2.44 (0.23)*							1.26 (0.14)*
21	2.19 (0.16)*							0.95 (0.11)*

*Does not contain zero within its Highest Density Interval

Table 5.11: Exercise dependence scale intercept parameter means (and standard deviations) estimated from a bifactor model with all possible estimated interactions

Item	β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}
1	-1.83 (0.15)	-0.79 (0.12)	-0.09 (0.11)	0.95 (0.12)	1.86 (0.15)
2	-1.27 (0.16)	-0.50 (0.14)	0.38 (0.14)	1.53 (0.17)	2.67 (0.24)
3	-1.89 (0.19)	-0.74 (0.16)	0.02 (0.15)	1.14 (0.17)	2.54 (0.24)
4	-1.18 (0.17)	-0.22 (0.15)	0.43 (0.15)	1.70 (0.19)	2.61 (0.25)
5	-0.38 (0.13)	0.55 (0.14)	1.33 (0.16)	2.16 (0.19)	3.20 (0.27)
6	-0.43 (0.16)	0.45 (0.16)	1.30 (0.18)	2.29 (0.22)	3.89 (0.35)
7	-2.86 (0.21)	-1.84 (0.16)	-0.84 (0.14)	0.36 (0.13)	1.81 (0.17)
8	-2.54 (0.21)	-1.35 (0.17)	-0.30 (0.16)	0.91 (0.16)	2.57 (0.23)
9	-1.96 (0.20)	-1.18 (0.17)	0.01 (0.16)	1.32 (0.19)	3.14 (0.28)
10	-0.71 (0.13)	0.47 (0.12)	1.28 (0.14)	2.30 (0.20)	2.97 (0.27)
11	-0.41 (0.15)	0.88 (0.16)	1.94 (0.19)	2.64 (0.22)	3.39 (0.29)
12	-0.38 (0.15)	0.87 (0.16)	1.74 (0.18)	2.69 (0.24)	3.33 (0.30)
13	-0.18 (0.12)	1.00 (0.14)	1.79 (0.17)	2.56 (0.23)	3.32 (0.33)
14	-0.37 (0.11)	0.63 (0.11)	1.16 (0.12)	1.80 (0.15)	2.93 (0.26)
15	0.51 (0.12)	1.66 (0.16)	2.33 (0.20)	2.85 (0.24)	3.63 (0.36)
16	-2.24 (0.19)	-0.84 (0.13)	0.44 (0.13)	1.55 (0.16)	2.69 (0.23)
17	-1.06 (0.15)	0.29 (0.15)	1.37 (0.16)	2.41 (0.20)	3.63 (0.31)
18	-1.14 (0.15)	0.27 (0.15)	1.39 (0.17)	2.51 (0.20)	4.09 (0.34)
19	-1.54 (0.17)	-0.15 (0.15)	0.97 (0.16)	2.10 (0.19)	3.61 (0.31)
20	-1.24 (0.17)	-0.04 (0.15)	1.28 (0.16)	2.63 (0.20)	4.06 (0.32)
21	-0.86 (0.15)	0.13 (0.14)	1.46 (0.16)	2.54 (0.19)	4.03 (0.36)

Table 5.12: Exercise dependence scale main effect parameter means (and standard deviations) estimated from a bifactor model with all possible estimated interactions

Item	General	Withdrawal	Continuance	Tolerance	Lack of Control	Reduction in Other Activities	Time	Intention Effects	Interaction
1	1.34 (0.12)*	0.92 (0.12)*							0.06 (0.12)
2	1.74 (0.17)*	1.39 (0.16)*							0.01 (0.15)
3	1.96 (0.19)*	1.41 (0.17)*							-0.07 (0.16)
4	1.57 (0.16)*		1.32 (0.17)*						-0.31 (0.17)
5	1.72 (0.16)*		0.93 (0.12)*						0.04 (0.15)
6	2.01 (0.19)*		1.28 (0.15)*						-0.39 (0.22)
7	1.78 (0.15)*			1.11 (0.13)*					0.18 (0.16)
8	2.18 (0.18)*			1.41 (0.16)*					0.06 (0.19)
9	2.32 (0.20)*			1.50 (0.15)*					0.20 (0.22)
10	1.42 (0.13)*				0.85 (0.14)*				0.27 (0.13)
11	2.20 (0.19)*				1.01 (0.16)*				0.48 (0.16)*
12	2.42 (0.26)*				0.87 (0.14)*				0.86 (0.19)*
13	1.57 (0.17)*					0.76 (0.16)*			-0.11 (0.18)
14	1.73 (0.15)*					0.24 (0.11)*			0.10 (0.17)
15	1.65 (0.17)*					0.66 (0.14)*			-0.36 (0.19)
16	1.97 (0.16)*						0.73 (0.14)*		-0.10 (0.19)
17	2.52 (0.16)*						0.69 (0.14)*		-0.24 (0.17)
18	2.51 (0.16)*						0.76 (0.12)*		-0.17 (0.25)
19	2.33 (0.18)*							1.12 (0.14)*	-0.37 (0.16)*
20	2.40 (0.16)*							1.23 (0.12)*	-0.24 (0.16)
21	2.28 (0.17)*							0.95 (0.12)*	-0.19 (0.19)

*Does not contain zero within its Highest Density Interval

Table 5.13: Exercise dependence scale intercept parameter means (and standard deviations) estimated from a bifactor model with reduced estimated interactions

Item	β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}
1	-1.81 (0.15)	-0.77 (0.11)	-0.07 (0.11)	0.97 (0.12)	1.89 (0.16)
2	-1.24 (0.16)	-0.47 (0.14)	0.41 (0.14)	1.57 (0.17)	2.71 (0.25)
3	-1.81 (0.18)	-0.68 (0.15)	0.08 (0.14)	1.19 (0.17)	2.56 (0.24)
4	-1.09 (0.15)	-0.14 (0.13)	0.50 (0.14)	1.73 (0.21)	2.59 (0.28)
5	-0.37 (0.12)	0.56 (0.12)	1.34 (0.14)	2.19 (0.18)	3.23 (0.27)
6	-0.29 (0.14)	0.55 (0.13)	1.33 (0.15)	2.26 (0.19)	3.73 (0.31)
7	-2.83 (0.21)	-1.84 (0.16)	-0.86 (0.13)	0.35 (0.12)	1.83 (0.17)
8	-2.61 (0.22)	-1.37 (0.16)	-0.30 (0.15)	0.96 (0.17)	2.69 (0.25)
9	-1.88 (0.17)	-1.14 (0.15)	-0.02 (0.14)	1.26 (0.17)	3.05 (0.25)
10	-0.73 (0.13)	0.47 (0.12)	1.30 (0.14)	2.36 (0.20)	3.04 (0.27)
11	-0.39 (0.14)	0.86 (0.13)	1.89 (0.14)	2.57 (0.18)	3.31 (0.24)
12	-0.36 (0.14)	0.92 (0.15)	1.81 (0.16)	2.77 (0.20)	3.42 (0.24)
13	-0.13 (0.11)	1.04 (0.13)	1.83 (0.16)	2.60 (0.21)	3.35 (0.32)
14	-0.34 (0.11)	0.66 (0.11)	1.20 (0.12)	1.84 (0.15)	2.96 (0.25)
15	0.58 (0.12)	1.67 (0.15)	2.30 (0.19)	2.79 (0.23)	3.53 (0.35)
16	-2.19 (0.19)	-0.79 (0.13)	0.49 (0.12)	1.59 (0.16)	2.72 (0.23)
17	-0.96 (0.15)	0.37 (0.14)	1.43 (0.16)	2.44 (0.19)	3.60 (0.28)
18	-1.06 (0.15)	0.33 (0.13)	1.42 (0.15)	2.51 (0.19)	4.05 (0.32)
19	-1.47 (0.18)	-0.07 (0.15)	1.05 (0.16)	2.17 (0.19)	3.66 (0.29)
20	-1.14 (0.17)	0.07 (0.15)	1.38 (0.16)	2.68 (0.20)	4.05 (0.32)
21	-0.77 (0.14)	0.21 (0.14)	1.51 (0.15)	2.57 (0.19)	4.00 (0.34)

Table 5.14: Exercise dependence scale main effect parameter means (and standard deviations) estimated from a bifactor model with reduced estimated interactions

Item	General	Withdrawal	Continuance	Tolerance	Lack of Control	Reduction in Other Activities	Time	Intention Effects	Interaction
1	1.29 (0.11)*	0.91 (0.11)*							
2	1.70 (0.18)*	1.38 (0.16)*							
3	1.90 (0.17)*	1.36 (0.15)*							
4	1.55 (0.17)*		1.22 (0.18)*						
5	1.63 (0.15)*		0.90 (0.11)*						
6	1.93 (0.16)*		1.16 (0.13)*						
7	1.74 (0.14)*			1.17 (0.13)*					
8	2.22 (0.17)*			1.51 (0.15)*					
9	2.15 (0.18)*			1.47 (0.13)*					
10	1.33 (0.13)*				0.92 (0.13)*				
11	2.00 (0.14)*				0.97 (0.11)*				0.29 (0.15)*
12	2.35 (0.13)*				0.90 (0.11)*				0.74 (0.14)*
13	1.53 (0.16)*					0.75 (0.14)*			
14	1.68 (0.15)*					0.22 (0.11)*			
15	1.58 (0.17)*					0.63 (0.14)*			
16	1.90 (0.15)*						0.74 (0.14)*		
17	2.41 (0.19)*						0.67 (0.14)*		
18	2.41 (0.16)*						0.71 (0.13)*		
19	2.25 (0.21)*							1.09 (0.13)*	-0.21 (0.15)
20	2.30 (0.18)*							1.18 (0.13)*	
21	2.14 (0.14)*							0.90 (0.11)*	

*Does not contain zero within its Highest Density Interval

Chapter 6

Discussion

This study examined the parameter recovery of bifactor models that include interaction effects between the general and specific latent dimensions and the potential consequences of omitting interactions in misspecified models. In doing so, a bifactor model was defined with interactions for constructs that involve multiple factors and the risk of fitting bifactor models with only main effects was judged for both item and respondent parameters. In this chapter, I present a summary of the major results of this study centered around the research questions outlined previously, followed with a discussion of study limitations, possible directions for future research, and implications in psychological and educational measurement research.

6.1 Review of Study Findings

The three primary questions can be summarized as follows. First, how well does the proposed bifactor model with interaction effects recover item and person parameters in realistic data generating conditions? Second, what is the impact of ignoring interaction effects on model recovery of item and person parameters when fitting a bifactor model with only main effects? Third, how well do model selection and parameter refinement strategies perform in identifying underlying interactions in a bifactor model? For each of these questions, the study findings are summarized in the following sections.

6.1.1 Parameter Recovery

The results from the simulation studies indicated that the bifactor model with interactions had little bias and adequate root mean squared error (RMSE) for the estimated parameters, and few non-converged solutions. Respondent parameters for all data-generating and estimation conditions were recovered well and did not vary much among the conditions. This finding suggests that the correct specification of the interaction effect parameters may not be necessary for practical applications where the modeling purpose is to estimate the respondent parameters. However, if the researcher is interested in understanding the relationships between items and the measured latent dimensions, such as in assessing item dimensionality or violations of local independence, interaction effects should be evaluated in the model building process.

6.1.2 Model Misspecification

Typical model building procedures involve 1) the development of a prior model that identifies the hypothesized patterns of associations among manifest and latent variables, 2) a test of the hypothesized model with a sample of data, 3) evaluation of the solution in terms of model-data fit and parameter estimates, and 4) modification of the model to identify the most parsimonious model without sacrificing goodness of fit or substantive meaning (MacCallum, 1986). In confirmatory model building, particularly in factor analytic and item response theory traditions, only main effects are specified between items and latent variables, and a bifactor structure is often tested only after identifying multidimensionality in data. Given this tradition, this study investigated the impact of ignoring interactions when they are present in the underlying data.

As demonstrated by the simulation studies, the main effect parameters had a greater bias when the interactions were present in the underlying data but unmodeled. The main effects tended to have a negative bias when the interactions were omitted, which could potentially cue the researcher to remove the main effect parameter(s) to simplify the model in favor of parsimonious results. Thus, the new constrained model is estimated and potentially supported in model-data fit comparisons. In some situations, consistently underestimated general main effect parameters may indicate that a bifactor model is not necessary and instead a simpler model may be preferred instead. The same could be said for the specific dimension main effects. Subsequent inferences about the construct based on the invalid result, such as in tests of explanatory variables, may be incorrect because the true complexity of the data is masked when the interactions are omitted. However, bias, absolute bias, and RMSE were averaged in this study for each parameter type. It may be that parameter bias and RMSE varies depending on the size of the effect, such that smaller main effects may show a positive bias whereas large main effects may show a negative bias (or vice versa). Further research should evaluate the impact of model misspecification on parameter recovery, conditional on the values of the generated effects.

6.1.3 Model Selection and Parameter Refinement

As indicated by the high frequency of deviance information criteria (DIC) to select the true underlying interaction model, particularly with the 2-category data, these findings suggest that including interaction effects in bifactor models could provide a better fit to the data, even with the added model complexity. However, model selection based on information criteria is not devoid of error. The DIC is calculated based on two statistics, deviance and *pD*, that also have sampling variability and are measured with a certain degree of precision. Particularly in the graded response data simulation and empirical application, *pD* was non-estimable in some or all iterations of the Markov chains. For many replications in the graded response simulation, the sample mean *pD* was only calculated from a very small handful of iterations. This may be a function of the model estimated or arising from a glitch or miscalculation in JAGS software. Comparatively, there were no instances of infinite *pD* in the 2-category simulation. Researchers interested in using DIC with a penalty for the effective number of parameters should consider examining their performance using simulation with generating values based on their empirical data prior to using them to make model selection decisions.

The methods used in this study to uncover the true interactions in empirical data were unreliable in the simulation studies. Interaction effects were poorly identified using the highest density intervals and credible intervals for parameter refinement when estimating all interactions. Future research should compare alternate methods of model parameter refinement and provide clear guidelines for model parameter identification. As with any latent variable model, the items should be evaluated from a substantive perspective to consider the skills involved that elicit responses, defined by what construct represents the general and specific dimensions and the interpretations that would follow with the addition of an interaction for an item. For example, is it likely that the rate of increase in the probability of a correct response for a word problem item would be augmented with more math ability in respondents with below average reading comprehension? These results emphasize the importance of using multiple sources of substantive and statistical evidence to identify misfit in model-data fit and select an appropriate model.

6.2 Directions for Future Research

The simulation studies were not exhaustive with respect to relevant assessment design and misspecification conditions, and different outcomes were not investigated, including reliability and conditional bias or RMSE (e.g., differences in parameter bias as a function of θ). Generalizability of findings are dependent on the limitations in the design of the simulations, including the fixed test length, the number of specific dimensions, distributions of latent and observed variables, and types of misspecification. The choice of design factors and corresponding levels were selected for their relevance to psychological assessment and educational measurement research, and parameter generating values were selected for generalizability. However, the data simulated was generated under ideal conditions, where the latent variables were multivariate normal without skewness, the specific dimensions were uncorrelated, and bifactor models were estimated with accurate design matrices for main effects.

Recovery of the respondent parameters were robust to misspecification of the interaction effects in all data generating conditions, though this result may not be observed when the measurement model is misspecified in other ways (i.e., omitted main effects or latent dimensions). Future research should evaluate the performance of the bifactor model with interactions in unfavorable conditions, such as with dimensions measured with few items and respondents, in data with few respondents endorsing different categories in a graded response bifactor model, or in situations where monotonicity is violated for an item, such as when the values of the main effects and interactions indicate a change in the direction of the probability of a response with increasing levels of the latent variables. At points where this occurs, using conditional evaluation criteria may inform how the shift in direction from increasing to decreasing probabilities as a function of θ affects bias or RMSE. This strategy would help illuminate cases where bias at one end of the θ distribution is balanced at the other end, such that bias averages out to zero.

Given that estimating the number of effective parameters (i.e., *pD*) in the graded response data is not always possible, further research is needed to evaluate methods to estimate the penalty parameter when applied to different statistical distributions of the data and parameters. In the context of this study, other methods to estimate the effective number of parameters and other types of information criteria that may be better suited to make model comparisons should be further explored. Furthermore, alternative methods of parameter refinement at the item level should be investigated. This study used Bayesian estimation and therefore relied upon the highest density intervals and confidence intervals to identify whether an interaction parameter could be constrained in a reduced model. However, the results of this study suggest that over-specification by including all possible interactions to estimate a model does not adversely impact model parameter estimates of other item parameters, and in fact had a comparable bias with the model estimated to match the data-generating model for both graded response and binary response bifactor models.

An interesting area of future work would be investigating interaction effects for items that measure more than two dimensions, such as items that measure two specific dimensions in addition to the general dimension. This would entail a substantial increase in the number of estimated effects per item: a 3-way interaction, three 2-way interactions, three main effects, and an intercept. However, this increased complexity may cause a significant strain on model estimation and convergence. Furthermore, the complex model may have a greater propensity to overfit the data, resulting in a model that may not replicate in repeated samples.

6.3 Implications for Measurement Research

While previous studies have investigated parameter recovery of interaction effects in traditional multidimensional IRT models with correlated factors, this study extended this research to bifactor models that may be more suitable for multifaceted data with correlated secondary dimensions alongside a common general dimension. The benefit of the bifactor model as applied to partially compensatory interpretations is that it is structured to have uncorrelated dimensions because the commonality between specific dimensions is due to their common core reflected in the general dimension. In previous research, one of the contributors of poor parameter recovery of item and respondent parameters and difficulty in estimation in partially compensatory product models was due to correlated dimensions (Chalmers & Flora, 2014; Babcock, 2011; DeMars, 2016; Bolt & Lall, 2003). Thus, this research expands on these studies to investigate response processes that can vary depending on the levels of the general and specific latent variables in a bifactor model that can accommodate highly correlated secondary dimensions by way of a common general dimension. Adding an interaction effect to allow diminishing and augmented rates of returns in the probabilities of endorsed responses to items (conditional on the levels of the latent dimensions) could have interesting insight into alternative conceptual interpretations, or at the very least give and a more accurate measure of the relationships between items and latent variables without the added estimation burden inherent in traditional MIRT models.

The simulation studies and empirical analyses in this study informs practitioners interested in using bifactor models with interactions between specific and general dimensions about 1) how item main effect parameters may be underestimated when interactions are omitted but not respondent parameters, and 2) the risk of using traditional approaches of model selection and parameter refinement in Bayesian analysis. Larger sample sizes are ideal for having sufficient accuracy to estimate effects, but the simulations showed that having a larger sample size actually resulted in a higher rate of misidentification of interacting items. The root mean squared errors, on the other hand, were smaller with the larger sample size, supporting increased accuracy. Instead, estimating all interactions compared to matched interactions did not severely impact model parameter estimates.

In fact, bias and absolute bias was often smaller when all interactions were estimated compared to the matched interaction model for main effect parameters. Thus, though the highest density intervals and credible intervals were often inaccurate for interaction items, estimating all interactions may be advantageous, but at a cost to model parsimony.

In summary, this study expands the researcher's toolbox with new strategies to explore different patterns of associations between items and latent variables that could result in a more accurate solution and resulting inference. As with any research, bifactor models with interactions applied to empirical data should be cross-validated in repeated samples for replicability and in different populations for generalizability (MacCallum, 2003).

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13(2), 113–127.

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

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

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV*. Washington, DC: Author, 4th ed. edition.

Babcock, B. (2011). Estimating a noncompensatory IRT model using Metropolis within Gibbs sampling. *Applied Psychological Measurement*, 35(4), 317–329.

Babcock, B. G. E. (2009). *Estimating a Noncompensatory IRT Model Using a Modified Metropolis Algorithm*. Unpublished Doctoral Dissertation, University of Minnesota.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.

Bolt, D. M. & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, 27(6), 395–414.

Brooks, S., P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, 80(4), 796–846.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75(4), 581–612.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16(3), 221–248.

Camilli, G. (1994). Teacher's corner: Origin of the scaling constant D = 1.7 in item response theory. *Journal of Educational Statistics*, 19(3), 293–295.

Carver, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology*, 56(4), 577.

Chalmers, R. P. & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement*, 38(5), 339–358.

Chen, F. F., West, S., & Sousa, K. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41(2), 189–225.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327–335.

Cudeck, R. & Harring, J. R. (2009). Marginal maximum likelihood estimation of a latent variable model with interaction. *Journal of Educational and Behavioral Statistics*, 34(1), 131–144.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, 36(2), 104–121.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, 13(4), 354–378.

DeMars, C. E. (2016). Partially compensatory multidimensional item response theory models: Two alternate model forms. *Educational and Psychological Measurement*, 76(2), 231–257.

Downs, D. S., Hausenblas, H. A., & Nigg, C. R. (2004). Factorial validity and psychometric examination of the Exercise Dependence Scale-Revised. *Measurement in Physical Education and Exercise Science*, 8(4), 183–201.

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7(2), 189–199.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, 49, 175–186.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics*, volume 4 (pp. 169–193). Oxford: Oxford University Press.

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

Gibbons, R. D. & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423–436.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science*, 50(1), 21.

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, 55, 57–68.

Gustafsson, J.-E. & Blake, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, 28(4), 407–434.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dosage is subject to error*. Technical Report 15, Applied Mathematics and Statistics Laboratory, Stanford University., Stanford, CA.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE.

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

Hausenblas, H. A. & Downs, D. S. (2002). How much is too much? The development and validation of the Exercise Dependence Scale. *Psychology and Health: An International Journal*, 17, 387–404.

Hausenblas, H. A. & Symons Downs, D. (2002). Exercise dependence: A systematic review. *Psychology of Sport and Exercise*, 3, 89–123.

Holzinger, K. J. & Harman, H. H. (1941). *Factor Analysis: A Synthesis of Factorial Methods*. Chicago: University of Chicago Press.

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

Jennrich, R. I. & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, 76(4), 537–549.

Jennrich, R. I. & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, 77(3), 442–454.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2012). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, 38(1), 32–60.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2018). CFA models with a general factor and multiple sets of secondary factors. *Psychometrika*, 4(83), 785–808.

Kieftenbeld, V. & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement*, 36(5), 399–419.

Kim, J.-S. & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, 26(4), 38–51.

Koehler, E., Brown, E., & Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2), 155–162.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30(1), 3–21.

Li, Y. & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, 36(1), 3–20.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, Monograph 140.

Longo, Y., Coyne, I., Joseph, S., & Gustavsson, P. (2016). Support for a general factor of well-being. *Personality and Individual Differences*, 100, 68–72.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100(1), 107–120.

MacCallum, R. C. (2003). 2001 Presidential address: Working with imperfect models. *Multivariate Behavioral Research*, 38(1), 113–139.

Mansolf, M. & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research*, 51(5), 698–717.

Mansolf, M. & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. *Intelligence*, 61, 120–129.

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, 60(4), 523–547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2), 187–212.

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

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika*, 27(4), 397–415.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, R. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1091.

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.

Murray, A. L. & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41(5), 407–422.

Neal, R. (2003). Slice sampling. *Annals of Statistics*, 31(3), 705–767.

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

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

Plummer, M. (2002). Discussion on the paper by Spiegelhalter et al. *Journal of the Royal Statistical Society Series B*, 64(Part 4), 620–621.

Plummer, M. (2017). *JAGS Version 4.3.0 user manual.* Retreived from https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1), 7–11.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer New York.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667–696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95(2), 129–140.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92(6), 544–559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(SUPPL. 1), 19–31.

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

Rizopoulos, D. & Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, 61(2), 415–438.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. New York: Psychometric Society.

Schmid, J. & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1), 53–61.

Seo, D. G. & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement*, 75(6), 954–978.

Simpson, M. A. (2005). *Use of a Variable Compensation Item Response Model to Assess the Effect of Working-Memory Load On Noncompensatory Processing in an Inductive Reasoning Task*. Unpublished Doctoral Dissertation, University of North Carolina at Greensboro.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet based tests. *Journal of Educational Measurement*, 28(3), 237–247.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201–293.

Spiegelhalter, D. J., Best, N., Carlin, B., & van der Linde., A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64(Part 4), 583–639.

Sympson, J. B. (1977). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota.

Takane, Y. & Leeuw, J. D. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393–408.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological Assessment*, 24(1), 101–113.

Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago Press.

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409–426.

Weiss, D. J. & Gibbons, R. D. (2007). *Computerized Adaptive Testing With the Bifactor Model*. Paper presented at the New CAT Models session at the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved from http://iacat.org/sites/default/files/biblio/cat07weiss&gibbons.pdf.

Whitley, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479–494.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement*, 26(3), 339–352.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64(2), 113–128.

Appendix A

JAGS Model Code for 2-Category Ordinal Data

```
model{
  for (j in 1:n) { ## Persons
    for (m in 1:nItems) { ## Items
      for (d in 1:D) { ## All Dimensions
        pdim[j,m,d] <- beta1[m,d]*theta[j,d] +
          beta2[m,d]*theta[j,1]*theta[j,d]
      }
      probit(p[j,m]) <- (sum(pdim[j,m,1:D])+beta0[m])
      dat[j,m]~dbern(p[j,m])
    }
  }
  for (j in 1:n){
    theta[j,1:D] ~ dmnorm(mu, SIG)
  }
  for(m in 1:nItems){
    beta0[m] ~ dnorm(0,1)
    for(d in 1:D){
      b_star[m,d] ~ dnorm(0,1) T(0,)
      beta1[m,d] <- designMain[m,d]*b_star[m,d]
```

```
         x_star[m,d] ~ dnorm(0,1)
         beta2[m,d] <- designInt[m,d]*x_star[m,d]
      }
   }
}
```

Appendix B

JAGS Model Code for K-Category Ordinal Data

```
model
{
  for (j in 1:n) { ## Persons
    for (m in 1:nItems) { ## Items
      dat[j,m] ~ dcat(prob[j,m,1:K])
    }
    for (m in 1:nItems) {
      for (d in 1:D) { ## All Dimensions
        pdim[j,m,d] <- beta1[m,d]*theta[j,d] +
            beta2[m,d]*theta[j,1]*theta[j,d]
      }
    }
    for (m in 1:nItems) {
      for(k in 1:(K-1)) {
        probit(P[j,m,k]) <- (sum(pdim[j,m,1:D]) + beta0[m, k])
      }
      P[j,m,K] <- 1.0
    }
    for (m in 1:nItems) {
      prob[j,m,1] <- P[j,m,1]
      for(k in 2:K){
        prob[j,m,k] <- P[j,m,k] - P[j,m,(k-1)]
      }
    }
    theta[j,1:D] ~ dmnorm(mu, SIG)
  }
  for (m in 1:nItems){
    for(k in 1:(K-1)){  ## Thresholds
      k_star[m,k] ~ dnorm(0,1)
    }
    beta0[m, 1:(K-1)] <- sort(k_star[m, 1:(K-1)])
  }
  for (m in 1:nItems) {
    for(d in 1:D){  ## Main effects and Interactions
      b_star[m,d] ~ dnorm(0,1) T(0,)
      beta1[m,d] <- designMain[m,d]*b_star[m,d]

      x_star[m,d] ~ dnorm(0,1)
      beta2[m,d] <- designInt[m,d]*x_star[m,d]
    }
  }
}
```