A Bayesian MIMIC Model for Testing Non-uniform DIF in Two and Three Groups

By

Jared K. Harpole

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

Committee members

Wei Wu, Chairperson

Carol Woods, Co-Chairperson

Pascal DeBoeck

William Skorupski

Jonathan Templin

David Johnson

Date defended: October 26, 2015

The Dissertation Committee for Jared K. Harpole certifies
that this is the approved version of the following dissertation :

A Bayesian MIMIC Model for Testing Non-uniform DIF in Two and Three Groups

Wei Wu, Chairperson

Carol Woods, Co-Chairperson

Date approved: November 9, 2015

Abstract

Multiple-indicator multiple cause (MIMIC) models have become a popular latent variable method to detect differential item functioning (DIF) by practitioners. The ease of including groups for DIF testing and the implementation of MIMIC models in structural equation modeling software have helped drive the use of MIMIC models by applied researchers. However, there are several shortcomings within the methodological literature that are important questions yet to be addressed. First, only the case of two groups have been studied in simulations studies, yet practitioners are increasingly utilizing MIMIC models on more than two groups (e.g. Fleishman, Spector, & Altman, 2002; Sacco, Casado, & Unick, 2011; Sacco, Torres, Cunningham-Williams, Woods, & Unick, 2011; Woods, Oltmanns, & Turkheimer, 2009; Yang, Tommet, & Jones, 2009). Second, MIMIC models can be parameterized to test for non-uniform DIF (e.g. Woods & Grimm, 2011), but in current implementations Type I error rates were too high possibly due to assumption violations in the estimation of the latent interaction. Third, almost all previous simulations for MIMIC models have not considered the MIMIC model's robustness to violations of the homogeneity of variance assumption (see Carroll, 2014 for an exception). A Monte Carlo simulation study was conducted to address these three shortcomings utilizing a 2 (number of groups) x 3 (latent variance differences) x 3 (sample size imbalance) factorial design and comparing the proposed Bayesian MIMIC model with an improved version of Lord's (1980) χ^2. Results of the simulation study indicated that when the assumption of homogeneity of latent variances held the Bayesian MIMIC model was a competitive method for assessing DIF. However, when the assumption was not met the Bayesian MIMIC

model would not be recommended due to poor parameter recovery. Overall, this research provides evidence that practitioners should not use MIMIC models for testing DIF.

Acknowledgements

I would first like to thank God for giving me the abilities and strengths to complete my dissertation. Second, I would like to thank my beautiful wife Rebecca for all her support and love over the last 6 years in deciding to go back to school and completing my PhD. I would also like to thank my son Isaiah who came on the scene at the latter part of my studies and helped me keep a positive outlook and be balanced in completing this dissertation. I also want to extend thanks to my parents Ken and Carolyn Harpole who raised me and provided support for me to be able to do what I do today. I also want to thank my Grandparents Ruth and Murray Harpole who provided me with a wonderful example of how to live life.

I would like to thank my Ph.D. advisor, Dr. Carol Woods, who provided unwavering support throughout my studies and was always willing to meet with me to discuss various questions or concerns that came up. I would also like to thank other members of my dissertation committee, Drs. Wei Wu, Pascal DeBoeck, Billy Skorupski, Jonathan Templin, and David Johnson. I wish to thank Dr. Skorupski in particular for introducing me to Bayesian statistics and for helpful discussions throughout my years in the psychometric work group. I would also like to thank Dr. Jonathan Templin for his help in showing me how to use the ACF cluster to run my Bayesian MIMIC simulation. Lastly, I want to thank Drs. Wei Wu and Pascal DeBoeck in particular for all their support and instruction throughout my graduate studies.

Finally, I want to acknowledge the contribution of the Quantitative Psychology program and the Center for Research Methods and Data Analysis. I am especially grateful to Dr. Paul Johnson at the CRMDA for teaching me how to use R and cluster computing. I want to further thank Mauricio Garnier-Villareal for helping me with the Stan code used in this dissertation. Lastly, I want to offer further thanks to Terry Jorgenson, Kyle Lang, and Ian Carroll for the many stimulating discussions and helpful suggestions toward this project.

Contents

Figures

Tables

Chapter 1

Background

1.1 Introduction

Item fairness has been an important consideration in educational and psychological testing for the past 40 years. This issue continues to be an active area of research today. Items on educational and psychological tests should function the same way across participants to ensure the validity of an assessment. Measurement invariance of items refers to assessment questions functioning the same way for all participants across a variety of conditions assuming the conditions are irrelevant to the items being measured (Millsap, 2011). If a psychological instrument measures depression and males and females have been matched on their level of depression then the assessment should function the same way for both males and females. When this condition fails depression is related to some other nuisance dimension for a given item. When items do not function the same way for all participants these items are said to be exhibiting differential item functioning (DIF) (Holland & Wainer, 1993; Mellenbergh, 1989; Thissen, Steinberg, & Wainer, 1993).

Item response theory (IRT) provides a framework for testing DIF that has been utilized in practice. One of the downsides of DIF testing in an IRT framework are that larger sample sizes are often required than may be feasible in some educational and psychological contexts. However, Woods (2009b) found that testing for uniform DIF using an IRT based multiple-indicator multiple

cause (MIMIC; Jöreskog & Goldberger, 1975; B. O. Muthén, 1985) model provided adequate power, Type I error control, and parameter recovery compared with a multiple group IRT DIF testing procedure at smaller sample sizes (i.e. 25, 50, 100, and 200) for the focal group. Further, MIMIC models seamlessly allow for the inclusion of multiple grouping variables for DIF testing and are able to be fit in popular structural equation modeling software. For these reasons MIMIC models have been widely implemented by practitioners (e.g. Fleishman et al., 2002; Sacco, Casado, & Unick, 2011; Sacco, Torres, et al., 2011; Woods et al., 2009; Yang et al., 2009; Yu, Yu, & Ahn, 2007) and studied by methodologists (Carroll, 2014; Finch, 2005; E. S. Kim, Yoon, & Lee, 2012; MacIntosh & Hashim, 2003; Shih & Wang, 2009; W. C. Wang & Shih, 2010; Woods, 2009b; Woods & Grimm, 2011).

Currently, all of the applied research involving MIMIC models with categorical data are limited to the case of testing only uniform DIF. However, there has been some movement within the methodological community to test for non-uniform DIF on categorical data with MIMIC models (e.g. Woods & Grimm, 2011) and continuous data with MIMIC models (e.g. Barendse, Oort, & Garst, 2010; Barendse, Oort, Werner, Ligtvoet, & Schermelleh-Engel, 2012) by adding a latent interaction into the model. Yet, the studies by Barendse et al. (2010), Barendse et al. (2012), and Woods and Grimm (2011) all noted inflated Type I error rates for non-uniform MIMIC DIF models possibly due to the assumption violation inherent in the estimation of the latent interaction term.

The current body of MIMIC DIF testing research has three limitations. First, for the case of uniform DIF methodological research has only considered the cased of two groups in simulation studies (e.g. Finch, 2005; Woods, 2009b). However, the methodological research on two groups is lagging behind applications of MIMIC DIF models in more than two groups (e.g. Fleishman et al., 2002; Sacco, Casado, & Unick, 2011; Sacco, Torres, et al., 2011; Woods et al., 2009; Yang et al., 2009). Within the two group case for uniform DIF testing MIMIC models have shown better performance than multiple group IRT with small focal group sample sizes (Woods, 2009b). Methodological work with respect to testing non-uniform DIF has only considered the two group case and it is unknown how this method will perform with more than two groups. Further, whether the small

focal group benefits found by Woods (2009b) of the uniform DIF MIMIC model will carry over to the non-uniform DIF MIMIC model is an open question.

A second limitation of previous studies on MIMIC DIF testing involve the estimation of the latent interaction term. The assumption for estimating a latent interaction as implemented in testing for non-uniform DIF with MIMIC models by Barendse et al. (2010), Barendse et al. (2012), and Woods and Grimm (2011) is that both variables in the interaction are normally distributed. Clearly, this is not true as binary variables are not normally distributed. All previous studies using a latent interaction term to test for non-uniform DIF noted inflated Type I error rates (Barendse et al., 2010; Barendse et al., 2012; Woods & Grimm, 2011). Woods and Grimm (2011) noted that the IRT based MIMIC model could be reparameterized as a non-linear logistic mixed model to more appropriately estimate the latent interaction. Thus, there is a need to utilize an estimation method of the latent interaction that is more theoretically justified in order to determine the utility of testing non-uniform DIF in MIMIC models.

A third limitation of previous studies is a failure to assess the MIMIC models robustness to a critical assumption: Equal latent variances across the groups. Some studies of IRT based DIF testing with real data in more than two groups show that this assumption may not hold (e.g. Harpole et al., 2014; Langer, 2008). At the time of this writing only Carroll (2014) has manipulated this assumption in MIMIC DIF testing for uniform DIF and this assumption has not been manipulated for the case of testing non-uniform DIF with MIMIC models. Carroll (2014) found that violating this assumption resulted in inflated Type I errors and parameter bias for the case of testing uniform DIF. However, Carroll (2014) only considered the two-group case and only uniform DIF.

The purpose of the present research was to conduct a Monte Carlo simulation and applied example to address these three limitations of the present body of MIMIC DIF research and provide an illustration of how one might implement this model in practice. This dissertation consists of several sections: Theoretical background, a review of unidimensional IRT based DIF methods, method section for the simulation, results of the simulation, an applied example, and discussion. First, a brief overview of DIF, IRT, and Bayesian estimation will be given. Next, a review of several

types of unidimensional IRT based DIF methods will be given followed by the methods and results of the simulation study. Lastly, an applied example and overall discussion will be given.

1.2 Theoretical Background

1.2.1 Background on DIF

DIF can be defined as the difference in the probability of endorsing a given response to an item differs for one manifest group compared to another after controlling for true mean differences on the target trait. In DIF testing the reference group is the group to which other groups denoted focal groups are compared. Typically, the reference group consists of a larger representative group which a given psychological or educational test is posited to favor or may have been the norming group. A more precise definition of DIF can be given within a mathematical framework. Let the response to a particular item on a test be denoted by Y and the response to that item is determined solely by the latent variable θ. The conditional probability of an item response given the latent trait is denoted as $f(Y|\theta)$. If we are interested in comparing the conditional probability of Y for $h = 1,\ldots,H$ focal groups to a reference group then an item is said to be unbiased if the following expression holds

$$f(Y|\theta,G = R) = f(Y|\theta,G = F_1),\ldots,f(Y|\theta,G = R) = f(Y|\theta,G = F_H), \tag{1.1}$$

where, G denotes the group membership (R = reference group and F_h = focal group h), $f(Y|\theta,G = R)$ represents the conditional probability of item response given the latent trait and being in the reference group and $f(Y|\theta,G = F_h)$ represents the conditional probability of item response given the latent trait and being a member of focal group h ($h = 1\ldots,H$). If Equation 1.1 does not hold then the item is said to show DIF for focal group h relative to the reference group. Equation 1.1 gives a formal definition to pairwise measurement invariance where the conditional probability of an item response given group membership for a reference group is compared with each focal

group. A more general definition which includes Equation 1.1 as a special case would be

$$f(Y|\theta, G = R) = f(Y|\theta, G = F_1) = \cdots f(Y|\theta, G = F_H), \tag{1.2}$$

suggesting that the equality of all conditional response probabilities must hold regardless of group for measurement invariance to hold, otherwise an item would show DIF. This expression evaluates if the item parameters are invariant across all groups and does not contrast the groups with a specified reference group. For the current study only the definition in Equation 1.1 will be considered.

According to Holland and Wainer (1993), Mellenbergh (1989), Thissen, Steinberg, and Wainer (1988), and Thissen et al. (1993) there are two main types of DIF: uniform and non-uniform. Uniform DIF occurs when the difference in the probability of the reference group responding in a given category compared with the probability of a focal group(s) response in that category is constant over the range of the latent construct. Non-uniform DIF can be thought in terms of an interaction effect where the probability of the reference group responding in a given category compared with the probability of the focal group responding in that category depends on the level of the latent construct. In other words at some levels of the latent variable the reference group may be favored and at other levels the focal group may be favored.

A plethora of methods have been proposed to test for DIF (see Holland & Wainer, 1993; Millsap, 2011; R. D. Penfield & Camilli, 2007; R. D. Penfield & Lam, 2000 for a review of popular methods). Theses methods can be divided into two main classes: Latent variable methods and observed variable methods (Millsap, 2011; R. D. Penfield & Camilli, 2007). The notion of latent and observed methods stems from the matching criteria that is used to align persons at different ability levels for the purpose of DIF testing. Observed variable methods use the examinees summed score as a proxy for their level of the latent trait being assessed. Latent variable methods typically postulate a continuous normally distributed latent variable that gives rise to the manifest item responses of participants.

The advantages of observed variable methods are that they typically do not require specialized software, can perform well in small samples, and sometimes make fewer assumptions than latent variable methods (R. D. Penfield & Camilli, 2007). The disadvantages of observed variables methods are that measurement error is not modeled and that the summed score may be an inappropriate proxy for the latent variable (Holland & Wainer, 1993; Millsap, 2011; R. D. Penfield & Camilli, 2007). The advantages of latent variable methods are that measurement error is modeled and more realistic assumptions about the latent variable distribution can be made. The disadvantages of latent variable methods are that they often require larger sample sizes and may require more specialized software (Millsap, 2011).

In both observed variable methods and latent variable methods the scales of the two groups must be aligned in order to test the items for DIF. There are many different ways to do this (see W. C. Wang, 2004 for an overview). The two main approaches discussed here are using all-other items as anchor (AOAA) items or designating a set of anchor items. Anchor items refer to those items which are known or presumed to be DIF free either by subject matter experts or more commonly by empirical analysis. Anchor items function to align the scales between the groups so that item parameters can meaningfully be compared for DIF testing.

The AOAA method involves assuming that all items other than the tested item are invariant and the tested item is evaluated for DIF. This process is repeated for each item on the test. The AOAA method has shown inflated Type I error rates when the anchor set is contaminated with DIF items and is not a recommended method (e.g. Finch, 2005; W. C. Wang, 2004; Woods, 2009a). The more accepted method for using anchor items is to empirically select the items using a rank based procedure (e.g. Woods, 2009a), purification procedure (e.g. W. C. Wang & Shih, 2010), or combination of both (e.g. W. C. Wang, Shih, & Sun, 2012). Once the anchor items are empirically selected then these items are used to link the scale across the groups.

1.2.2 Background on IRT

IRT consists of a set of latent variable models that define one way of establishing the correspondence between a latent variable and a categorical item response (e.g. De Ayala, 2009). Unidimensional IRT typically makes three assumptions: (1) responses to the manifest variables are accounted for by a single latent trait, (2) the data exhibit a pattern consistent with the model's specified form, and (3) conditional independence of item responses (D. Kim, De Ayala, Ferdous, & Nering, 2011). Assumption one is self-explanatory. The second assumption, the functional form assumption, states that the data should follow the form of the unidimensional IRT model (e.g. a logistic or normal ogive function). The third assumption states that once we condition on a single latent trait, the item responses are statistically independent. Another assumption made in the current study was that the latent trait is normally distributed, however, this assumption can be relaxed (e.g. Woods & Thissen, 2006).

There are many different types of IRT models for binary, ordinal, and nominal response data. In the present study only binary IRT models were considered. The most general binary IRT model is the three parameter logistic (3PL; Birnbaum, 1968)

$$P(Y_{ij} = 1|\theta_j) = c_i + \frac{(1 - c_i)}{1 + exp[-\alpha_i(\theta_j - \beta_i)]}, \tag{1.3}$$

where Y_{ij} is the binary response of person j ($j = 1, \ldots, J$) to item i ($i = 1, \ldots, I$), θ_j is the latent trait of person j, c_i is the pseudo-guessing parameter of item i, α_i is the discrimination parameter of item i and β_i is the difficulty parameter of the ith item. The two-parameter logistic (2PL) and one-parameter logistic (1PL) models are special cases of Equation 1.3. The 2PL model can be obtained by fixing c_i to zero for all items, the 1PL can be obtained from the 2PL by adding the constraint that α_i be the same for all items.

According to R. D. Penfield and Camilli (2007) there are two related interpretations of DIF within the IRT framework: (1) Between group differences in item characteristic curves (ICCs) and (2) between group differences in item parameters. The first interpretation of DIF is related

to the second, because if the item parameters differ across groups then the ICCs will also differ. In reviewing the existing unidimensional IRT DIF methods literature the first interpretation in relation to area measures of DIF will not be addressed in this dissertation. Area measures evaluate the amount of DIF based on calculating the area between the ICCs among various groups (Raju, 1989; 1990; Raju, Van der Linden, & Fleer, 1995; Rudner, Getson, & Knight, 1980) and while significance tests are available, they are more commonly used to evaluate effect sizes (e.g. Millsap, 2011).

1.2.3 Background on Bayesian Estimation

The treatment of Bayesian statistics within psychometrics is becoming more popular (e.g. Levy, 2009) but many applied researchers are still less familiar with this framework. A brief introduction to Bayesian estimation will be given here and further details may be found in Patz and Junker (1999a) and Patz and Junker (1999b) for IRT specifically or Gelman et al. (2013) for a general overview of Bayesian statistics. To estimate IRT models within a Bayesian framework, Baye's Theorem is applied as follows

$$p(\zeta|Y) = \frac{p(Y|\zeta)p(\zeta)}{p(Y)}, \tag{1.4}$$

where ζ is a vector containing the parameters of interest and Y is the data that has been observed. The posterior distribution is given as $p(\zeta|Y)$ and represents the probability of the parameters given the observed data. The likelihood is given by $p(Y|\zeta)$ and is identical to the likelihood used in maximum likelihood estimation, $p(\zeta)$ is the prior distribution of the parameters, and $p(Y)$ is the marginal likelihood. In practice $p(Y)$ can be ignored when computing the posterior and Equation 1.4 is written as

$$p(\zeta|Y) \propto p(Y|\zeta)p(\zeta), \tag{1.5}$$

indicating the posterior is proportional to the likelihood times the prior (see Gelman et al., 2013 for details). The two main differences between Equation 1.5 and maximum likelihood estimation are the specification of the prior distribution and maximum likelihood attempts to find a point

estimate while Bayesian estimates a posterior distribution.

Computation of Equation 1.5 for various IRT models can be accomplished by means of Markov Chain Monte Carlo (MCMC) methods (see Gelman et al., 2013; Jackman, 2009 for an overview). The benefits of the Bayesian framework are the extreme flexibility in estimating models that may not be practically viable in maximum likelihood or have not yet been implemented in available software. If researchers can specify the full probability model as in Equation 1.5 of the likelihood and prior(s) for their question of interest then it is possible to estimate their model in a Bayesian framework. Some of the existing methods in the literature have only been or may currently only be implemented in a Bayesian framework due to the lack of available software for maximum likelihood estimation. The disadvantages of the Bayesian framework using MCMC are the diligence and mathematical understanding of the researcher to monitor convergence and various model diagnostics, use of specialized software to fit complex models, and the slow computation time required (Gelman et al., 2013).

In what follows a review of unidimensional IRT-based DIF methods is given. Special emphasis will be placed on how these procedures may generalize beyond the two group case. The methods are likelihood ratio tests, Lagrange multiplier tests, Wald χ^2, latent classes over persons, latent classes over items, logistic mixed models, machine learning methods, and MIMIC models. This review will cover MIMIC models and the Wald χ^2 in somewhat more detail than other methods as these two methods were used in the current study.

Chapter 2

Review of the Literature

2.1 Review of Unidimensional IRT-Based DIF Methods

2.1.1 Likelihood Ratio (LR) DIF Testing

The LR DIF test was originally conceived by Thissen, Steinberg, and Gerrard (1986) and was further described in Thissen et al. (1988; 1993). Within the IRT literature the LR test for DIF is associated with IRTLRDIF but LR tests for DIF are more general than IRTLRDIF. IRTLRDIF is a software package implemented by David Thissen in 2001 that automates the process of many model fittings required to run the LR DIF tests in two groups (Thissen, 2001). The idea of LR tests for DIF is not specific for use in multiple group IRT models and can be used with a variety of procedures such as MIMIC models and logistic mixed models. Here LR tests for DIF are described in the context of multiple group IRT as implemented by Thissen (2001) but generalized to N groups. To illustrate the method the multiple group 2PL IRT Model is used

$$P(y_{ij} = 1|\theta_j) = \frac{1}{1 + exp[-\alpha_{ig}(\theta_j - \beta_{ig})]}, \tag{2.1}$$

where $g = 1 \ldots, G$ refers to the group, θ_j refers to the latent trait of person $j = 1, \ldots, J$ and α_{ig} and β_{ig} are the discrimination and threshold parameters for item $i = 1, \ldots, I$ in group g, respectively.

It is assumed that the reference group will be denoted as $g = 1$. Further $\theta_j|\lambda_{g(j)} \sim N(\mu_{g(j)}, \sigma^2_{g(j)})$, where $\lambda_g = (\mu_{g(j)}, \sigma^2_{g(j)})^T$ with $\lambda_1 = (\mu_1 = 0, \sigma^2_1 = 1)^T$ for the reference group to identify the scale. Note the notaion $g(j)$ refers to person j in group g.

Let two multiple group 2PL IRT models be denoted as $\mathcal{M}_0$ and $\mathcal{M}_1$. $\mathcal{M}_1$ has necessary constraints on item parameters across groups only to identify the model with all other item parameters being freely estimated. $\mathcal{M}_0$ adds certain constraints to $\mathcal{M}_1$ to achieve invariance across the groups for a given item or set of items. Note that $\mathcal{M}_0$ is nested within $\mathcal{M}_1$. Then the LR DIF statistic can be defined as

$$\chi^2_{LR} = -2 * ln\left(\frac{L_0}{L_1}\right), \tag{2.2}$$

where L_0 is the likelihood of model $\mathcal{M}_0$ and L_1 is the likelihood of model $\mathcal{M}_1$. Equation 2.2 is very general and can be applied to binary, ordinal, and nominal IRT models. The test statistic is asymptotically distributed as chi-square with the degrees of freedom equal to the number of constraints on $\mathcal{M}_1$ to obtain $\mathcal{M}_0$. The LR DIF test allows for both an omnibus test of the α_{ig} and β_{ig} parameters and individual tests for each parameter. To conduct these tests a series of nested models (i.e. $\mathcal{M}_0$ and $\mathcal{M}_1$) are conducted for each item and in the case of individual tests; each parameter as illustrated in Equation 2.2.

The procedure requires many model fittings and becomes cumbersome for a large number of groups. For example, assuming the 2PL model with three groups and 20 items tested for DIF this requires 20 + 1 = 21 model fittings for the omnibus test. More generally, it requires $|I_D| + 1$ model fittings where $|I_D|$ denotes the number of DIF items (i.e. non-anchor items). Assuming that omnibus DIF has already been tested the number of model fittings required to test for pairwise DIF in the 2PL model is $[2 * (G - 1) * |I_D|]$ (where G is the number of groups). For example if we have a 20 items tested for DIF ($|I_D| = 20$) and three groups, then this would involve $2*(2)*20 = 80$ model fittings to test for individual parameter DIF. Hence, this is probably the reason that the LR DIF test has not been well studied beyond two groups (see Woods, Cai, & Wang, 2013 for three groups). For cases involving two groups the LR DIF test has been shown to have high power and well controlled Type-I errors when the model and anchor sets are well specified (Finch, 2005;

Woods, 2009b; Woods et al., 2013). However, when moving beyond two groups the LR DIF testing procedure becomes cumbersome and does not scale well computationally when compared to other procedures such as the improved Wald χ^2 (Cai, 2015; Cai, Thissen, & du Toit, 2013; Langer, 2008) or the LM tests described below.

2.1.2 Lagrange Multipliers for DIF Testing

Lagrange Multiplier (LM) tests were first introduced by Silvey (1959) and are general to any maximum-likelihood based estimation model. Glas (1998; 1999) first introduced the notion of using LM tests for testing DIF. The idea of the LM test is to compare the fit of a general unrestrictive model to a restricted model that is a special case of the unrestrictive model. The LM test is based on computing the first-order partial derivatives with respect the the log likelihood of the general model evaluated at the maximum likelihood estimates obtained from the restricted model (Glas & Falcón, 2003). A special feature of the LM test is that only the restricted model is estimated to compute the statistic making the LM much more efficient than the LR test. The reason for this is that the first-order partial derivatives that are not restricted will be zero because they were solved using maximum likelihood estimation. However, the magnitudes of the first-order partial derivatives of the restrictions on the general model are only zero when the constraints hold exactly (Bollen, 1989; Glas & Falcón, 2003). Thus, the size of the first-order partial derivatives dictate the magnitude of the LM statistic with higher values indicating worse model fit.

To better illustrate the ideas assume $i = (1,\ldots,I)$ items, $j = (1,\ldots,J)$ persons, and $g = (1,\ldots,G)$ groups with the 2PL model be represented as

$$P(y_{ij} = 1|\theta_j) = \frac{1}{1 + exp\{-(\alpha_i + \delta_{ig(j)})[\theta_j - (\beta_i + \omega_{ig(j)})]\}}, \tag{2.3}$$

where $\delta_{ig(j)}$ and $\omega_{ig(j)}$ represent non-uniform and uniform DIF for item i with respect to person j in group g, respectively. Note that $\delta_{i1} = 0$ and $\omega_{i1} = 0$ indicating the reference group and

$g = (2, \ldots, G)$ the $(G-1)$ focal groups. Further, it is assumed that $\theta_j|\lambda_{g(j)} \sim N(\mu_{g(j)}, \sigma^2_{g(j)})$ with $g(j)$ indicating that participant j is in group g and $\lambda_g = (\mu_g, \sigma^2_g)$ $\forall g = (1, \ldots, 2)$. To identify the scale $\lambda_1 = (\mu_1 = 0, \sigma^2_1 = 1)$. Using notation from Glas and Falcón (2003) let the parameters of a general multiple group 2PL IRT be denoted by $\boldsymbol{\eta}$. The parameters of the restricted model are a subset of $\boldsymbol{\eta}$ by adding in the appropriate constraints. From this $\boldsymbol{\eta}$ can be partitioned into $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)^T$, where $\boldsymbol{\eta}_1$ is a vector of unrestricted parameters and $\boldsymbol{\eta}_2$ is a vector of restricted parameters placed on the general model. Here $\boldsymbol{\eta}_1 = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \lambda_2, \ldots, \lambda_G)^T$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)^T$ and $\boldsymbol{\eta}_2 = (\boldsymbol{\delta_1}, \ldots, \boldsymbol{\delta}_I, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_I)^T$ where $\boldsymbol{\delta}_i = (\delta_{i2}, \ldots, \delta_{ig})^T$ and $\boldsymbol{\omega}_i = (\omega_{i2}, \ldots, \omega_{iG})$. Under the null hypothesis that the restricted model holds then $\boldsymbol{\eta}_2 = \mathbf{0}$. Letting $\boldsymbol{h(\eta)}$ be the first-order partial derivatives of the unrestricted such that

$$\boldsymbol{h(\eta)} = \frac{\partial LogL(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}, \tag{2.4}$$

where $LogL(\boldsymbol{\eta})$ is the log likelihood of the unrestricted or general model. $\boldsymbol{h(\eta)}$ is also known as the score function (Rao, 1948) which gives the change in the log-likelihood for local changes in $\boldsymbol{\eta}$ (Glas & Falcón, 2003).

Letting the vector of partial derivatives $\boldsymbol{h(\eta)}$ be partitioned as $\boldsymbol{[h(\eta_1) = 0, h(\eta_2)]}^T$ then the LM statistic is given as

$$LM = \boldsymbol{h(\eta_2)}^T \Sigma^{-1} \boldsymbol{h(\eta_2)}. \tag{2.5}$$

In Equation 2.5 Σ is given as

$$\Sigma = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, \tag{2.6}$$

where

$$\Sigma_{pq} = \frac{\partial^2 LogL(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_p \partial \boldsymbol{\eta}_q^T}, \tag{2.7}$$

with $p = 1, 2$ and $q = 1, 2$. The specific details of computing these derivatives is beyond the scope of this paper and the interested reader is directed to Glas (1998; 1999) for details. The LM statistic in Equation 2.5 has an asymptotic χ^2 distribution with degrees of freedom equal to the number of

parameters in $\boldsymbol{\eta}_2$ (Silvey, 1959). In practice the LM statistic in Equation 2.5 is not typically used, instead the individual tests for each parameter are used. For the discrimination parameters the individual tests are given as

$$LM_{\alpha_{ig}} = h(\eta_{\alpha_{ig}})^T \Sigma^{-1}_{\alpha_{ig}} h(\eta_{\alpha_{ig}}) \tag{2.8}$$

and for the threshold parameters the individual tests are given by

$$LM_{\beta_{ig}} = h(\eta_{\beta_{ig}})^T \Sigma^{-1}_{\beta_{ig}} h(\eta_{\beta_{ig}}), \tag{2.9}$$

where the subscripts α_{ig} and β_{ig} refer to the discrimination and threshold parameters, respectively of item i of group g. The LM statistics in Equations 2.8 and 2.9 are asymptotically χ^2 distributed with one degree of freedom.

The advantages of using LM tests for DIF detection in two or more groups are that LM procedures only require fitting the restricted model to obtain estimates as compared to the LR or Wald χ^2 tests. Further, multiple aspects of model fit and assumptions in addition to DIF testing may also be specified Glas (1998; 1999). Several simulation studies have found the LM tests to perform well for the two group case when the proportion of DIF items was not too large (Glas, 1998; Glas, 1999; Khalid & Glas, 2014). However, the LM statistics have been shown to perform inadequately when the proportion of DIF items is large and/or the model being tested is grossly violated by the imposed constraints (Glas, 1999; Khalid & Glas, 2014). Another disadvantage of the LM tests are their lack of availability in current IRT software.

2.1.3 Wald Chi-Square Tests

The original idea of testing for DIF using Wald (Wald, 1943) χ^2 tests was from Lord (1980, pp. 212-224). Lord believed the way to detect DIF was to compare the ICCs or item response functions (IRFs) between the groups. He noted that it is difficult or impossible to determine whether a meaningful difference between IRFs of various groups exist just by looking at pictures. To better address this issue he proposed the Wald χ^2 for DIF testing. In Lord's 1980 original implementation

he noted three estimation steps to test for DIF in the 3PL model (see Equation 1.3):

1. Estimate the item parameters for all groups combined, standardizing on the difficulty parameters (β_i).
2. Fix the pseudo-guessing parameters (c_i) equal to the values obtained in step one for all groups, reestimate the α_i and β_i parameters separately in each group, and standardize on the β_i estimates.
3. For each item compare the item parameters (α_i and β_i) for differences across the groups using the χ^2 statistic given by Lord (1980) (see below).

Lord noted that the c_i parameters should be constrained equal between the groups and not tested given the known problems with estimation of the pseudo-guessing parameters. The implementation of Lord's 1980 Wald χ^2 statistic for comparing two groups is given by

$$\chi_i^2 = \boldsymbol{v}_i^T \Sigma_i^{-1} \boldsymbol{v}_i, \tag{2.10}$$

where $\boldsymbol{v}_i^T = [\hat{\alpha}_{iR} - \hat{\alpha}_{iF}, \hat{\beta}_{iR} - \hat{\beta}_{iF}]$ is the vector of the difference between the reference (R) and focal (F) group difficulty and discrimination parameters for item i and Σ_i^{-1} is the inverse of the asymptotic covariance matrix for $\hat{\alpha}_{iR} - \hat{\alpha}_{iF}$ and $\hat{\beta}_{iR} - \hat{\beta}_{iF}$. Note that the $\hat{}$ over the item parameters refers to the fact that these are maximum likelihood estimates. The contents of Σ_i for two groups are given as

$$\Sigma_i = \begin{pmatrix} var(\hat{\alpha}_{iR}) + var(\hat{\alpha}_{iF}) & cov(\hat{\alpha}_{iR}, \hat{\beta}_{iR}) + cov(\hat{\alpha}_{iF}, \hat{\beta}_{iF}) \\ cov(\hat{\alpha}_{iR}, \hat{\beta}_{iR}) + cov(\hat{\alpha}_{iF}, \hat{\beta}_{iF}) & var(\hat{\beta}_{iR}) + var(\hat{\beta}_{iF}). \end{pmatrix} \tag{2.11}$$

Asymptotically χ_i^2 follows a chi-square distribution with two degrees of freedom. Lord's (1980) Wald Chi-square requires less model fittings than IRTLR-DIF making it a much more computationally efficient procedure (Langer, 2008; Millsap, 2011).

As originally implemented Lord's 1980 Wald χ^2 has four main criticisms: (1) Use of joint maximum-likelihood (JML) estimation of person and item parameters versus a more modern

estimation such as marginal maximum likelihood (MML; Bock & Aitkin, 1981), (2) linking the groups based on standardizing the βs, (3) failure to test the pseudo-guessing parameter for DIF, and (4) the estimation of the asymptotic covariance matrix. The first criticism (use of JML) appears to have been resolved in the early to mid 1990s. When using JML to estimate person and item parameters the item parameter estimates may not be consistent and the item parameter estimates may not be asymptotically efficient (e.g. De Ayala, 2009). These two consequences undermine the asymptotic χ^2 distribution that Lord's 1980 Wald χ^2 test rests on for statistical inference in DIF analyses. McLaughlin and Drasgow (1987) conducted a simulation study using Lord's Wald χ^2 with JML estimation to compare the power and Type I error rates and found that Type I error inflation was up to 11 times above the nominal level. Both S. H. Kim and Cohen (1995) and Lim and Drasgow (1990) used MML for Lord's Wald χ^2 and found that Type I error rates were much improved over using JML.

The last three criticisms have been addressed more recently albeit the linking criticism (2) was partially addressed in the mid 1990s. S. H. Kim, Cohen, and Kim (1994) addressed criticism (2) by utilizing the Stocking and Lord (1983) method of equating to link the item parameters between groups. S. H. Kim et al. (1994) notes that this method has been shown to perform well and is recommended over standardizing on the β coefficients. More recently, Langer (2008) and Woods et al. (2013) recommended using designated anchor items to link the scales between the groups which tends to perform better on average than the Stocking and Lord (1983) method. The third and fourth criticisms (testing pseudo-guessing parameters and estimation of the asymptotic covariance matrix, respectively) were addressed by an improved version of the Wald-χ^2 test (Cai, 2015; Cai et al., 2013; Langer, 2008). The remainder of this section explains how Equation 2.10 can be generalized to more than two groups. Then a description of the improved version of Lord's Wald χ^2 test is given.

2.1.3.1 Lord's (1980) Wald Chi-Square in More than Two Groups

S. H. Kim, Cohen, and Park (1995) were the first to use Lord's (1980) Wald χ^2 test in more than two groups. They presented an illustrative example using real data. A generalization of Equation 2.10 is presented to more than two groups following the notation provided in S. H. Kim et al. (1995). The illustration presented by S. H. Kim et al. (1995) is extended by illustrating the 3PL model[1] (see Equation 1.3). First let $\boldsymbol{v}_i = (\hat{\alpha}_{i1}\ \hat{\beta}_{i1}\ \hat{c}_{i1} \cdots \hat{\alpha}_{ik}\ \hat{\beta}_{ik}\ \hat{c}_{ik})^T$ is a $3 * K$ by 1 dimensional vector where K is the number of groups ($k = 1, \ldots, K$). Let Σ_i by a $3 * K$ by $3 * K$ block diagonal matrix given by

$$\Sigma_i = \begin{pmatrix} \boldsymbol{\Phi}_{i1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{i2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Phi}_{\boldsymbol{ik}}, \end{pmatrix} \tag{2.12}$$

where $\boldsymbol{\Phi}_{ik}$ is given as

$$\boldsymbol{\Phi}_{ik} = \begin{pmatrix} var(\hat{\alpha}_{ik}) & cov(\hat{\alpha}_{ik}, \hat{\beta}_{ik}) & cov(\hat{\alpha}_{ik}, \hat{c}_{ik}) \\ cov(\hat{\alpha}_{ik}, \hat{\beta}_{ik}) & var(\hat{\beta}_{ik}) & cov(\hat{\beta}_{ik}, \hat{c}_{ik}) \\ cov(\hat{\alpha}_{ik}, \hat{c}_{ik}) & cov(\hat{\beta}_{ik}, \hat{c}_{ik}) & var(\hat{c}_{ik}), \end{pmatrix} \tag{2.13}$$

for item i and group k. Note each $\mathbf{0}$ in Σ_i is a 3 x 3 null matrix.

In order to make comparisons across multiple groups a contrast matrix $\boldsymbol{C}$ is introduced such that $\boldsymbol{C}$ has p rows which contain contrast vectors and $3 * K$ columns (three refers to the number of parameters tested). Here p is the rank of $\boldsymbol{C}$ and in the present example for the 3PL model $p = (3 * K - 3)$. Thus, $\boldsymbol{C}$ is a $(3 * K - 3)$ by $3 * K$ contrast matrix. For the case of K groups the

[1] Testing for DIF in the pseudo-guessing parameter (c_i) should be done with great care. Typically a very large sample size is needed or the use of prior distributions is recommended (see Langer, 2008 for details)

contrast matrix is given by

$$\boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & \ddots & \cdots & \cdots & 0 \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & -1 & 0 & 0 \\ 0 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 0 \\ 0 & 0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 & -1 \end{bmatrix}. \tag{2.14}$$

The first three columns of the contrast matrix refer to $\hat{\alpha}_{i1}$, $\hat{\beta}_{i1}$, and $\hat{c}_{i1}$, respectively. In the first row of the matrix the fourth column refers to $\hat{\alpha}_{i2}$, in the second row the fifth column refers to $\hat{\beta}_{i2}$, and in the third row the sixth column refers to $\hat{c}_{i2}$. In the third to last row of the contrast matrix the -1 refers to $\hat{\alpha}_{iK}$ parameter, in the second to last row the -1 refers to $\hat{\beta}_{iK}$, and in the last row the -1 refers to the $\hat{c}_{iK}$ parameter.

S. H. Kim et al. (1995) define a test statistic they call Q_i that is used to test for multiple group DIF. Keeping with S. H. Kim et al. (1995) testing for DIF in K groups for the 3PL model using $\boldsymbol{v}_i$ and elements from Equations 2.12, 2.13, and 2.14 the multiple group DIF statistic for item i is given by

$$Q_i = (\boldsymbol{C}\boldsymbol{v}_i)^T(\boldsymbol{C}\Sigma_i\boldsymbol{C}^T)^{-1}(\boldsymbol{C}\boldsymbol{v}_i). \tag{2.15}$$

Equation 2.15 is asymptotically chi-square distributed with $(3 * K - 3)$ degrees of freedom under the null hypothesis that there is no DIF for any of the item parameters across the K groups for item i. One thing to note is that in Equation 2.15 $\boldsymbol{C}\boldsymbol{v}_i = (\hat{\alpha}_{i1} - \hat{\alpha}_{i2}\ \hat{\beta}_{i1} - \hat{\beta}_{i2}\ \hat{c}_{i1} - \hat{c}_{i2} \cdots \hat{\alpha}_{i1} - \hat{\alpha}_{iK}\ \hat{\beta}_{i1} - \hat{\beta}_{iK}\ \hat{c}_{i1} - \hat{c}_{iK})^T$, which is an omnibus test of DIF for all parameters of the 3PL model with respect to groups. The omnibus test refers to a difference among the parameters of the 3PL model for a reference group ($k = 1$) against each of the focal groups ($k = 2, \ldots, K$). Practitioners may

also be interested in specific pairwise or other types of complex contrasts between the groups. These types of contrasts can be accommodated by specifying a different contrast matrix C for the problem of interest (see Cai, 2015; S. H. Kim et al., 1995; Langer, 2008 for details).

Equation 2.15 shows the flexibility of testing for DIF in more than two groups is straightforward. This method requires less model fittings than a likelihood ratio based DIF testing approach (Langer, 2008; Woods et al., 2013). The method also can be adapted to conduct pairwise and complex contrasts such as an interaction between two groups (S. H. Kim et al., 1995; Langer, 2008). However, Equation 2.15 as described in S. H. Kim et al. (1995) still suffers from issues related to estimation of Σ and use of an ad-hoc linking method of Stocking and Lord (1983). In the next section improvements to Lord's (1980) Wald χ^2 are described that address these issues.

2.1.3.2 Improved Wald Chi-square DIF Test

Langer (2008) describes three shortcomings of Lord's (1980) Wald χ^2 test: (1) estimation of the asymptotic covariance matrix, (2) ad-hoc linking procedure, and (3) allowing for tests of the pseudo-guessing parameter (c_i). Langer (2008) notes that the major pitfall with Lord's (1980) Wald χ^2 test is estimation of the asymptotic variance covariance matrix. This issue was also raised by Millsap (2011), and Thissen and Wainer (1982). To circumvent this problem Lord's (1980) Wald χ^2 was improved by use the supplemental expectation-maximization (Cai, 2008; Meng & Rubin, 1991) algorithm to obtain the asymptotic variance covariance matrix of the item parameter estimates (Σ_i from Equation 2.15) (Cai, 2015; Cai et al., 2013; Langer, 2008). The supplemental expectation-maximization algorithm has been shown to provide a convenient computational procedure for latent variable models such as IRT and categorical confirmatory factor analysis for estimating the information matrix of item parameters (Cai, 2008). This allows for more accurate standard error estimates of the item parameters than has been previously used with various implementations of Lord's (1980) Wald χ^2 (S. H. Kim et al., 1994; S. H. Kim et al., 1995; Lim & Drasgow, 1990).

The linking procedure as originally implemented by Lord (1980) standardized on the location

parameters (βs). S. H. Kim et al. (1994) and S. H. Kim et al. (1995) noted the use of Stocking and Lord (1983) to link had been shown to perform better than the standardizing of the location parameters as proposed by Lord (1980). However, although an improvement over standardizing on the βs, the Stocking and Lord (1983) method is an ad-hoc linking method that placed participants estimates on the same scale for comparison. As an improvement to the method of Stocking and Lord (1983), Langer (2008) suggested the use of concurrent calibration, an IRT-based linking procedure to anchor groups to a common metric for DIF testing. The main benefit of concurrent calibration is that a participants ability estimate is independent of the set of items that the participant answers. In addition, concurrent calibration provides conversions of parameters that are independent of the group or groups to obtain them and allows for greater accuracy of linking along the entire score scale (Langer, 2008).

Langer (2008) mentioned two means of linking groups for testing DIF with the Wald χ^2. The first was to designate anchor items known a priori or by empirical selection using a method such as W. C. Wang and Shih (2010), W. C. Wang et al. (2012), or Woods (2009a). The second method involves a two-stage method. The first stage constrains the item parameters equal in all groups and estimates the population means and standard deviations of the focal groups relative to the reference group. In the second stage the estimates of the population means and standard deviations of the focal groups are fixed to those obtained in stage one and the item parameters are allowed to differ for DIF detection. These estimates are then tested for statistical significance using the same framework as in Equation 2.15.

Lord (1980) mentioned that the pseudo-guessing parameter (c_i) should be constrained equal between the groups and was not directly tested in his original implementation. Previous researchers have also not extended the Wald χ^2 DIF test to accommodate the pseudo-guessing parameter (S. H. Kim et al., 1994; S. H. Kim et al., 1995; Lim & Drasgow, 1990). Langer (2008) proposed the use of conditional tests of DIF to deal with the case of testing for all item parameters in the 3PL model. These two group tests can be easily generalized to pairwise tests in more than two groups. Here group $k = 1$ is the reference group and group $k = 2$ is the focal group The

unconditional DIF test of the c_i parameter for the 3PL model is given by

$$\chi^2_{c_i} = \frac{(\hat{c}_{i1} - \hat{c}_{i2})^2}{(\hat{\sigma}^2_{c_{i_1}} + \hat{\sigma}^2_{c_{i2}})}, \tag{2.16}$$

where $\hat{\sigma}^2_{c_ik}$ is the variance for the c parameter for item i in group k. $\chi^2_{c_i}$ is asymptotically chi-square distributed with 1 degree of freedom. The conditional test of the α_i parameter given an equal c_i parameter is given by

$$\chi^2_{\alpha_i|c_i} = \frac{y_i^2}{\hat{\sigma}^2_{\alpha_i|c_i}}, \tag{2.17}$$

where

$$y_i = (\hat{\alpha}_{i1} - \hat{\alpha}_{i2}) - \frac{(\hat{\rho}_{\alpha_{i1}c_{i1}}\hat{\sigma}_{\hat{\alpha}_{i1}}\hat{\sigma}_{c_{i1}} + \hat{\rho}_{\alpha_{i2}c_{i2}}\hat{\sigma}_{\alpha_{i2}}\hat{\sigma}_{c_{i2}})}{(\hat{\sigma}^2_{c_{i1}} + \hat{\sigma}^2_{c_{i2}})}(\hat{c}_{i1} - \hat{c}_{i2}) \tag{2.18}$$

and

$$\hat{\sigma}^2_{\alpha_i|c_i} = (\hat{\sigma}^2_{\alpha_{i1}} + \hat{\sigma}^2_{\alpha_{i2}}) - \frac{(\hat{\rho}^2_{\alpha_{i1}c_{i1}}\hat{\sigma}^2_{\alpha_{i1}}\hat{\sigma}^2_{c_{i1}} + \hat{\rho}^2_{\alpha_{i2}c_{i2}}\hat{\sigma}^2_{\alpha_{i2}}\hat{\sigma}^2_{c_{i2}})}{(\hat{\sigma}^2_{c_{i1}} + \hat{\sigma}^2_{c_{i2}})}, \tag{2.19}$$

where $\rho_{..}$ refers to the correlation between the respective item parameters for a given group. $\chi^2_{\alpha_i|c_i}$ is asymptotically distributed as chi-square with 1 degree of freedom. Lastly, the conditional test of DIF for the β_i parameters given equal α_i and c_i parameters is simply the difference between the overall χ^2_i test for two groups and the sum of the unconditional test of c_i and the conditional test of $\alpha_i|c_i$ given as

$$\chi^2_{\beta_i|\alpha_i,c_i} = \chi^2_i - [\chi^2_{c_i} + \chi^2_{\alpha_i|c_i}]. \tag{2.20}$$

Note $\chi^2_i = \boldsymbol{v}_i^T \Sigma_i^{-1} \boldsymbol{v}_i$, where $\boldsymbol{v}_i = (\hat{\alpha}_{i1} - \hat{\alpha}_{i2}\ \hat{\beta}_{i1} - \hat{\beta}_{i2}\ \hat{c}_{i1} - \hat{c}_{i2})^T$ and Σ_i is the asymptotic covariance matrix between the parameters of the two groups.

Langer's (2008) improved Wald χ^2 test was shown to perform well in simulation studies against IRTLR-DIF (Langer, 2008; Woods et al., 2013). Both Langer (2008) and Woods et al. (2013) found that the improved Wald χ^2 had adequate power and well-controlled Type I error rates. Further, the improved Wald χ^2 test was almost as powerful as IRTLR-DIF in simulations by Langer (2008) and Woods et al. (2013) found in certain situations the improved Wald χ^2 was as powerful as IRTLR-DIF. Both studies mention that the improved Wald χ^2 test is much more computationally

efficient especially when testing for more than two groups. Further, the improved Wald χ^2 allows for more complex contrasts that are not currently feasible with implementations of IRTLR-DIF.

2.1.4 Latent Class DIF Models Over Persons

DIF testing can be divided into two types: manifest groups and latent groups. Examples of manifest group DIF testing are DIF with respect to gender and/or ethnicity. The implicit assumption of using manifest groups for DIF testing is that participants in a given manifest group are assumed to behave or respond in a similar manner compared to participants in another manifest group. For example, when testing for DIF related to gender it is assumed that all males share the same IRT model and in turn all females share the same IRT model. If for example we have two manifest groups and there are two latent classes then if the latent classes and manifest groups perfectly overlap this assumption makes sense. Perfect overlap means all members of manifest group one belong to one latent class and all members of manifest group two belong to the other latent class. However, if this is not true then bias can occur (e.g. De Ayala, Kim, Stapleton, & Dayton, 2003).

The assumption of perfect overlap is probably suspect, because it asserts that all members of a given manifest group are more similar than members of another manifest group. De Ayala et al. (2003) give an example citing that some African Americans may come from families that have lived in the United States for 100 or more years and other African Americans may come from families that could have just moved from Africa. In this case it is clear that these two subgroups may not share the same IRT model. Further, De Ayala et al. (2003) and Cohen and Bolt (2005) argue that the manifest approach does not provide information about underlying causes for DIF. For example when testing for DIF using latent class analysis and manifest methods with respect to gender Cohen and Bolt (2005) found that some members of each gender group responded differently than other members even though gender DIF (manifest DIF) was identified on some items. In addition, De Ayala et al. (2003) found in a Monte Carlo study with two manifest groups and two latent classes that when 50% of manifest group one were in latent class one and 50% of manifest group two were in latent class two the power to detect DIF decreased. To address these

concerns the use of latent class models for DIF can be used.

To illustrate the ideas of a latent class over person DIF model the Bayesian approach was chosen for this illustration given the straight-foward testing of DIF across multiple latent classes and the ability to easily generalize the model to more complicated IRT models. This paradigm is an extension of Samuelsen (2005) for the 2PL model. Let $i = (1, \ldots, I)$ indicate the number of items, $j = (1, \ldots, J)$ indicate persons, and $g = (1, \ldots, G)$ indicate latent classes. The latent class 2PL DIF model is given as

$$P(y_{ij} = 1|g, \theta_{jg}) = \frac{1}{1 + exp[-\alpha_{ig}(\theta_{jg} - \beta_{ig})]}, \tag{2.21}$$

where α_{ig} is the ith item's discrimination parameter in latent class g, β_{ig} is the ith item's threshold parameter in latent class g, and θ_{jg} is the latent trait for person j in latent class g. The marginal probability of a correct response for person j and item i is given by

$$P(y_{ij}) = \sum_{g=1}^{G} \pi_g P(y_{ij}|g, \theta_{jg}), \tag{2.22}$$

where π_g is the mixture probability for latent class g and $\sum_{g=1}^{G} \pi_g = 1$.

Prior distributions for the parameters in Equations 2.21 and 2.22 must be specified as part of the Bayesian analysis. An overview of some priors that could be used is given in Cho and Cohen (2010). Note that π_g can be given a prior or can be predicted using a multinomial logistic regression (Bilir, 2009; Cho & Cohen, 2010). When a multinomial logistic regression is used to predict π_g then manifest covariates can be included that help explain why individuals are classified into a given latent class. The addition of manifest covariates can assist with explanation of DIF. Assuming the scale has been anchored via anchor items or with appropriate constraints testing for DIF involves defining a new parameter that is the difference between the item parameters. For example in the case of thresholds $\beta_{i1} - \beta_{i2}, \beta_{i1} - \beta_{i2}, \ldots, \beta_{i1} - \beta_{iG}$. This same process can be repeated for the discrimination parameters to obtain the DIF statistics. DIF can be assessed using the 95 % credible intervals and/or highest posterior density intervals by assessing whether the difference

between the parameters for the reference and focal groups contain zero (Cho & Cohen, 2010).

While latent class DIF models provide new and exciting insights into DIF testing there are several issues that must be addressed when using them in practice. First, it is necessary to select the number of latent classes as it is unlikely to be known a priori. This is typically accomplished using the information criteria: Akaike information criterion (AIC; Akaike, 1974) and/or the Bayesian information criterion (BIC; Schwarz, 1978) (Bilir, 2009; Cho & Cohen, 2010; Cohen & Bolt, 2005; Samuelsen, 2005). The second issue with using latent class models for DIF detection involves dealing with the issue of label switching. According to Cho (2007) there are two types of label switching researchers should be aware of: (1) The latent class labels or membership switch based on different initial values and (2) only relevant to Bayesian estimation, the labels of latent classes switching across iterations within a specific Markov chain. It is important that practitioners check for these issues when running latent class DIF models (recommendations can be found in Cho, 2007; Cho & Cohen, 2010).

A final issue practitioners should be aware of involves anchoring the scale for DIF testing. According to Cho (2007) finding anchor items in latent class DIF models is a difficult problem that needs additional research. In typical applications the following constraint is used for anchoring the thresholds: $\sum_{i=1}^{I} \beta_{ig} = 0$. This indicates that the threshold parameters in a given latent class must sum to 0 (Bilir, 2009; Cho & Cohen, 2010; Samuelsen, 2005). In order to link the discrimination parameters it is possible to use the following constraint: $\prod_{i=1}^{I} \alpha_{ig} = 1$. This constraint has been used in other types of DIF testing across multiple groups as a way to identify the α_{ig} parameters (e.g. De Jong & Steenkamp, 2010; Verhagen & Fox, 2013). The implications of these anchor methods to cases of asymmetric and symmetric DIF have not been fully tested in latent class DIF models. However, borrowing from simulation research in the manifest DIF testing literature W. C. Wang (2004) found that when using the $\sum_{i=1}^{I} \beta_{ig} = 0$ constraint in the presence of asymmetric DIF Type I error rates were inflated. It is likely this would occur in the latent class DIF case but this is an open question for future research.

2.1.5 Latent Class DIF Models over Item Parameters

Most latent class DIF analyses focus on the latent class being defined on persons, however, recently several researchers have looked at DIF detection by defining the latent classes over items (e.g. De Boeck, 2008; De Jong & Steenkamp, 2010; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010; Goncalves, Gamerman, & Soares, 2013; Soares et al., 2009). The ideas of DIF latent class models over persons described above apply to latent classes over items. The difference is that with DIF latent class models over items the persons are manifest but the item classes are latent. In the simplest case items would be classified into either a DIF class or non-DIF class. More complex cases can be constructed by having multiple latent classes to compare each group with a reference group.

At the present there are three main variations that have been proposed and studied for DIF latent class models on item parameters. DIF latent class model detection methods proposed by Frederickx et al. (2010) and Soares et al. (2009) address the idea of testing for DIF amongst multiple groups directly, while De Jong and Steenkamp (2010) proposed a model that models the pertinent non-invariance that may be present across a large number of groups. The method by De Jong and Steenkamp (2010) is designed to permit non-invariant items to be modeled so that latent constructs and their parameters (e.g. latent means) can be compared across multiple groups as in cross-cultural research.

At the present time the method of De Jong and Steenkamp (2010) is not designed to test for DIF amongst multiple groups, however, adopting a procedure such as that described in Verhagen and Fox (2013) could allow this to be done. For these reasons, the present section only covers the ideas by Frederickx et al. (2010) and Soares et al. (2009) in detail and directs interested readers to De Jong and Steenkamp (2010) for a more detailed account of this method. Further, only the Soares et al. (2009) method is described in detail here for the following reasons. First, Goncalves et al. (2013) compared a model based on Soares et al. (2009) and found that the Soares et al. (2009) performed better than Frederickx et al. (2010). Second, the basic idea of item mixtures is similar in the two cases. Third, the implementation by Soares et al. (2009) is more general than the

Frederickx et al. (2010) model.

2.1.5.1 Soares et al. (2009) Model

The model proposed by Soares et al. (2009) and further improved by Goncalves et al. (2013) gives a DIF representation of the 3PL IRT model in a Bayesian framework in contrast to Frederickx et al. (2010) who only demonstrate their model in the case of the Rasch model. The notation from Soares et al. (2009) is used to illustrate the model. Letting y_{ij} be the binary response to item i by person j, then letting $P(y_{ij} = 1) = p_{ij}$ and $\Delta_{ij} = logit(p_{ij})$. Then $p_{ij} = logit^{-1}(\Delta_{ij}) = ln(\frac{1}{1+exp[-\Delta_{ij}]})$. From the previous notation the $P(y_{ij} = 1|\theta_j, \alpha_{ig}, \beta_{ig}, c_{ig})$ is given as

$$P(y_{ij} = 1|\theta_j, \alpha_{ig}, \beta_{ig}, c_{ig}) = c_{ig} + (1 - c_{ig}) * logit^{-1}(\Delta_{ij}), \tag{2.23}$$

where $\Delta_{ij} = D\alpha_{ig}(\theta_j - \beta_{ig})$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $g = 1, \ldots, G$ where g denotes the manifest group of a participant. To separate the DIF items from the non-DIF items the threshold parameters are represented as $\beta_{ig} = \beta_i - d_{ig}^{\beta}$, the discrimination parameter as $\alpha_{ig} = \alpha_i * exp(-d_{ig}^{\alpha})$, and pseudo-guessing parameters (c_{ig}) as $c_{ig} = c_i (\in [0, 1]), \forall g$. Although theoretically possible that DIF can be tested in the c_{ig} parameters this was not done in Soares et al. (2009) and Goncalves et al. (2013) based on the known estimation difficulties of the pseudo-guessing parameters.

From the above notation d_{ig}^{h} is the DIF parameter for the threshold and discrimination parameters for $h = \beta, \alpha$, respectively. For identification $d_{i1}^{h} = 0$ for α and β with $g = 1$ to denote the reference group. When $d_{ig}^{h} \neq 0$ for $g = 2, \ldots, G$ and $h = \beta, \alpha$ then this is indicative of DIF in the item parameters for the thresholds and/or discrimination parameters. It is also assumed a priori that $\theta_j|\lambda_{g(j)} \sim N(\mu_{g(j)}, \sigma^2_{g(j)})$, where $g(j)$ is the group participant j belongs to and $\lambda_g = (\mu_{g(j)}, \sigma^2_{g(j)})$. To identify the model for the reference group ($g = 1$) $\lambda_g = (\mu_1, \sigma_1^2) = (0, 1)$. The means and variances of the focal group(s) are unknown (λ_g for $g = 2, \ldots, G$) and must be estimated. Letting N be a normal distribution, LN be a log-normal distribution, and IG be the inverse gamma distribution

then the priors of the structural parameters given by Soares et al. (2009) are

$$\alpha_i \sim LN(\mu_{\alpha_i}, \sigma^2_{\alpha_i}),\ \beta_i \sim N(\mu_{\beta_i}, \sigma^2_{\beta_i}),\ and c_i \sim Beta(a_{c_i}, b_{c_i}),\ \text{for } i = 1, \ldots, I. \tag{2.24}$$

The latent trait distribution priors given in Soares et al. (2009) are

$$\mu_g \sim N(\mu_{0g}, \sigma^2_{0g}) \text{ and } \sigma_g \sim IG(a_g, b_g)\ \forall g = 1, \ldots, G. \tag{2.25}$$

In order to perform DIF analyses the reference and focal group(s) must be linked. Soares et al. (2009) suggest that if a set of items can be determined a priori then the d^h_{ig}s can be set to zero for those items. It is also possible to have the model select a set of anchor items in the estimation (see Goncalves et al. (2013); Soares et al. (2009) for details). The mixture model is formed by having a latent indicator variable Z^h_{ig} for $h = \alpha, \beta$, item i, and group g. If $Z^h_{ig} = 1$ then item i shows DIF in group g for parameter h and $Z^h_{ig} = 0$ otherwise. If Z^h_{ig} is fixed a priori then this involves fixing a set of anchor items otherwise Z^h_{ig} is given a prior distribution in order to select a set of anchor items during model estimation. In either case when $Z^h_{ig} = 1$ a regression structure is applied to allow explanation of the potential sources of DIF. The regression structure is given as

$$d^h_{ig} = \gamma^h_{0g} + \sum_{k=1}^{K_h} \gamma^h_{kg} W^h_{ik} + \eta^h_{ig},\ \text{if } Z^h_{ig} = 1, \tag{2.26}$$

where γ^h_{ig} refer to the regression coefficients for item i in group g for parameter h, $k = 1, \ldots, K_h$ refers to the index of the kth covariate of item parameter h (α, β), W^h_{ik} refers the kth covariate for item i and parameter h, and η^h_{ig} is the item specific random factor for group g and parameter h. If $d^h_{ig} = 0$ then $Z^h_{ig} = 0$. The prior for $\eta^h_g \sim N(0, T_g)$, where $T_g = (\tau^h_g)^2 * I$ for all $g = 1, \ldots, G$. The regression structure of Equation 2.26 pertains to all items deemed to show DIF.

Letting $\gamma^h_g = (\gamma^h_{0g}, \ldots, \gamma^h_{K^h g})^T$ and $W^h_i = (1, W^h_{i1}, \ldots, W_{iK^h})^T$ when $Z^h_{ig} = 1$ the conditional distribution of d^h_{ig} is given as $d^h_{ig} | \gamma^h_g, W^h_i, (\tau^h_g)^2 \sim N(W^h_i \gamma^h_g, (\tau^h_g)^2)$. In the case when $Z^h_{ig} = 0$ Soares et al. (2009) use the idea of stochastic search and variable selection (SSVS) proposed by George

and McCulloch (1993). The idea of SSVS is to concentrate the coefficient around zero by reducing the variance of a prior on $d_{ig}^h | Z_{ig} = 0, (\tau_g^h)^2 \sim N(0, s_i^2(\tau_g^h)^2)$, where s_i^2 is chosen to be small enough to ensure that d_{ig}^h is concentrated around zero. The conditional distribution of $d_{ig}^h | \gamma_g^h, W_i^h, Z_{ig}^h, (\tau_g^h)$ is given as follows

$$d_{ig}^h | \gamma_g^h, W_i^h, Z_{ig}^h, (\tau_g^h) \sim N\{(W_i^h \gamma_g^h) Z_{ig}^h, [s_i^2]^{1-Z_{ig}^h} (\tau_g^h)^2\}. \tag{2.27}$$

Soares et al. (2009) recommend the following priors for the parameters in Equation 2.27: $\gamma_g^h \sim N(\gamma_0^h, S_0^h)$, $(\tau_g^h)^2 \sim IG(a_g^h, b_g^h)$, and $Z_{ig}^h \sim Ber(\pi_{ig}^h)$, where Ber is the Bernoulli distribution.

The item mixture model of Soares et al. (2009) is quite complex and very general. This model allows for the simultaneous detection and explanation of DIF in a single model as opposed to a two step process. An anchor set does not have to be specified a priori and anchors can be selected within the estimation of the model. Soares et al. (2009) noted that the simulation study showed how the model had good parameter recovery and that the empirical example showed the viability of the model in practical situations. Goncalves et al. (2013) compared a variation of the Soares et al. (2009) against the model proposed by Frederickx et al. (2010) and noted that the Soares et al. (2009) performed better than the Frederickx et al. (2010) model. At the present time this model is not available in mainstream software such as Mplus, Stata, or R and can be fit by careful programming in JAGS, WinBUGS, OpenBUGS, or Stan.

2.1.6 Logistic Mixed Model Methods

It has been shown that IRT models can be parameterized as logistic mixed models (LMM) (Adams, Wilson, & Wu, 1997; Kamata, 2001; Mellenbergh, 1994). The connection of IRT with LMM allows for a number of ways to test for differential item functioning (DIF) in two or more groups and in some models to help explain the source of DIF. LMM have many uses and their utility extends beyond DIF testing (see De Boeck & Wilson, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003 for details). Only the case of categorical predictor variables for DIF testing is considered

here, but the LMM can easily incorporate continuous predictors as well. Noortgate and De Boeck (2005) outline a taxonomy of four types of DIF models that can easily accomodate more than two groups: (1) Fixed items and fixed groups; (2) Random items and fixed groups; (3) Fixed items and random groups; (4) Random items and random groups. First, a brief discussion of parameterizing an IRT model as an LMM is given, followed by an explanation of the taxonomy of DIF models mentioned above, and last a brief overview of estimation methods. In all cases the binary Rasch model was used for illustrative purposes in this section. The models can be extended to the 2PL and GRM and the interested reader is directed to De Boeck and Wilson (2004) for details.

2.1.6.1 IRT Model as a LMM

This section illustrates the parameterization of a Rasch IRT model as a LMM. Let the responses to items be independent and identically distributed (iid) Bernoulli trials and assume persons are independent. The Rasch IRT model is given as

$$P(y_{ij} = 1|\theta) = \frac{1}{1 + exp[-(\theta_j - \beta_i)]}, \tag{2.28}$$

where y_{ij} is the response of item i by person j, θ_j is the ability of person j, and β_i is the difficulty of item i. The nesting structure of a LMM is responses nested in persons unless otherwise noted. Assuming $j = 1, \ldots, n$ for the jth person and $i = 1, \ldots, I$ for the ith item, then the probability π_{ij} for item i and person j can be modeled as

$$\begin{aligned} log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) &= \eta_{ij} \\ \pi_{ij} &= logit^{-1}(\eta_{ij}), \end{aligned}$$

where

$$\eta_{ij} = u_j + \sum_{k=1}^{I} \beta_k X_{ki}. \tag{2.29}$$

X is an $(I*n)$ by I design matrix such that $X_{ki} = 1$ when $k = i$ and 0 otherwise and β_k equals minus the difficulty of the kth item. Further, u_j is equivalent to θ_j or the ability estimate of person j as defined in Equation 2.28. Equation 2.29 represents level one of the hierarchy which are responses. Level two is represented as $u_j \sim N(0, \sigma_\theta^2)$ with the reduced form equation given as

$$\eta_{kj} = u_j + \beta_{k0} \; \forall k \in \{1, \ldots, I\}. \tag{2.30}$$

Equation 2.30 gives the logit of the probability of a 1 response for the kth item with β_k being minus the difficulty of item k and u_j the ability of person j.

2.1.6.2 Fixed Items Fixed Groups DIF Models

In the fixed items and fixed groups (FIFG) framework both the item parameters and grouping covariates are assumed fixed across persons as is traditionally done in DIF testing with Wald Chi-square (Langer, 2008; Lord, 1980), IRTLRDIF (Thissen et al., 1993), and most MIMIC models (B. O. Muthén, 1985; Woods, 2009b). The FIFG DIF model adds an item by group interaction term for each item being tested for DIF to the model in Equation 2.29. The model is defined as

$$\eta_{ij} = u_j + \sum_{k=1}^{I} \beta_k X_{ki} + \sum_{h=2}^{H} \alpha_h G_{hj} + \sum_{h=2}^{H} \gamma_{kh} G_{hj} X_{ki}, \tag{2.31}$$

where $G_{hj} = 1$ if person j belongs to group h and 0 otherwise; α_h is the difference between focal group h and the reference group; and γ_{kh} is the kth specific item effect of belonging to focal group h compared with the reference group. Note that $h = 1, \ldots, H$, where H is the total number of groups and $h = 1$ corresponds to the reference group. The γ_{kh} coefficients in Equation 2.31 indicate the DIF between focal group h and the reference group on item k. These coefficients can be tested for statistical significance to determine if item k exhibits DIF for focal group h using a t-test or likelihood ratio (LR) test (Noortgate & De Boeck, 2005). Further, exponentiating the γ_{kh} coefficients provides an effect size measure based on odds ratios and can be interpreted in a similar way as the Mantel-Haenszel (MH; Holland & Thayer, 1988; Mantel & Haenszel, 1959)

statistic.

According to Noortgate and De Boeck (2005) there are two main criticisms of the model: Capitalization on chance and estimation of a large number of parameters. Given that each item suspected of DIF in Equation 2.31 includes a person by item interaction term there are a large number of statistical tests which can inflate Type I error rates. This phenomenon is consistent with previous methods and is inherent in the nature of DIF testing which is predominately an exploratory procedure. To address this criticism a p-value correction such as the Bonferroni or Benjamini-Hochberg (BH; Benjamini & Hochberg, 1995) could be used (see Thissen, Steinberg, & Kuang, 2002 for applications to DIF testing). The second problem of estimating a large number of parameters is a tougher problem. Although it is straightforward to extend the model in Equation 2.31 to a large number of groups each additional group beyond two adds $I + 1$ parameters to the model. Thus, a large sample size is needed to obtain precise estimates of the parameters when there are a myriad of parameters. Given the highly parameterized nature of Equation 2.31 when adding many groups, it is not possible to explain the sources of DIF in a single model. However, it is possible to conduct a second analysis and add in covariates to explain the potential sources of DIF.

2.1.6.3 Random Items and Fixed Groups DIF Models

To address the criticisms of the FIFG DIF model from the previous section Noortgate and De Boeck (2005) introduce random item effects to assess DIF. The random item fixed group (RIFG) DIF model is a cross-classified random effect model (Noortgate, De Boeck, & Meulders, 2003). Unlike the FIFG DIF model the RIFG DIF model has responses nested within persons and items. Because responses are nested within both persons and items the person and item random effects are crossed at level two. It is important to note that the RIFG DIF model is appropriate when items on the instrument can be considered a random sample from a population of items and primary interest is not in the specific items but in the category/population they represent (Noortgate & De Boeck, 2005).

The RIFG model is more parsimonious than Equation 2.31, because only the parameters of the random effects are estimated instead of the individual main effects and person by item interactions of each item. This allows for inclusion of item property covariates to help explain the potential sources of DIF (Noortgate et al., 2003; Noortgate & De Boeck, 2005). The RIFG DIF model is given as

$$\eta_{ij} = u_j + \beta_0 + r_{0i} + \sum_{h=2}^{H} \alpha_h G_{hj} + \sum_{h=2}^{H} r_{hi} G_{hj}, \tag{2.32}$$

where β_0 is the expected negative difficulty of an average item in the reference group, r_{0i} is the random main effect of item i, r_{hi} is the random ith specific effect of belonging to group h (this is the DIF effect), r_{0i} and r_{hi} follow a multivariate normal distribution with mean vector 0 and a covariance matrix, and α_h, G_{jh}, and u_j are the same as described in Equation 2.31.

The $r_{hi}G_{hj}$ term in Equation 2.32 represents the DIF effect for item i in group h. If the variance of r_{hi} is different from zero then this indicates that there is DIF present on the set of items in group h. Testing this parameter for statistical significance can be done using a LR test[2]. If r_{0i} and r_{hi} are allowed to correlate, a positive association between these terms indicates that controlling for overall ability the most difficult items are especially difficult for group h.

2.1.6.4 Fixed Item Random Group DIF Models

The models in this section are classified as multilevel IRT (MLIRT) models as they denote a three or more level hierarchy (reponses nested within persons and persons nested in level three units) (J.-P. Fox & Glas, 2001; J. P. Fox, 2005; Kamata, 2001). Typically in educational or psychological settings students or participants may be nested in classrooms and/or schools. If the researcher is interested in inference about the specific items on the measure and wants to regard the groups as a sample from a population of groups from which to make inferences about then fixed item random group (FIRG) DIF models are an option. If the number of groups (level three units) is

[2]When testing a single variance of a random effect for statistical significance using a LR test there is a zero boundary condition (e.g. variances are non-negative). Thus, it is recommended that the p-value of the random effects be tested against a mixture of Chi-squares with $df_1 = p + 1$, $df_2 = p$, and mixture proportion 1/2 (see Snijders & Bosker, 2012, pp. 98-99 for details)

large and the researcher is interested in the population these groups represent then these models are a more parsimonious alternative to dummy coding the grouping variables. This is because only the variance of the random effects are estimated instead of the group main effects and group by item interactions as in the FIFG DIF model.

FIRG DIF models are used when researchers want to assess the impact of whether items function differently across schools or level three units. The question could be posed just as easily by allowing classrooms to be level three units and items nested within persons nested within classrooms (Noortgate & De Boeck, 2005). Further, a fourth level could be added with classrooms nested within schools. Here only the three level case for the Rasch model is described but this can be generalized to an arbitrary number of levels and other IRT models (see Noortgate & De Boeck, 2005; De Boeck & Wilson, 2004 for details).

The FIRG DIF model is given by

$$\eta_{ij} = u_j + \sum_{k=1}^{I} \beta_k X_{ki} + v_{0h} + v_{kh} X_{ki}, \tag{2.33}$$

where v_{0h} is the random effect of group h on overall performance, v_{kh} is the random effect of group h on the difficulty of item k, and β_k and u_j are as described previously. If the variance of v_{kh} is different from zero then item k is exhibiting DIF for group h. Both v_{0h} and v_{kh} follow a multivariate normal distribution with mean vector 0 and a covariance matrix. A positive association between v_{0h} and v_{kh} indicates that the estimated difficulty of item k is greater for the less able group versus the more able group after controlling for overall group ability. Effect sizes and statistical significance for the variances of the random effects can be calculated as mentioned for the RIFG model.

2.1.6.5 Random Item Random Group DIF Models

Lastly, it is possible to have random effects over items and groups. In this situation items and group random effects are crossed at level three which implies the random effects are also crossed

at level two (person level) (Noortgate et al., 2003). In the random item random group (RIRG) DIF model both item and person covariates can be introduced to explain the source of the DIF. For cases when DIF could be a function of both item properties and groups the RIRG model provides a framework to decompose these DIF effects. The RIRG DIF model without covariates is given by

$$\eta_{ij} = \beta_0 + r_{0i} + v_{0h} + t_{ih} + u_j, \tag{2.34}$$

where β_0 is the mean negative difficulty in the reference group for an average item, r_{0i} is the random main effect of item i, v_{0h} is the random effect of group h (level three ability for someone in group h), and t_{ih} is the random interaction effect of the ith item and group h. The t_{ih} parameter is the DIF estimand which is normally distributed with mean zero and variance σ_t^2. If t_{ih} is different from zero then there is sufficient evidence that there is uniform DIF present. Effect sizes and statistical significance of the random effects can be calculated as mentioned in the RIFG section.

2.1.7 DIF Methods from Machine Learning

The term "big data" is a widely known term synonymous with the massive amounts of data that are common in today's world of technology. According to Murphy (2012, p. 1) there are more than one trillion web pages, approximately one hour of video is uploaded to YouTube every second, and some corporations have databases with petabytes (1 petabyte = $1x10^{15}$ bytes) of information. In order to deal with this huge amount of information the field of machine learning has developed a plethora of sophisticated algorithms to automate data analysis. Much of the research on machine learning stems from computer science, yet the methods can provide opportunities for interesting research in psychometric applications when applied appropriately. Recently, in the context of DIF testing two ideas have been proposed that were inspired by methods and problems faced in machine learning applications: Rasch trees (Strobl, Kopf, & Zeileis, 2013) and the DIF least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996; Tutz & Schauberger, 2013). Each of these methods is described briefly in the following sections.

2.1.7.1 Rasch Trees

Rasch trees were proposed by Strobl et al. (2013) for binary items and El-Komboz, Zeileis, and Strobl (2014) for polytomous items. They introduced Rasch trees as a compromise between manifest DIF methods and latent class (over persons) DIF methods. According to Strobl et al. (2013) the manifest DIF approaches such as IRTLRDIF, Wald χ^2 test, LM tests, and MIMIC models require researchers to explicitly define the manifest groups to be tested. The advantages of these approaches are that they give very specific and interpretable results with respect to the items that have DIF. The downside of these approaches are that if certain groups are not included in the DIF analyses then this may cause ambiguity later. The advantage of the latent class (over persons) approach is that the model for DIF is tested over all possible combinations of groups of subjects regardless of covariates included in the model. However, the disadvantages of this approach are that the number of classes must be selected as it is unknown apriori and often another analysis must be conducted to compare the manifest groups to the LCs (see Cohen & Bolt, 2005 for an example).

To overcome the limitations by manifest DIF methods and LC methods mentioned above Strobl et al. (2013) proposed Rasch trees which allow researchers to test pre-defined easily interpretable manifest groups versus dealing with the limitations of solely using LC or manifest DIF approaches. Rasch trees are based on model-based recursive partitioning (MBRP), which is a technique inspired from the machine learning and data mining literature. MBRP is closely related to classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984) but overcomes the limitations of selecting splitting variables in classification and regression trees (see Hothorn, Hornik, & Zeileis, 2006 for details). MBRP is a semi-parameteric approach that employs structural change tests to detect differences in parameters of a statistical model across groups of subjects defined over combinations of manifest covariates (Hothorn et al., 2006; Kopf, Augustin, & Strobl, 2014; Strobl et al., 2013). The main difference between MBRP and classification and regression trees is that in MBRP the parameters of a parametric model rather than values of an outcome variable are allowed to vary over groups (Kopf et al., 2014; Strobl et al., 2013).

Structural change tests are applied in econometrics for detecting parameter instabilities in time series models along a time dimension (Merkle & Zeileis, 2013). This same methodology can be used for detecting parameter instabilities or changes over person covariates for use in DIF testing. According to Strobl et al. (2013) there are four steps to test for DIF using Rasch trees:

1. Estimate the item parameters for a joint Rasch model in the sample of interest.
2. Assess the parameter instability of the joint Rasch model in the full sample with respect to each covariate of interest.
3. If significant parameter instability is found split the sample along the covariate with the highest instability and at the cutpoint leading to the greatest improvement in model fit.
4. Repeat steps 1-3 until convergence (i.e. no more significant parameter instabilities or the sub-samples become too small).

The structural change tests used to test for significant parameter instability are generalized M-fluctuation tests (Zeileis & Hornik, 2007). The Rasch tree procedure can involve multiple statistical tests when a large number of covariates are needed for splitting and also when selecting the optimal cutpoint. To control for Type I errors the procedure uses a Bonferroni adjustment when assessing these criteria for statistical significance. Identifying the optimal cutpoint is straightforward with a binary covariate. If the covariate is ordinal or continuous the method assess the parameter instability across all possible cutpoints for the given covariate by maximizing a partitioned log-likelihood (see Strobl et al., 2013 for details). The cutpoint with the strongest parameter instability is selected as the optimal cutpoint.

The performance of the Rasch tree was tested in simulation by Strobl et al. (2013) and found to perform well in detecting relevant covariates and complex interactions that may occur in practice. Further, the method allows for a data-based way to find the optimal cutpoint on a continuous covariate without resorting to median splits. When using Rasch trees the number of groups that are assessed for DIF can grow quickly. For example, if the algorithm is passed covariates age and gender, it is quite possible that the algorithm could find four groups if the tree split on gender and then split on age in both men and women sub-samples. If more covariates were available

and the sample size was sufficiently large it is possible to have an even larger number of groups. One issue with the method is that it is unclear how to test the specific items for DIF in the splits of the tree. This concern was also raised by Tutz and Schauberger (2013) in their comparison with Rasch trees and the DIF LASSO (discussed below). Another limitation of Rasch trees are that they are only implemented for the Rasch family of models. Strobl et al. (2013) note that it is possible to extend this framework to other IRT models, however doing so may be computationally prohibitive due to the many model fittings required when dealing with selecting the optimal cut point for ordinal and continuous covariates.

2.1.8 DIF LASSO

Tibshirani (1996) first introduced the LASSO in the context of linear regression and noted that this method performs regularization, variable selection, and can be applied to many other models such as generalized linear models and trees. The idea of the LASSO is to maximize a cost function subject to an L_1 penalty term. In the case of linear regression this can be expressed as

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.35}$$

where λ is a penalty parameter and $|\beta_j|$ is the L_1 norm of the jth coefficient. The first term in Equation 2.35 refers to the normal least squares term common in linear regression optimization problems and the second term corresponds to a penalty. Thus, the optimization is to minimize the sum of squared residuals subject to the constraint of the L_1 norm. The L_1 penalty term encourages sparsity in the solution by performing regularization and variable selection (Tibshirani, 1996).

The LASSO was implemented in the context of DIF testing using the binary Rasch model with unconditional or joint maximum likelihood estimation by Tutz and Schauberger (2013). Let $p = 1 \ldots, P$ denotes persons and $i = 1 \ldots, I$ denotes items. The Rasch model is given as

$$log \left[\frac{P(Y_{pi} = 1|\theta_p)}{1 - P(Y_{pi} = 1|\theta_p)} \right] = \theta_p - \beta_i = \mathbf{1}_{P_{(p)}}^T \boldsymbol{\theta} - \mathbf{1}_{I_{(i)}}^T \boldsymbol{\beta}, \tag{2.36}$$

where $\mathbf{1}^T_{P_{(p)}} = (0,\ldots,0,1,0,\ldots,0)$ of length $P-1$ with 1 at position p, $\mathbf{1}^T_{I_{(i)}} = (0,\ldots,0,1,0,\ldots,0)$ has length I with 1 at position i, $\boldsymbol{\theta}^T = \left(\theta_1,\ldots,\theta_{(P-1)}\right)$, and $\boldsymbol{\beta}^T = (\beta_1,\ldots,\beta_I)$ giving a total parameter vector $\boldsymbol{\alpha}^T = \left(\boldsymbol{\theta}^T,\boldsymbol{\beta}^T\right)$. Note to identify the model $\theta_P = 0$. A general DIF model that allows for estimation of an arbitrary number of groups and also allows for the inclusion of continuous covariates can be estimated by replacing β_i by $\beta_i + \boldsymbol{x}_p^T\boldsymbol{\gamma}_i$. Note $\boldsymbol{x}_p^T$ is the m by 1 person specific covariate vector containing the covariates of interest for person p and $\boldsymbol{\gamma}_i$ is a m by 1 vector of item specific parameters corresponding to the covariates in $\boldsymbol{x}_p^T$.

By making this substitution all person specific factors that may induce DIF for a given item i are included. The form of the model for estimation is

$$log\left[\frac{P(Y_{pi}=1|\theta_p)}{1-P(Y_{pi}=1|\theta_p)}\right] = \mathbf{1}^T_{P_{(p)}}\boldsymbol{\theta} - \mathbf{1}^T_{I_{(i)}}\boldsymbol{\beta} - \boldsymbol{x}_p^T\boldsymbol{\gamma}_i, \tag{2.37}$$

with the parameter vector of Equation 2.37 is given by $\boldsymbol{\alpha}^T = \left(\boldsymbol{\theta}^T,\boldsymbol{\beta}^T,\boldsymbol{\gamma}_1^T,\ldots,\boldsymbol{\gamma}_I^T\right)$. Tutz and Schauberger (2013) noted that the model in Equation 2.37 is not identified and recommended two conditions for model identification. First, set $\beta_i = 0$ and $\boldsymbol{\gamma}_i^T = \mathbf{0}$ for any $i \in \{1,\ldots,I\}$. Second, ensure that the matrix $\boldsymbol{X}$ with rows $(1,\boldsymbol{x}_1^T),\cdots,(1,\boldsymbol{x}_P^T)$ has full rank. A proof of these identification conditions is given in Tutz and Schauberger (2013). The model in Equation 2.37 can become problematic to estimate when a large number of covariates are present in $\boldsymbol{x}_p^T$ for each item i. To overcome these problems regularization or penalized maximum likelihood estimation can be used. The penalized log-likelihood is given by

$$\mathcal{L}(\boldsymbol{\alpha})_{Penalized} = \mathcal{L}(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}), \tag{2.38}$$

where $\mathcal{L}(\cdot)$ is the log-likelihood of the Equation 2.37, λ is the penalty parameter designated by the user, and $J(\boldsymbol{\alpha})$ is the penalty term. Note that $J(\boldsymbol{\alpha}) = \sum_{i=1}^{I}||\boldsymbol{\gamma}_i||$, which indicates that only the person specific covariate terms that may induce DIF are penalized. The $||\boldsymbol{\gamma_i}||$ penalty term was recommend by Tutz and Schauberger (2013) and is a modification of the group LASSO (Yuan & Lin, 2006).

A critical component of penalized maximum likelihood estimation is selection of λ the penalty parameter. Tutz and Schauberger (2013) recommend using the BIC by calculating the degrees of freedom as proposed by Yuan and Lin (2006) (see Tutz & Schauberger, 2013; Yuan & Lin, 2006 for details). The procedure works as follows:

1. Specify a grid of values for λ such as a grid of 20 equally spaced values from 0.5 to 20 for example (this does not have to be 20 values between 20 and 0.5 it can be other values as specified by the user).
2. Fit the model by optimizing Equation 2.38 for the first λ value in the grid from step 1.
3. Calculate the BIC using the degrees of freedom as specified by Tutz and Schauberger (2013) for the first λ value from steps 1 and 2.
4. Repeat steps 2 - 3 for all remaining λ values in the grid and select the λ value with the lowest BIC as the model to test for DIF.

According to Tutz and Schauberger (2013) DIF is indicated by the number of non-zero $\hat{\boldsymbol{\gamma}}$ coefficients using the optimal λ value from the procedure above.

Tutz and Schauberger (2013) found in simulations that the DIF LASSO method had adequate parameter recovery of both person and item parameters. Further, they compared the DIF LASSO method to three other methods: Lord's (1980) Wald χ^2, logistic regression, and the Mantel-Haenszel. They found that for a large number of groups with medium to large DIF the DIF LASSO approach performed well and was competitive with the other methods as far as power and was superior to the other methods in controlling Type I errors. However, for few groups and weak DIF the DIF LASSO should probably not be preferred to the other approaches.

2.2 MIMIC Models for DIF

MIMIC models are a type of structural equation model (SEM) where a latent variable is regressed on a set of covariates/exogenous variables. Jöreskog and Goldberger (1975) first introduced the MIMIC model for the case of continuous manifest variables. The application of MIMIC models in

the IRT framework for DIF testing was introduced later by Múthen (1985; 1988; 1989) and Múthen, Kao, and Burstein (1991). MIMIC models easily allow for the inclusion of multiple covariates for testing DIF across many groups and at the same time allow for controlling the influence of other groups and/or additional continuous covariates (e.g. age).

MIMIC models for DIF testing in an IRT framework make all the traditional assumptions IRT methods make, but they also assume that the variance of the latent variable is the same across the groups being compared (typically the latent variance is fixed at one). Further, when testing for uniform DIF, invariance of the discrimination parameters across groups is also assumed; however, this assumption can be relaxed making testing of non-uniform DIF possible (Woods & Grimm, 2011). In order to test for DIF a common scale for the item parameters between the groups must be established. Typically, this is accomplished by empirically selecting anchor items using a rank based technique as in Woods (2009a), a purification based approach like W. C. Wang and Shih (2010), or a combination of a rank and purification based approach as in W. C. Wang et al. (2012).

As mentioned previously, MIMIC DIF models can be parameterized as a logistic IRT model and estimated using available software for structural equation modeling such as Mplus (Muthén & Muthén, 1998–2012) or GLLAMM (Rabe-Hesketh, Skrondal, & Pickles, 2004) by using a full-information maximum likelihood estimator. The 3PL model cannot be estimated in a MIMIC model framework but it is possible to estimate the 1PL, 2PL, GRM, GPCM, and Rasch variants of the aforementioned models. It is also possible to estimate MIMIC DIF models in a Bayesian framework although research and applications in this framework are scarce (see Samuelsen, 2005; Bilir, 2009 for applications to latent class DIF testing). In the sections that follow two types of MIMIC DIF models will be discussed. First, MIMIC models that test for uniform DIF as Múthen (1985; 1988; 1989) originally proposed are presented, an overview of latent interactions necessary for non-uniform MIMIC DIF testing are reviewed, and an explanation of testing non-uniform DIF using an IRT based MIMIC model will be given as illustrated by Woods and Grimm (2011).

2.2.1 MIMIC Uniform DIF Models

MIMIC models can be parameterized as IRT models and estimated with Bock and Aitkin (1981)'s expectation-maximization marginal maximum likelihood (EM-MML) algorithm. Here the 2PL model within the Mplus framework is used to illustrate the ideas of EM-MML. Mplus was chosen because it is a widely used software package within the social and behavioral sciences and is probably very familiar to many researchers. When specifying a MIMIC model in Mplus if the estimator is set to robust maximum likelihood (MLR) then Mplus estimates an IRT model using EM-MML. The equation of the 2PL model in Mplus is given as

$$P(y_{ij} = 1|\theta) = \frac{1}{1 + exp[\tau_i - \alpha_i\theta_j]}, \tag{2.39}$$

where $\tau_i = \alpha_i b_i$.[3] Note that in Mplus τ_i is parameterized slightly different than is presented in an IRT framework but the parameterizations are equivalent with a change of sign (see L. K. Muthén & Muthén, 1998–2012 for details). To test for uniform DIF using a MIMIC model item i is regressed on a latent variable (θ) and binary grouping covariate G_h where h denotes the focal groups and $h = 0$ denotes the reference group. For item i the underlying continuous response process y_i^* underlying a binary response y_i equals 1 if $y_i^* > \tau_i$ and 0 otherwise. For a MIMIC DIF model on a given item i

$$y_i^* = \alpha_i\theta + \beta_{i1}G_1 + \beta_{i2}G_2$$

and

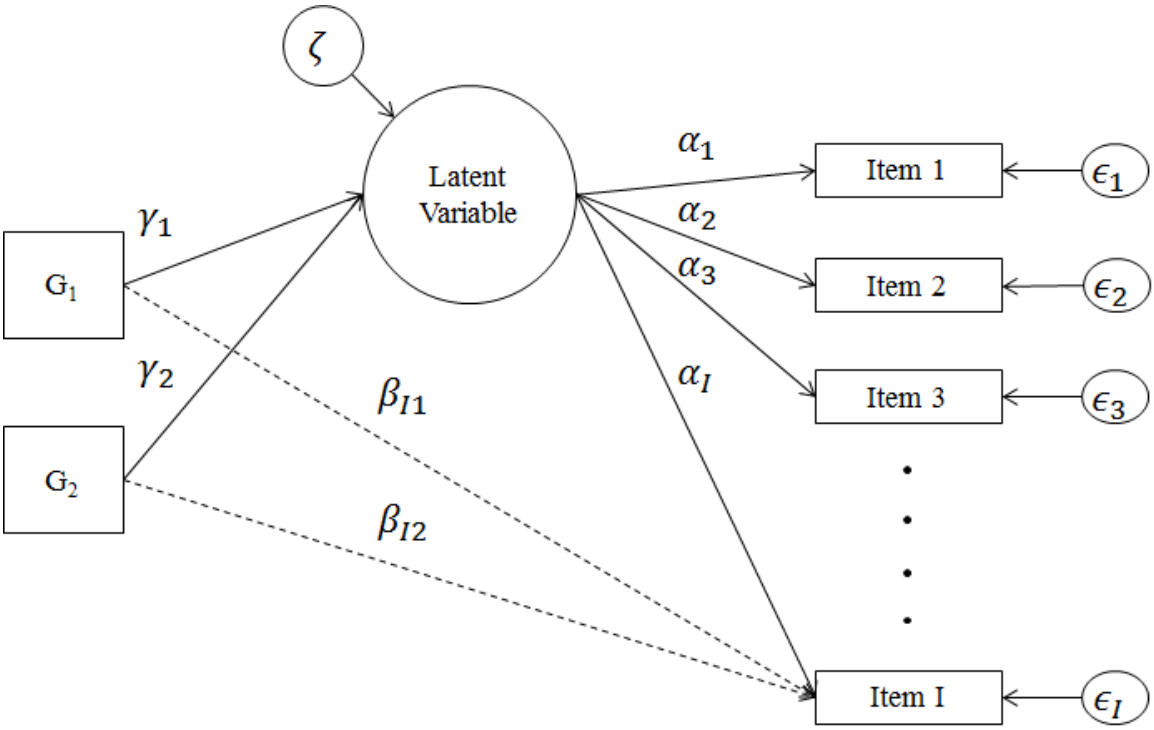
$$\theta = \gamma_1 G_1 + \gamma_2 G_2 + \zeta, \tag{2.40}$$

where α_i is the discrimination parameter for item i, β_{ih} is the regression coefficient showing the DIF effect between the reference and focal group h, γ_h represents the mean difference between the reference and focal group h on θ (note that G_0 represents the reference group), and ζ is the residual variance of the latent variable. Equation 2.40 is illustrated in Figure 2.1 for testing DIF in

[3]Note that b_i is used in place of β_i for the location parameter to be consistent with previous literature on DIF testing with MIMC models. β is reserved for the regression coefficient of group on item response or uniform DIF

item I with three groups [G_0 (reference group), G_1 (focal group one), and G_2 (focal group two)]. In Figure 2.1 items $1, \ldots, (I-1)$ can be thought of as anchor items when testing for DIF for item I.

Figure 2.1: MIMIC Uniform DIF Path Diagram.



Note. G_1 and G_2 correspond to the grouping covariate for group 1 and group 2 with group 0 (G_0) being the reference group. θ represents the latent variable, ζ represents the residual variance of the latent variable, γ represents the mean difference on θ between the given group and reference group, β represents the uniform DIF, α represents the discrimination parameter for the given item, and ϵ represents the residual for the given item.

Woods et al. (2009) notes four steps researchers should take when conducting DIF testing using MIMIC models. First, anchor items must be selected either empirically using a rank-based, iterative, or other purification method or using a known DIF free anchor set. Woods (2009a) recommends approximately 10 to 20 percent of the items be used as anchor items. Second, using the anchor items selected from the first step, conduct DIF testing for each item. This second phase can be done two ways: (1) using likelihood ratio tests or (2) using Wald z-tests. The results of these tests are asymptotically equivalent in large samples. However, using a Wald z-test requires

less model fittings than the likelihood ratio test and would probably be preferred.

Testing for uniform DIF with Wald z-tests is straightforward. Fit a model with all β_{ih} paths freely estimated for all items except the anchor items. If any of the β_{ih} paths are statistically significant then these items are flagged as showing uniform DIF. The use of likelihood ratio tests requires more tests and was described in Woods (2009b). According to Woods (2009b) the anchor items won't have any β_{ih} paths estimated as with items 1 to $(I-1)$ in Figure 2.1. The β_{ih} paths of all non-anchor items are freely estimated as with item I in Figure 2.1 and the $-2 * Log\ Likelihood$ value is computed (this serves as the baseline model). Then for the ith studied item fit a model with the β_{ih} path constrained to zero and obtain the $-2 * Log\ Likelihood$ and conduct a likelihood ratio test with the degrees of freedom equal to the difference between the constrained and less constrained model.[4] Repeat this for all studied items. If the LR test is statistically significant then there is evidence of uniform DIF in item i.

Step three involves fitting a final model with all the β_{ih} paths estimated for items that showed significant uniform DIF. The fourth step is to report the α_is, τ_is, group mean difference on θ, and a measure of DIF effect size. The group mean difference on θ are given by the γ_h coefficients. A measure of effect size can be obtained by exponentiating the β_{ih} coefficients for the DIF items giving an odds ratio. These measures of effect can be used to demonstrate if the DIF effect is practically significant. It should be noted that not all previous researchers' have fit a final model when conducting DIF testing (e.g. Barendse et al., 2010; Barendse et al., 2012; Carroll, 2014; Finch, 2005). However, fitting of a final model allows for researchers to observe the scale after setting all non-significant paths to zero and provides a better representation of the scale after DIF testing.

MIMIC models are a popular tool for testing for uniform DIF given their relative simplicity and ability to be illustrated clearly with a path diagram as in Figure 2.1. Applied researchers have utilized MIMIC models for uniform DIF testing in the case of more than two groups (e.g. Fleishman et al., 2002; Sacco, Casado, & Unick, 2011; Sacco, Torres, et al., 2011; Woods et al., 2009).

[4]According to Woods et al. (2009) computation of the LR statistic in Mplus must be divided by a term that is a function of the number of estimated parameters in each model and scaling correction factors given in the Mplus output. See the example on the Mplus website (http://www.statmodel.com/chidiff.shtml) for details.

Yet, methodological research for MIMC DIF models has lagged behind applied researchers and has been limited to the two-group case, which is briefly reviewed here.

Finch (2005) compared MIMIC models with several other DIF methods for binary responses (i.e. IRTLR DIF, SIBTEST, and Mantel-Haenszel) and found that in certain conditions the MIMIC model for uniform DIF performed favorably compared to the other methods. Specifically, Finch (2005) notes that when the data generating model was not the 3PL and/or the test length was 50 items Type I error rates were consistent with the other DIF methods. Finch also noted that with respect to power the MIMIC model was as powerful at detecting DIF as the other three approaches when the test was long and/or the data did not have pseudo-guessing present (i.e. 2PL model). Lastly, Finch noted that under certain simulation conditions (e.g. longer tests or no pseudo-guessing) the MIMIC model was more robust to anchor item contamination than the other three methods.

According to Woods (2009b) MIMIC models exhibited good power to detect uniform DIF for lower focal group sample sizes (50, 100, 200, 400) and also had reasonable parameter recovery for these sample sizes in both the 2PL and GRM models. Further, Woods noted that at all values of the focal group sample sizes (i.e. 25, 50, 100, 200, 400) Type I errors were well controlled and power was greater for the MIMIC model compared to IRTLRDIF. Another advantage of MIMIC models is the ease of adding additional groups to test for DIF. Two disadvantages of MIMIC models for uniform DIF are there inability to account for differences in the variance of θ between the groups and the inability to test for non-uniform DIF.

Recently, Carroll (2014) assessed the MIMIC DIF model in a categorical factor analytic framework to look at the impact of violating the assumption of homogeneity of latent variances on the power, Type I error rate, and parameter recovery. He found that in terms of Type I error rate the violation of this assumption tended to cause inflated Type I error rates and impacted power to detect uniform DIF. Further, Carroll (2014) noted that the factor loadings, thresholds, latent group mean difference estimates, and DIF effects were all adversely impacted to varying degrees when this assumption was violated.

2.2.2 MIMIC Non-uniform DIF Models

Up until fairly recently researchers did not associate MIMIC models with the ability to to test for non-uniform DIF. The first mention of this idea was from Barendse et al. (2010) in the context of using continuous indicators (non-IRT) in a restricted factor analysis (RFA) and adding an interaction term between a grouping variable and the latent variable by adding a latent interaction predicting item response. Barendse et al. (2012) extended the simulation conditions of Barendse et al. (2010) for the continuous indicator case of MIMIC DIF models and Woods and Grimm (2011) provided an application of IRT MIMIC DIF models.

According to Marsh, Wen, Nagengast, and Hau (2012) typically the interaction of a latent variable on group(s) would be accomplished in SEM by fitting a multiple group model (note that Marsh et al., 2012 talk about this in the context of SEM but the logic applies to IRT as well). Given that the traditional recommendation of an interaction between an observed categorical and continuous latent variable can be handled by multiple group analysis (e.g. Marsh et al., 2012; Rigdon, Schumacker, & Wothke, 1998) most of the work on latent interactions within SEM has focused on interactions between two continuous latent variables[5] (e.g. Marsh et al., 2012; Klein & Moosbrugger, 2000; Klein & Muthén, 2007). Several researchers have noted that fitting a multiple group model when wanting to test a categorical observed by continuous latent variable interaction typically requires a larger sample size to obtain good parameter estimates and power in DIF settings compared with using a MIMIC model for uniform DIF (e.g. Barendse et al., 2010; Barendse et al., 2012; B. O. Muthén, 1989; Woods, 2009b; Woods & Grimm, 2011). If the assumptions of the MIMIC model hold the ability to provide good parameter recovery and power compared with multiple group approaches has been one reason for the popularity of MIMIC models for uniform DIF testing within the social and behavioral sciences. It is an open question whether the benefits of using MIMIC models for uniform DIF when assumptions hold carry over to the non-uniform DIF case. Thus, it makes sense to review some of the ways that latent variable interactions are

[5]Some work has discussed the case of an observed continuous variable by latent variable interaction (see Rigdon et al., 1998 for details)

constructed in order to assess the viability of using these methods for non-uniform DIF testing with MIMIC models.

2.2.2.1 Overview of Latent Interactions and MIMIC DIF Models

Marsh et al. (2012) note three main ways of dealing with latent interactions within an SEM framework: (1) Product indicator approaches, (2) distribution-analytic approaches, and (3) Bayesian approaches. In what follows each of these three approaches will be reviewed and relevant research or lack thereof describing implementations of these approaches will be discussed. These approaches are presented in the continuous by continuous latent variable interaction context as per Marsh et al. (2012).

Product indicator approaches form a latent interaction by creating product terms of the manifest variables or indicators. These methods were first introduced by Kenny and Judd (1984), but were later refined by Marsh, Wen, and Hau (2004) and Wall and Amemiya (2001) to address the cumbersome implementation and non-robustness to violations of normality of the latent variable product terms. Woods and Grimm (2011) point out two issues with using product indicator approaches to DIF testing in IRT MIMIC models that need to be addressed. First, most of the research on product indicator variables involves continuous indicators (e.g. Marsh et al., 2004; Wall & Amemiya, 2001) and the performance of the method for categorical indicators has not been well-established and more research is needed. Second, it is unclear which items should be included in the interaction: only anchor items, anchors and the studied items, all items, and so on. For these reasons and others this is probably why at the time of this writing product indicator approaches have not been used in IRT MIMIC models to test for non-uniform DIF.

The second method mentioned by Marsh et al. (2012) were the distribution analytic approaches. Distribution analytic approaches explicitly model the non-normality implied by the indicator variable distributions that occur within the latent interactions. Currently there are two main methods: Latent moderated structural equations (LMS; Klein & Moosbrugger, 2000) and quasi-maximum likelihood (QML; Klein & Muthén, 2007). LMS models the non-normal distribution

caused by the latent variable product terms by approximating it using a finite mixture of Gaussian distributions (Klein & Moosbrugger, 2000). LMS assumes that both variables in the latent interaction are normally distributed. QML also models the non-normality caused by the latent product terms but has less stringent distributional assumptions than LMS. QML reduces the number of components used to approximate the non-normality caused by the latent product terms to a normal and conditionally normal distribution versus a potentially large number of normal mixture components used by LMS (Klein & Muthén, 2007; Marsh et al., 2012).

Several studies have looked at non-uniform DIF testing with MIMIC or related RFA models using the LMS approach (Barendse et al., 2010; Barendse et al., 2012; Woods & Grimm, 2011). These studies can be classified into those treating indicators as continuous (Barendse et al., 2010; Barendse et al., 2012) versus using an IRT parameterization (Woods & Grimm, 2011). Both Barendse et al. (2010) and Barendse et al. (2012) used RFA (equivalent to MIMIC in the two group case) to test for non-uniform DIF with continuous indicators. In both studies they considered only two groups, made the assumption that the grouping variable was latent for LMS, and used all-other items as anchors for DIF testing. In both studies Type I error rates were inflated possibly due to the use of all-other items as anchors (Stark, Chernyshenko, & Drasgow, 2006; W. C. Wang, 2004) and that the group variable was not normally distributed as assumed by LMS (Klein & Moosbrugger, 2000).

Woods and Grimm (2011) tested non-uniform DIF with a MIMIC model as parameterized in an IRT framework within Mplus. Although LMS assumes that the latent variables involved in the interaction are normally distributed, at the time of publication of Woods and Grimm (2011), the Mplus user's guide recommended using the variation of the LMS method with an observed categorical and continuous latent variable. Thus, Woods and Grimm (2011) wanted to test the utility of LMS as recommended by the Mplus user's guide despite the assumption violation. The results from Woods and Grimm (2011) showed greater power for detecting non-uniform DIF when using Equation 2.41 compared with Equation 2.40. However, use of Equation 2.41 showed inflated Type I error rates. At the present time estimating MIMIC interaction models with Mplus is not recom-

mended. Woods and Grimm (2011) note that the MIMIC interaction model could be parameterized as a nonlinear mixed model and estimated in a program such as SAS PROC NLMIXED.

The third and final approach to estimating latent interactions mentioned by Marsh et al. (2012) is a Bayesian approach. In the Bayesian approach the creation of product indicators apriori were not needed. Instead, similar to the distribution-analytic approaches the Bayesian approach samples the cross-product terms within the MCMC estimation, properly modeling the latent interaction. Simulation studies using continuous indicators and binary categorical indicators (non-MIMIC DIF models) indicated that Bayesian methods perfomed well especially in small sample sizes (Lee, Song, & Cai, 2010; Lee, Song, & Tang, 2007). At the time of this writing I could find no publications or presentations involving research on latent interactions with IRT MIMIC models for DIF testing. As noted by Marsh et al. (2012) the Bayesian approach is extremely flexible and allows for easy extensions of multiple interaction effects, higher order interaction effects, and other polynomial effects of the latent variables. One downside of this great flexibility is the sound statistical knowledge and thought required to specify the distributions and priors for the parameters. However, software programs such as Mplus, WinBUGS, OpenBUGS, JAGS, and Stan are gaining popularity which may ease the difficulty for applied researchers.

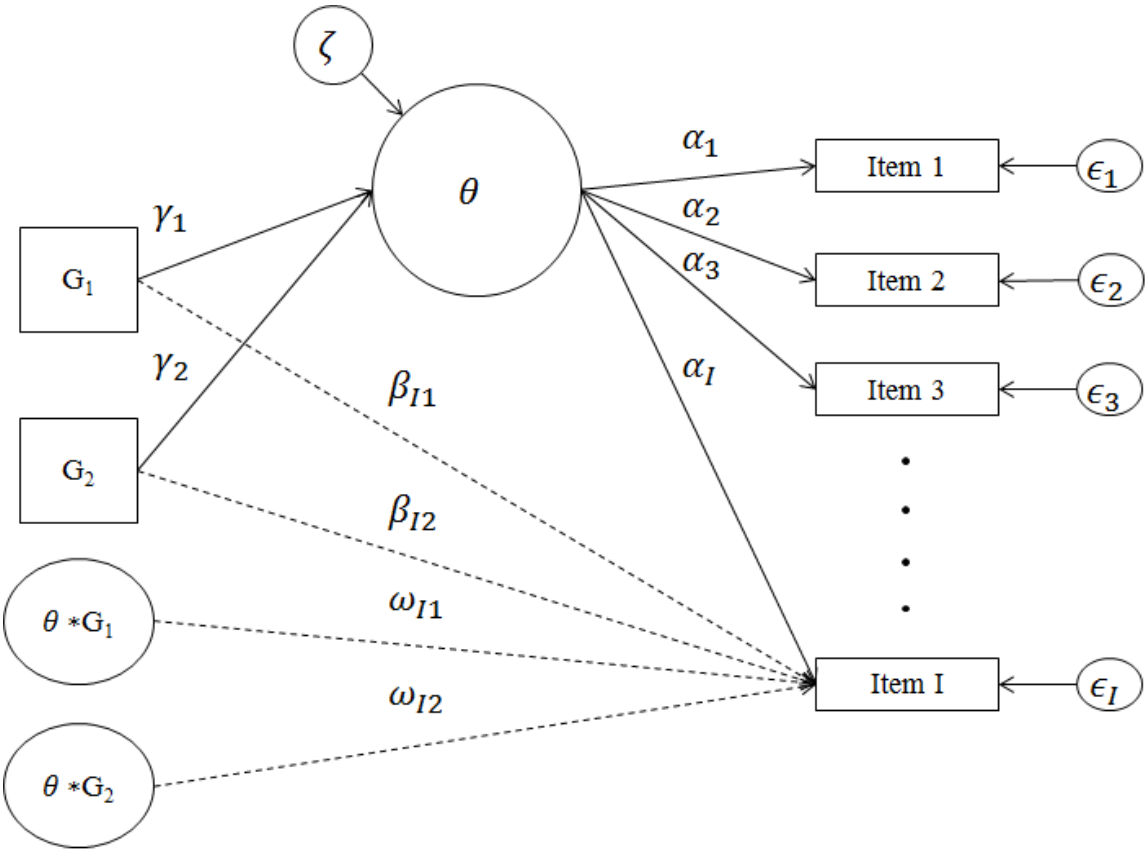
2.2.2.2 Illustration of How to Test for Non-uniform DIF with MIMIC Models

In this section an overview of how to test for non-uniform DIF with IRT based MIMIC models is given as described in Woods and Grimm (2011). Woods and Grimm (2011) were the first to use MIMIC models for testing non-uniform DIF in an IRT MIMIC model. Woods and Grimm (2011) noted that if an interaction between a grouping variable G and the latent variable θ was added in Equation 2.40 then non-uniform DIF could be tested. The equation for testing non-uniform DIF in a MIMIC model for item i is given by

$$y_i^* = \alpha_i\theta + \beta_{i1}G_1 + \beta_{i2}G_2 + \omega_{i1}G_1\theta + \omega_{i2}G_2\theta, \tag{2.41}$$

where ω_{ih} represents the interaction between group G_h and the latent variable θ and all other terms are as previously defined in Equation 2.40. Figure 2.2 gives the path diagram corresponding to Equation 2.41 for testing uniform and non-uniform DIF in item I. If ω_{ih} is significantly different than zero this indicates that the relationship between item response i and focal group h depends on θ, which is the definition of non-uniform DIF.

Figure 2.2: MIMC Non-uniform DIF Path Diagram.



Note. G_1 and G_2 correspond to the grouping covariate for group 1 and group 2 with group 0 (G_0) being the reference group. The $\theta * G$ circles represent the latent interaction between group and the latent variable. θ represents the latent variable, ζ represents the residual variance of the latent variable, γ represents the mean difference on θ between the given group and reference group, β represents the uniform DIF, ω represents non-uniform DIF, α represents the discrimination parameter for the given item, and ϵ represents the residual for the given item.

To conduct DIF testing a similar set of four steps as described for uniform DIF testing are carried out for the LR tests. Steps one and two differ slightly with the addition of an omnibus DIF test and individual tests of non-uniform and uniform DIF. To perform an omnibus test of DIF

using LR tests (uniform, non-uniform, or both) an augmented model is fitted where all studied items are regressed on G_h and the interaction $G_h\theta$. Then for each item a constrained model is fitted where the studied item is neither regressed on the grouping variables or the interactions (i.e. all β_{ih} and ω_{ih} are zero) and a LR test is conducted. If the LR test is statistically significant then this indicates that item i shows significant uniform, non-uniform, or both types of DIF. Follow-up tests for non-uniform and uniform DIF can be conducted using the same logic as the omnibus test with LR tests.

Note that use of Wald tests for an omnibus test of overall DIF could be slightly more complicated in the case of non-uniform DIF with a MIMIC model. The reason is that most software programs don't allow you to specify a contrast matrix out of the box to perform multivariate tests as was described in the Wald χ^2 DIF test section of this paper. It may still be possible but the user may have to specify these contrasts separately as model constraints and obtain good estimates of the asymptotic covariance matrix as described in the section Wald χ^2 test.

2.3 Current Study

The purpose of this research was to propose a new implementation of the MIMIC model for testing uniform and non-uniform DIF, conduct a Monte Carlo simulation to address the three limitations of the present body of MIMIC DIF research mentioned previously, and provide an empirical example for applied researchers. Specifically, these three limitations were only considering two groups in simulation studies, a lack of assessing the MIMIC models robustness to violations of the homogeneity of latent variance assumption, and not properly modeling the latent interactions needed to estimate non-uniform DIF with MIMIC models. In all conditions the proposed method will be compared to the improved Wald χ^2 approach discussed previously (see Equations 2.10 and 2.15).

First, to address the issue of only considering two groups the present study for apparently the first time considers the case of three groups in a methodological study with MIMIC DIF models

for non-uniform DIF. Second, the issue of estimating the latent interaction will be handled by utilizing Bayesian estimation to more appropriately model the latent interaction necessary for non-uniform DIF. Use of Bayesian estimation for estimating this model is a new implementation that has not currently been done. Third, the latent variances of the groups will be allowed to differ from the reference group and be manipulated in certain conditions and held to equality in other conditions as a comparison. Lastly, this new approach will be applied to a real data set to demonstrate the utility of the method to applied researchers.

Chapter 3

Methods for the Simulation Study

3.1 Overview

A simulation study was carried out to address the aforementioned shortcomings within the IRT-based MIMIC non-uniform DIF literature. The simulation study consisted of 18 crossed factors (described below) with 100 replications in each condition. Pilot tests indicated each replication will take approximately 1 to 5 hours to run. This number of replications was chosen based on computational feasibility and to improve upon the modest number of replications typically seen with computationally intensive MCMC simulations (e.g. Jiao, Kamata, Wang, & Jin, 2012; W. C. Wang, Liu, & Wu, 2013). An R program was used to generate, analyze, and process the data (R Core Team, 2014). The Bayesian MIMIC model was run on the advanced computing facility (ACF) cluster at the University of Kansas to expedite computation time. On the ACF cluster the Bayesian MIMIC model was run with R (v. 3.1.0). All improved Wald runs were run on a single PC using flexMIRT™(v. 3.0.3) using R (v. 3.1.1) and results were summarized using R (v. 3.2.2). R code for simulating the Bayesian MIMIC model is available by request from the author.

This section is organized into five subsections. First, a discussion of the fixed factors in the simulation will be given. Second, a discussion of the varying factors will be described. Third, a description of Bayesian estimation, prior selection, and software will be provided. Lastly, a

description of the procedure and the outcomes will be provided.

3.2 Fixed Factors in the Simulation

There were six factors that remained fixed throughout the simulation study (scale length and item parameters, magnitude of DIF, type of DIF, proportion of DIF items, proportion of anchor items, and latent means) each of these is described below. First, the scale length was fixed at 20 binary items to represent an appropriate scale length that may be commonly used in educational and psychological settings and has also been used in previous DIF simulation studies (e.g. Finch, 2011; French & Finch, 2010; W. C. Wang & Shih, 2010). Further, French and Finch (2010) note that it is not uncommon to fix the test length in DIF studies (e.g. Finch & French, 2007; French & Maller, 2007; Hidalgo-Montesinos & Gómez-Benito, 2003; R. Penfield, 2007), especially when there are other factors that may have a greater influence on DIF (e.g. latent variances, sample size imbalance, number of groups) and to keep the simulation study size manageable.

The true values of the generating item parameters were based on item parameters found in Woods et al. (2009) on the negative temperament scale of the Schedule of Nonadaptive and Adaptive Personality (SNAP; Clark, 1996). This allowed for a more realistic depiction of the item parameters seen in practical psychological assessments being used within the simulation. The negative temperament scale consists of 28 items so, a random sample of 20 items was chosen. The true values of the 20 randomly drawn item parameters from the 28 items on the negative temperament scale are presented in Table 3.1. Item parameters for the focal group(s) that are not DIF items (i.e. items seven through 20) will be the same as those in the reference group and item parameters that are DIF items (i.e. items one through six) are described in the next section.

The second fixed factor, the type of DIF consists of three types: Uniform DIF (only), non-uniform DIF (only), and mixed DIF (both uniform and non-uniform). These three types are illustrated in Figure 3.1. The three types of DIF were chosen to assess the DIF detection methods (i.e. Wald χ^2 and proposed MIMIC model) ability to identify not only whether a given item shows

Table 3.1: True Item Parameter Estimates Used for Data Generation

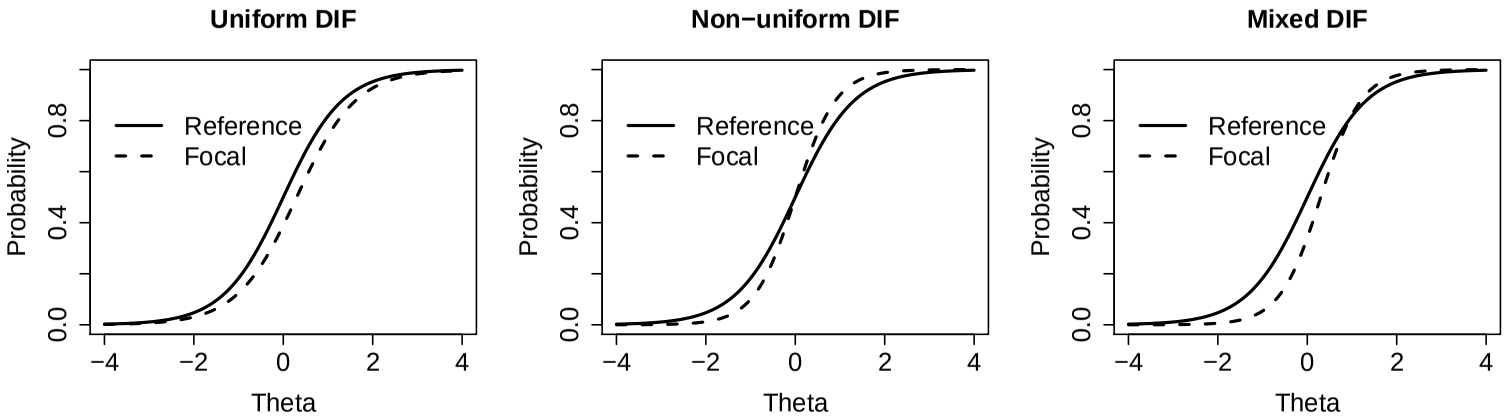
Item Number	α parameter	b parameter
1(311)DIF_b	1.41	1.18
2(269)DIF_α	2.42	1.30
3(250)$DIF_{\alpha,b}$	1.96	-0.09
4(281)DIF_b	1.50	-0.60
5(273)DIF_α	1.75	1.04
6(301)$DIF_{\alpha,b}$	1.92	0.33
7(290)	1.68	0.15
8(277)	1.64	1.13
9(294)	1.86	1.14
10(274)	1.32	1.83
11(259)	2.03	0.72
12(244)	1.55	0.18
13(245)	1.66	0.88
14(323)	1.15	-1.00
15(316)	1.55	0.43
16(333)	2.10	1.14
17(260)*Anchor*	1.98	1.74
18(325)*Anchor*	1.60	0.07
19(312)*Anchor*	0.95	2.02
20(331)*Anchor*	0.56	1.21

Note. DIF = differential item functioning. Numbers in parentheses refer to the item number on the negative temperament scale in Table 4 of Woods et al. (2009). $b = \tau/\alpha$ in Table 4 of Woods et al. (2009). DIF_α = non-uniform DIF only, DIF_b = uniform DIF only, $DIF_{\alpha,b}$ = mixed DIF, and *Anchor* denotes an anchor item.

DIF but also to assess the correct classification as to the type of DIF. These procedures are similar to those conducted in Lopez (2012). Each of the types of DIF will constitute two items out of the total number of six DIF items. For example, two items will be uniform DIF (i.e. items 1 and 4 in Table 3.1), two will be non-uniform DIF (i.e. items 2 and 5 in Table 3.1), and two will be mixed DIF (i.e. items 3 and 6 in Table 3.1). This framework allows for a deeper understanding as to the classification accuracy of the two DIF detection methods by examining whether a given DIF method is able to detect the correct type of DIF.

The third fixed factor, the magnitude of DIF, was constant throughout the simulation but

Figure 3.1: Illustrations of DIF Types



Note. DIF = differential item functioning and Theta = latent variable

allowed to vary randomly among the DIF items (i.e. items one through six) on the assessment among three values that are considered small, medium, and large amounts of DIF: $\delta = .3, .5, .7$ (Woods et al., 2013). These values reflect those commonly seen in practice and those that have been used in previous simulation studies (e.g. Woods, 2009b; Woods et al., 2013). Items with uniform DIF only (items 1 and 4) had $b_{iR} < b_{iF_h}$, items with non-uniform DIF only (items 2 and 5) had $\alpha_{iR} > \alpha_{iF_h}$ and, items with mixed DIF (items 3 and 6) had both $\alpha_{iR} > \alpha_{iF_h}$ and $b_{iR} < b_{iF_h}$. Note R denotes the reference group and F_h focal group h (i.e. $h = \{1, 2\}$). The focal group(s) discrimination parameter was calculated as $\alpha_{iF_h} = \alpha_{iR} - \delta_{\alpha_{ih}}$ and threshold parameter as $b_{iF_h} = b_{iR} + \delta_{b_{ih}}$. Values of δ will be determined separately for each α_i and b_i (and also separately for each focal group). To determine the specific δ a random draw from a $U(0, 1)$ distribution was taken and the δ determined such that:

$$\text{if}(x \leq .33) \text{ then } \delta = .3$$
$$\text{else if}((x > .33) \ \& \ (x \leq .66)) \text{ then } \delta = .5$$
$$\text{else if}((x > .66) \ \& \ (x \leq 1)) \text{ then } \delta = .7.$$

The fourth fixed factor, the proportion of DIF items, was constant at 30 percent (i.e. six items) which is consistent with previous simulation studies (e.g. W. C. Wang & Shih, 2010) and a reasonable amount of DIF items that may be seen in practice on a scale. Further, this allows for two items from each of the three DIF types to be considered. The fifth fixed factor, the proportion of anchor items, was 20 percent (i.e. four items). The proportion of anchor items was selected based on recommendations from previous methodological research (e.g. W. C. Wang, 2004; Woods, 2009a) and at the same time balancing the difficulty of anchor selection that may occur in practice as the number of groups increases beyond two.

The final fixed factor, the latent impact or true mean difference on the latent trait, was $\mu_{\theta_{F_1}} = -0.20$ for focal group one (two group and three group conditions) and $\mu_{\theta_{F_2}} = -0.40$ for focal group two (three group conditions only). These values reflect reasonable values that may be seen

in DIF applications (e.g. Harpole et al., 2014; Langer, 2008). The latent mean of the reference group was fixed at zero (i.e. $\mu_{\theta_R} = 0$) for model identification purposes. Thus, for the reference group $\theta_R \sim N(0,1)$ and for the focal group(s) $\theta_{F_1} \sim N(-0.20, \sigma_{F_1})$ and $\theta_{F_2} \sim N(-0.40, \sigma_{F_2})$.

3.3 Varying Factors in the Simulation

Three factors within this simulation study were chosen to vary; combined these constitute 18 crossed conditions [2 (number of groups) x 3 (latent standard deviation differences) x 3 (reference/focal group sample sizes)]. The number of groups will consist of two (reference and one focal group) or three (reference and two focal groups). This was be done to compare the findings from the present simulation study to other studies testing non-uniform DIF with MIMIC models (e.g. Woods & Grimm, 2011) as well as generalize the results to the case of three groups.

The second varying factor was the latent standard deviation (SD) differences across the groups. There are three levels of this factor: (1) smaller focal group(s) latent SDs compared with the reference group (LSDS), (2) equal latent SDs among the reference and focal group(s) (LSDE), and (3) higher focal group(s) latent SDs compared with the reference group (LSDH). Within the LSDE condition all the latent SDs will be fixed at one to have a condition where the MIMIC model assumption of homogeneity of latent SDs across groups holds. In these conditions (i.e. LSDE) the latent traits were simulated as follows: $\theta_R \sim N(0,1)$ and $\theta_{F_1} \sim N(-0.20,1)$ for the two-group conditions and $\theta_R \sim N(0,1)$, $\theta_{F_1} \sim N(-0.20,1)$, and $\theta_{F_2} \sim N(-0.40,1)$ for the three-group conditions. For the LSDS conditions the latent SDs were fixed at 0.50 and for the LSDH conditions the latent SDs were fixed at 1.5. These values were chosen based on reasonable values that may be seen in practical applications of DIF studies (e.g. Harpole et al., 2014; Langer, 2008).

The third varying factor was the reference group to focal group sample sizes ratio. For the two group condition the total sample size was 1000. The three ratios were 500:500 (equal), 750:250 (moderately unequal), and 900:100 (highly unequal). For the three group condition the total sample size was 1500 with ratios 500:500:500 (equal), 1000:250:250 (moderately unequal), and

1300:100:100 (highly unequal). These sample sizes and ratios were selected to show what may be seen in practice with both balanced and unbalanced group sizes. The lower bound of 100 for the small sample sizes in both two group and three group conditions (i.e. 100) was chosen based on results from Woods (2009b) where sample sizes lower than 100 were more likely to have convergence problems and had higher parameter bias. The proposed Bayesian MIMIC model is more complicated than that in Woods (2009b) and going lower than 100 would likely cause poor convergence and parameter recovery.

A within condition factor in this study was DIF detection method. The two methods were the non-uniform DIF MIMIC model and the improved Wald χ^2 (see Equation 2.15). The Improved Wald χ^2 was chosen due to its asymptotic equivalence to the likelihood ratio test, better performance than the likelihood ratio test in a recent simulation study in the case of three groups (Woods et al., 2013), and much smoother implementation computationally than the likelihood ratio test as discussed previously. In all simulation conditions unless noted otherwise the improved Wald was used with the supplemental EM algorithm setting the maximum number of E-steps to 4000 and the maximum number of M-steps to 1000 (the defaults were 500 and 100, respectively). Further, the SmartSEM option was set to No in order to utilize the full EM history when assessing convergence which may lead to better convergence results (Houts & Cai, 2013).

3.4 Bayesian Estimation

In order to extend findings from previous research with respect to estimating the latent interaction term a Bayesian MIMIC model was used. From here out the Bayesian MIMIC model will be used to denote the non-uniform DIF Bayesian MIMIC model in the present study. As noted in Woods and Grimm (2011) the MIMIC model can be reparameterized as an equivalent non-linear mixed model to test for non-uniform DIF. All Bayesian analyses were run using Stan (v. 2.6.0) (Stan Development Team, 2014b), with the RStan interface (Stan Development Team, 2014a) for the R platform (R Core Team, 2014) with two Markov chains and random starting values.

Stan uses Hamiltonian Monte Carlo (HMC) sampling with the No-U-Turn (NUTS; Hoffman & Gelman, 2011; Hoffman & Gelman, 2013) sampler. HMC using NUTS borrows ideas from physics and Hamiltonian dynamics to explore the parameter space of the joint posterior distribution (Hoffman & Gelman, 2011; Hoffman & Gelman, 2013). The idea is to suppress the local random walk behavior in the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) allowing for much more rapid movement through the parameter space. This is accomplished by utilizing a momentum variable along with the parameter vector that are jointly updated together although inferences are taken only from the parameter vector. The momentum variable acts as an auxiliary variable to allow the algorithm to move faster through the parameter space (see Gelman et al., 2013, pp. 300-302 for further discussion). Stan was chosen because in pilot studies Stan performed superior to JAGS which uses a variation of Metropolis within Gibbs sampling and slice sampling depending on the model.

As with Gibbs sampling (Geman & Geman, 1984), and Metropolis-Hastings (Hastings, 1970) HMC using NUTS requires a full probability model as per Baye's rule (see Equation 1.4 and 1.5) to be specified. Letting $\boldsymbol{S} = \{\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{\beta}\}$ the joint distribution of the parameters given the data for the Bayesian MIMIC model will be specified as follows:

$$P(\boldsymbol{S}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{\beta})P(\boldsymbol{\theta}|\boldsymbol{\mu})P(\boldsymbol{\mu})P(\boldsymbol{\alpha})P(\boldsymbol{b})P(\boldsymbol{\omega})P(\boldsymbol{\beta}). \tag{3.1}$$

Note in Equation 3.1 $\boldsymbol{\theta}$ is a vector of latent traits for persons, $\boldsymbol{\mu}$ is a vector of latent means for the latent trait, $\boldsymbol{\alpha}$ is a vector of discrimination parameters, $\boldsymbol{b}$ is a vector of location parameters[1], $\boldsymbol{\omega}$ is a vector of non-uniform DIF parameters (coefficient for interaction between group and the latent trait), and $\boldsymbol{\beta}$ is a vector of uniform DIF parameters.

3.4.0.3 Priors

To estimate the Bayesian MIMIC model, prior distributions must be specified for $\boldsymbol{S} = \{\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{\beta}\}$. Note all priors are presented with SDs below and not precisions. In the present study the follow-

[1]Note that b is used instead of τ to better align with the MIMIC IRT parameterization as $\tau = -\alpha * b$

ing prior distributions were used:

$$\theta_{jh} \sim N(\mu_h, 1)\ j = 1, \ldots, J\ \&\ h = 1, \ldots, G \tag{3.2}$$

$$\mu_h \sim N(0, 2) \text{ Note } \mu_1 = 0\ \&\ h = 1, \ldots, G \tag{3.3}$$

$$\alpha_i \sim LN(0, 1)\ i = 1, \ldots, 20 \tag{3.4}$$

$$b_i \sim N(0, 2)\ i = 1, \ldots, 20 \tag{3.5}$$

$$\omega_{ih} \sim N(0, 0.15)\ i = 1, \ldots, (20 - A)\ \&\ h = 1, \ldots, G \tag{3.6}$$

$$\beta_{ih} \sim N(0, 1)\ i = 1, \ldots, (20 - A)\ \&\ h = 1, \ldots, G. \tag{3.7}$$

G denotes the number of groups (either two or three), J the number of persons, LN denotes a log-normal distribution, A denotes the number of anchor items (i.e. four items in the present study). Thus, the first 16 items were tested for DIF and the last four are anchor items. Note that $\mu_1 = 0$ and $\sigma^2_{\theta_{jg}} = 1$ in order to identify the model. These prior distributions have been used previously in other research and practical applications (e.g. Curtis, 2010; J. Fox, 2010; Patz & Junker, 1999a; Patz & Junker, 1999b).

All prior distributions were chosen to ensure that the Bayesian MIMIC model had reasonable convergence across all conditions. More diffuse priors were originally used but they resulted in 50 to 70 percent convergence rates. The priors for ω_{ih} were the most informative due to problematic convergence issues related to label switching caused by the discrimination parameters. The priors for β_{ih} were chosen to have a mean of 0 and SD of 1 in order to balance non-informativeness and convergence issues with more diffuse priors.

3.4.1 Burn-in and Convergence Diagnostics

As mentioned above the model was fit using two Markov chains. Three important considerations when conducing Bayesian estimation using MCMC are selecting the burn-in, assessing approximate convergence of the Markov chains, and the amount of thinning. To determine the appropriate burn-in for the Markov chains several pilot runs were conducted. In all cases the

pilot runs indicated that the Markov chains reached stationarity around 100 to 300 draws. For the simulation a conservative burn-in of 1000 draws was used.

The convergence check was determined using the potential scale reduction factor (PSRF) also know as univariate r-hat (see Gelman & Rubin, 1992). As mentioned by Gelman and Rubin (1992) when all parameters have an r-hat less than 1.10 this indicates approximate convergence has been reached and was the cutoff used in the present study. Several pilot runs were conducted to provide insight into how many post-burn-in runs were needed. The pilot runs indicated that 1000 post burn-in draws per Markov chain were sufficient. A conservative 2000 post-burn-in draws per chain were used in the simulation study.

3.5 Procedure

The procedure of the simulation study was as follows for each replication within a given condition. First, a single data set was generated in R from the multiple group 2PL model (i.e. Equation 2.1) with the appropriate varying and fixed factors discussed above. Next, the rstan package was used to call Stan and analyze the data with the Bayesian MIMIC model. Note this series of steps was calculated on the ACF cluster. On a single PC the R program called flexMIRT™to analyze the same data set and conducted the DIF testing using the improved Wald χ^2 test. Upon completion of the DIF testing and parameter estimation, several R functions processed, organized, and saved relevant output for later analysis. Additionally, the r-hat values (PSRF; Gelman & Rubin, 1992) were monitored to assess the approximate convergence to the posterior for the Bayesian MIMIC model. Convergence was also monitored for the improved Wald method as well.

3.6 Outcomes

The outcomes in this simulation study were power, Type I error rate, and parameter recovery. In order to provide more insight and clarify how to measure power and Type I error for the Bayesian MIMIC model the methods used for evaluating power and Type I error for the improved

Wald χ^2 (non-Bayesian) tests will be described first. Understanding choices in this context (i.e. Frequentist) is probably more familiar to most researchers and will help in understanding how the choices for evaluating power and Type I error were made for the Bayesian MIMIC model.

To conduct the DIF tests using the improved Wald χ^2 contrasts were carried out for all items except the anchor items (i.e. items 17-20). The same contrasts used in Woods et al. (2013) for testing two group and three group DIF were considered here. That is for the two group case the contrast matrix for a given item was

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \tag{3.8}$$

and for the three group case

$$\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}. \tag{3.9}$$

The improved Wald χ^2 test is a Frequentist method and uses an asymptotic statistical test to assess the classification of DIF in a given item. In the present study in contrast to previous research (e.g. S. H. Kim et al., 1994; Woods et al., 2013; Woods & Grimm, 2011) only the pairwise DIF tests for each group were analyzed and the omnibus test of DIF was not considered. The reasons for doing this are two-fold. First, in the Bayesian framework a multivariate omnibus test for DIF in the discrimination and threshold parameters with similar properties to the Frequentist omnibus test is not straightforward to compute. One possible avenue for a multivariate test would be to use a Bayes factor (see Verhagen, 2012 for univariate application). However, a multivariate Bayes factor although theoretically possible would be quite complicated and computationally expensive to compute (Morey, Rouder, Pratte, & Speckman, 2011, p. 371). One goal of the present study is to provide a framework for DIF testing in Bayesian models that is straightforward to implement by practitioners and other researchers. The second reason for considering only the pairwise tests is that to the best of my knowledge this is the first time the properties of the pairwise tests have been considered for the Bayesian MIMIC model and the improved Wald χ^2. Thus, only the pairwise tests are considered in the proposed study for both the improved Wald χ^2 and Bayesian MIMIC model.

For the improved Wald χ^2 the unconditional test of the α parameters is

$$\chi^2_{\alpha_{ih}} = \frac{(\hat{\alpha}_{F_{ih}} - \hat{\alpha}_{R_i})}{\sigma^2_{\hat{\alpha}_{ih}}}, \tag{3.10}$$

where $\hat{\alpha}_{F_{ih}}$ and $\hat{\alpha}_{R_i}$ are the maximum likelihood estimates of the ith item discrimination parameter for focal group h and reference groups, respectively. $\sigma^2_{\hat{\alpha}_{ih}}$ is the asymptotic variance of the difference between focal group h and reference group discrimination parameters. Equation 3.10 is asymptotically distributed as χ^2 with one degree of freedom. In the case of two groups with 16 DIF candidate items in the present study there were 16 tests for non-uniform DIF and in the three group case 32 tests. The improved Wald χ^2 also conducts a conditional test of the b parameters[2]. The conditional test conditions on equal α parameters and is given as

$$\chi^2_{b_{ih}|\alpha_i} = \chi^2_i - \chi^2_{\alpha_{ih}}, \tag{3.11}$$

where χ^2_i is given by Equation 2.10. Equation 3.11 is also asymptotically distributed as χ^2 with one degree of freedom. For the case of two groups in the present study there were 16 tests and in the case of three groups 32 tests for uniform DIF. Thus, the total number of tests (i.e. both uniform and non-uniform DIF tests) in the two group case was 32 and 64 for the three group case.

In the case of a Bayesian MIMIC model an asymptotic statistical test is not needed since the appropriate values from the posterior distribution can be calculated by integration (in practice summation). For the Bayesian MIMIC model a contrast matrix will not be used and instead the analogous dummy codes for Equations 3.8 and 3.9 were used to test for DIF (see Equation 2.41 and Figure 2.2 for details). As with the improved Wald χ^2 pairwise DIF tests were used to examine the power and Type I error rates.

To determine whether an item had DIF or not a Bayesian estimation method needs to be used to detect DIF as well. In this regard the β and ω parameters in Equation 2.41 can be tested

[2]In Langer (2008) b = -c/α where c is the intercept parameter.

with Bayesian credible intervals. In the context of Bayesian hypothesis testing Box and Tiao (1973) describe this as the highest posterior density (HPD) interval. The HPD credible interval containing $1-\alpha$ percent of the probability under the posterior distribution is referred to the $1-\alpha$ percent HPD interval.[3] Assuming a specified α level for rejection of the null hypothesis that a parameter value is zero, the HPD interval can be used to test for both uniform and non-uniform DIF.

In the present study if an HPD interval for β and/or ω (see Equation 2.41) does not contain 0 then this item would be flagged as either having uniform (β), non-uniform (ω), or mixed DIF (both β and ω). For two groups there will be 32 total HPD tests (i.e. 16 for β and 16 for ω) and in the three group case 64 total HPD tests (i.e. 32 for β and 32 for ω). Also, to better align with the improved Wald χ^2 test for both two and three groups a conditional test of equal α parameters will be used for the Bayesian MIMIC model. This involves fitting the model with all DIF candidate items (i.e. items 1 to 16) having β and ω freely estimated (unconditional DIF test of αs). Then non-uniform DIF will be assessed by examining the $1-\alpha$ percent HPD intervals for only the discrimination parameters. A second model will then be fit where all the discrimination parameters for a given item are set equal across the groups and the test for uniform DIF will be evaluated on this model using the HPDs for the threshold parameters (conditional DIF test of the bs).

3.6.1 Issues with Multiple Testing

The proposed study unlike past DIF research (e.g. S. H. Kim et al., 1994; Woods et al., 2013; Woods & Grimm, 2011) does not use an omnibus test for DIF and instead utilizes the individual item parameter tests. This introduces a concern of inflated Type I errors. This issue is complicated by the fact that a Bayesian estimation method (the Bayesian MIMIC model) and a Frequentist method (improved Wald χ^2) are being compared. In Bayesian estimation many authors note that one should not worry too much about Type I error rates (e.g. Gelman et al., 2013; Gelman & Hill, 2006; Gelman, Hill, & Yajima, 2012; Kruschke, 2010). However, other people within the DIF

[3]Note that α here is referring to the Type I error rate used in notation for hypothesis testing and not the the IRT discrimination parameter

literature have noted that even in a Bayesian paradigm multiple testing can become problematic (e.g. Cho, 2007). Thus, this issue is somewhat complex to navigate. To address this issue I turn to the simulation research on DIF testing.

Within the MIMIC DIF simulation literature Woods (2009b) used a BH p-value correction as recommended by Thissen et al. (2002). Type I error results from Woods (2009b) indicated that the BH p-value correction seemed to be an overcorrection and may not be necessary. Woods and Grimm (2011) also noted that the BH p-value correction used in Woods (2009b) seemed to be an over correction and did not use the BH p-value correction for evaluating non-uniform DIF in MIMIC models. Although Woods and Grimm (2011) used the omnibus test to detect DIF they still conducted 16 tests without a p-value correction and noted no issue of Type I error inflation under ideal conditions.[4]

More recently, Woods et al. (2013) tested the improved Wald χ^2 with designated anchors (denoted as Wald-I) and the LR DIF test in both two and three groups. In this study there were also 16 omnibus DIF tests per group (i.e. 16 in two groups and 32 in three groups) for both methods. Woods et al. (2013) did not use a p-value correction and for both Wald-I and LR DIF methods there was no inflation of Type I error rates even in the case of 32 tests (i.e. three groups). Further, J. Kim and Oshima (2012) conducted a simulation study to directly address the multiple comparisons issue in uniform DIF testing. Within this study they found that Lord's (1980) χ^2 (not the improved Wald χ^2) did not require p-value adjustments to control Type I errors using Bonferroni, Holm's or the BH correction for the conditions in their study. J. Kim and Oshima (2012) noted that p-value adjustments may not be necessary for some IRT based DIF methods.

Given the above findings from simulation research it is possible that a p-value adjustment may not be necessary for either the Bayesian MIMIC model and/or the improved Wald χ^2. Nevertheless, to be on the safe side given that the present study will conduct 32 and 64 comparisons and that the previous evidence was on the omnibus tests or tests of uniform DIF I propose the following strategy. The strategy involves adjusting the p-values (in the Frequentist case) and ad-

[4]Woods and Grimm (2011) did find inflated Type I errors for the non-uniform DIF MIMIC model but this was likely due to the latent interaction as the uniform DIF MIMIC model did not experience these issues.

justing the α level in the HPD case. Given that the optimal correction value is unknown (since a correction may or may not be needed) a priori I am opting to use several levels of correction for both the Frequentist case and the Bayesian case in order to guard against the possibility that Type I errors become a problem. Further, using this strategy will also provide some additional evidence to the insights that p-value adjustments may not be necessary for some IRT based DIF methods (e.g. J. Kim & Oshima, 2012; Woods & Grimm, 2011).

The following α levels used were as follows: (1) 0.05, (2) 0.01, (3) 0.005 (3) 0.0030/0.0015 (4) .0015/.0008. The first level corresponds to no correction. The second and third level corresponds to adjustments that may be used in practice when practitioners want to be slightly more conservative in multiple testing scenarios and have also been used in DIF simulations (e.g. S. H. Kim et al., 1994). The fourth level corresponds to a Bonferroni type correction. This correction is based on adjusting for the 16 tests (two group case) or 32 tests (three group case) for each parameter (i.e. 16 tests for ω and 16 tests for β). The fifth level corresponds to a full Bonferroni correction among all the tests for a given group condition (i.e. 32 in two groups and 64 in three groups). Each of these corrections would be computed for each replication within a condition. Taken together this strategy allows for the possibility that no or a slight correction may be necessary, but also will guard against problematic Type I error rates should they become a problem.

3.6.2 Power and Type I Error

Power for the improved Wald χ^2 was calculated for items known to show DIF (i.e. items one through six). For both the uniform and non-uniform tests of DIF the flag values of the χ^2 statistic for considering if an item has DIF or not will be based on the five p-value rules from the previous section. In the case of two groups if $\alpha = 0.05$ then the flag value was 3.85, if $\alpha = 0.01$ the value was 6.63, if $\alpha = 0.005$ then the flag value was 7.88, if $\alpha = 0.003$ the flag value was 8.75, and if $\alpha = 0.0015$ the flag value was 10.08. In the case of three groups the $\alpha = 0.05,\ 0.01$ and 0.005 are the same values as the two group case, if $\alpha = 0.0015$ then the flag value was 10.08, if $\alpha = 0.0008$ then the flag value was 11.24. If the χ^2 statistic from the Wald test is greater than or equal to the

appropriate flag value from above then this is considered a hit (correctly identified DIF item). If an item known to show DIF is correctly identified then this was given a one and zero otherwise. This coding creates a Bernoulli random variable from which the average of the six items for a given replication was the overall power for that replication. The overall power for a given condition was the average of all mean power computations over replications to obtain the average power for that condition.

In the case of the Bayesian MIMIC model the appropriate HPDs for either β and/or ω will have α adjusted based on the five values from above to reflect whether the appropriate HPD interval includes zero. In the case of $\alpha = 0.05$ if the 95 percent HPD interval does not include zero this would be a hit. Likewise if $\alpha = 0.0015$ then if the 99.85 percent HPD interval does not include zero then this would also be considered a hit. This same logic applies to the other α levels. The coding of a correctly identified DIF item was the same as that for the Wald χ^2 test. Given that DIF items are either categorized as uniform DIF only, non-uniform DIF only, and mixed DIF several additional pieces of information were evaluated for both the improved Wald χ^2 and the Bayesian MIMIC model.

In order to evaluate the concordance of the DIF type predicted by a given method and the type of DIF simulated a confusion matrix was constructed for the power outcomes. A confusion matrix allows for more granular information on classification error than typically reported in DIF studies. For example, the overall classification hit rate for items 1 through 6 regardless of DIF type were reported. Additionally, the classification hit rate for uniform only, non-uniform only, and mixed DIF were also reported to provide a more rich understanding of the classification accuracy of each DIF method.

For the improved Wald χ^2 overall Type I error rates were computed for items known to be DIF free (i.e. items 7-16). If an item known to not show DIF (i.e. items 7-16 as anchors are not tested) was incorrectly identified as showing DIF this was scored as one and if the DIF-free item was correctly identified as not showing DIF scored as zero. Similar to the power the average of all known DIF free items in a given replication was computed. Then the average of averages over

the replications will be the overall Type I error rate for each condition. For the Bayesian MIMIC model if the appropriate HPD for a given item known to not show DIF is correctly identified then this will be coded as zero and if the item is incorrectly identified as having DIF this will be coded as a one.

3.6.3 Parameter Recovery

The parameter outcomes of interest were the latent means, item parameters for the reference group, and item parameters for the focal group(s). To evaluate parameter recovery from these three outcomes bias will be used as computed by estimate minus true. As described previously, Woods (2009b), Woods et al. (2009), and Woods and Grimm (2011) note that when computing parameter recovery for some parameter estimates (i.e. items with DIF) a final model was used. Other researchers computing parameter recovery did not fit a final DIF model when computing parameter recovery (e.g. Barendse et al., 2010; Barendse et al., 2012; Carroll, 2014). In the present study a final model will not be fit and parameter recovery will be determined by the unconditional Bayesian MIMIC model used to test non-uniform DIF and the IRT model in flexMIRT™. The reason for this is that the main interest of the present study is in DIF classification and not in fitting a final model to determine what the parameters of the overall scale would be.

The item parameter recovery for the reference group was computed as the respective estimate from the Bayesian MIMIC or improved Wald method in flexMirt™minus the true value in the given replication. For the Bayesian MIMIC model the focal group location parameters were computed by subtracting the δ value of the posterior mean of the β_{ih} from the posterior mean value of b_i for the reference group for a given item and focal group h. Additionally, the b_i value of the reference group was multiplied by -1 to put the parameters on the IRT metric from which they were simulated. For the discrimination parameters the posterior mean value of ω_{ih} was added to the posterior mean value of the reference group α_i for a given item i and focal group h. Then these focal group values (i.e. the posterior mean values of the focal group) were subtracted from the true values to obtain bias. For each method (i.e. improved Wald and the Bayesian MIMIC

model) the average bias in a replication was calculated and then the mean of the average bias was taken over the replications to obtain the overall estimate of bias.

The latent means for each focal group in the improved Wald were calculated as the estimate minus true to obtain an estimate of bias. For the Bayesian MIMIC model the posterior mean value of the latent means for each focal group were used as the estimate to calculate bias. In the same way as with the improved Wald the latent mean minus the true value will be the bias. For each method (i.e. improved Wald and the Bayesian MIMIC model) the average bias in a replication was calculated and then the mean of the average bias was taken over the replications to obtain the overall estimate of bias.

Chapter 4

Results of the Simulation Study

4.1 Overview

The results of the simulation study are presented below. The outcomes are organized into two main sections: 2 group and 3 group results. Note that the 3 group results include the results for the reference group, focal group 1, and focal group 2. For both the 2 group and 3 group results there are four main subsections: Overall Type I error rate, overall power, confusion matrices for classification accuracy, and parameter recovery. Note that when interpreting Type I error Wilson's score interval (Wilson, 1927) was used to compute a confidence interval for the proportion p to aid interpretation. Wilson's interval was chosen because of its good performance compared with other methods (see Brown, Cai, & DasGupta, 2001). The formula for Wilson's score interval is

$$\frac{1}{1 + 1/n * z^2} * \left[\hat{p} + 1/(2 * n) * z^2 \pm z * \sqrt{1/n * \hat{p} * (1 - \hat{p}) + 1/(4 * n^2) * z^2}\right]. \tag{4.1}$$

where $z = 1.96$ here, $\hat{p} = 0.05$ for nominal Type I error rate, and n denotes the number of converging replications.

If the number of converging replications was 100 then the 95% CI bounds would be $[0.022, 0.112]$. So if a method had a Type I error rate above 0.112 then this was flagged as inflated Type I error and if the Type I error was below 0.022 too low of Type I error. When contrasting power among

the two methods within a given condition a paired t-test was used with a Bonferroni correction for nine planned comparisons to control the family wise error rate. If a given test was significant at the $0.005/9 = 0.0056$ α level the methods were deemed to have significantly different power. The following subsections discuss the results of the convergence criteria, Type I error controls, and parameter recovery that informed the analyses that follow.

4.1.1 Convergence

Convergence diagnostics were assessed for both the Bayesian MIMIC model and the improved Wald as described in the method section. For the Bayesian MIMIC model convergence was quite good overall with only 2 replications out of the 1800 condition reps (0.111%) not converging. These replications were both from the 2 group, latent SD of focal groups (LSD) of 1, and equal (500:500) sample size condition. All other conditions for the Bayesian MIMIC model had 100% convergence. The improved Wald had a total of 7 out of the 1800 condition reps (0.389%) not converge. All of the non-converging replications occurred in the condition corresponding to an LSD of 0.50, and sample size of 1300:100:100 for the 3 group condition.

For the convergence issues of the improved Wald I explored the category frequencies of the non-converging data sets. In almost all cases there were situations were there may have only been 2 successes or 3 successes on certain items out of 100. Moreover, I hypothesize that the root cause of this problem was the fact that a maximum likelihood routine with a highly parameterized model was fit under less than ideal conditions which was why there were convergence problems in this condition. In essence this illustrates the breaking point of using the 2PL model with highly unbalanced data under less than ideal circumstances. For the purposes of analyzing the results, only those conditions for which 100% convergence was obtained for both methods were used in the analyses of the outcomes. For example in the LSD of 0.50, sample size of 1300:100:100, and 3 group condition only the 93 converged reps in both methods were used in the analyses.

4.1.2 Issues with Multiple Testing and Parameter Recovery

Upon reviewing the results for the various Type I error guards described in the method section only results of the critical value of $p = 0.05$ are presented across all conditions. In conditions where the Bayesian MIMIC model should not see excessive Type I error rates (i.e. with the LSDs fixed at 1) there were no inherent problems. In looking at the other values for the 2 and 3 group cases lowering the p-value below 0.05 provided more Type I error control but at a loss in power and too low of Type I error rates. Thus, only results with $\alpha = 0.05$ were reported.

When calculating bias some values were extreme and these outliers for both discrimination and difficulty parameters were recoded. This practice of recoding is similar to that used by Woods (2009b). Item discrimination parameters that were greater than 4 were recoded to 4 and also item difficulty parameters that were less than -4 and/or greater than 4 were recoded to -4 and 4 respectively. When the results were analyzed without these controls the results of parameter recovery had more bias than what was reported. All item parameter recovery results in this study are presented with the aforementioned recoding of the discrimination and difficulty parameters that were flagged as outliers.

4.2 Results of 2 Group Conditions

4.2.1 Overall Type I Error and Power

Figure 4.1 shows the Type I error rates for the nine 2 group conditions using an $\alpha = 0.05$ threshold and the improved Wald with SEM SEs. The three labels on the left side of the figure represent the sample sizes (i.e. equal [500:500], moderately unequal [750:250], and highly unequal [900:100]) and the columns on the x-axis represent the three LSD conditions (i.e. lower = 0.5, equal = 1, and higher = 1.5). Note that in the top middle of Figure 4.1 (i.e. LSD of 1 and equal sample size condition) there were only 98 converging reps and the 95% CI was [0.021, 0.113], whereas all other conditions had 100 converging replications with a 95% CI of [.022, 0.112]. In general for both methods as the sample size imbalance increased the Type I error decreased. Looking at

Figure 4.1 both the Bayesian MIMIC model and improved Wald had well controlled Type I errors across all conditions.

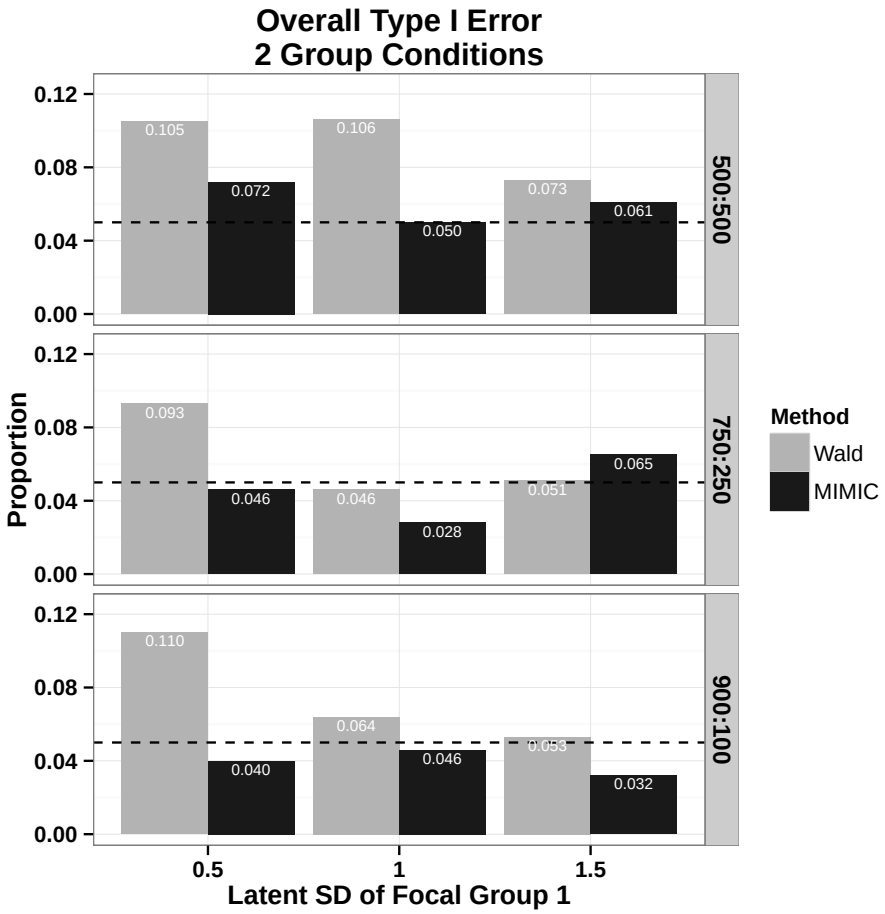
Power for the improved Wald with supplemental EM SEs and Bayesian MIMIC model using an $\alpha = 0.05$ are presented in Figure 4.2. In general as the sample size imbalance increased the power tended to decrease for both methods. Given the results from Figure 4.1 power can be interpreted unambiguously across all conditions for both methods. Looking at Figure 4.2 the power for the Bayesian MIMIC model was lower than the improved Wald in the equal sample size (500:500) and LSD of 0.5 and 1 conditions. In all other conditions the power results were not different between the two methods.

4.2.2 Confusion Matrices

The confusion matrices for the improved Wald using SEM SEs and the Bayesian MIMIC model are presented in Tables 4.1 and 4.2 respectively. Confusion matrices provide a way to see how well each DIF classifier performs in terms of accuracy and mis-classifications. For the purposes of this dissertation these confusion matrices will be used in a descriptive nature based on the converging replications within this simulation study to gain insight into the mis-classification of DIF. For both methods the confusion matrices are set up in a grid pattern that maps directly onto the Type I error (Figure 4.1) and power (Figure 4.2) plots. For each individual subtable the left hand flap gives the condition information, (True) refers to the true type of DIF (i.e. U = uniform DIF, NU = non-uniform DIF, M = mixed DIF, and N = no DIF) and (Predicted) refers to the DIF that was predicted by the respective method (i.e. improved Wald or Bayesian MIMIC model). The diagonal elements are highlighted in bold as these represent the accuracy for a given DIF type. The sum of the diagonal elements give the method's overall classification accuracy. Note that all elements in each confusion matrix sum to 1.

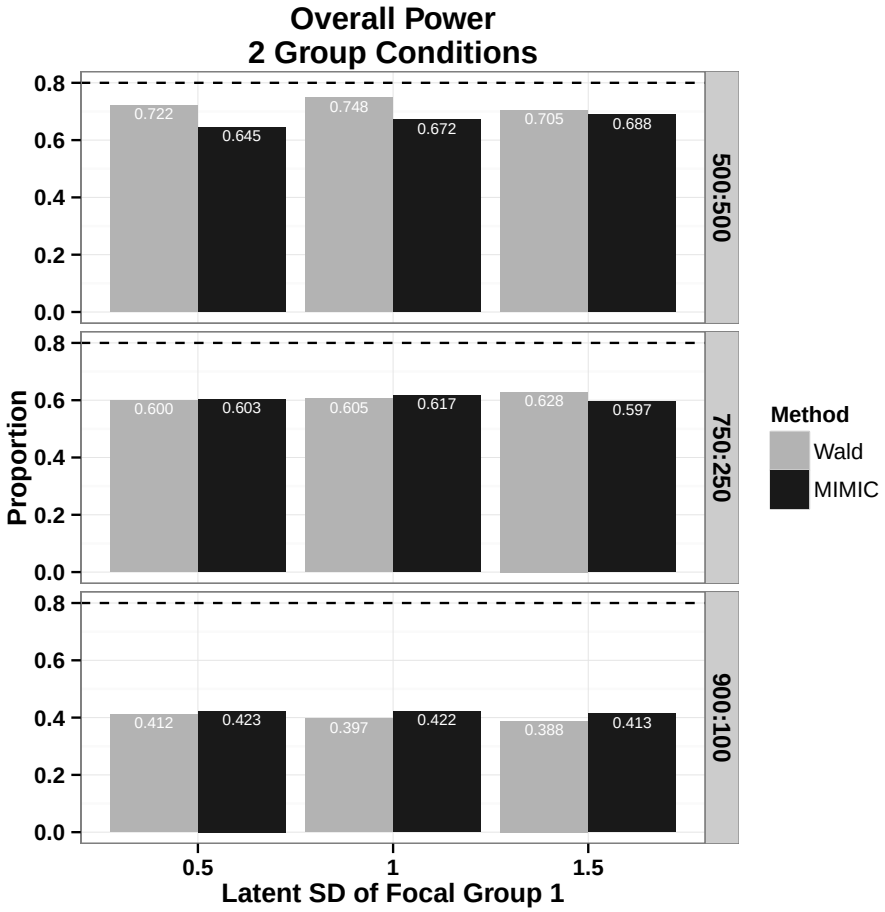
For the (U,U), (NU,NU), and (M,M) diagonal elements a perfect accuracy would be reflected by each diagonal cell being 0.125. For (N,N) perfect accuracy would be reflected as the diagonal element being 0.625. Any discrepancy between these values indicates that the method did not

Figure 4.1: Overall Type I Error for 2 Group Conditions



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500 corresponds to a reference group sample size of 500 and a focal group 1 sample size of 500. The horizontal black dashed line represents p = 0.05. The shading of the bars is for Wald = improved Wald and MIMIC = Bayesian MIMIC model. The numbers on each bar give the specific Type I error rate for that condition and method.

Figure 4.2: Overall Power for 2 Group Conditions



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500 corresponds to a reference group sample size of 500 and a focal group 1 sample size of 500. The horizontal black dashed line represents p = 0.80. The shading of the bars is for Wald = improved Wald and MIMIC = Bayesian MIMIC model. The numbers on each bar give the specific power rate for that condition and method.

correctly predict the true DIF classification in some or all instances. When interpreting the results of the confusion matrices I will refer to a given element in the grid by it's row and column position. For example, confusion matrix (1,1) will be denoted as element (1,1) and corresponds to the upper left cell with sample size (SS) of 500:500 and LSD of 0.5. Element (3, 2) would be the third row and second column corresponding to SS of 900:100 and LSD of 1.

In what follows the confusion matrices for the reference group versus focal group 1 are presented for the improved Wald and Bayesian MIMIC model. For each table the minimum and maximum accuracy are reported, the table with the maximum accuracy is interpreted, and the overall trends in the table are summarized. Note that the term DIF misclassification refers to items that are DIF items (i.e. U, NU, or M) and looks at the highest classification rate that was not N (i.e. no DIF).

4.2.2.1 Improved Wald Confusion Matrices

For the improved Wald in Table 4.1, the highest accuracy was 0.750 corresponding to element (1, 3) [SS: 500:500 and LSD: 1.5] and the lowest accuracy was 0.617 corresponding to element (3, 1) [SS: 900:100 and LSD: 0.5]. For element (1, 3) (i.e. matrix with maximum accuracy) the accuracy for uniform DIF (U,U) was 0.092 and when U was misclassified the most common misclassification was N (i.e. no DIF). For element (1, 3) the accuracy for non-uniform DIF (NU,NU) was 0.036 and the most common misclassification was N. For element (1, 3) the accuracy for mixed DIF (M,M) was 0.042 with the most common misclassification being U (i.e. uniform DIF). Lastly, for no DIF (N,N) the accuracy was 0.579 with the most common misclassification being NU (i.e. non-uniform DIF).

For all confusion matrices in Table 4.1 the most common misclassification for U and NU was N (i.e. no DIF). For M (i.e. mixed DIF) the most common misclassification was U in the equal (500:500) and moderately unequal (750:250) sample sizes [i.e elements (1,1)-(1,3) and (2,1)-(2,3)]. However, the most common misclassification for M when the sample sizes were highly unequal (900:100) [i.e. elements (3,1)-(3,3)] was N. For N the most common misclassification rate was NU.

In the moderately unequal (750:250) with LSD of 1.5 condition [i.e. element (2, 3)] the highest misclassification rate for N was U. When an item was a DIF item (i.e. U, NU, or M) the most common DIF misclassification for U was M. However, in elements (2, 3) and (3, 1) the most common DIF misclassification for U was NU. The most common DIF misclassification for both NU and M was U.

4.2.2.2 Bayesian MIMIC Model Confusion Matrices

For the Bayesian MIMIC model in Table 4.2, the highest accuracy was 0.708 corresponding to element (2, 2) [SS: 750:250 and LSD: 1] and the lowest accuracy was 0.668 corresponding to element (3, 2) [SS: 900:100 and LSD: 1]. For element (2, 2) (i.e. matrix with maximum accuracy) the accuracy for uniform DIF (U,U) was 0.100 and when U was misclassified the most common misclassification was N. For element (2, 2) the accuracy for non-uniform DIF (NU,NU) was 0.000 and the most common misclassification was N. For element (1, 2) the accuracy for mixed DIF (M,M) was 0.000 with the most common misclassification being U. Lastly, for no DIF (N,N) the accuracy was 0.608 with the most common misclassification being U.

For all confusion matrices in Table 4.2 the most common misclassification for both U and NU was N. For M the most common misclassification was U except in element (3, 1) where it was N. For N the most common misclassification was U in all conditions. When an item was a DIF item (i.e. U, NU, or M) for U the most common DIF misclassification was M. However, in five of the nine conditions neither NU or M were picked [i.e. elements (1, 2), (2, 2), (2, 3), (3,1), (3, 2), (3, 3)]. For both NU and M the most common DIF misclassification was U.

4.2.3 Parameter Recovery

4.2.3.1 Discrimination Parameters

Figure 4.3 shows the bias in the discrimination parameter estimates by condition. Looking down the middle column corresponding to a LSD of 1 we see that the Bayesian MIMIC model had good parameter recovery for both the reference and focal groups across all sample size condi-

Table 4.1: Confusion Matrices for 2 Group Conditions: Improved Wald for Reference vs Focal Group 1

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500 LSD: 0.5**
(True)	**U**	**0.086**	0.003	0.010	0.026	
	NU	0.039	**0.022**	0.009	0.055	
	M	0.057	0.014	**0.031**	0.023	
	N	0.019	0.044	0.002	**0.559**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500 LSD: 1**
(True)	**U**	**0.095**	0.001	0.004	0.024	
	NU	0.018	**0.027**	0.017	0.063	
	M	0.055	0.016	**0.047**	0.007	
	N	0.022	0.043	0.001	**0.559**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500 LSD: 1.5**
(True)	**U**	**0.092**	0.001	0.003	0.029	
	NU	0.013	**0.036**	0.009	0.067	
	M	0.052	0.016	**0.042**	0.015	
	N	0.013	0.032	0.001	**0.579**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 750:250 LSD: 0.5**
(True)	**U**	**0.072**	0.001	0.008	0.044	
	NU	0.029	**0.018**	0.008	0.070	
	M	0.052	0.015	**0.021**	0.036	
	N	0.013	0.044	0.001	**0.567**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 750:250 LSD: 1**
(True)	**U**	**0.079**	0.002	0.003	0.041	
	NU	0.013	**0.024**	0.007	0.081	
	M	0.052	0.010	**0.037**	0.026	
	N	0.005	0.023	0.001	**0.596**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 750:250 LSD: 1.5**
(True)	**U**	**0.081**	0.003	0.001	0.041	
	NU	0.008	**0.029**	0.007	0.081	
	M	0.065	0.009	**0.034**	0.017	
	N	0.018	0.014	0.000	**0.593**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 900:100 LSD: 0.5**
(True)	**U**	**0.039**	0.005	0.004	0.077	
	NU	0.028	**0.009**	0.010	0.077	
	M	0.034	0.012	**0.013**	0.066	
	N	0.017	0.044	0.007	**0.556**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 900:100 LSD: 1**
(True)	**U**	**0.046**	0.001	0.003	0.076	
	NU	0.007	**0.019**	0.001	0.098	
	M	0.041	0.008	**0.024**	0.052	
	N	0.016	0.024	0.000	**0.585**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 900:100 LSD: 1.5**
(True)	**U**	**0.038**	0.001	0.003	0.083	
	NU	0.006	**0.019**	0.006	0.094	
	M	0.049	0.007	**0.017**	0.052	
	N	0.012	0.021	0.001	**0.592**	

Note. U = uniform DIF, NU = non-uniform DIF, M = mixed DIF, N = no DIF, SS = sample size, LSD = latent focal group SD.

Table 4.2: Confusion Matrices for 2 Group Conditions: Bayesian MIMIC Model for Reference vs Focal Group 1

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500 LSD: 0.5**
(True)	**U**	**0.094**	0.000	0.004	0.027	
	NU	0.033	**0.003**	0.007	0.083	
	M	0.059	0.016	**0.026**	0.024	
	N	0.026	0.016	0.003	**0.580**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500 LSD: 1**
(True)	**U**	**0.103**	0.000	0.000	0.022	
	NU	0.041	**0.000**	0.001	0.082	
	M	0.105	0.001	**0.001**	0.018	
	N	0.031	0.000	0.000	**0.594**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500 LSD: 1.5**
(True)	**U**	**0.102**	0.000	0.001	0.022	
	NU	0.046	**0.000**	0.000	0.079	
	M	0.110	0.000	**0.000**	0.015	
	N	0.037	0.001	0.000	**0.587**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 750:250 LSD: 0.5**
(True)	**U**	**0.096**	0.000	0.001	0.028	
	NU	0.037	**0.000**	0.001	0.087	
	M	0.089	0.001	**0.001**	0.034	
	N	0.028	0.001	0.000	**0.596**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 750:250 LSD: 1**
(True)	**U**	**0.100**	0.000	0.000	0.025	
	NU	0.039	**0.000**	0.000	0.086	
	M	0.092	0.000	**0.000**	0.033	
	N	0.018	0.000	0.000	**0.608**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 750:250 LSD: 1.5**
(True)	**U**	**0.095**	0.000	0.000	0.030	
	NU	0.026	**0.000**	0.000	0.099	
	M	0.102	0.000	**0.000**	0.022	
	N	0.040	0.001	0.000	**0.584**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 900:100 LSD: 0.5**
(True)	**U**	**0.071**	0.000	0.000	0.054	
	NU	0.026	**0.000**	0.000	0.099	
	M	0.062	0.000	**0.000**	0.063	
	N	0.025	0.000	0.000	**0.600**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 900:100 LSD: 1**
(True)	**U**	**0.072**	0.000	0.000	0.053	
	NU	0.019	**0.000**	0.000	0.106	
	M	0.068	0.000	**0.000**	0.058	
	N	0.029	0.000	0.000	**0.596**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 900:100 LSD: 1.5**
(True)	**U**	**0.068**	0.000	0.000	0.057	
	NU	0.017	**0.000**	0.000	0.108	
	M	0.070	0.000	**0.000**	0.055	
	N	0.020	0.000	0.000	**0.605**	

Note. U = uniform DIF, NU = non-uniform DIF, M = mixed DIF, N = no DIF, SS = sample size, LSD = latent focal group SD.

tions. This was to be expected as the assumption of homogeneity of latent SDs was met for the Bayesian MIMIC model. Additionally, the improved Wald also had good parameter recovery of the discrimination parameters across the LSD of 1 conditions.

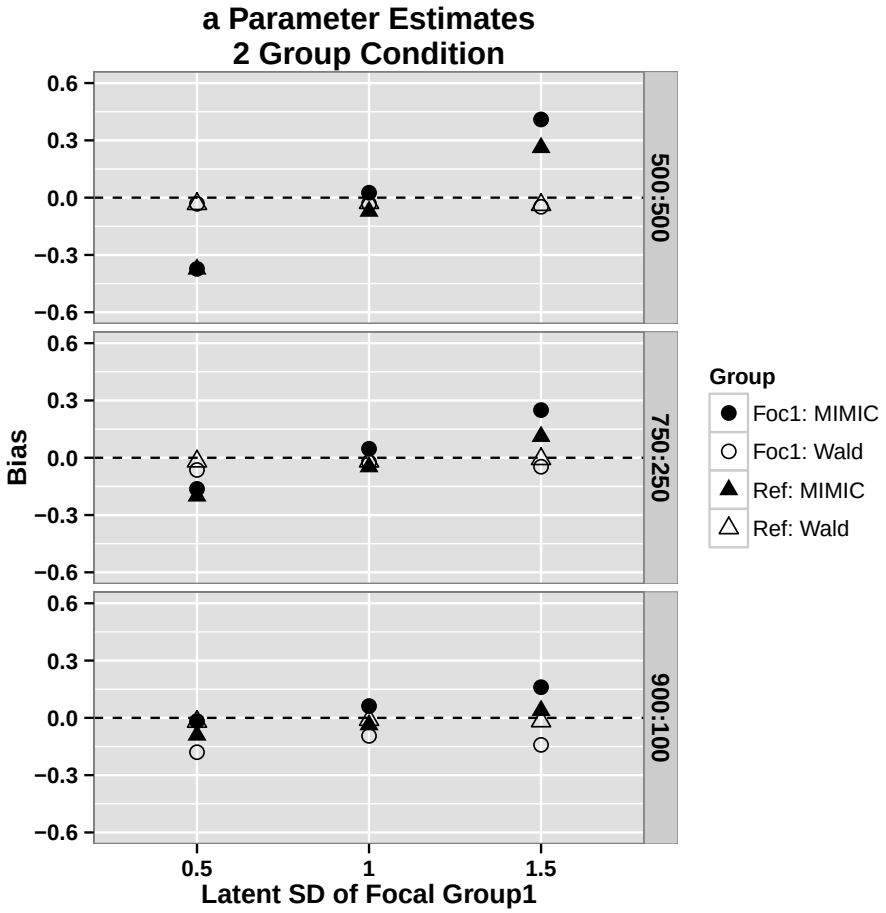
Looking at the far right column corresponding to a LSD of 1.5 the Bayesian MIMIC model parameters were consistently overestimating the discrimination parameters and the improved Wald discrimination parameters were generally well recovered. The Bayesian MIMIC model discrimination parameters were more biased than the improved Wald parameters in the equal (500:500) and moderately unequal (750:250) sample size conditions. In the highly unequal sample size condition (900:100) the two methods were about equally biased, however, the Bayesian MIMIC model overestimated the discrimination parameters and the improved Wald underestimated the true parameters.

For the LSD of 0.5 conditions the improved Wald had good parameter recovery in the equal (500:500) and moderately unequal (750:250) sample size conditions. In the highly unequal (900:100) sample size condition the reference group parameters were well recovered and the focal group parameters were slightly underestimated. When the sample sizes were equal (500:500) the Bayesian MIMIC model underestimated the discrimination parameters. In the moderately unequal (750:250) sample size condition this underestimation gap lessened in comparison to the equal sample size condition. When the sample sizes were highly unequal (900:100) the Bayesian MIMIC model had reasonable parameter recovery of the discrimination parameters for both reference and focal group 1.

4.2.3.2 Difficulty Parameters

Figure 4.4 shows the bias in the difficulty parameter estimates by condition. Looking down the middle column corresponding to a LSD of 1 we see that the Bayesian MIMIC model had good parameter recovery in all three sample sizes. The improved Wald also had adequate parameter recovery across all three sample size conditions. In the LSD of 0.5 conditions the Bayesian MIMC model overestimated the reference and focal group parameters. This effect was most severe in the

Figure 4.3: Discrimination Parameter Bias for 2 Group Conditions



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500 corresponds to a reference group sample size of 500 and a focal group 1 sample size of 500. The dashed line indicates no bias. Foc1 = focal group 1, Ref = reference group. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

equal sample size (500:500) condition with the overestimation becoming less severe as the sample sizes became more unbalanced. For the improved Wald method parameter recovery was adequate in all sample size conditions. For the LSD of 1.5 conditions the improved Wald method had good parameter recovery of the difficulty parameters in all three sample size conditions. For the Bayesian MIMIC model the difficulty parameters were underestimated in the equal sample size condition and this overestimation effect essentially disappeared in the moderately unbalanced (750:250) and highly unbalanced (900:100) sample size conditions. Note that the degree of bias for the difficulty parameters was less striking than those of the discrimination parameters in Figure 4.3.

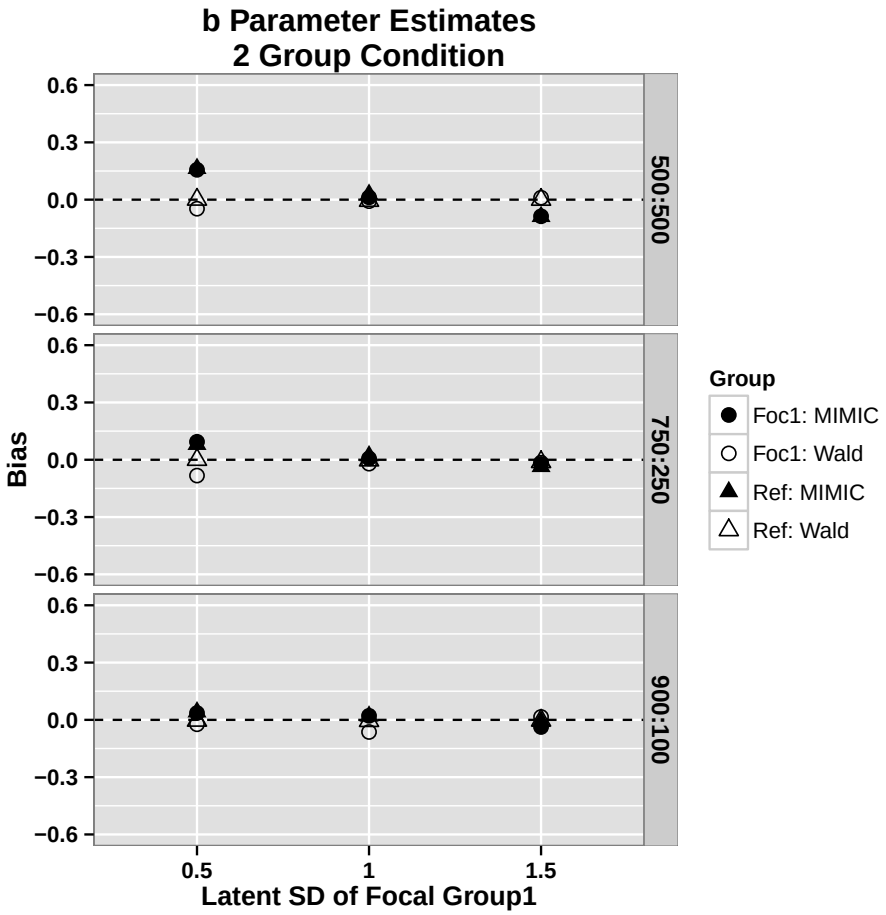
4.2.3.3 Latent Means

Figure 4.5 illustrates the bias in the latent means for focal group 1 by method. The plot shows that overall the improved Wald test does a good job of recovering the true mean difference on the latent variable in all conditions. The Bayesian MIMIC model underestimated the latent mean of focal group 1 when the LSD was 0.5. This underestimation effect slightly lessened as the sample size imbalance increased. When the assumption of homogeneity of LSDs held the Bayesian MIMIC model had good parameter recovery of the latent mean of focal group 1 across all sample size conditions. When the LSD was 1.5 the Bayesian MIMIC model slightly overestimated the difficulty parameters. This overestimation slightly lessened as the sample size imbalance increased.

4.3 Results of 3 Group Conditions

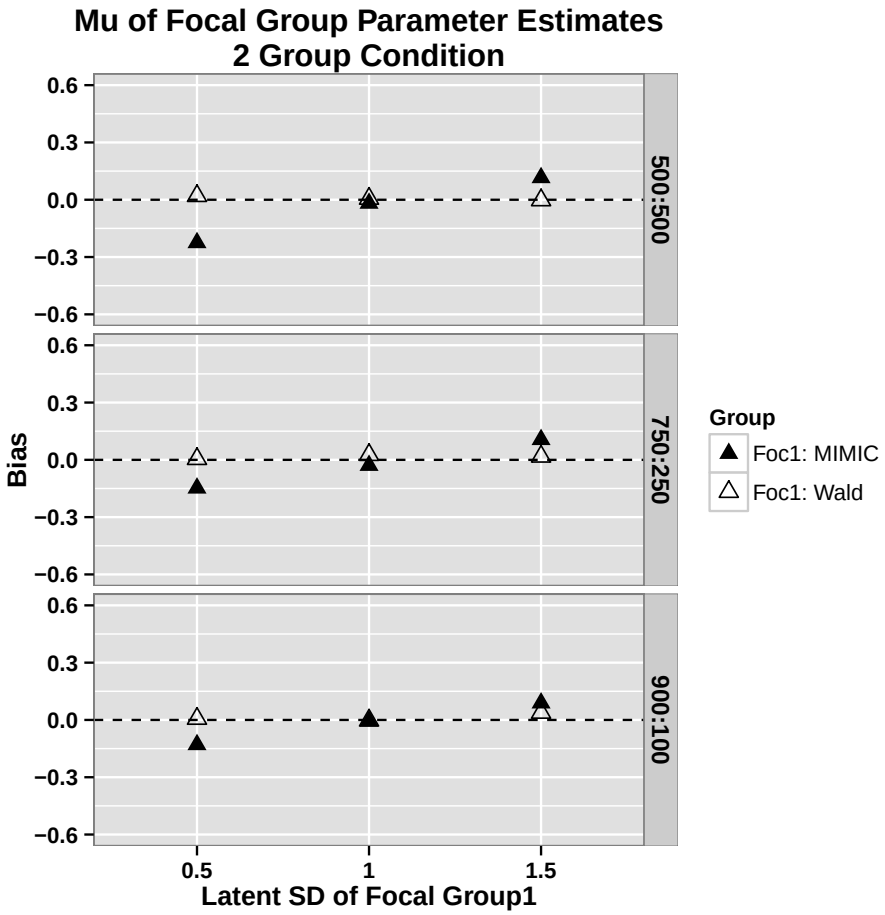
In general the pattern of results from the 2 group case and 3 group case were similar. As in the two group case a 95% CI was used to flag items for excessive Type I error rates. For conditions having 100 converging reps the 95% CI was $[0.022, 0.112]$. Note that in the LSD of 0.5 and highly unequal sample size condition the improved Wald with SEM SEs had 93 converging replications, thus the 95% CI for this condition was [0.021, 0.115].

Figure 4.4: Difficulty Parameter Bias for 2 Group Conditions



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500 corresponds to a reference group sample size of 500 and a focal group 1 sample size of 500. The dashed line indicates no bias. Foc1 = focal group 1, Ref = reference group. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

Figure 4.5: Latent Mean Parameter Bias for 2 Group Conditions



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500 corresponds to a reference group sample size of 500 and a focal group 1 sample size of 500. The dashed line indicates no bias. Foc1 = focal group 1. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

4.3.1 Overall Type I Error: Reference vs. Focal Groups 1 and 2

Figure 4.6 shows the Type I error rates for the nine 3 group conditions for reference group versus focal group 1 with the supplemental EM SEs. The three labels on the left represent the sample sizes (i.e. equal [500:500:500], moderately unequal [1000:250:250], and highly unequal [1300:100:100]) and the columns on the x-axis represent the three focal group latent SD conditions (i.e. lower = 0.5, equal = 1, and higher = 1.5). From Figure 4.6 both methods had well controlled Type I errors across all conditions for reference versus focal group 1. Additionally, these same conclusions can be seen for reference versus focal group 2 in Figure 4.7.

4.3.2 Overall Power: Reference vs. Focal Groups 1 and 2

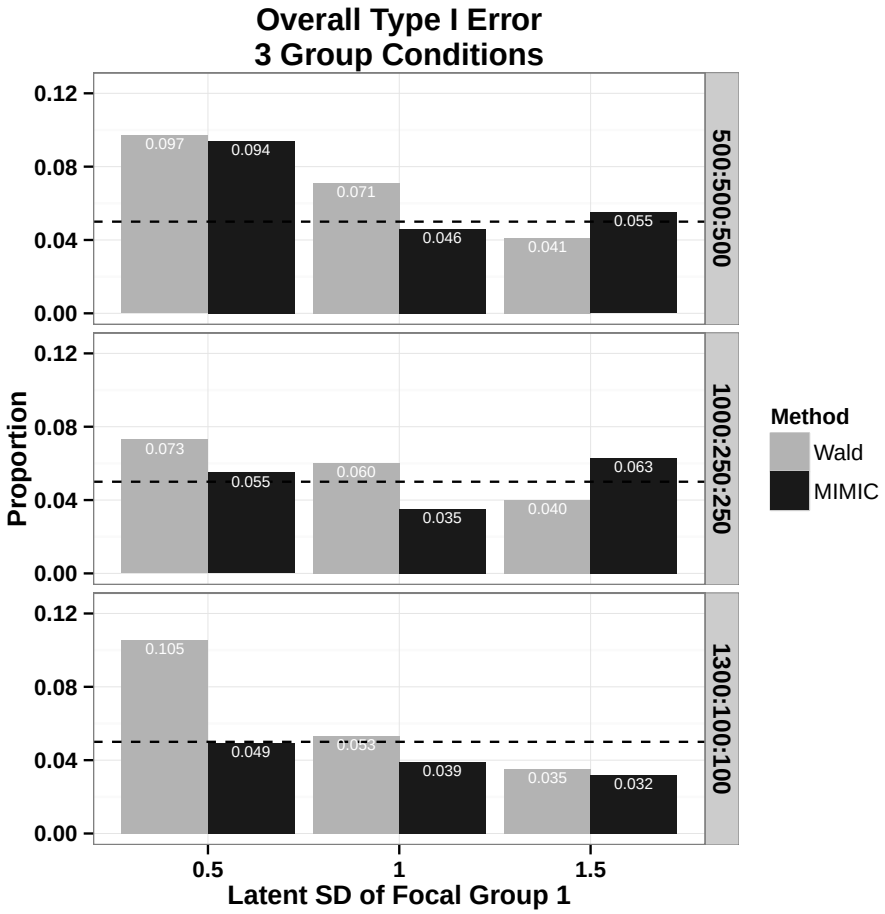
Figures 4.8 and 4.9 illustrate the power for the 3 group conditions of reference versus focal group 1 and reference versus focal group 2 with SEM SEs and an $\alpha = 0.05$, respectively. As with the two group case power was lower as the sample size imbalance increased. Power can be interpreted in all conditions for both methods. Looking at Figure 4.8 (i.e. reference vs. focal group 1) the Bayesian MIMIC model and the improved Wald did not show any differences in power across all conditions. Further, Figure 4.9 (i.e. reference versus focal group 2) also showed that power was not different for both methods.

4.3.3 Confusion Matrices

The confusion matrices for the improved Wald and the Bayesian MIMIC model are presented in Tables 4.3, A.1, 4.4, and A.2. The layout and interpretation of the confusion matrices is the same as described previously for the 2 group case. For the purposes of this dissertation these confusion matrices will be used in a descriptive nature based on the converging replications within this simulation study to gain insight into the mis-classification of DIF.

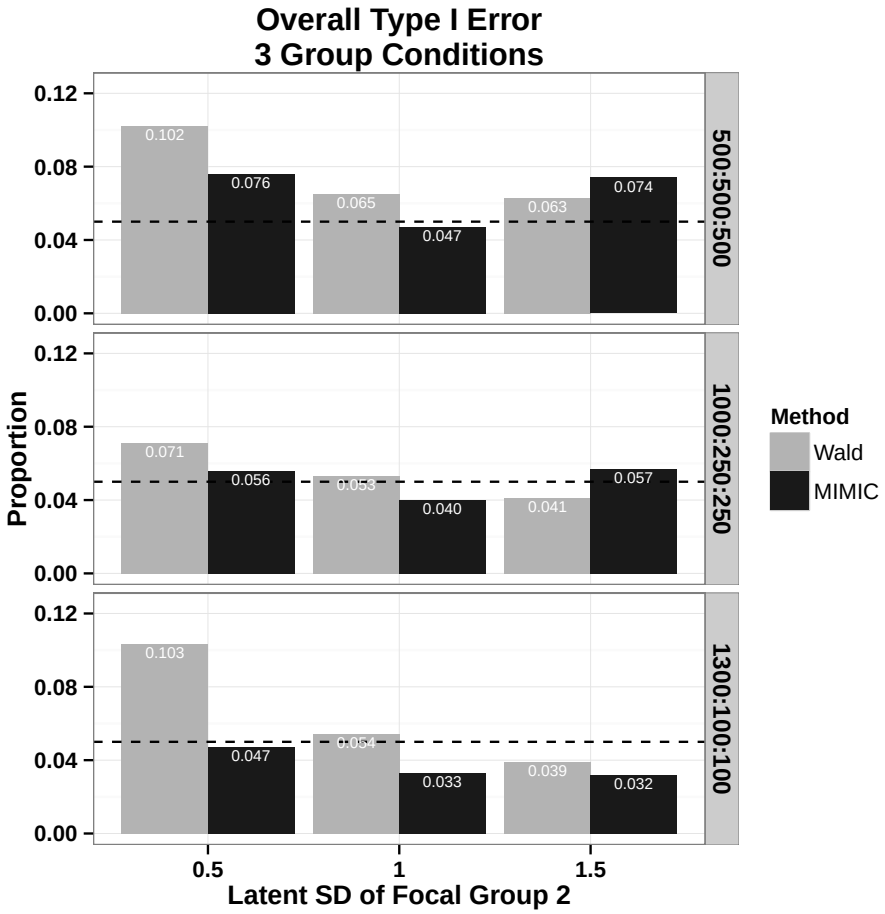
In what follows the confusion matrices for the reference group versus focal group 1 are presented. The results for reference and focal group 2 generally yielded the same general pattern as

Figure 4.6: Overall Type I Error for 3 Group Conditions: Reference vs. Focal Group 1



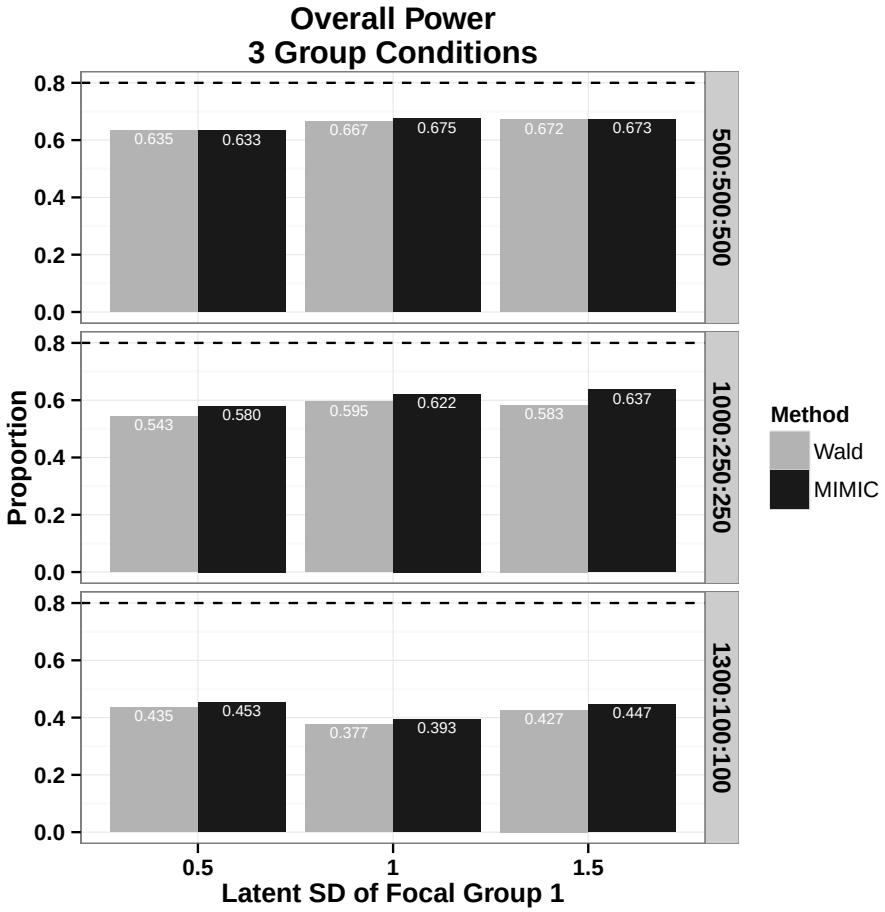
Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. The horizontal black dashed line represents p = 0.05. The shading of the bars is for Wald = improved Wald and MIMIC = Bayesian MIMIC model. The numbers on each bar give the specific Type I error rate for that condition and method.

Figure 4.7: Overall Type I Error for 3 Group Conditions: Reference vs. Focal Group 2



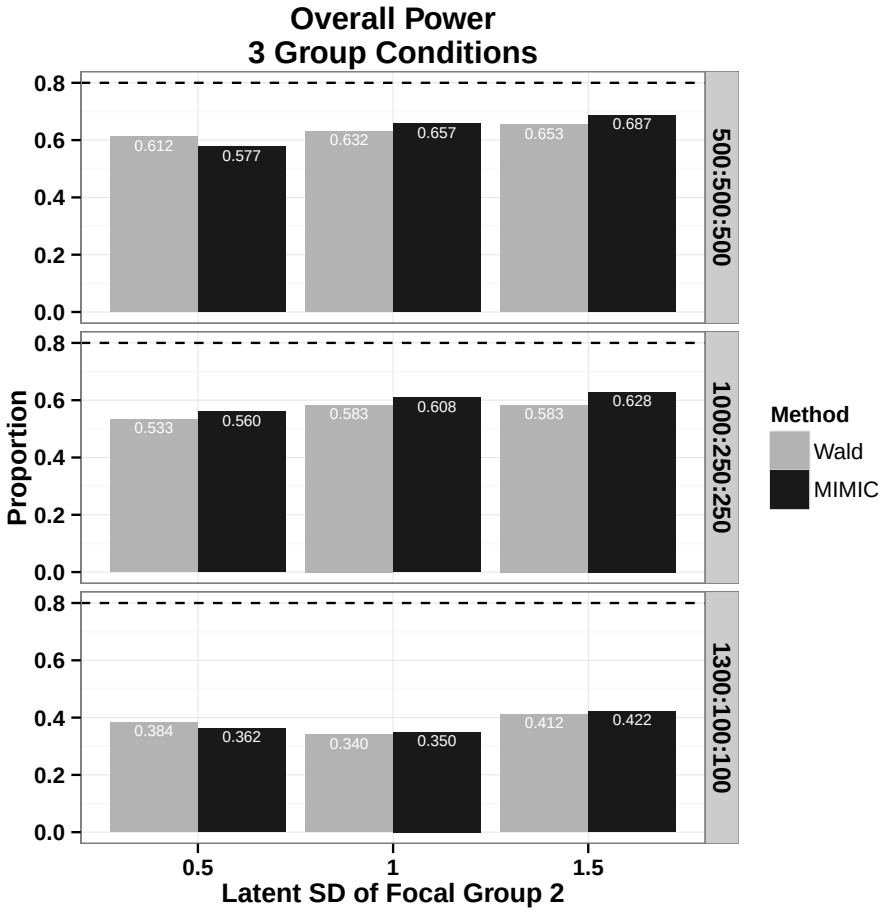
Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. The horizontal black dashed line represents p = 0.05. The shading of the bars is for Wald = improved Wald and MIMIC = Bayesian MIMIC model. The numbers on each bar give the specific Type I error rate for that condition and method.

Figure 4.8: Overall Power for 3 Group Conditions: Reference vs. Focal Group 1



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. The horizontal black dashed line represents p = 0.80. The shading of the bars is for Wald = improved Wald and MIMIC = Bayesian MIMIC model. The numbers on each bar give the specific power rate for that condition and method.

Figure 4.9: Overall Power for 3 Group Conditions: Reference vs. Focal Group 2



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. The horizontal black dashed line represents p = 0.80. The shading of the bars is for Wald = improved Wald χ^2 and MIMIC = Bayesian MIMIC model. The numbers on each bar give the specific power rate for that condition and method.

reference versus focal group 1 and are not presented here. The confusion matrices for reference versus focal group 2 are available in the appendix. For each table the minimum and maximum accuracy are reported, the table with the maximum accuracy is interpreted, and the overall trends in the table are summarized. Note that the term DIF misclassification refers to items that are DIF items (i.e. U, NU, or M) and looks at the highest misclassification rate that was not N (i.e. no DIF). Note that the results of using $\alpha = 0.01$ for both methods and using the cross-product SEs for the Wald method yielded the same differences as described in the section for confusion matrices in the 2 group case.

4.3.3.1 Improved Wald Confusion Matrices: Reference vs. Focal Group 1

The confusion matrices for the reference group versus focal group 1 are presented in Table 4.3. The maximum accuracy in Table 4.3 was 0.756 occuring in element (1, 3) [SS: 500:500:500 and LV: 1.5] and the minimum accuracy was 0.621 occurring in element (3, 1) [SS: 1300:100:100 and LV: 0.5]. Looking at element (1, 3) [SS: 500:500:500 and LV: 1.5] the accuracy of predicting uniform DIF (U,U) was 0.094 with the majority of mis-classifications being N (i.e. no DIF). For non-uniform DIF (NU, NU) the accuracy was 0.029 with the majority of mis-classifications being N. For mixed DIF (M, M) the accuracy was 0.034 with the majority of mis-classifications being U (i.e. uniform DIF). Lastly, the accuracy of no DIF (N, N) was 0.599 with the majority of mis-classifications being NU (i.e. non-uniform DIF).

For all confusion matrices in Table 4.3 the majority of mis-classifications for both U and NU were N. For M (i.e. mixed DIF) the majority of mis-classifications were U for all conditions except elements (3, 1) [SS: 1300:100:100 and LV: 1.5] and (3, 2) [SS: 1300:100:100 and LV: 1] where the majority was N. The majority of mis-classifications for N was NU except in element (2, 3) [SS: 1000:250:250] where it was U. When an item was a DIF item the most common DIF misclassification for uniform DIF (U) was mixed DIF (M) except in elements (3, 1) and (3, 2) where it was non-uniform DIF (NU). For non-uniform DIF (NU) and mixed DIF (M) the most common DIF misclassification was as uniform DIF (U).

Table 4.3: Confusion Matrices for 3 Group Conditions: Improved Wald for Reference vs Focal Group 1

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500:500 LSD: 0.5**
(True)	**U**	**0.078**	0.001	0.009	0.037	
	NU	0.029	**0.019**	0.011	0.065	
	M	0.050	0.016	**0.024**	0.035	
	N	0.014	0.043	0.004	**0.564**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500:500 LSD: 1**
(True)	**U**	**0.087**	0.004	0.006	0.028	
	NU	0.013	**0.030**	0.005	0.077	
	M	0.055	0.007	**0.043**	0.020	
	N	0.018	0.027	0.000	**0.581**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500:500 LSD: 1.5**
(True)	**U**	**0.094**	0.001	0.002	0.028	
	NU	0.012	**0.029**	0.006	0.078	
	M	0.060	0.014	**0.034**	0.017	
	N	0.011	0.014	0.000	**0.599**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1000:250:250 LSD: 0.5**
(True)	**U**	**0.069**	0.003	0.003	0.050	
	NU	0.033	**0.011**	0.011	0.070	
	M	0.051	0.011	**0.013**	0.051	
	N	0.005	0.039	0.001	**0.579**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1000:250:250 LSD: 1**
(True)	**U**	**0.074**	0.002	0.004	0.045	
	NU	0.012	**0.026**	0.006	0.081	
	M	0.062	0.007	**0.029**	0.026	
	N	0.015	0.022	0.000	**0.588**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1000:250:250 LSD: 1.5**
(True)	**U**	**0.069**	0.001	0.004	0.051	
	NU	0.006	**0.032**	0.002	0.085	
	M	0.064	0.007	**0.033**	0.020	
	N	0.013	0.012	0.001	**0.600**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1300:100:100 LSD: 0.5**
(True)	**U**	**0.044**	0.007	0.003	0.071	
	NU	0.024	**0.007**	0.011	0.083	
	M	0.047	0.009	**0.011**	0.058	
	N	0.021	0.040	0.005	**0.559**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1300:100:100 LSD: 1**
(True)	**U**	**0.039**	0.003	0.001	0.082	
	NU	0.012	**0.020**	0.001	0.092	
	M	0.041	0.012	**0.013**	0.060	
	N	0.008	0.022	0.003	**0.592**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1300:100:100 LSD: 1.5**
(True)	**U**	**0.048**	0.001	0.001	0.076	
	NU	0.005	**0.021**	0.002	0.097	
	M	0.060	0.006	**0.016**	0.043	
	N	0.009	0.011	0.002	**0.603**	

Note. U = uniform DIF, NU = non-uniform DIF, M = mixed DIF, N = no DIF, SS = sample size, LSD = latent focal group SD.

4.3.3.2 Bayesian MIMIC Model Confusion Matrices: Reference vs. Focal Group 1

The confusion matrices for the reference group versus focal group 1 are presented in Table 4.4. The maximum accuracy in Table 4.4 was 0.709 occuring in element (1, 2) [SS: 500:500:500 and LV: 1] and the minimum accuracy was 0.665 occurring in element (3, 2) [SS: 1300:100:100 and LSD: 1]. Looking at element (1, 2) [SS: 500:500:500 and LSD: 1] the accuracy of predicting uniform DIF (U,U) was 0.106 with the majority of mis-classifications being N. For non-uniform DIF (NU, NU) the accuracy was 0.001 with the majority of mis-classifications being N. For mixed DIF (M, M) the accuracy was 0.006 with the majority of mis-classifications being U. Lastly, the accuracy of no DIF (N, N) was 0.596 with the majority of mis-classifications being U.

For all confusion matrices in Table 4.4 the majority of mis-classifications for both U and NU were N. For M and N the majority of mis-classifications were U for all conditions. When an item was a DIF item the most common misclassification for uniform DIF (U) was as mixed DIF (M) although this only occured in element (1,1) and in all other elements neither M or NU were a majority. For non-uniform DIF (NU) and mixed DIF (M) the most common misclassification was uniform DIF (U).

4.3.4 Parameter Recovery

4.3.4.1 Discrimination Parameters: Reference vs. Focal Groups 1 and 2

Figures 4.10 (i.e. reference versus focal group 1) and 4.11 (i.e. reference versus focal group 2) show the bias in the discrimination parameter estimates by condition for reference versus focal group 1 and reference versus focal group 2 respectively. The pattern of results in Figure 4.10 (i.e. reference versus focal group 1) was nearly identical to Figure 4.11 (i.e. reference versus focal group 2) and only Figure 4.10 will be interpreted. Looking down the middle column corresponding to a LV of 1 we see that the Bayesian MIMIC model had good parameter recovery for both the reference and focal group 1. This was to be expected as the assumption of homogeneity of latent SDs was met for the Bayesian MIMIC model. Additionally, the improved Wald also had good parameter

Table 4.4: Confusion Matrices for 3 Group Conditions: Bayesian MIMIC Model for Reference vs Focal Group 1

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.094**	0.000	0.003	0.028	**SS: 500:500:500 LSD: 0.5**
	NU	0.034	**0.002**	0.006	0.083	
	M	0.075	0.012	**0.012**	0.026	
	N	0.048	0.011	0.000	**0.566**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.106**	0.000	0.000	0.019	**SS: 500:500:500 LSD: 1**
	NU	0.041	**0.001**	0.001	0.083	
	M	0.098	0.001	**0.006**	0.021	
	N	0.029	0.000	0.000	**0.596**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.107**	0.000	0.000	0.018	**SS: 500:500:500 LSD: 1.5**
	NU	0.039	**0.000**	0.001	0.085	
	M	0.105	0.000	**0.000**	0.020	
	N	0.034	0.000	0.000	**0.591**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.092**	0.000	0.000	0.033	**SS: 1000:250:250 LSD: 0.5**
	NU	0.043	**0.000**	0.000	0.083	
	M	0.080	0.002	**0.001**	0.042	
	N	0.034	0.001	0.000	**0.591**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.101**	0.000	0.000	0.024	**SS: 1000:250:250 LSD: 1**
	NU	0.036	**0.001**	0.001	0.087	
	M	0.094	0.001	**0.000**	0.031	
	N	0.022	0.000	0.000	**0.603**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.099**	0.000	0.000	0.026	**SS: 1000:250:250 LSD: 1.5**
	NU	0.036	**0.001**	0.000	0.088	
	M	0.102	0.001	**0.001**	0.022	
	N	0.039	0.000	0.001	**0.586**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.072**	0.000	0.000	0.053	**SS: 1300:100:100 LSD: 0.5**
	NU	0.028	**0.000**	0.000	0.097	
	M	0.070	0.000	**0.000**	0.055	
	N	0.031	0.000	0.000	**0.594**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.064**	0.000	0.000	0.061	**SS: 1300:100:100 LSD: 1**
	NU	0.020	**0.000**	0.000	0.105	
	M	0.064	0.000	**0.000**	0.061	
	N	0.024	0.000	0.000	**0.601**	

		(Predicted)				
		U	**NU**	**M**	**N**	
(True)	**U**	**0.071**	0.000	0.000	0.054	**SS: 1300:100:100 LSD: 1.5**
	NU	0.019	**0.000**	0.000	0.106	
	M	0.077	0.000	**0.000**	0.048	
	N	0.020	0.000	0.000	**0.605**	

Note. U = uniform DIF, NU = non-uniform DIF, M = mixed DIF, N = no DIF, SS = sample size, LSD = latent focal group SD.

recovery in the equal and moderately unequal sample sizes. In the highly unequal sample sizes the improved Wald reference group parameters were well recovered and the focal group 1 parameters were slightly underestimated.

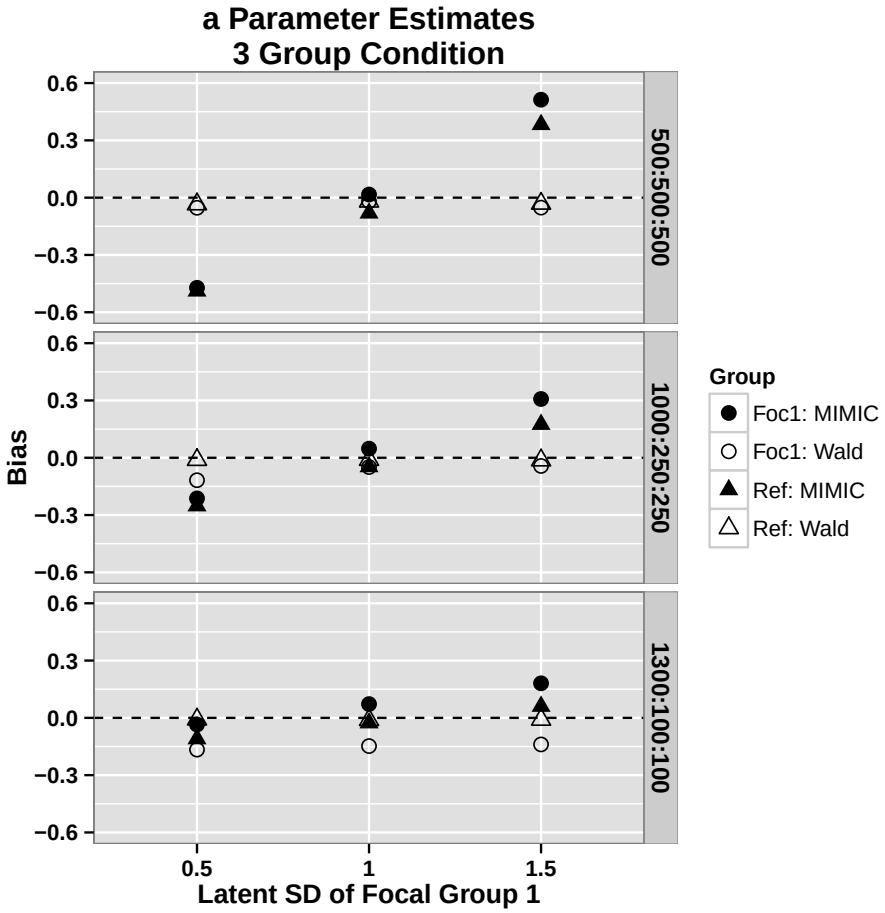
Looking at the far right column corresponding to a LV of 1.5 the Bayesian MIMIC model parameters were consistently overestimating the discrimination parameters. The Bayesian MIMIC model discrimination parameters were more biased than the improved Wald parameters in the equal (500:500:500) and moderately unequal (100:250:250) sample size conditions. In the highly unequal sample size condition (1300:100:100) the two methods were about equally biased, however, the Bayesian MIMIC model overestimated the discrimination parameters and the improved Wald underestimated the true parameters. The improved Wald had good parameter recovery in the equal and moderately unequal sample size conditions. In the highly unequal sample size condition the improved Wald had good parameter recovery of the reference group and slightly underestimated the focal group discrimination parameters.

For the LV of 0.5 conditions the improved Wald had good parameter recovery for the reference group across all sample size conditions. The improved Wald had good parameter recovery for the focal group in the equal sample size condition and slightly underestimated the discrimination parameters in the moderately unequal and highly unequal sample size conditions. For the Bayesian MIMIC model the discrimination parameters were severely underestimated in the equal sample size condition. The underestimation of the discrimination parameters attenuated as the sample size imbalance increased.

4.3.4.2 Difficulty Parameters: Reference vs. Focal Groups 1 and 2

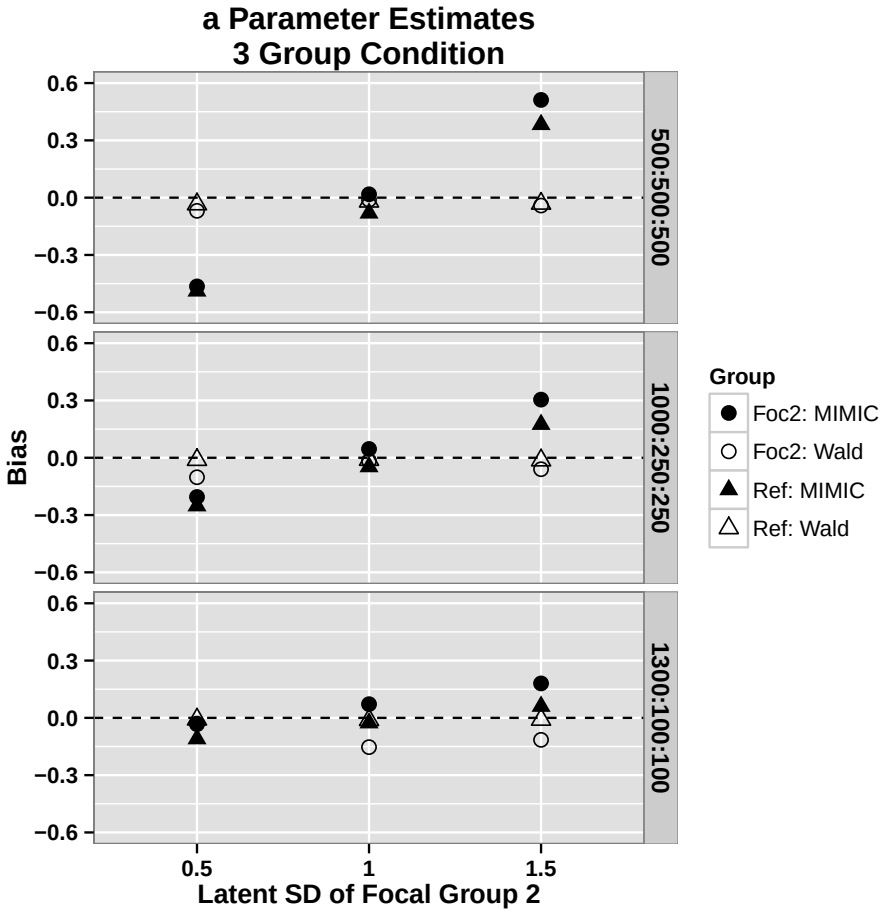
Figures 4.12 (i.e. reference versus focal group 1) and 4.13 (i.e. reference versus focal group 2) show the bias in the difficulty parameter estimates by condition. The pattern of results in Figure 4.12 (i.e. reference versus focal group 1) was nearly identical to Figure 4.13 and only Figure 4.12 will be interpreted. Looking down the middle column corresponding to a LV of 1 we see that the Bayesian MIMIC model had good parameter recovery for the reference group in all three sample

Figure 4.10: Discrimination Parameter Bias for 3 Group Conditions: Reference vs. Focal Group 1



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. Foc1 = focal group 1, Ref = reference group. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

Figure 4.11: Discrimination Parameter Bias for 3 Group Conditions: Reference vs. Focal Group 2



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. Foc2 = focal group 2, Ref = reference group. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

size conditions. The improved Wald also had good parameter recovery across all three sample size conditions when the LSD was 1.

In the focal group latent variance of 0.5 conditions the Bayesian MIMC model overestimated the reference and focal group parameters. This effect was most severe in the equal sample size (500:500:500) condition with the underestimation becoming less severe as the sample sizes become more unbalanced. For the improved Wald method the reference and focal group parameters were well recovered across all sample size conditions.

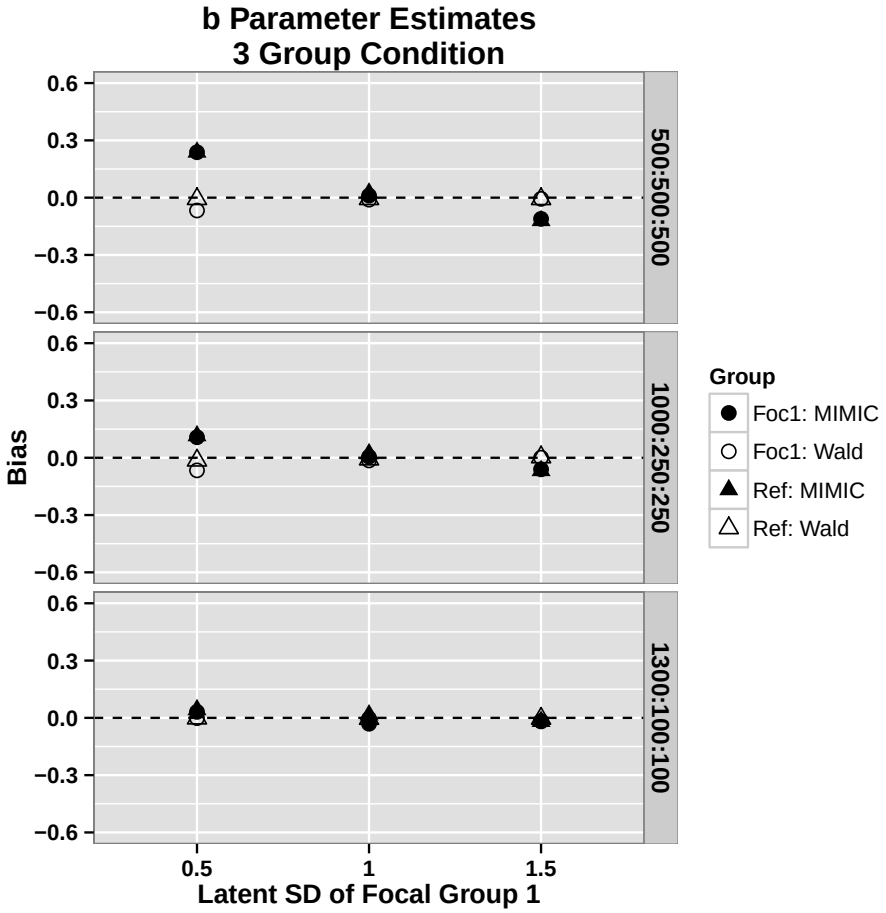
For the LV of 1.5 conditions the improved Wald method had good parameter recovery across all sample size conditions. The Bayesian MIMIC model slightly underestimated the difficulty parameters in the equal sample size condition and this underestimation decreased as the sample size imbalance increased. When the sample size was highly unequal the Bayesian MIMIC model difficulty parameters were well estimated.

4.3.4.3 Latent Means: Focal Groups 1 and 2

Figure 4.14 presents the results of the latent mean bias for focal group 1 and 2 by method. The Bayesian MIMIC model focal group 1 is represented by a dark circle and focal group 2 by a dark triangle. For the improved Wald focal group 1 is represented by an open circle and focal group 2 by an open triangle. The improved Wald had good parameter recovery of the latent means for both focal groups in all conditions.

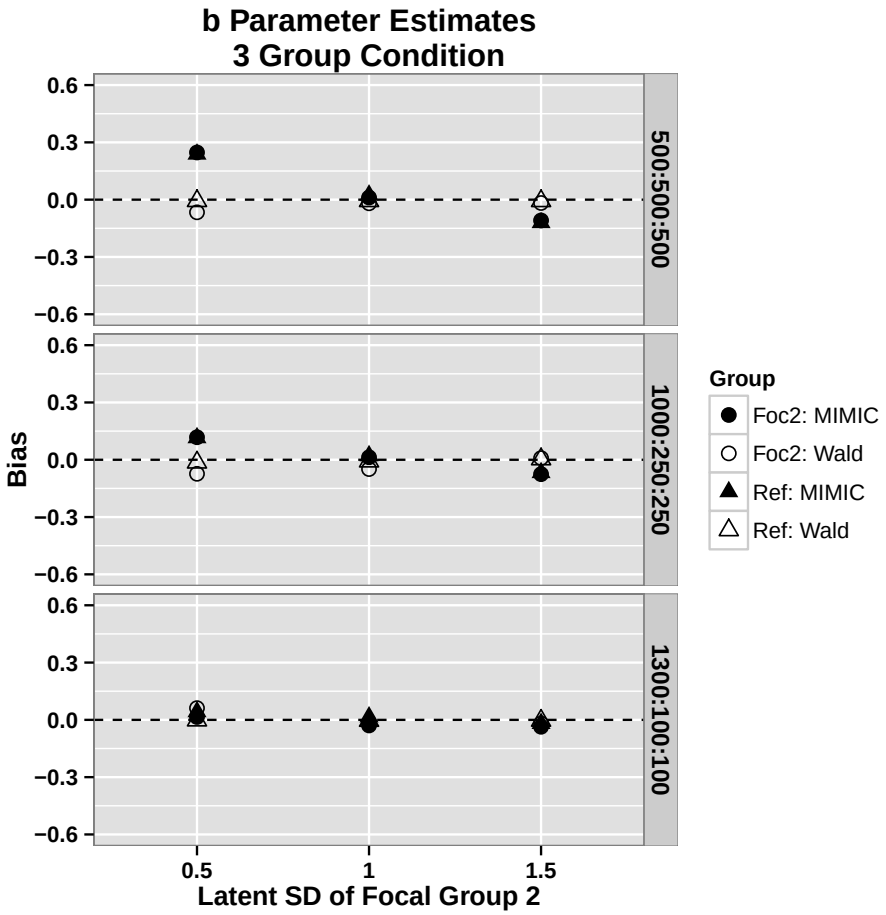
Within the LSD of 1 condition the Bayesian MIMIC model had good parameter recovery of the latent means for both focal groups. When the LSD was 0.5 the Bayesian MIMIC model underestimated the latent means when the sample sizes were equal and this effect slightly lessened as the sample size imbalanced increased. When the LSD was 1.5 the Bayesian MIMIC model tended to overestimate the latent means although this overestimation was not as strong as the underestimation when the LSD was 0.5. The overestimation effect also slightly decreased as the sample size imbalance increased.

Figure 4.12: Difficulty Parameter Bias for 3 Group Conditions: Reference vs. Focal Group 1



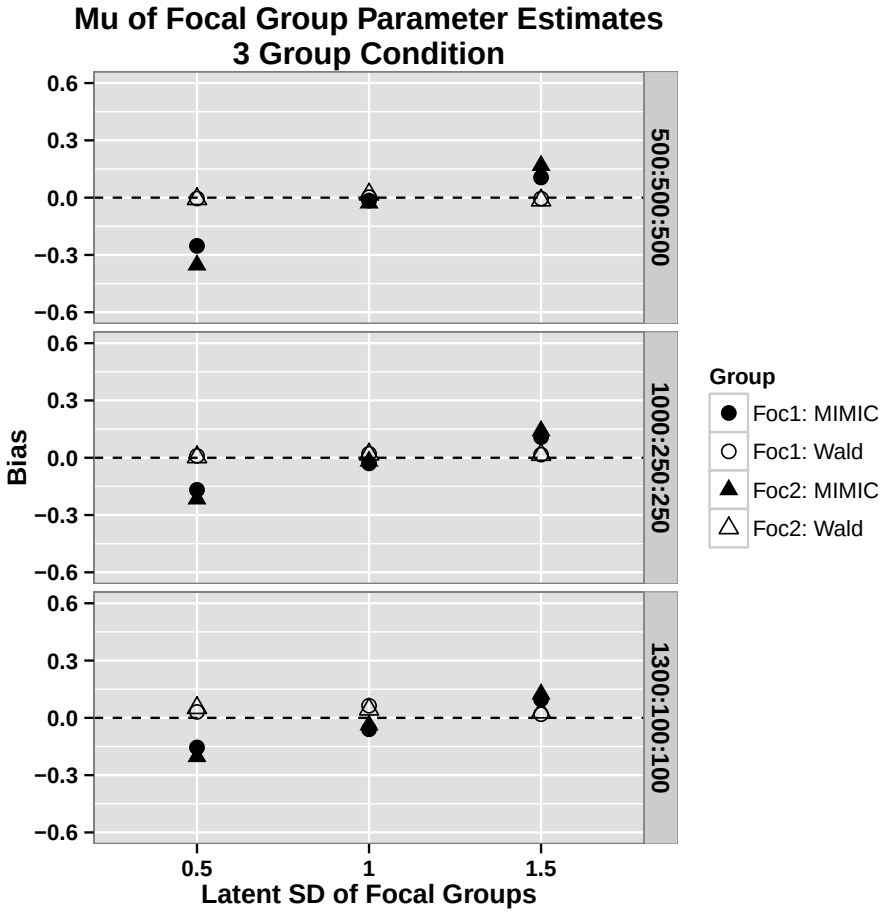
Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. Foc1 = focal group 1, Ref = reference group. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

Figure 4.13: Difficulty Parameter Bias for 3 Group Conditions: Reference vs. Focal Group 2



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500:500 corresponds to a reference group sample size of 500, a focal group 1 sample size of 500, and a focal group 2 sample size of 500. Foc2 = focal group 2, Ref = reference group. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

Figure 4.14: Latent Mean Parameter Bias for 3 Group Conditions



Note. The labels on the right three partitions correspond to the sample size conditions. For example, 500:500 corresponds to a reference group sample size of 500 and a focal group 1 sample size of 500. Foc1 = focal group 1. Wald = improved Wald and MIMIC = Bayesian MIMIC model.

Chapter 5

Applied Example of the Bayesian MIMIC Model

5.1 Overview

An illustration of using the Bayesian MIMIC model to test for both uniform and/or non-uniform DIF is presented on a real data set. The data were collected on the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1996). These data have been analyzed with a uniform MIMIC DIF model by Woods et al. (2009). All Stan and R code for these analyses are presented in the appendix.

The sample in the entire data set consisted of 2026 Air Force personnel (1265 men, 761 women) completing basic training at Lackland Air Force Base in San Antonio, Texas. The racial composition was Caucasian (1305), African-American (348), Hispanic (75), Asian (68), Native American (17), and 213 classified as "other". More specific details about the sample can be obtained from Oltmanns and Turkheimer (2006). For the purposes of this applied example only three of the seven groups presented by Woods et al. (2009) were used. The groups were Caucasian (1305), African American (348), and other (213) for a total of 1866 participants (1165 mean and 701 women).

The SNAP is a self-report questionnaire that was originally developed to assess personal-

ity disorders in terms of trait dimensions (Clark, 1996). The SNAP consists of 375 True/False questions on three temperament scales (i.e. Negative, Positive, and Disinhibition) and 12 trait scales (i.e. Aggression, Exhibitionism, Manipulativeness, Mistrust, Self-harm, Eccentric Perceptions, Entitlement, Dependency, Impulsivity, Detachment, Workaholism, and Propriety). In the present application only the Negative temperament scale was used. The negative temperament scale consists of 28 binary items. The negative temperament scale was chosen because it was the scale from which the generating parameters of the simulation were taken.

5.2 Data Analysis

The following procedures were used in the assessment of DIF using the Bayesian MIMIC model. First, referent items (i.e. anchor items) were empirically selected in order to determine items to link the scales across the groups. All remaining items not selected as anchor items were tested for DIF. Second, each item tested for DIF was evaluated for both uniform and/or non-uniform DIF using the same methodology described in the simulation study but using an α level for the HPD intervals of 0.05. Third, a final model was fit which allowed group differences in all items showing DIF for both the African-American and other groups. All parameter estimates from the final model and latent mean differences were reported. A conservative burn-in of 1000 iterations per chain was used with 2000 post burn-in draws used to sample from the posterior. The 1000 burn-in number was selected to be consistent with the simulation study as pilot runs indicated a burn-in of 100-200 draws would suffice. All convergence and thinning criteria were the same as those described in the simulation study.

In the present applied example the following prior distributions were proposed to be used:

$$\begin{aligned}
\theta_{jh} &\sim N(\mu_h, 1)\ j = 1,\ldots,J\ \&\ h = 1,\ldots,G \\
\mu_h &\sim N(0,10) \text{ Note } \mu_1 = 0\ \&\ h = 1,\ldots,G \\
\alpha_i &\sim LN(0,1)\ i = 1,\ldots,20 \\
b_i &\sim N(0,10)\ i = 1,\ldots,20 \\
\omega_{ih} &\sim N(0,4)\ i = 1,\ldots,22\ \&\ h = 1,\ldots,G \\
\beta_{ih} &\sim N(0,4)\ i = 1,\ldots,22\ \&\ h = 1,\ldots,G. \qquad (5.1)
\end{aligned}$$

G denotes the number of groups (three here), J the number of persons, LN denotes a log-normal distribution, and h indexes the group number. Note that both ω_{ih} and β_{ih} are only for 22 items as the six anchor items are held equal for identification purposes. Additionally, $\mu_1 = 0$ and $\sigma^2_{\theta_{jh}} = 1$ in order to identify the model. These prior distributions have been used previously in other research and practical applications (e.g. Curtis, 2010; J. Fox, 2010; Patz & Junker, 1999a; Patz & Junker, 1999b) and were selected to be mildly non-informative. If the use of these prior distributions failed to get approximate convergence according to the r-hat criteria then the priors would be adjusted as necessary.

5.2.1 Empirical Selection of Anchor Items

To empirically select anchor items the procedure described in Woods (2009a) was used. The procedure was conducted using flexMIRT™ to expedite the computations and use an established anchor selection technique that has been used in practical applications (e.g. Harpole et al., 2014). This anchor selection strategy consists of four steps. First, test an item for DIF using a likelihood-ratio test with all-others as anchors. Second, compute the likelihood-ratio statistics for the given item. Repeat steps one and two for all items. Third, rank the items based on the likelihood-ratio magnitudes in ascending order. Fourth, designate n (i.e. the number of items) items with the n smallest likelihood-ratios.

5.2.2 Testing for DIF and Fitting the Final Model

The final model was fit using the results from the uniform and non-uniform DIF tests of the Bayesian MIMIC models. If an item had uniform and/or non-uniform DIF for a given group then either the difficulty and/or discrimination parameter(s) were allowed to be freely estimated for that group. If there was no evidence of uniform and/or non-uniform DIF then the parameters for that item were set equal across the three groups. Details of the implementation in R and Stan can be seen in the appendix. For both the non-uniform and uniform DIF effect the posterior mean and SE of the posterior mean of each effect was reported. Additionally, the 95% HPD intervals were also reported for the DIF effects. The posterior mean of the difficulty, discrimination, and latent mean parameters of the final model and there standard errors were also reported.

5.3 Results

5.3.1 Anchor Selection and Convergence Criteria

For all models the r-hat values were less than 1.02 indicating approximate convergence to the posterior distribution. Thus, the prior distributions in Equation 5.1 were used in all analyses. Items 3, 9, 10, 13, 21, and 22 were selected as anchor items. Six items were chosen based on the recommendation of Woods (2009a) that approximately 10-20% of the test items be used as anchor items. These items corresponded to SNAP items 245, 264, 269, 277, 311, and 312. Results of the DIF tests for Caucasian versus African-American and Caucasian versus other are presented in Tables 5.1 and 5.2, respectively. In each table any item that was flagged as having either uniform and/or non-uniform DIF is in bold text. Further, the specific posterior mean DIF effect, the SE of the posterior mean DIF effect, and the 95% HPD intervals are in bold for each type of DIF. The posterior mean uniform DIF effect is denoted as $\bar{\beta}$ and the posterior mean non-uniform DIF effect as $\bar{\omega}$. Results of the final model are presented in Table 5.3.

5.3.2 DIF Tests and Final Model

Table 5.1 shows the results of DIF tests for Caucasian versus African-American. There were eight items flagged for DIF (items 1, 2, 5, 6, 11, 19, 23, and 26). Item 5 was flagged for both uniform and non-uniform DIF (i.e. mixed DIF) and items 1, 2, 6, 11, 19, 23, and 26 were flagged for uniform DIF only. Item 5 was less discriminating for African-Americans as compared to Caucasians ($\bar{\omega} = -0.528$, $SE = 0.006$) and item 5 required a higher amount of negative temperament for African Americans to select true compared with Caucasians ($\bar{\beta} = 0.348$, $SE = 0.004$). Uniform DIF was present in items 1 ($\bar{\beta} = -0.391$, $SE = 0.003$), 6 ($\bar{\beta} = -1.074$, $SE = 0.003$), 11 ($\bar{\beta} = -0.546$, $SE = 0.003$), 23 ($\bar{\beta} = -0.534$, $SE = 0.002$), and 26 ($\bar{\beta} = -0.623$, $SE = 0.003$) indicating that a lower amount of negative temperament for African Americans was needed to select true compared with Caucasians. Uniform DIF was also present in items 2 ($\bar{\beta} = 0.325$, $SE = 0.002$) and 19 ($\bar{\beta} = 0.936$, $SE = 0.004$) showing that a higher amount of negative temperament was required for African Americans to select true compared with Caucasians.

Table 5.2 shows the results of DIF tests for Caucasian versus other. There were seven items flagged for DIF (items 7, 16, 19, 20, 23, 26, and 28). Items 7, 16, and 26 were flagged as mixed DIF items (i.e. both uniform and non-uniform DIF). Item 7 was less discriminating for others ($\bar{\delta}_\alpha = -0.753$, $SE = 0.006$) than for Caucasians and also required more negative temperament for others ($\bar{\beta} = 0.483$, $SE = 0.006$) to respond true compared with Caucasians. Item 16 was more discriminating for others ($\bar{\omega} = 0.802$, $SE = 0.009$) compared with Caucasians and also required less negative temperament ($\bar{\beta} = -0.435$, $SE = 0.005$) for others to respond true compared with Caucasians. Item 26 was also more discriminating for others ($\bar{\omega} = 0.770$, $SE = 0.009$) than for Caucasians and also required less negative temperament for others ($\bar{\beta} = -0.451$, $SE = 0.003$) to respond true compared with Caucasians. Items 20 and 28 both showed only non-uniform DIF. Both items 20 ($\bar{\omega} = -0.564$, $SE = 0.003$) and 28 ($\bar{\omega} = -0.734$, $SE = 0.009$) were less discriminating for others compared with Caucasians. Items 19 and 23 both only showed uniform DIF. Item 19 required more negative temperament for others ($\bar{\beta} = 0.552$, $SE = 0.006$) to respond true compared with Caucasians. Item 23 required less negative temperament for others ($\bar{\beta} = -0.690$, $SE = 0.003$)

to respond true compared with Caucasians.

Results of the final model are presented in Table 5.3. Items that are in bold are those that were deemed non-invariant for either the discrimination, difficulty, or both for a given group. If an item parameter is not bold then this indicates the item parameter was invariant. For example, in Table 5.3 item 26 is bold for both the discrimination and difficulty parameter of Caucasians. This indicates that either African Americans, other, or both have non-invariant parameters. For item 26 the difficulty parameters for both African American and other participants were non-invariant. Further, the discrimination parameter for other was also non-invariant, whereas the discrimination parameter for African Americans was invariant. The true mean difference on negative temperament for African Americans was -0.09 ($SE = 0.01$) and for other was -0.03 ($SE = 0.01$). This indicates that both African American and other participants have slightly less negative temperament than Caucasian participants albeit not by much.

Table 5.1: Bayesian MIMIC Model DIF Caucasian versus African American

	Non-Uniform DIF				Uniform DIF			
item	$\bar{\omega}$	$SE_{\bar{\omega}}$	95% LL	95% UL	$\bar{\beta}$	$SE_{\bar{\beta}}$	95% LL	95% UL
1 (SNAP 241)	0.180	0.007	-0.486	0.795	**-0.391**	**0.003**	**-0.741**	**-0.002**
2 (SNAP 244)	0.388	0.006	-0.115	0.945	**0.325**	**0.002**	**0.028**	**0.633**
4 (SNAP 248)	0.055	0.007	-0.574	0.725	0.301	0.004	-0.062	0.643
5 (SNAP 250)	**-0.528**	**0.006**	**-1.016**	**-0.028**	**0.348**	**0.004**	**0.028**	**0.720**
6 (SNAP 252)	-0.167	0.007	-0.796	0.447	**-1.074**	**0.003**	**-1.481**	**-0.655**
7 (SNAP 259)	0.084	0.008	-0.658	0.738	0.096	0.004	-0.254	0.449
8 (SNAP 260)	0.289	0.010	-0.556	1.081	-0.412	0.004	-0.840	0.035
11 (SNAP 273)	-0.098	0.006	-0.668	0.477	**-0.546**	**0.003**	**-0.925**	**-0.211**
12 (SNAP 274)	0.052	0.007	-0.485	0.656	-0.112	0.003	-0.517	0.253
14 (SNAP 281)	-0.116	0.005	-0.529	0.337	0.185	0.003	-0.146	0.497
15 (SNAP 288)	0.091	0.004	-0.312	0.549	0.214	0.002	-0.093	0.510
16 (SNAP 290)	-0.055	0.004	-0.487	0.394	-0.193	0.003	-0.504	0.116
17 (SNAP 294)	-0.247	0.006	-0.768	0.345	-0.222	0.003	-0.600	0.154
18 (SNAP 298)	-0.147	0.003	-0.503	0.205	0.065	0.002	-0.213	0.347
19 (SNAP 301)	-0.175	0.007	-0.778	0.373	**0.936**	**0.004**	**0.576**	**1.304**
20 (SNAP 309)	0.139	0.004	-0.248	0.572	-0.055	0.003	-0.337	0.236
23 (SNAP 316)	0.088	0.005	-0.370	0.530	**-0.534**	**0.002**	**-0.825**	**-0.217**
24 (SNAP 320)	0.329	0.008	-0.341	0.984	0.247	0.004	-0.112	0.585
25 (SNAP 323)	-0.195	0.004	-0.583	0.172	-0.033	0.003	-0.348	0.304
26 (SNAP 325)	-0.301	0.003	-0.739	0.093	**-0.623**	**0.003**	**-0.945**	**-0.302**
27 (SNAP 331)	-0.216	0.002	-0.490	0.084	0.002	0.002	-0.259	0.249
28 (SNAP 333)	-0.415	0.008	-1.052	0.308	-0.294	0.003	-0.686	0.106

Note. SNAP = Schedule of Nonadaptive and Adaptive Personality, $\bar{\omega}$ = posterior mean DIF parameter for α, $\bar{\beta}$ = posterior mean DIF parameter for b, 95% LL is the lower limit of the 95% highest posterior density interval, 95% UL is the upper limit of the 95% highest posterior density interval.

Table 5.2: Bayesian MIMIC Model DIF Caucasian versus Other

	Non-Uniform DIF				Uniform DIF			
item	$\bar{\omega}$	$SE_{\bar{\omega}}$	95% LL	95% UL	$\bar{\beta}$	$SE_{\bar{\beta}}$	95% LL	95% UL
1 (SNAP 241)	0.062	0.007	-0.727	0.748	-0.298	0.006	-0.711	0.148
2 (SNAP 244)	-0.220	0.004	-0.699	0.324	0.131	0.005	-0.229	0.497
4 (SNAP 248)	0.402	0.011	-0.459	1.270	0.043	0.006	-0.381	0.478
5 (SNAP 250)	-0.323	0.007	-0.937	0.376	0.237	0.005	-0.188	0.638
6 (SNAP 252)	-0.443	0.008	-1.146	0.336	-0.398	0.005	-0.951	-0.166
7 (SNAP 259)	**-0.753**	**0.006**	**-1.349**	**-0.196**	**0.483**	**0.006**	**0.028**	**0.915**
8 (SNAP 260)	-0.306	0.009	-1.093	0.436	-0.457	0.004	-0.961	0.057
11 (SNAP 273)	-0.410	0.007	-1.014	0.204	-0.086	0.004	-0.535	0.367
12 (SNAP 274)	0.401	0.010	-0.435	1.273	0.439	0.004	-0.107	0.961
14 (SNAP 281)	0.226	0.007	-0.386	0.874	0.036	0.003	-0.330	0.450
15 (SNAP 288)	0.461	0.007	-0.143	1.091	-0.144	0.003	-0.528	0.227
16 (SNAP 290)	**0.802**	**0.009**	**0.002**	**1.640**	**-0.435**	**0.005**	**-0.823**	**-0.030**
17 (SNAP 294)	0.628	0.010	-0.272	1.548	-0.186	0.004	-0.647	0.273
18 (SNAP 298)	0.347	0.004	-0.147	0.866	0.131	0.003	-0.202	0.478
19 (SNAP 301)	-0.176	0.006	-0.826	0.523	**0.552**	**0.006**	**0.128**	**0.980**
20 (SNAP 309)	**-0.564**	**0.003**	**-0.984**	**-0.153**	0.005	0.003	-0.346	0.360
23 (SNAP 316)	0.327	0.007	-0.288	0.976	**-0.690**	**0.003**	**-1.045**	**-0.298**
24 (SNAP 320)	-0.323	0.006	-0.952	0.279	0.085	0.006	-0.366	0.492
25 (SNAP 323)	0.095	0.005	-0.419	0.619	0.252	0.003	-0.101	0.635
26 (SNAP 325)	**0.770**	**0.009**	**0.031**	**1.615**	**-0.451**	**0.003**	**-0.821**	**-0.073**
27 (SNAP 331)	-0.133	0.003	-0.527	0.268	0.258	0.003	-0.065	0.578
28 (SNAP 333)	**-0.734**	**0.009**	**-1.529**	**-0.006**	0.443	0.006	-0.102	0.992

Note. SNAP = Schedule of Nonadaptive and Adaptive Personality, $\bar{\omega}$ = posterior mean DIF parameter for α, $\bar{\beta}$ = posterior mean DIF parameter for b, 95% LL is the lower limit of the 95% highest posterior density interval, 95% UL is the upper limit of the 95% highest posterior density interval.

Table 5.3: Bayesian MIMIC Model Final Model

	Caucasians				African Americans				Other			
	$\bar{\alpha}$	$SE_{\bar{\alpha}}$	$\bar{b}$	$SE_{\bar{b}}$	$\bar{\alpha}$	$SE_{\bar{\alpha}}$	$\bar{b}$	$SE_{\bar{b}}$	$\bar{\alpha}$	$SE_{\bar{\alpha}}$	$\bar{b}$	$SE_{\bar{b}}$
1	1.87	0.03	**0.95**	**0.02**	1.87	0.03	**0.61**	**0.02**	1.87	0.03	0.95	0.02
2	1.49	0.01	**0.10**	**0.00**	1.49	0.01	**0.38**	**0.01**	1.49	0.01	0.10	0.00
3	1.62	0.02	0.71	0.01	1.62	0.02	0.71	0.01	1.62	0.02	0.71	0.01
4	2.08	0.02	0.47	0.00	2.08	0.02	0.47	0.00	2.08	0.02	0.47	0.00
5	**2.01**	**0.01**	**-0.18**	**0.00**	**1.55**	**0.02**	**0.13**	**0.01**	2.01	0.01	-0.18	0.00
6	1.51	0.07	**2.43**	**0.41**	1.51	0.07	**1.45**	**0.43**	1.51	0.07	2.43	0.41
7	**2.10**	**0.02**	**0.64**	**0.01**	2.10	0.02	0.64	0.01	**1.34**	**0.02**	**0.77**	**0.02**
8	1.88	0.07	1.74	0.25	1.88	0.07	1.74	0.25	1.88	0.07	1.74	0.25
9	1.52	0.01	0.04	0.00	1.52	0.01	0.04	0.00	1.52	0.01	0.04	0.00
10	2.34	0.03	1.33	0.04	2.34	0.03	1.33	0.04	2.34	0.03	1.33	0.04
11	1.73	0.02	**1.09**	**0.02**	1.73	0.02	**0.55**	**0.01**	1.73	0.02	1.09	0.02
12	1.29	0.07	2.20	0.51	1.29	0.07	2.20	0.51	1.29	0.07	2.20	0.51
13	1.57	0.02	1.12	0.02	1.57	0.02	1.12	0.02	1.57	0.02	1.12	0.02
14	1.45	0.01	-0.65	0.01	1.45	0.01	-0.65	0.01	1.45	0.01	-0.65	0.01
15	1.30	0.01	-0.55	0.01	1.30	0.01	-0.55	0.01	1.30	0.01	-0.55	0.01
16	**1.54**	**0.01**	**0.13**	**0.01**	1.54	0.01	0.13	0.01	**2.23**	**0.03**	**-0.31**	**0.02**
17	1.80	0.02	1.14	0.01	1.80	0.02	1.14	0.01	1.80	0.02	1.14	0.01
18	0.98	0.01	0.47	0.01	0.98	0.01	0.47	0.01	0.98	0.01	0.47	0.01
19	1.92	0.03	**0.35**	**0.01**	1.92	0.03	**1.28**	**0.01**	1.92	0.03	**0.90**	**0.01**
20	**1.26**	**0.01**	-0.17	0.01	1.26	0.01	-0.17	0.01	**0.69**	**0.02**	-0.17	0.01
21	1.35	0.02	1.20	0.03	1.35	0.02	1.20	0.03	1.35	0.02	1.20	0.03
22	0.93	0.01	2.04	0.03	0.93	0.01	2.04	0.03	0.93	0.01	2.04	0.03
23	1.47	0.01	**0.45**	**0.01**	1.47	0.01	**-0.10**	**0.01**	1.47	0.01	**-0.27**	**0.02**
24	1.88	0.02	0.57	0.01	1.88	0.02	0.57	0.01	1.88	0.02	0.57	0.01
25	1.11	0.03	-1.35	0.19	1.11	0.03	-1.35	0.19	1.11	0.03	-1.35	0.19
26	**1.50**	**0.02**	**0.07**	**0.01**	1.50	0.02	**-0.58**	**0.01**	**2.21**	**0.04**	**-0.48**	**0.02**
27	0.55	0.02	1.31	0.14	0.55	0.02	1.31	0.14	0.55	0.02	1.31	0.14
28	**2.17**	**0.04**	1.08	0.04	2.17	0.04	1.08	0.04	**1.48**	**0.03**	1.08	0.04

Note. $\bar{\alpha}$ = posterior mean discrimination parameter for α, $\bar{b}$ = posterior mean difficulty parameter for b. Items in bold are those that had non-invariant discrimination, difficulty, or both parameters from the reference group. If an item is not bold this indicates the item was invariant across the respective groups.

Chapter 6

Discussion

6.1 Overview

The purpose of this research was to propose a new implementation of the MIMIC model for testing uniform and non-uniform DIF, conduct a Monte Carlo simulation to address the three limitations within the present body of MIMIC DIF research, and provide an applied example on a real data set. With regards to the Monte Carlo study the present research sought to extend the findings of non-uniform MIMIC DIF research from the 2 group case to the 3 group case, assessing the robustness of the non-uniform MIMIC DIF model to violations of homogeneity of latent SDs, and properly estimating the latent interaction term by utilizing Bayesian estimation. Additionally, the proposed Bayesian MIMIC model was compared to the improved Wald χ^2 throughout all conditions within the Monte Carlo simulation. In what follows I will summarize the main findings from the 2 group simulations, 3 group simulations, applied example, and discuss limitations and directions for future research.

6.2 2 Group Simulation

6.2.1 Overall Power and Overall Type I Error

Within the two group conditions the Bayesian MIMIC model did not experience problematic Type I error rates using an $\alpha = 0.05$ criterion. Overall the Bayesian MIMIC model had less problematic Type I error rates in comparison to those found in Carroll (2014) within similar conditions. There are a couple of explanations for this. First, the Bayesian MIMIC model estimates uniform and non-uniform DIF whereas the model from Carroll (2014) only estimated uniform DIF. Second, the use of Bayesian estimation and mildly informative priors on the difficulty and informative priors on the discrimination DIF parameters may have attenuated the DIF effects. Given that the priors were more informative than typically used in DIF studies the DIF effect from the likelihood would need to be more pronounced to be flagged as DIF than if the priors were less informative. The present findings were a vast improvement over using LMS in Woods and Grimm (2011) and shows that when more appropriately estimating the latent interaction Type I error problems vanished.[1]

Power for both the Bayesian MIMIC model and the improved Wald with SEM SEs can be interpreted unambiguously across all conditions. Overall the improved Wald edged out the Bayesian MIMIC model in terms of power given the higher power in two of the nine conditions and equivalent power in the remaining seven conditions. These findings were somewhat different than those in Woods and Grimm (2011) within similar conditions. The likely reason for this phenomenon was probably that in Woods and Grimm (2011) the non-uniform DIF MIMIC model had severely inflated Type I error rates and interpretation of power would be misleading, whereas in the present study power was interpretable in all conditions.

[1]Note that Woods and Grimm (2011) did not blithely fit LMS for testing non-uniform DIF in the MIMIC model. They did so because that is what the Mplus manual recommended at the time of their publication and they posited that this would result in Type I error inflation as is what occurred.

6.2.1.1 Conclusion of Bayesian MIMIC vs. Improved Wald Confusion Matrices

The general conclusions of the confusion matrices for the improved Wald with SEM SEs and the Bayesian MIMIC model are presented below. As mentioned in the results section the confusion matrices are presented to provide a more granular description of the overall power and Type I error results. In this light the interpretations that follow are based on the current simulation study and are descriptive and not meant to be generalized beyond this study. In what follows the summary for the confusion matrices using $\alpha = 0.05$ with the Bayesian MIMIC model and the SEM SE estimation for the improved Wald are discussed.

Overall there were both similarities and differences between the improved Wald and Bayesian MIMIC model confusion matrices for the 2 group conditions. Both the improved Wald and the Bayesian MIMIC model misclassified uniform DIF (U) and non-uniform DIF (NU) as N (no DIF) in the majority of cases and mixed DIF (M) as uniform DIF (U) in the majority of cases. For no DIF (N) the Bayesian MIMIC model misclassified the majority of cases as uniform DIF (U), whereas, the improved Wald misclassified the majority of cases as non-uniform DIF (NU). Thus, the majority misclassification rates were almost the same. Additionally, the majority DIF misclassification rates were also similar in that when an item was a uniform DIF item (U) the majority of DIF mis-classifications were as mixed DIF (M) and when an item was either a non-uniform DIF item (NU) or a mixed DIF item (M) the majority of DIF mis-classifications were uniform DIF (U).

Both methods struggled to correctly classify mixed DIF (M), with the improved Wald having a higher accuracy than the Bayesian MIMIC model. However, in most conditions these misclassified items were misclassified as either uniform (U) or non-uniform (NU) DIF. Thus, while the correct type of DIF was not identified the items were still flagged as being DIF items. Further, both methods also struggled to identify non-uniform (NU) DIF as well. The Bayesian MIMIC model had more trouble with non-uniform (NU) DIF compared with the improved Wald. When misclassifying non-uniform DIF the majority of mis-classifications for both methods were no DIF (N), which partially explains why overall power was low in Figure 4.2.

In all conditions the improved Wald always had higher accuracy for mixed (M) and non-

uniform (NU) DIF compared to the Bayesian MIMIC model. This finding was probably due to the informative priors that were placed on the ω parameters for non-uniform DIF. In a majority of cases the Bayesian MIMIC model had higher accuracy for N compared with the improved Wald. The discrepancy here was an artifact of the larger Type I errors for the improved Wald in a majority of conditions (see Figure 4.1). The Bayesian MIMIC model always had higher accuracy to detect uniform DIF compared with the improved Wald method. This can be partially explained by the slightly less information priors placed on the β parameters compared with the informative priors placed on the ω.

6.2.2 Parameter Recovery

In general the improved Wald method had good parameter recovery for the discrimination, difficulty, and latent mean parameters across all conditions. When the assumption of homogeneity of latent SDs held (LSD = 1) the Bayesian MIMIC model also had good parameter recovery for the discrimination, difficulty, and latent mean parameters across all sample size conditions. When the assumption of homogeneity of latent SDs was violated things were not always as favorable for the Bayesian MIMIC model.

When the latent SD was lower than predicted by the model (LSD = 0.5) the discrimination parameters were underestimated. The most severe underestimation occurred in the equal sample size condition and the underestimation attenuated as the sample size imbalance increased. The reason for the underestimation is probably due to less variance in the latent variable than being predicted by the model. Having less variance would directly impact the discrimination parameters as they are also known as scale parameters. Given less variance than predicted this would cause an underestimation of the discrimination parameters. For the LSD of 1.5 conditions the discrimination parameters were overestimated compared with the true values. This overestimation effect attenuated as the sample size imbalance increased. This overestimation effect was likely due to the fact that there was more latent variance than predicted by the model and this excess variance drove the values of the discrimination parameters higher than they should have been.

When the LSD was 0.5 the latent mean parameters were also underestimated. This underestimation was most severe when the sample sizes were equal and the effect attenuated as the sample size imbalance increased. As with the discrimination parameters when there was less variability than predicted by the MIMIC model this tended to cause an underestimation of the latent mean parameters of the MIMIC model. For the LSD of 1.5 conditions the latent means were overestimated compared with the true values. This overestimation effect slightly lessened as the sample size imbalance increased. This was likely caused by more variance than predicted by the model driving up the latent mean parameters.

When the LSD was 0.5 the difficulty parameters were overestimated. The overestimation lessened as the sample size imbalance increased. This overestimation was probably caused by the fact that the latent variance was less than predicted by the model the latent means were underestimated (see above) causing the difficulty parameters to be overestimated. This effect flipped for the LSD of 1.5 conditions. Here the latent means were overestimated which caused the difficulty parameters to be underestimated. This underestimation lessened as the sample size imbalance increased.

One key observation across all conditions when the assumption of homogeneity of LSDs did not hold for the Bayesian MIMIC model was that the bias attenuated as the sample size imbalance increased. This result seems counterintuitive at first glance but makes perfect sense upon looking at the mildly informative and informative priors in Equation 3.7. The reason for the attenuation with the discrimination and difficulty parameters has to do with more reliance on the priors when there is not much information from the likelihood for focal group 1. In a Bayesian analysis when there is a lack of information present in the likelihood and the priors are mildly informative or informative this will cause the parameters to borrow more information from the priors to estimate the parameters. Likewise as the imbalance increased this caused more people to be present in the reference group which would provide more information in the likelihood and less reliance on the prior distributions. Further, given that the reference group was used to scale the analysis and latent variables were always simulated from a $N(0,1)$, there was little bias in the reference group

parameters across all conditions.

6.2.3 General Conclusions 2 Group Simulation

Overall when the assumption of homogeneity of latent SD held (LSD = 1) the Bayesian MIMIC model had competitive Type I error control, power, and parameter recovery compared with the improved Wald with SEM SEs. However, the improved Wald would be slightly favored in these conditions given the excellent parameter recovery, Type I error control, and higher power of the improved Wald in the LSD of 1 and equal sample size condition and equal power in all other conditions. When the assumption of latent homogeneity of variance did not hold the Bayesian MIMIC model was competitive with the improved Wald in terms of Type I error control and power, albeit the improved Wald had a slight edge with higher power in the LSD of 1.5 and equal sample size condition. However, the Bayesian MIMIC model performed rather poorly in many conditions involving parameter recovery. Thus, the Bayesian MIMIC model would not be a recommended method to use in terms of DIF testing.

6.3 3 Group Simulation

The results of the 3 group simulation conditions were consistent with the 2 group conditions in terms of parameter recovery and Type I error control. In essence as with the 2 group conditions the Bayesian MIMIC model would not be recommended for general DIF testing given the poor performance in terms of parameter recovery compared with a method such as the improved Wald. The one difference between the 2 group and 3 group conditions was in terms of power. In the 3 group conditions power was equivalent for both methods in both the reference versus focal group 1 and reference versus focal group 2 conditions. Whereas in the 2 group conditions the improved Wald had higher power than the Bayesian MIMIC model in two of the nine conditions.

6.4 General Conclusions from the Simulation Study

Overall, several conclusions can be reached regarding the Bayesian MIMIC model and improved Wald method. First, when the assumption of homogeneity of LSD was met (i.e. LSD = 1) the Bayesian MIMIC model performed competitively with the improved Wald in the 3 group conditions and slightly below the improved Wald the 2 group conditions. Additionally, when the assumption of LSDs did not hold in terms of DIF classification (i.e. Type I error and power) the Bayesian MIMIC model performed on par with the improved Wald in the 3 group conditions and slightly below the improved Wald in the 2 group conditions. However, when the assumption did not hold parameter recovery was poor in many conditions for the Bayesian MIMIC model. Thus, overall the Bayesian MIMIC model would not be recommended for use in general DIF testing. Further, given the findings in the present study along with those from Carroll (2014), I would not recommend practitioners use Bayesian MIMIC models or MIMIC models for DIF testing. My reasons are that if the assumption holds then things go well. However, it is not possible to test this assumption with a MIMIC model and some practical applications of DIF testing show that these assumptions may not hold (e.g. Harpole et al., 2014; Langer, 2008). In order to test the assumption a multiple group model would need to be fit at which point why not just use the multiple group model and reap the benefits of excellent DIF classification and parameter recovery?

6.5 Applied Example

The illustration of the applied example showed how practitioners might implement the Bayesian MIMIC model on their own data sets. There are a couple of important points to note about the practical implementation and the simulation study. First, in the simulation study mildly informative and informative priors were placed on the DIF effects in order to improve the convergence across the 1800 condition reps. Initially, more diffuse priors were used in the simulation study but convergence was unacceptable and the source of the issue was sign switching due to more diffuse priors. In the applied example more diffuse priors were used in comparison to the sim-

ulation study because convergence was not an issue. That is one luxury that an application has over a simulation study because a single data set may not have convergence issues and if it does you only need to tweak the model to fit this single data set and not 1799 more.

Another thing to note was anchor selection. In the applied example I empirically selected anchors using the method proposed by Woods (2009a) in order to expedite the anchor selection procedure. It is entirely possible to do anchor selection in a fully Bayesian way however, that is beyond the scope of the present study and would add additional unnecessary complexity to the Bayesian MIMIC model. That being said one could implement a procedure of posterior anchoring for the Bayesian MIMIC model as discussed in Frederickx et al. (2010), Soares et al. (2009), and Goncalves et al. (2013). More research is needed for anchor selection using the Bayesian MIMIC model.

A final note on a practical application of this method is computational speed. In the applied example I used similar burn-in, thinning, and post-burn-in draws to summarize the posterior of each parameter. I probably could have used less draws than I did but doing more will never hurt you and can only help. The speed for the applied example took approximately 5.5 hours to run the unconditional test of the discrimination parameters and conditional test of the difficulty parameters on a single processor using the ACF cluster at the University of Kansas. Then the final model took a little over two hours running a single processor on the ACF cluster at the University of Kansas. Clearly, the computational time is high for these models and future research could look at shoring up this gap.

6.6 Limitations and Future Directions

6.6.1 Limitations

The findings of the present study should be viewed in light of several limitations. First, and most important is that generalization of the findings to conditions outside of the scope of the simulated conditions should be done with caution. Care was taken to provide as much practical application

when choosing conditions, however, it is impossible to take into account every possible combination that could occur.

Second, the results of the Bayesian MIMIC model should be viewed in light of having mildly informative and informative prior distributions on the DIF effects. This was a bi-product of sign switching problems and other convergence related issues when more diffuse priors were used. Ideally it would be better to use more diffuse priors in certain circumstances. However, this was not possible due to some replications within conditions experiencing problematic non-convergence rates. Specifically, when conducting pilot studies to assess convergence, in general the most problematic conditions were with priors similar to those in the applied example and with equal sample sizes amongst the groups. This was likely due to more information from the focal group being available that would challenge the violation of assumption from the model. As the sample size imbalance increased there was less information in the likelihood and more information was borrowed from the priors. In these conditions convergence issues were much less of a problem.

A third limitation was the computational requirements and level of sophistication required to fit this type of model. Currently, the models take almost 6 to 8 hours to run a full analysis for DIF and a final model as discussed in the applied example section, this is probably too long for most practitioners to want to reasonably wait. However, one purpose of this study was to assess how well the MIMIC model would perform under adverse conditions to determine if implementing a faster method would be a fruitful future direction. Another computational concern was the amount of programming required to fit these models is quite high for many practitioners as can be seen by the code in the appendix for the applied example. Shoring up these two issues would be necessary in order to see more wide spread use of the Bayesian MIMIC model under conditions were it would be appropriate to use.

A fourth limitation was the number of replications used in the simulation. 100 replications were chosen given the hypothesis that this number would be adequate to detect Type I error problems that might arise with the Bayesian MIMIC model. However, the Bayesian MIMIC model

did not experience inflated Type I errors according to the criteria that was used to flag inflation. Yet, in the LSD of 0.5 and equal sample size condition for reference versus focal group 1 in the 3 group condition there was a pattern that suggested Type I error inflation might be present albeit there was not sufficient statistical evidence in the present study to make this claim. Further, a more striking unanticipated pattern was the Type I error rates that the improved Wald with SEM SEs showed. Although it was not possible to conclude that there was Type I error inflation in the present study it appears that there may be something going on with this method. In order to fully explore these pattens for both methods a larger number of replications would be needed.

6.6.2 Future Directions

Given the peculiar patterns of Type I errors mentioned in the limitation section involving the improved Wald it would be interesting to run a simulation with different SE estimators for the pairwise tests. Currently, in flexMIRT™ there are several options for SE estimation: the cross-product approximation, the supplemental EM algorithm, the expected Fisher information matrix, the sandwich covariance matrix, the forward difference method, and the Richardson extrapolation method (Houts & Cai, 2013). In a follow-up study it would be interesting to compare the performance of these different SE estimators on the pairwise DIF tests under a range of circumstances to determine the optimal method.

Another avenue for future research would be continuously varying the variances of the latent variables in order to get a more realistic sense of when things start to break down. This was not currently implemented in the present study because it would be prudent to have 1000 replications per condition to get adequate estimates of the variances, but attempting something like this in the future would be useful. Another area for further development for the Bayesian MIMIC model but also other Bayesian IRT models for DIF testing would be to develop fully Bayesian anchor techniques. One possible idea would be to implement a fully Bayesian version of the two-stage Wald anchor strategy proposed by Langer (2008) combined with the ranked based strategy in Woods (2009a) similar to what M. Wang and Woods (2015) proposed.

A final recommendation for future research would be to explore the possibility of implementing a variational Bayesian implementation for MIMIC models in order to expedite computation. Recently, Rijmen and Jeon (2013) implemented a variational Bayesian type method to test for measurement invariance of item parameters across countries. They concluded that in their circumstance the variational method offered a computationally tractable alternative to maximum likelihood or MCMC methods.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76. Retrieved from `http://www.jstor.org/stable/1165238`

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

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *Advances in Statistical Analysis*, *94*(2), 117–127. doi: 10.1007/s10182-010-0126-1

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling*, *19*(4), 561–579. doi: 10.1080/10705511.2012.713261

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. Retrieved from `http://www.jstor.org/stable/2346101`

Bilir, M. K. (2009). *Mixture item response theory-mimic model: Simultaneous estimation of differential item functioning for manifest groups and latent classes* (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinees ability. In F. Lord & M. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 395–479). Addison Wesley, Reading, MA.

Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi: 10.1007/BF02293801

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison–Wesley.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York, NY: Chapman and Hall.

Brown, L. D., Cai, T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133. Retrieved from `http://projecteuclid.org/download/pdf_1/euclid.ss/1009213286`

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *Journal of the American Statistical Association*, *61*(2), 309–329. doi: 10.1348/000711007X249603

Cai, L. (2015). *flexMIRT™version 3: A numerical engine for multilevel item factor analysis and test scoring*. [Computer Software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. (2013). IRTPRO™user's manual version 2.1: Flexible professional item response theory modeling for patient-reported outcomes [Computer software manual]. Skokie, IL: Scientific Software International, Inc..

Carroll, I. A. (2014). *MIMIC DIF testing when the latent variable variance differs between groups* (Unpublished Master's Thesis). University of Kansas, Lawrence, KS.

Cho, S. J. (2007). *A multilevel mixture IRT model for DIF analysis* (Unpublished doctoral dissertation). University of Georgia, Athens, GA.

Cho, S. J., & Cohen, A. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336–370. doi: 10.3102/1076998609353111

Clark, L. (1996). *SNAP manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning.

Journal of Educational Measurement, *42*(2), 133–148. doi: 10.1111/j.1745-3984.2005.00007

Curtis, S. (2010). Bugs code for item response theory. *Journal of Statistical Software*, *36*(1), 1–34. Retrieved from `http://www.jstatsoft.org/v36/c01`

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* Springer-Verlag, New York.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford Press.

De Ayala, R. J., Kim, S., Stapleton, L. M., & Dayton, C. M. (2003). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*(3&4), 243–276. doi: 10.1080/15305058.2002.9669495

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. doi: 10.1007/s11336-008-9092-x

De Jong, M. G., & Steenkamp, J. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika*, *75*(1), 3–32. doi: 10.1007/s11336-009-9134-z

El-Komboz, B. A., Zeileis, A., & Strobl, C. (2014). *Detecting differential item and step functioning with rating scale and partial credit trees* (Tech. Rep. No. 152). Department of Statistics at the University of Munich.

Finch, W. H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel–Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. doi: 10.1177/0146621605275728

Finch, W. H. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*, *71*(4), 663–683. doi: 10.1177/0013164410385226

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *67*(4), 565–582. doi: 10.1177/0013164406296975

Fleishman, J., Spector, W., & Altman, B. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(4), S275–S284. doi: 10.1093/geronb/57.5.S275

Fox, J. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 145–172. Retrieved from `http://onlinelibrary.wiley.com/doi/10.1348/000711005X38951/abstract`

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika*, *66*(2), 271–288. Retrieved from `http://link.springer.com.www2.lib.ku.edu:2048/article/10.1007/BF02294839`

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*(4), 432–457. doi: 10.1111/j.1745-3984.2010.00122.x

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*(3), 299–317. doi: 10.1111/j.1745-3984.2010.00115.x

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for DIF detection. *Educational and Psychological Measurement*, *67*(3), 373–393. doi: 10.1177/0013164406294781

Gelman, A., Carlin, J., Stern, H., Dunson, D. B., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (Third ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211. doi: 10.1080/19345747.2011.618213

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences.

Statistical Science, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741. doi: 10.1080/02664769300000058

George, E., & McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. doi: 10.1080/01621459.1993.10476353

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*, 647–667. Retrieved from `http://www3.stat.sinica.edu.tw/statistica/oldpdf/A8n32.pdf`

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*(3), 273–294. doi: 10.1007/BF02294296

Glas, C. A. W., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106. doi: 10.1177/0146621602250530

Goncalves, F. B., Gamerman, D., & Soares, T. M. (2013). Simultaneous multifactor DIF analysis and detection in item response theory. *Computational Statistics and Data Analysis*, *59*(4), 144–160. doi: 10.1016/j.csda.2012.10.011

Harpole, J. K., Levinson, C. A., Woods, C. M., Rodebaugh, T. L., Weeks, J. W., Brown, P. J., … Liebowitz, M. (2014). Assessing the straightforwardly-worded Brief Fear of Negative Evaluation Scale for differential item functioning across gender and ethnicity. *Journal of Psychopathology and Behavioral Assessment*, Advance online publication. doi: 10.1007/s10862-014-9455-9

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. doi: 10.1093/biomet/57.1.97

Hidalgo-Montesinos, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment*, *19*(1), 1–11. doi: 10.1027//1015-5759.19.1.1

Hoffman, M. D., & Gelman, A. (2011). The no–U–turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv*, *1111*(4246). Retrieved from http://arxiv.org/abs/1111.4246

Hoffman, M. D., & Gelman, A. (2013). The no–U–turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, in press*.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. doi: 10.1198/106186006X133933

Houts, C. R., & Cai, L. (2013). flexMIRT™user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software manual]. Chapel Hill, NC: Vector Psychometric Group. Retrieved from http://flexmirt.vpgcentral.com/Downloads/flexMIRTUserManual.pdf

Jackman, S. (2009). *Bayesian analysis for the social sciences*. West Sussex, PO19 8SQ, United Kingdom: John Wiley and Sons.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82–100.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639. doi: 10.1080/01621459.1975.10482485

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93. Retrieved from http://www.jstor.org/stable/1435439

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*(1), 201–210. Retrieved from http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy\&uid=1984-27738-001

Khalid, M. N., & Glas, C. A. W. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186–197. doi: 10.1016/j.measurement.2013.12.019

Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement*, *35*(6), 447–471. doi: 10.1177/0146621611407909

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*(3), 469–492. doi: 10.1177/0013164411427395

Kim, J., & Oshima, T. C. (2012). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458–470. doi: 10.1177/0013164412467033

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's Chi-square, Raju's area measures and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*(4), 291–312. doi: 10.1207/s15324818ame0804_2

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, *18*(3), 217–228. doi: 10.1177/014662169401800303

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*(3), 261–276. doi: 10.1111/j.1745-3984.1995.tb00466.x

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*(4), 457–474. doi: 10.1007/BF02296338

Klein, A., & Muthén. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, *42*(4),

647–673. doi: 10.1080/00273170701710205

Kopf, J., Augustin, T., & Strobl, C. (2014). The potential of model-based recursive partitioning in the social sciences: Revisiting Ockham's razor. In J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 75–95). Routledge.

Kruschke, J. (2010). *Doing Bayesian Data Analysis: A Tutorial Introduction with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill, NC.

Lee, S. Y., Song, X. Y., & Cai, J. H. (2010). A Bayesian approach for nonlinear structural equation models with dichotomous variables using logit and probit links. *Structural Equation Modeling*, *17*(2), 280–302. doi: 10.1080/10705511003659425

Lee, S. Y., Song, X. Y., & Tang, N. S. (2007). Bayesian methods for analyzing structural equation models with covariates, interaction, and quadratic latent variables. *Structural Equation Modeling*, *14*(3), 404–434. doi: 10.1080/10705510701301511

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics*, *2009*, 1–18. doi: 10.1155/2009/537139

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, *75*(2), 164–174. doi: 10.1037/0021-9010.75.2.164

Lopez, G. E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-SIBTEST, and logistic regression* (Unpublished doctoral dissertation). Florida State University, Tampa, FL.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372–379. doi:

10.1177/0146621603256021

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. doi: 10.1093/jnci/22.4.719

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*(3), 275–300. Retrieved from `http://psycnet.apa.org/journals/met/9/3/275/`

Marsh, H. W., Wen, Z., Nagengast, B., & Hau, K. T. (2012). Structural equation models of latent interaction. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. New York: The Guilford Press.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's Chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement*, *11*(2), 161–173. doi: 10.1177/014662168701100205

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. doi: 10.1016/0883-0355(89)90002-5

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300–307. doi: 10.1037/0033-2909.115.2.300

Meng, X., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*(416), 899–909. doi: 10.1080/01621459.1991.10475130

Merkle, E., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*(1), 59–82. doi: 10.1007/s11336-012-9302-4

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087. doi: 10.1063/1.1699114

Millsap, R. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*(5), 368–378. doi: 10.1016/j.jmp.2011.06.004

Murphy, K. (2012). *Machine learning: A probabalistic perspective*. Cambridge, MA: MIT Press.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, *10*, 121–132. doi: 10.3102/10769986010002121

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Erlbaum.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. doi: 10.1007/BF02296397

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: An application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1–22. doi: 10.1111/j.1745-3984.1991.tb00340.x

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Múthen & Múthen. Retrieved from `http://www.statmodel.com`

Noortgate, W. V. d., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443–464. Retrieved from `http://www.jstor.org/stable/3701298`

Noortgate, W. V. d., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*(4), 369–386. Retrieved from `http://www.jstor.org/stable/3701341`

Oltmanns, T. F., & Turkheimer, E. (2006). Perceptions of self and others regarding pathological personality traits. In R. F. Krueger & J. Tackett (Eds.), *Personality and psychopathology: Building bridges* (pp. 71–111). New York: Guilford.

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*(4), 342–366. doi: 10.3102/10769986024004342

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178. doi: 10.3102/10769986024002146

Penfield, R. (2007). Assessing differential step functioning in polytomous items using the common odds ratio estimator. *Journal of Educational Measurement*, *44*(3), 187–210. doi: 10.1111/j.1745-3984.2007.00034.x

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125–167). Elsevier B.V. doi: 10.1016/S0169-7161(06)26005-X

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), 5–15.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. doi: 10.1007/BF02295939

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. doi: 10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207. doi: 10.1177/014662169001400208

Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368. doi: 10.1177/014662169501900405

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters

with applications to problems of estimation. In *Cambridge philosophical society* (Vol. 44, pp. 50–507).

Rigdon, E. E., Schumacker, R. E., & Wothke, W. (1998). A comparative review of interaction and nonlinear modeling. In R. E. Schumacker & G. A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation modeling* (pp. 1–16). Mahwah, NJ: Erlbaum.

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research*, *206*(1), 647–662. doi: 10.1007/s10479-012-1181-7

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185–205. Retrieved from `http://search.proquest.com.www2.lib.ku.edu:2048/docview/614385388?accountid=14556`

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, *5*(3), 213–233. doi: 10.2307/1164965

Sacco, P., Casado, B. L., & Unick, G. J. (2011). Differential item functioning across race in aging research: An example using a social support measure. *Clinical Gerontologists*, *34*(1), 57–70. doi: 10.1080/07317115.2011.524818

Sacco, P., Torres, L. R., Cunningham-Williams, R. M., Woods, C. M., & Unick, G. J. (2011). Differential item functioning of pathological gambling criteria: An examination of gender, race/ethnicity, and age. *Journal of Gambling Studies*, *27*(2), 317–330. doi: 10.1007/s10899-010-9209-x

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. Retrieved from `http://projecteuclid.org/euclid.aos/1176344136`

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Mea-*

surement, *33*(3), 184–199. doi: 10.1177/0146621608321758

Silvey, S. D. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics*, *30*(2), 389–407. doi: 10.1214/aoms/1177706259

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications.

Soares, T. M., Goncalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*(3), 348–377. doi: 10.3102/1076998609332752

Stan Development Team. (2014a). *Rstan: the r interface to stan, version 2.5.0.* Retrieved from `http://mc-stan.org/rstan.html`

Stan Development Team. (2014b). Stan modeling language users guide and reference manual, version 2.5.0 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. doi: 10.1037/0021-9010.91.6.1292

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. doi: 10.1177/014662168300700208

Strobl, C., Kopf, J., & Zeileis, A. (2013). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, Advance online publication. doi: 10.1007/s11336-013-9388-3

Thissen, D. (2001). IRTLRDIF user's manual version 2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software manual]. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118–128.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-

Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77–83. doi: 10.3102/10769986027001077

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412. doi: 10.1007/BF02293705

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288. doi: 10.1111/j.1467-9868.2011.00771.x

Tutz, G., & Schauberger, G. (2013). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, Advance online publication. doi: 10.1111/j.1745-3984.1992.tb00368.x

Verhagen, A. J. (2012). *Bayesian item response theory models for measurement variance* (Unpublished doctoral dissertation). University of Twente, Enschede, The Netherlands.

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383–401. doi: 10.1111/j.2044-8317.2012.02059.x

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is larger. *Transactions of the American Mathematical Society*, *54*(3), 426–482. doi: 10.2307/1990256

Wall, M. M., & Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics*, *26*(1), 1–29. doi: 10.3102/10769986026001001

Wang, M., & Woods, C. W. (2015). Anchor selection using Wald test anchor-all-test-all procedure.

Manuscript submitted for publication.

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261. Retrieved from http://www.jstor.org/stable/20157369

Wang, W. C., Liu, C., & Wu, S. (2013). The random–threshold generalized unfolding model and its application of computerized adaptive testing. *Applied Psychological Measurement*, *37*(3), 179–200. doi: 10.1177/0146621612469720

Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166–180. doi: 10.1177/0146621609355279

Wang, W. C., Shih, C. L., & Sun, G. W. (2012). The DIF-Free-Then-DIF Strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209–212. doi: 10.1080/01621459.1927.10502953

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42–57. doi: 10.1177/0146621607314044

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavorial Research*, *44*(1), 1–27. doi: 10.1080/00273170802620121

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532–547. doi: 10.1177/0013164412464875

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361. doi: 10.1177/0146621611405984

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing

with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment*, *31*(4), 320–330. doi: 10.1007/s10862-008-9118-9

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*(2), 281–301. doi: 10.1007/s11336-004-1175-8

Yang, F. M., Tommet, D., & Jones, R. N. (2009). Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: An extension of the bi-factor item response theory model for use in differential item functioning. *Journal of Psychiatric Research*, *43*, 1025–1035. doi: 10.1016/j.jpsychires.2008.12.007

Yu, F. Y., Yu, A. P., & Ahn, J. (2007). Investigating differential item functioning by chronic diseases in the SF–36 Health Survey: A latent trait analysis using MIMIC models. *Medical Care*, *45*(9), 851–859. doi: 10.1016/j.jpsychires.2008.12.007

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, *68*(1), 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508. doi: 10.1111/j.1467-9574.2007.00371.x

Appendix A

Confusion Matrices: Reference vs. Focal Group 2

Table A.1: Confusion Matrices for 3 Group Conditions: Improved Wald for Reference vs Focal Group 2

SS: 500:500:500 LSD: 0.5

(True)	(Predicted) U	NU	M	N
U	**0.076**	0.004	0.006	0.039
NU	0.029	**0.018**	0.009	0.070
M	0.051	0.016	**0.022**	0.036
N	0.018	0.043	0.003	**0.561**

SS: 1000:250:250 LSD: 0.5

(True)	(Predicted) U	NU	M	N
U	**0.064**	0.004	0.004	0.052
NU	0.037	**0.009**	0.013	0.066
M	0.044	0.015	**0.009**	0.057
N	0.007	0.037	0.000	**0.581**

SS: 1300:100:100 LSD: 0.5

(True)	(Predicted) U	NU	M	N
U	**0.038**	0.007	0.001	0.079
NU	0.034	**0.005**	0.015	0.071
M	0.024	0.014	**0.007**	0.081
N	0.021	0.036	0.008	**0.560**

SS: 500:500:500 LSD: 1

(True)	(Predicted) U	NU	M	N
U	**0.079**	0.000	0.004	0.041
NU	0.008	**0.028**	0.009	0.080
M	0.061	0.014	**0.033**	0.017
N	0.016	0.024	0.001	**0.584**

SS: 1000:250:250 LSD: 1

(True)	(Predicted) U	NU	M	N
U	**0.066**	0.002	0.005	0.052
NU	0.008	**0.029**	0.007	0.081
M	0.054	0.011	**0.037**	0.024
N	0.007	0.025	0.001	**0.592**

SS: 1300:100:100 LSD: 1

(True)	(Predicted) U	NU	M	N
U	**0.031**	0.001	0.002	0.091
NU	0.011	**0.021**	0.004	0.089
M	0.031	0.015	**0.011**	0.068
N	0.012	0.021	0.001	**0.591**

SS: 500:500:500 LSD: 1.5

(True)	(Predicted) U	NU	M	N
U	**0.089**	0.001	0.002	0.033
NU	0.007	**0.033**	0.007	0.078
M	0.052	0.014	**0.040**	0.019
N	0.015	0.024	0.000	**0.586**

SS: 1000:250:250 LSD: 1.5

(True)	(Predicted) U	NU	M	N
U	**0.069**	0.001	0.002	0.053
NU	0.012	**0.029**	0.006	0.078
M	0.062	0.007	**0.031**	0.025
N	0.015	0.011	0.000	**0.599**

SS: 1300:100:100 LSD: 1.5

(True)	(Predicted) U	NU	M	N
U	**0.041**	0.000	0.002	0.082
NU	0.007	**0.022**	0.004	0.091
M	0.055	0.009	**0.014**	0.048
N	0.012	0.013	0.000	**0.601**

Note. U = uniform DIF, NU = non-uniform DIF, M = mixed DIF, N = no DIF, SS = sample size, LSD = latent focal group SD.

Table A.2: Confusion Matrices for 3 Group Conditions: Bayesian MIMIC Model for Reference vs Focal Group 2

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500:500 LSD: 0.5**
(True)	**U**	**0.091**	0.001	0.002	0.031	
	NU	0.028	**0.001**	0.003	0.092	
	M	0.072	0.008	**0.010**	0.035	
	N	0.041	0.006	0.001	**0.578**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500:500 LSD: 1**
(True)	**U**	**0.106**	0.000	0.000	0.019	
	NU	0.037	**0.000**	0.001	0.087	
	M	0.100	0.000	**0.002**	0.023	
	N	0.029	0.000	0.000	**0.596**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 500:500:500 LSD: 1.5**
(True)	**U**	**0.107**	0.000	0.000	0.018	
	NU	0.043	**0.000**	0.000	0.083	
	M	0.107	0.001	**0.001**	0.017	
	N	0.046	0.000	0.000	**0.579**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1000:250:250 LSD: 0.5**
(True)	**U**	**0.086**	0.000	0.001	0.039	
	NU	0.045	**0.000**	0.002	0.078	
	M	0.076	0.001	**0.000**	0.048	
	N	0.034	0.001	0.000	**0.590**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1000:250:250 LSD: 1**
(True)	**U**	**0.097**	0.000	0.000	0.028	
	NU	0.039	**0.000**	0.000	0.086	
	M	0.091	0.001	**0.001**	0.033	
	N	0.025	0.000	0.000	**0.600**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1000:250:250 LSD: 1.5**
(True)	**U**	**0.098**	0.000	0.000	0.028	
	NU	0.038	**0.000**	0.000	0.087	
	M	0.100	0.000	**0.000**	0.025	
	N	0.036	0.000	0.000	**0.589**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1300:100:100 LSD: 0.5**
(True)	**U**	**0.054**	0.000	0.000	0.071	
	NU	0.035	**0.000**	0.000	0.090	
	M	0.046	0.000	**0.000**	0.079	
	N	0.030	0.000	0.000	**0.595**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1300:100:100 LSD: 1**
(True)	**U**	**0.060**	0.000	0.000	0.065	
	NU	0.022	**0.000**	0.000	0.103	
	M	0.049	0.000	**0.000**	0.076	
	N	0.021	0.000	0.000	**0.604**	

		(Predicted)				
		U	**NU**	**M**	**N**	**SS: 1300:100:100 LSD: 1.5**
(True)	**U**	**0.064**	0.000	0.000	0.061	
	NU	0.022	**0.000**	0.000	0.103	
	M	0.072	0.000	**0.000**	0.053	
	N	0.020	0.000	0.000	**0.605**	

Note. U = uniform DIF, NU = non-uniform DIF, M = mixed DIF, N = no DIF, SS = sample size, LSD = latent focal group SD.

Appendix B

R Code for Applied Example

Code Listing B.1: Stan Syntax to Fit Non-uniform DIF

```
stan3GrpFit1 <- "
data{
int<lower=0>NumGrp;
int<lower=1>N;
int<lower=1>nItems;
int<lower=1>nAnchors;
int<lower=0>Grp[N];
int<lower=0, upper=1>G1[N];
int<lower=0, upper=1>G2[N];
int<lower=0, upper=1>y[N,nItems];
}
parameters{
real theta[N];
real<lower=0>alpha[nItems];
real b[nItems];
real mu_theta[(NumGrp - 1)]; //Means for focal groups
//Items tested for DIF
real beta1[(nItems - nAnchors)];
real beta2[(nItems - nAnchors)];
real omega1[(nItems - nAnchors)];
real omega2[(nItems - nAnchors)];
}
transformed parameters{
//Creates latent means
vector[NumGrp] mu_main;
//Creates the DIF parameters and anchor items;
```

```
vector[nItems] beta1_main;
vector[nItems] beta2_main;
vector[nItems] omega1_main;
vector[nItems] omega2_main;
for(i in 1:NumGrp){
if(i == 1){
//Set latent mean to zero for reference group
mu_main[i] <- 0;
} else {
//Estimate latent means for focal groups
mu_main[i] <- mu_theta[(i-1)];
}
}

for(i in 1:nItems){
if(nItems - i >= nAnchors){
//Uniform DIF
beta1_main[i] <- beta1[i];
beta2_main[i] <- beta2[i];
//Non-uniform DIF
omega1_main[i] <- omega1[i];
omega2_main[i] <- omega2[i];
//Else deals with anchor items
} else {
beta1_main[i] <- 0;
beta2_main[i] <- 0;
omega1_main[i] <- 0;
omega2_main[i] <- 0;

}
}
}
model{
//Sample item parameters
b ~ normal(0, 10); //2
alpha ~ lognormal(0, 1);
//Sample DIF parameters
beta1 ~ normal(0, 4); //2.25
beta2 ~ normal(0, 4); //2.25
omega1 ~ normal(0, 4); //.2
omega2 ~ normal(0, 4); //.2
mu_theta ~ normal(0, 10); //2
//Sample latent traits
for(i in 1:N){
theta[i] ~ normal(mu_main[Grp[i] + 1], 1);
```

```
}
//Model
for(i in 1:N){
for(j in 1:nItems){
y[i, j] ~ bernoulli_logit(alpha[j]*theta[i] + omega1_main[j]*theta[i]*G1[i]
   +
omega2_main[j]*theta[i]*G2[i] + alpha[j]*b[j] + beta1_main[j]*G1[i]
+ beta2_main[j]*G2[i]);
}
}
}
"
```

Code Listing B.2: Stan Syntax to Fit Uniform DIF

```
stan3GrpFit2 <- "
data{
int<lower=0>NumGrp;
int<lower=1>N;
int<lower=1>nItems;
int<lower=1>nAnchors;
int<lower=0>Grp[N];
int<lower=0, upper=1>G1[N];
int<lower=0, upper=1>G2[N];
int<lower=0, upper=1>y[N,nItems];
}
parameters{
real theta[N];
real<lower=0>alpha[nItems];
real b[nItems];
real mu_theta[(NumGrp - 1)]; //Means for focal groups
//Items tested for DIF
real beta1[(nItems - nAnchors)];
real beta2[(nItems - nAnchors)];
}
transformed parameters{
//Creates latent means
vector[NumGrp] mu_main;
//Creates the DIF parameters and anchor items;
vector[nItems] beta1_main;
vector[nItems] beta2_main;
for(i in 1:NumGrp){
if(i == 1){
```

```
//Set latent mean to zero for reference group
mu_main[i] <- 0;
} else {
//Estimate latent means for focal groups
mu_main[i] <- mu_theta[(i-1)];
}
}

for(i in 1:nItems){
if(nItems - i >= nAnchors){
//Uniform DIF
beta1_main[i] <- beta1[i];
beta2_main[i] <- beta2[i];
//Else deals with anchor items
} else {
beta1_main[i] <- 0;
beta2_main[i] <- 0;

}
}
}
model{
//Sample item parameters
b ~ normal(0, 10); // 10
alpha ~ lognormal(0, 1);
//Sample DIF parameters
beta1 ~ normal(0, 4); // 2.25
beta2 ~ normal(0, 4); //2.25
mu_theta ~ normal(0, 10); // 10
//Sample latent traits
for(i in 1:N){
theta[i] ~ normal(mu_main[Grp[i] + 1], 1);
}
//Model
for(i in 1:N){
for(j in 1:nItems){
y[i, j] ~ bernoulli_logit(alpha[j]*theta[i] + alpha[j]*b[j] +
beta1_main[j]*G1[i] + beta2_main[j]*G2[i]);
}
}
}
"
```

Code Listing B.3: Code to Prepare the Data Set for Analyses

```
## This stuff creates the stan data set necessary to run analyses.
stan_data <- subset(data1, RACE == 1 | RACE == 2 | RACE == 6)

stan_data$G1 <- ifelse(stan_data$RACE == 2, 1, 0)
stan_data$G2 <- ifelse(stan_data$RACE == 6, 1, 0)
stan_data$Grp <- ifelse(stan_data$RACE == 1, 0, ifelse(stan_data$RACE == 2,
    1, 2))

stan_data <- stan_data[, -grep("^GENDER$|^RACE$", colnames(stan_data))]

# stan_data <- select(stan_data, starts_with('it'), G1, G2, Grp)

itemI <- grep("^it\\d{1,}$", colnames(stan_data))

## Order anchor items at the end
anchors <- c('it3', 'it9', 'it10', 'it13', 'it21', 'it22')

not_anchors <- setdiff(paste0('it', 1:length(itemI)), anchors)

## Now order anchor items at the end
stan_data <- stan_data[, c(not_anchors, anchors, 'G1', 'G2', 'Grp')]

## Create stan input
stanInput <- list(y = stan_data[, itemI],
                  NumGrp = 3, G1 = stan_data$G1, G2 = stan_data$G2, N = nrow(
                     stan_data),
                  Grp = stan_data$Grp, nItems = length(itemI),
                  nAnchors = 6)
```

Code Listing B.4: Code to Run Stan for DIF Tests

```
library(rstan)
## Fit models and save the fitted object as RDS
## fit 1 is unconditional test of discriminations
fit1 <- stan(model_code = stan3GrpFit1, model_name = paste0("fit", 1, '_3
   grp'),
           data = stanInput, iter = 3000, chains = 2, warmup = 1000,
           verbose = FALSE, seed = 123456L)
saveRDS(fit1, file = '3grp_fit1.rds')
rm(fit1)
## fit 2 is conditional test of difficulties
```

```
fit2 <- stan(model_code = stan3GrpFit2, model_name = paste0('fit', 2, '_3
   grp'),
             data = stanInput, iter = 3000, chains = 2, warmup = 1000,
             verbose = FALSE, seed = 123456L)
saveRDS(fit2, file = '3grp_fit2.rds')
rm(fit2)
```

Code Listing B.5: Fitting the Final Model

```
stan3GrpFinalModel <- "
  data{
    int<lower=0>NumGrp;
    int<lower=1>N;
    int<lower=1>nItems;
    int<lower=0>Grp[N];
    int<lower=0, upper=1>G1[N];
    int<lower=0, upper=1>G2[N];
    int<lower=0, upper=1>y[N,nItems];
  }

  parameters{
    real theta[N];
    real<lower=0>alpha[nItems];
    real b[nItems];
    real mu_theta[(NumGrp - 1)]; //Means for focal groups
    //Items tested for DIF
    real beta1[(nItems)];
    real beta2[(nItems)];
    real omega1[(nItems)];
    real omega2[(nItems)];
  }

  transformed parameters{
    //Creates latent means
    vector[NumGrp] mu_main;
    //Creates the DIF parameters and anchor items;
    vector[nItems] beta1_main;
    vector[nItems] beta2_main;
    vector[nItems] omega1_main;
    vector[nItems] omega2_main;
    for(i in 1:NumGrp){
      if(i == 1){
      //Set latent mean to zero for reference group
```

```
  mu_main[i] <- 0;
  } else {
  //Estimate latent means for focal groups
  mu_main[i] <- mu_theta[(i-1)];
  }
} // End of loop over NumGrp

// This code puts the necessary constraints
// on the appropriate parameters for the final
// model.
for(i in 1:nItems){
  if(i == 1|| i == 2){
    // G1
    beta1_main[i] <- beta1[i];
    omega1_main[i] <- 0;
    // G2
    beta2_main[i] <- 0;
    omega2_main[i] <- 0;
  } else if(i == 5){
    // G1
    beta1_main[i] <- beta1[i];
    omega1_main[i] <- omega1[i];
    //G2
    beta2_main[i] <- 0;
    omega2_main[i] <- 0;
  } else if(i == 6){
    //G1
    beta1_main[i] <- beta1[i];
    omega1_main[i] <- 0;
    // G2
    beta2_main[i] <- 0;
    omega2_main[i] <- 0;
  } else if(i == 7){
    // G1
    beta1_main[i] <- 0;
    omega1_main[i] <- 0;
    // G2
    beta2_main[i] <- beta2[i];
    omega2_main[i] <- omega2[i];
  } else if(i == 11){
    //G1
    beta1_main[i] <- beta1[i];
    omega1_main[i] <- 0;
    //G2
    beta2_main[i] <- 0;
```

```
  omega2_main[i] <- 0;
} else if(i == 16){
  //G1
  beta1_main[i] <- 0;
  omega1_main[i] <- 0;
  //G2
  beta2_main[i] <- beta2[i];
  omega2_main[i] <- omega2[i];
} else if(i == 19){
  // G1
  beta1_main[i] <- beta1[i];
  omega1_main[i] <- 0;
  // G2
  beta2_main[i] <- beta2[i];
  omega2_main[i] <- 0;
} else if(i == 20){
  // G1
  beta1_main[i] <- 0;
  omega1_main[i] <- 0;
  // G2
  beta2_main[i] <- 0;
  omega2_main[i] <- omega2[i];
} else if(i == 23){
  // G1
  beta1_main[i] <- beta1[i];
  omega1_main[i] <- 0;
  // G2
  beta2_main[i] <- beta2[i];
  omega2_main[i] <- 0;
} else if(i == 26){
  // G1
  beta1_main[i] <- beta1[i];
  omega1_main[i] <- 0;
  // G2
  beta2_main[i] <- beta2[i];
  omega2_main[i] <- omega2[i];
} else if(i == 28){
  // G1
  beta1_main[i] <- 0;
  omega1_main[i] <- 0;
  // G2
  beta2_main[i] <- 0;
  omega2_main[i] <- omega2[i];
} else {
  // G1
```

```
      beta1_main[i] <- 0;
      omega1_main[i] <- 0;
      // G2
      beta2_main[i] <- 0;
      omega2_main[i] <- 0;
    }
  }// End of loop for final model constraints
}// End of transfomed parameters

model{
  //Sample item parameters
  b ~ normal(0, 10); //10
  alpha ~ lognormal(0, 1);
  //Sample DIF parameters
  beta1 ~ normal(0, 4); //4
  beta2 ~ normal(0, 4); //4
  omega1 ~ normal(0, 4); //4
  omega2 ~ normal(0, 4); //4
  mu_theta ~ normal(0, 10); //10
  //Sample latent traits
  for(i in 1:N){
    theta[i] ~ normal(mu_main[Grp[i] + 1], 1);
  } // End of loop over N

  //Model
    for(i in 1:N){
      for(j in 1:nItems){
        y[i, j] ~ bernoulli_logit(alpha[j]*theta[i] + omega1_main[j]*theta[
            i]*G1[i] +
                                  omega2_main[j]*theta[i]*G2[i] + alpha[j]*b[
                                      j] +
                                  beta1_main[j]*G1[i] + beta2_main[j]*G2[i]);
       } // End of loop over nItems
    } // End of loop over N
} // End of model block

generated quantities{
  // Focal group items
  vector[nItems] b_f1;
  vector[nItems] alpha_f1;
  vector[nItems] b_f2;
  vector[nItems] alpha_f2;

  // Reference Group items
  vector[nItems] b_ref;
```

```
  vector[nItems] alpha_ref;

  //Loop to create the item parameters for focal grps
  for(i in 1:nItems){
    // Focal Group 1
    b_f1[i] <- -1*b[i] - beta1_main[i];
    alpha_f1[i] <- alpha[i] + omega1_main[i];
    // Focal Group 2
    b_f2[i] <- -1*b[i] - beta2_main[i];
    alpha_f2[i] <- alpha[i] + omega2_main[i];
    // Reference Group
    b_ref[i] <- -1*b[i];
    alpha_ref <- alpha[i];
  } // End of loop over nItems
} // End of generated quantities
"

## Read in the data set and save the key
data1 <- read.csv('appliedEx.csv', stringsAsFactors = FALSE)

## Create key for later
item_ind <- grep('^SNAP', colnames(data1))

datakey <- names(data1)[item_ind]
## Define key for flexMIRT for anchor selection
flexkey <- paste0('it', 1:length(datakey))
## Write out masterkey to file for reference later
masterkey <- cbind(datakey, flexkey)
colnames(masterkey) <- c('datakey', 'flexkey')
write.table(masterkey, file = 'masterkey.txt', row.names = FALSE)
## Set data1 item_index as flexkey
colnames(data1)[item_ind] <- flexkey
library(rstan)

## This stuff creates the stan data set necessary to run analyses.
## Get the necessary groups from data1
stan_data <- subset(data1, RACE == 1 | RACE == 2 | RACE == 6)
## Do the recoding
stan_data$G1 <- ifelse(stan_data$RACE == 2, 1, 0)
stan_data$G2 <- ifelse(stan_data$RACE == 6, 1, 0)
stan_data$Grp <- ifelse(stan_data$RACE == 1, 0, ifelse(stan_data$RACE == 2,
    1, 2))
## Remove Gender and race columns from stan_data
stan_data <- stan_data[, -grep("^GENDER$|^RACE$", colnames(stan_data))]
## Get the item index for stan_data for use in stanInput
```

```
itemI <- grep("^it\\d{1,}$", colnames(stan_data))
## Create stanInput object
stanInput <- list(y = stan_data[, itemI],
                  NumGrp = 3, G1 = stan_data$G1, G2 = stan_data$G2, N = nrow(
                     stan_data),
                  Grp = stan_data$Grp, nItems = length(itemI))
## Select parameters to track
params <- c('alpha', 'b', 'b_f1', 'b_f2', 'alpha_f1', 'alpha_f2',
            'beta1_main', 'beta2_main', 'omega1_main', 'b_ref', 'alpha_ref',
            'omega2_main', 'mu_theta', 'theta')
## Run stan
final_model <- stan(model_code = stan3GrpFinalModel, model_name = paste0("
   fit", 1, '_3grp'),
            data = stanInput, iter = 100, chains = 2, warmup = 10,
            verbose = FALSE, seed = 123456L, pars = params, include = TRUE)
saveRDS(final_model, file = '3grp_finalmodel.rds')
rm(fit1)
```