Comparison of Differential Item Functioning Detection Procedures under the Multidimensional Graded Response Model Framework

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

John Phillip Walker

August 2021

For Chelsea and our growing family

"The mystery of human existence lies not in just staying alive, but in finding something to live for." – Fyodor Dostoyevsky, The Brothers Karamazov

Acknowledgements

To my dissertation committee, Drs. McCallum, Morrow, and Nugent, thank you for your insightful comments and suggestions to help make my dissertation exceptional. I am grateful for your expertise, time, and patience over the past year. A special thanks to Dr. Morrow for her kindness and professionalism throughout my time in the Evaluation, Statistics, and Measurement (ESM) program. You were instrumental in my decision to enter the program. To my advisor and committee chair, Dr. Rocconi, I am appreciative for your mentorship and faith in me throughout the dissertation process and during my time in the ESM program. You were pivotal to my growth and development as a researcher. To Dr. Skolits, thank you for your trust and the opportunity to come to UT as a staff member. To current and past members of Teaching and Learning Innovation, I am thankful to be surrounded by intelligent and caring colleagues who pushed me to be the best version of myself. To Dr. Lavan, I am fortunate to have a supervisor who not only cares about me as a professional, but as a person. I could not have achieved my many accomplishments at UT without your support. To past colleagues and mentors at Schoolcraft College, University of Sydney, and Grand Valley State University, thank you for preparing me for my future professional and academic pursuits. To my family, I am blessed to have supportive loved ones who inspire me to pursue my goals. To my wife, Chelsea, thank you for your encouragement throughout this process, instilling my drive to be better, and for your patience as I "just finish this last sentence…"

Abstract

Construct validity is necessary to confirm psychometric instruments measure their intended construct and limit measurement error. Differential item functioning (DIF) helps to identify bias and error within measurement instruments and support construct validity. DIF detection procedures have been thoroughly scrutinized under several item response theory models (e.g., Bulut & Suh, 2017; Cohen et al., 1996; Suh & Cho, 2014; Elosua & Wells, 2013; Wang & Shih, 2010; Woods & Grimm, 2011), but no study has explored DIF detection procedures for data fitted to polytomous multidimensional item response theory models. This study aims to identify the optimal DIF procedures for data that fits the multidimensional graded response model.

Comparisons were made on three statistical and psychometric DIF detection procedures, the multidimensional IRT likelihood ratio (MIRT-LR) test, the multidimensional extension of the logistic discriminant function analysis (MLDFA) method, and the multidimensional multiple causes, multiple indicators interaction (MIMIC-interaction) model. Multidimensional graded response data were generated for twenty items with five response options through a Monte Carlo simulation with varied constraints on sample size, DIF type, percentage of DIF, correlations between latent traits, and latent mean differences between groups to determine the effect of the three DIF detection methods on type I error and rejection rates.

Results indicated type I error rates were inflated (i.e., greater than .05) for all three DIF detection procedures. The MLDFA method produced the highest rejection rates, but also displayed the type I error rates between .06 and .15. The MIRT-LR test indicated poor ability to detect nonuniform DIF and greatly inflated type I error rates when latent mean differences were unbalanced. The MIMIC-interaction model exhibited the lowest type I error rates and adequate

rejection rates, indicating strong statistical power under most conditions. General method selection recommendations are provided for psychometric and assessment professionals. Future research for DIF detection procedure for polytomous multidimensional item response theory models is also discussed.

Table of Contents

Chapter I: Literature Review 1
Background of the Problem 2
DIF under the 2-PL IRT Model 4
Uniform DIF 5
Nonuniform DIF 6
DIF Procedures under the 2-PL IRT Framework 7
MIMIC models 8
IRT-LR test 11
Logistic regression 12
Polytomous IRT Models 13
DIF under the GRM 16
DIF Procedures under the GRM 18
Logistic discriminant function analysis 19
Dimensionality 20
Multidimensional IRT 21
Multidimensional GRM 22
DIF Procedures under Multidimensional IRT Models 25
Multidimensional MIMIC-interaction model 25
Multidimensional extension of the IRT-LR test 26
Multivariate logistic regression 30
Factors that Affect DIF 31
Sample size 32
Type of DIF 34
Proportion of DIF 35
Correlation between latent traits 36
Latent mean differences across groups 37
Studies on Comparative DIF Procedures 37
DIF detection procedures for unidimensional dichotomous models 38
DIF detection procedures for unidimensional polytomous models 39
DIF detection procedures for multidimensional dichotomous models 40
Statement of the Problem 41
Significance of the Problem 42
Research Objectives 43

Chapter II: Methods 45
Data Generation 45
Data framework 46

Factor constraints 47
Analytical Procedures 52
MIRT likelihood ration test (MIRT-LR) 52
Multidimensional MIMIC-interaction model 54
Multidimensional extension of the logistic discriminant function analysis 57
Evaluation Criteria 59
Type I error rates 60
Rejection rates 61
Effect of factor constraints 62
Empirical Example 64
Instrument 65
Model fit procedures 68
Analytical procedures 69

Chapter III: Results 70
Type I Error Rates 70
Uniform DIF 71
Nonuniform DIF 72
Effect of factor constraints on type I error 76
Effect of DIF type 77
Effect of mean differences between groups 80
Effect of correlation between latent traits 83
Effect of sample size 84
Rejection Rates 87
Uniform DIF 88
Nonuniform DIF 92
Effect of factor constraints on rejection rates 96
Effect of percentage of DIF 98
Effect of DIF type 102
Effect of mean differences between groups 104
Effect of correlation between latent traits 105
Effect of sample size 106
Empirical Example 107
Model fit 108
Comparison of DIF detection methods 110

Chapter IV: Discussion, Limitations, and Conclusion 113
Type I Error Rates 114
MIRT-LR test 114

MIMIC-interaction model....115
MLDFA method....116
Type of DIF....117
Latent mean differences across groups....118
Correlation between latent traits....118
Sample sizes....119
Rejection Rates....119
MIRT-LR test....120
MIMIC-interaction model....121
MLDFA method....121
Percentage of DIF....122
Type of DIF....123
Latent mean differences across groups....124
Correlation between latent traits....125
Sample sizes....125
Empirical and Practical Implications....127
Limitations and Future Research....129
Recommendations and Conclusion....131

References....134

Appendices....153
Appendix A: Supplemental Tables....154
Appendix B: Supplemental Figures....160

Vita....164

List of Tables

Table 2.1: Anchor item parameter generation of multidimensional graded response model 49
Table 2.2: Item parameter generation of multidimensional graded response model for DIF conditions 51
Table 2.3: Alignment of survey items with Experience Learning SLOs and AAC&U VALUE rubrics 66
Table 3.1: Item descriptive statistics by gender 109
Table 3.2: Model fit statistics for comparison of 2-PL MGRM, MGPCM, and 1-PL MGRM models 109
Table 3.3: DIF detection method *p*-values for uniform and nonuniform DIF 112
Table A.1: Rejection and type I error rates for multidimensional item response theory likelihood ratio test (MIRT-LR) 154
Table A.2: Rejection and type I error rates for multiple indicator multiple causes interaction structural equation model (MIMIC) 155
Table A.3: Rejection and type I error rates for multidimensional extension of the logistic discriminant function analysis (MLDFA) 156
Table A.4: Item discrimination and threshold parameters for empirical data set 157
Table A.5: Item purification test for anchor and studied item selection 158
Table A.6: MGRM discrimination parameters for empirical data by gender 158
Table A.7: MGRM threshold parameters for empirical data by gender 159

List of Figures

Figure 1.1: Multiple indicator multiple causes (MIMIC) interaction DIF model.....10
Figure 1.2: Category response functions for four graded response model items.....17
Figure 1.3: Category response functions for a multidimensional graded response model item 24
Figure 1.4: Multiple indicator multiple causes (MIMIC) interaction DIF model for multidimensional data.....27
Figure 1.5: Flowchart for assessing DIF using the multidimensional extension of IRT-LR test ..29
Figure 2.1: Two-dimensional MIMIC-interaction model with 5% DIF and fixed variance 56
Figure 3.1: Type I error rates for uniform DIF results with 95% confidence intervals and α = .05 reference line.....73
Figure 3.2: Type I error rates for nonuniform DIF results with 95% confidence intervals and α = .05 reference line 74
Figure 3.3: Boxplot of interactions and significant simple main effects of factor constraints and DIF detection method for log-odds of type I error rate 79
Figure 3.4: Boxplot of DIF detection method main effects for log-odds of type I error rate 85
Figure 3.5: Rejection rates for 5% uniform DIF results with 95% confidence intervals and 1 – β = .80 reference line 89
Figure 3.6: Rejection rates for 10% uniform DIF results with 95% confidence intervals and 1 – β = .80 reference line 90
Figure 3.7: Rejection rates for 20% uniform DIF results with 95% confidence intervals and 1 – β = .80 reference line 91
Figure 3.8: Rejection rates for 5% nonuniform DIF results with 95% confidence intervals and 1 – β = .80 reference line 93
Figure 3.9: Rejection rates for 10% nonuniform DIF results with 95% confidence intervals and 1 – β = .80 reference line 94
Figure 3.10: Rejection rates for 20% nonuniform DIF results with 95% confidence intervals and 1 – β = .80 reference line 95
Figure 3.11: Boxplot of interactions and significant simple main effects of factor constraints and DIF detection method for log-odds of rejection rate 99
Figure 3.12: Boxplot of DIF detection method main effects for log-odds of rejection rate 101

Figure B.1: Kernel density distributions for type I error rates by DIF detection method.....160

Figure B.2: Kernel density estimate for type I error rates and logistic probability density function with scaling and mean adjustments.....161

Figure B.3: Kernel density distributions for rejection rates by DIF detection method.....162

Figure B.4: Kernel density estimate for rejection rates and logistic probability density function with scaling and mean adjustments.....163

Glossary of Terms

Augmented model: A statistical model that maximizes the full use of the available parameter space (i.e., allows all parameters to vary freely).

Compact model: A statistical model that constrains parameters and used to compare to the augmented model in a likelihood ratio test.

Compensatory model: A within-item multidimensional model that specifies the probability of endorsing an item based on the sum of the latent trait values. Thus, strong endorsement of one latent trait can compensate for low endorsement of another latent trait on the same item.

Dichotomous item: An item that contains only two possible response options.

Differential item functioning (DIF): A psychometric method that examines variability in measured responses between two or more groups.

Difficulty parameter (*b*): The trait level required for respondents to have a 50% chance of endorsing a dichotomous response.

Discrimination parameter (*a*): Describes how well items differentiate from one another on a latent trait.

Endorsed response: The selected response in an item from a respondent**.**

Focal group: A subgroup of interest from the population of a differential item functioning study.

Item Response Theory (IRT): A statistical model that represents the probability of a response to an item Y as a function of the latent ability of a person θ and one or more item-level parameters.

Latent ability/trait (θ): An unobservable variable used to express a person's underlying ability.

Likelihood ratio test: A statistical test that assesses the goodness-of-fit using the ratio of the augmented and compact models. The test provides evidence for statistical differences between the two models.

Logistic discrimination function analysis: A logistic regression analysis in which the dichotomous group variable is the dependent variable and the polytomous item response category is an independent variable.

Monte Carlo simulation: The generation of random data from specified conditions using multiple iterations of the data generation procedure to fit a certain model.

Multidimensional graded response model (MGRM): A two-step process that examines the relationship between the multiple latent abilities of a person and the probability of endorsing responses from two or more response categories.

Multidimensionality: A characterization of the relationship between a person and measured items on multiple latent traits.

Multiple indicator multiple cause (MIMIC) model: A special case within structural equation modeling in which latent factor(s) are regressed on covariates (i.e., multiple indicators) and manifest variables (i.e., the multiple causes) are regressed on the latent factor(s).

Nonuniform DIF: A case of differential item functioning where both the discrimination and difficulty parameters are statistically unequal between focal and reference groups.

Parameter: A statistical variable in a differential item functioning study.

Polytomous item: An item that contains more than two possible response options.

Reference group: A subgroup of the population of a differential item functioning study from which the focal group is compared.

Rejection rate: The rejection of a false null hypothesis (syn. Power).

Threshold parameter (*d*): the ability level needed for a 50% chance of endorsing at or above some category *j* for a polytomous response.

Type I error rate: the rejection of a true null hypothesis, sometimes referred to as a false positive**.**

Unidimensionality: A characterization of the relationship between a person and measured items on a single latent trait.

Uniform DIF: A case of differential item functioning where discrimination parameters are equal between groups, but difficulty parameters are statistically unequal between groups.

List of Acronyms

Abbreviation	Full Description
1-PL	One-parameter logistic model
2-PL	Two-parameter logistic model
AAC&U	Association of American Colleges and Universities
ANOVA	Analysis of variance
ASK	Application of knowledge and skills factor
BH	Benjamini-Hochberg correction
BRF	Boundary response function
CFI	Comparative fit index
COL	Collaboration factor
CRF	Category response function
DIF	Differential item functioning
EM	Expectation maximization
FIML	Full information maximum likelihood
GPCM	Generalized partial credit model
GRM	Graded response model
IRT	Item response theory
IRT-LR	Item response theory likelihood ratio test
LL	Lifelong learning factor
LMS	Latent moderated structural equation method
MGPCM	Multidimensional generalized partial credit model
MGRM	Multidimensional graded response model
MHRM	Metropolis-Hastings Robbins-Monro algorithm
MIMIC	Multiple indicators multiple causes

Abbreviation	Full Description
MIRT	Multidimensional item response theory
MIRT-LR	Multidimensional item response theory likelihood ratio test
MLDFA	Multidimensional extension of logistic discrimination function analysis
MLR	Maximum likelihood estimation with Huber-White robust standard errors
$RMSEA_2$	Root mean square error of approximation based on M_2 statistic
SA	Signed area
SEM	Structural equation modeling
SR	Structured reflection factor
SLO	Student learning outcome
SRMR	Standardized root mean square residual
TLI	Tucker-Lewis index
UA	Unsigned area
UIRT	Unidimensional item response theory
VALUE	Valid Assessment of Learning in Undergraduate Education
WLSMV	Weighted least square with adjusted mean and variance estimator

List of Symbols

Sym.	Full Description
A	Augmented model
a	Discrimination parameter
$\boldsymbol{a}$	Multidimensional discrimination parameter
b	Difficulty parameter
$\boldsymbol{b}$	Multidimensional difficulty parameter
C	Compact model
D	Normal ogive scaling constant
d	Multidimensional difficulty parameter
F	Focal group
G	Group
G^2	Likelihood ratio statistic
h	Number of intervals in dimensional set
i	Item set
J	Endorsed category set
j	Endorsed category
M	Sample mean
M_2	Limited-information fit statistic for dichotomous variables
M_2*	Limited-information fit statistic for polytomous variables
m	Number of latent abilities
N	Number of non-endorsed responses
N*	Number of endorsed responses
n	Multidimensional latent trait set
P	Probability of endorsed response category
P*	Probability of endorsed response category with condition
P*	Probability of endorsed response category with condition
p	Persons
R	Reference group
SD	Standard deviation
u	Category weights
v	Number of items in MIMIC model
Y	Dichotomous response
y	Discrete response
y*	Continuous latent response underlying discrete response
z	Linear exponent in logistic regression
α	Type I error
β	Type II error
β_i	Uniform DIF
γ	Latent mean
ε	Residual error of manifest variable
ζ	Disturbance of latent trait
θ	Continuous latent trait
$\boldsymbol{\theta}$	Multidimensional latent trait
λ	Factor loading
$\boldsymbol{\lambda}$	Multidimensional factor loading
μ	Arithmetic mean
π_i	Regression parameter
ρ	Correlation
τ	Threshold parameter
$\boldsymbol{\tau}$	Multidimensional threshold parameter
ω	Nonuniform effect
$\boldsymbol{\omega}$	Multidimensional nonuniform effect
ω^2	Effect size – omega squared

Chapter I

Literature Review

Differential item functioning (DIF) is a psychometric method that examines variability in measured responses between two or more groups (Holland & Wainer, 1993; Wu et al., 2016). The primary purpose of DIF is to identify the degree to which responses to an item statistically significantly differ as a result of a respondent's group membership (e.g., gender, race, marital status) and to which the item is unconnected to the instrument's core latent construct (Osterlind & Everton, 2009). The presence of DIF requires a review of the items in question to judge the adequacy of the item as a reasonable measure of the underlying latent construct. As a result, DIF assists in mitigating bias and ensuring respondents are measured on the same latent construct within an instrument; thus, DIF provides supportive evidence for construct validity (Walker, 2011).

There exists a notable distinction between DIF and common group-difference statistical tests (e.g., *t*-test, analysis of variance) that should be addressed. Inferential statistics, like independent-samples *t*-test, focus on the *impact* of an item, where group differences are examined without consideration of matching persons on the same latent attribute. Measurement bias (i.e., DIF) measures group differences with respect to the underlying latent ability of the persons being measured (Camilli, 2006; Holland & Wainer, 1993; Rogers & Swaminathan, 2016; Wu et al., 2016). That is, group differences are only comparable between individuals who share the same latent ability level. DIF examines the existence of measurement bias in instruments between individuals in the focal group, colloquially known as the group of interest, and individuals in the reference group, which refers to the group that is hypothesized to be advantaged. More

specifically, DIF is present if and only if individuals with the same ability from different groups have a statistically significantly different probability of endorsing a response within some item (Hambleton & Swaminathan, 1985).

Conversely, the absence of DIF, measurement invariance, suggests that the conditional probability distribution of an item is independent of group membership (Osterlind & Everton, 2009); that is, the likelihood of persons under the same ability endorsing a response is the same, regardless of which group they belong. Therefore, it would follow that DIF would be present for persons with corresponding ability levels, if measurement invariance failed to be established. Consider some construct θ which measures the ability of some persons p, then the null hypothesis of DIF is denoted as:

$$H_0: f(Y = j|\theta_p, G = R) = f(Y = j|\theta_p, G = F) \quad (1)$$

such that $j \in \{0, \ldots, J-1\}$ is some endorsed response in a set of indicators Y, and R is the reference group and F is the focal group in a set of groups G. Thus, rejection of the null hypothesis would indicate the presence of DIF. For example, a dichotomous item, where $j = 1$ is a correct or endorsed response and $j = 0$ is an incorrect or non-endorsed response, will display DIF if, after controlling for the ability θ, the probability distribution of Y is unequal between groups $G \in \{R, F\}$. In other words, DIF is present when the probability of a correct or endorsed response $j = 1$ for the reference group R is different than the focal group F and the ability θ is matched between individuals in the groups.

Background of the Problem

The psychometric community first examined DIF in response to public concern over unfair testing of minority students (Angoff, 1993). The rationale behind DIF is simple, “if the content

of some test items were unfamiliar to minority examinees, then these examinees would be more likely to answer incorrectly despite there being no true difference in their overall cognitive ability compared to majority examinees" (Walker, 2011, pp. 364-365). A similar argument can be extended to respondents of a survey or raters evaluating students with a rubric. The description of some items may be interpreted differently among some groups; respondents from one group may respond differently from other group despite there being no difference between two groups who hold the same latent trait(s). Thus, DIF provides valuable insight into the construct validity of a measurement instrument (Walker, 2011).

Several statistical methods have been adapted to examine DIF across many variants of item response theory (IRT)[1] data. For instance, researchers have developed and examined the likelihood ratio test, logistic (and probit) regression, Mantel-Haenszel method, SIBTEST, and structural equation models as practicable methods to detect and measure DIF in dichotomous IRT data (Holland & Thayer, 1988; Shealy & Stout, 1993; Swaminathan & Rogers, 1990; Thissen et al., 1993; Woods & Grimm, 2011). Most of these methods (e.g., likelihood ratio test, Liu-Agresti method, logistic discrimination function analysis, SIBTEST, structural equation models) were extended to detect DIF from polytomous[2] IRT data (Chang et al., 1996; Miller & Spray, 1993; Penfield & Algina, 2003; Wang & Shih, 2010) and multidimensional[3] IRT data (Lee et al., 2017; Mazor et al., 1998; Suh & Cho, 2014; Walker & Sahin, 2016). In general, researchers choose the optimal DIF detection method based on numerous factors that affect the likelihood of type I and II error occurring (e.g. sample size, the type of DIF, the proportion of

[1] Item response theory (IRT) is a statistical model that represents the probability of a response to an item Y as a function of the latent ability of a person θ and one or more item-level parameters.
[2] Polytomous item is an item that contains more than two possible response options.
[3] Multidimensionality refers to a characterization of the relationship between a person and measured items on multiple latent traits $\boldsymbol{\theta}$.

items with DIF in the instrument, the correlation between latent traits in multidimensional instruments). More precisely, DIF detection methods are selected to limit the rate of type I and II error.

Numerous studies have examined optimal conditions for DIF detection in unidimensional IRT, dichotomous data; unidimensional IRT, polytomous data; and multidimensional IRT, dichotomous data (Bulut & Suh, 2017; Cohen et al., 1996; Elosua & Wells, 2013; Finch, 2005; Kim & Cohen, 1998; Kristjansson et al., 2005; Lee et al., 2017; Su & Wang, 2005), but no study has examined optimal DIF detection methods for multidimensional IRT, polytomous data. Many instruments produce multidimensional IRT, polytomous data, including surveys, evaluation forms, and rubrics, that are used across an array of professional fields, such as medicine, education, psychology, and other social sciences (De Ayala, 1994; Depaoli et al., 2018; Immekus & Imbrie, 2008; Immekus et al., 2019; Jin & Chen, 2020; Walker, 2011). In fact, the multidimensional graded response model is one of the most widely produced multidimensional IRT, polytomous models (Reckase, 2009). Therefore, a thorough examination of the conditions that affect DIF detection helps researchers, who wish to ensure construct validity in instruments that produce data fitted to a multidimensional graded response model, select the most appropriate DIF detection method.

DIF under the 2-PL IRT Model

Item response theory sets a theoretical framework for which many DIF statistical procedures can be interpreted and compared (Penfield & Camilli, 2007). IRT models represent the probability of a response to an item Y as a function of the ability of a person θ and one or more item-level parameters (DeMars, 2010). One of the most common forms of IRT for dichotomous items is the two-parameter model (2-PL IRT). The discrimination and difficulty parameters

characterize the item-level parameters under the 2-PL IRT model. The discrimination parameter, denoted as *a*, describes how well items differentiate from one another on a latent trait (Furr, 2018). The difficulty parameter, represented as *b*, is the trait level required for participants to have a 50% chance of endorsing a response (Furr, 2018). The latent trait θ is used to express a person's underlying ability. The 2-PL IRT model can be expressed as a sigmoid function such that:

$$P(Y = 1|\theta) = \frac{1}{1 + e^{-Da(\theta - b)}} \quad (2)$$

where $P(Y = 1|\theta)$ is the probability of correct response given the latent trait and *D* represents a scaling constant equal to 1.7 that transforms the normal ogive model to the logistic model.

DIF is determined through the conditional between-group differences in the probability of an endorsed response. One method to determine DIF under the 2-PL IRT framework is to evaluate the between-group differences through the item response function (IRF) of the focal and reference groups (Penfield & Camilli, 2007). A second method is to examine the differences in item-level parameters; recall that the 2-PL IRT model is a function of two conditional item-level parameters, discrimination and difficulty (De Boeck & Wilson, 2004). Under these methods, the 2-PL IRT framework can infer DIF through graphical (i.e., IRF) and arithmetic (i.e., item-level analysis) interpretations. These interpretations have been used in the development of DIF statistics using 2-PL IRT methods for evaluating two types of DIF, uniform and nonuniform DIF (Penfield & Camilli, 2007). Both forms of DIF can occur depending upon the statistical and parameter conditions of the IRT model.

Uniform DIF. Uniform DIF occurs when the probability of endorsing a response remains constant between the focal and reference groups across the latent ability continuum (Osterlind &

Everton, 2009; Walker et al., 2001). Moreover, the probability of an endorsed response is statistically different between the two groups and retains the same degree of difference across all latent abilities. Uniform DIF is concluded if the discrimination parameters are equal between the reference group *R* and focal group *F*, but the difficulty parameters differ (i.e., $a_R = a_F \wedge b_R \neq b_F$; Walker et al., 2001). Uniform DIF effect size can be determined by finding the area between the functions of the reference and focal groups (Raju, 1988). This effect size model, the signed area (*SA*), can be expressed as:

$$SA = b_F - b_R \tag{3}$$

where the discrimination parameter *a* is identical between the focal and reference groups and b_R and b_F differ (Raju, 1988); therefore, DIF is present when $SA \neq 0$.

Nonuniform DIF. Nonuniform DIF occurs when an interaction between latent ability level and group membership exists; that is, the probability of endorsing a response changes between the focal and reference group at some point along the latent ability continuum (Osterlind & Everton, 2009; Walker et al., 2001). More notably, nonuniform DIF is determined if the discrimination parameters are unequal between the reference group *R* and focal group *F*. The relationship between the difficulty parameters is less of a concern, such that the parameter can either differ or be the same between groups (i.e., $a_R \neq a_F \wedge [b_R \neq b_F \vee b_R = b_F]$; Osterlind & Everton, 2009). Nonuniform DIF effect size can be estimated in a similar fashion to the uniform DIF effect size; however, the formula grows in complexity because the functions intersect at some point along the latent continuum (Raju, 1988). The nonuniform DIF effect size model, the unsigned area (*UA*), can be expressed as:

$$UA = \left| \frac{2(a_F - a_R)}{Da_F a_R} \ln\left(1 + e^{\frac{Da_F a_R(b_F - b_R)}{a_F - a_R}}\right) - (b_F - b_R) \right| \quad (4)$$

where *a* is the discrimination parameter, *b* is the difficulty parameter, and *D* is the scaling constant equal to 1.7 that conforms to the logistic model. The function *UA* in equation 4 can be simplified to equation 3 when the discrimination parameter *a* is equal between the focal and reference groups (i.e., when $a_F = a_R$, $UA = |SA|$); thus, arithmetically speaking, the function uniform DIF is nested in the nonuniform DIF function and both effect sizes can be determined from equation 4 (Raju, 1988).

DIF Procedures under the 2-PL IRT Framework

A 2-PL IRT model for dichotomous items, which contains the difficulty and discrimination parameters, can function as an ideal framework when comparing different DIF detection procedures, because of its ability to help identify uniform and nonuniform DIF. DIF can be detected through several commonly used procedures for dichotomous items, such as the multiple indicators multiple causes interaction (MIMIC-interaction) structural equation model (SEM), the IRT likelihood ratio (IRT-LR) test, and logistic regression model. Each procedure can be transformed into a IRT model (e.g. 2-PL IRT) with respect to the model parameters, discrimination and difficulty (De Boeck & Wilson, 2004; Penfield & Camilli, 2007; Swaminathan & Rogers, 1990; Woods, 2009a); thus, each procedure is introduced within the 2-PL IRT framework under a continuous Gaussian probability distribution for the latent trait, $\theta \sim N(0, 1)$.

MIMIC models. MIMIC models are special cases within SEM in which latent factor(s) are regressed on covariates (i.e., multiple indicators) and manifest variables (i.e., the multiple causes) are regressed on the latent factor(s). In its simplest form,

$$\boldsymbol{y}_i = \boldsymbol{\lambda}_{ji}\eta_j + \boldsymbol{\varepsilon}_i \tag{5}$$

$$\eta_j = \boldsymbol{\gamma}_{jk}\boldsymbol{x}_k' + \zeta_j \tag{6}$$

where $\boldsymbol{y}_i$ is a vector of endogenous indicators, $\boldsymbol{\lambda}_{ji}$ is the vector of factor loadings for items $\boldsymbol{y}_i$ regressed on η_j, $\boldsymbol{\varepsilon}_i$ is the vector of residuals of items $\boldsymbol{y}_i$, η_j is some latent construct, $\boldsymbol{\gamma}_{jk}$ is the vector of factor loadings of latent construct η_j regressed on $\boldsymbol{x}_k$, $\boldsymbol{x}_k$ (with a vector of residuals of $\boldsymbol{\xi}_k$) is some vector of covariates, and ζ_j is the disturbance of the latent construct η_j (Brown, 2015; Wang & Wang, 2012).

The MIMIC model can be modified to detect both uniform and nonuniform DIF under the IRT model (Finch, 2005). Only one item is examined at a time for DIF under the MIMIC model. The parameters of the item of interest are allowed to vary between the reference and focal groups, while the parameters of the other items in the model are constrained to be equal across groups, so that they serve as invariant anchors (Woods, 2009a; Woods et al., 2009). This process not only allows DIF to be easily detectible, but also allows the DIF effect size to be properly estimated (Woods, 2009a). For dichotomous items, a latent response variable conversion is used for MIMIC DIF models such that $y = 1$, if $y_i^* \geq \tau_i$, or $y = 0$, if $y_i^* < \tau_i$, where y_i^* is the continuous latent response that underlies the y_i discrete response and τ_i is the threshold parameter. Under the assumption that only one latent trait exists, the detection for a uniform DIF item, under the MIMIC model, can be expressed such that:

$$y_i^* = \lambda_i\theta + \beta_i G \ + \varepsilon_i \tag{7}$$

where λ_i is the factor loading of item *i* (synonymous to the discrimination parameter in IRT models; McDonald, 1997; Stark et al., 2006), θ is the continuous response latent construct, β_i is the uniform DIF effect, $G \in \{0, 1\}$ is the discrete observed covariate, and ε_i is the residual of item *i* (Finch, 2005). If group membership significantly predicts the item responses (i.e. $\beta_i \neq 0$), while controlling for group mean differences, then uniform DIF is present. The group mean difference γ between two groups G, as outlined in equation 6, is controlled by fixing the variable to 0 to eliminate any mean differences between the two groups (Finch, 2005).

The model can be expanded to detect both uniform and nonuniform DIF such that:

$$y_i^* = \lambda_i\theta + \beta_i G + \omega_i\theta G + \varepsilon_i \tag{8}$$

where ω_i is the nonuniform DIF effect or interaction between group membership and latent ability (Woods & Grimm, 2011). Figure 1.1 helps to illustrate this relationship across all items *k* in the newly formed, MIMIC-interaction model. The interaction effect, θG, which contains a continuous latent variable and discrete manifest covariate, cannot be simply multiplied because non-normality of the covariate produces a product that is not easily interpretable (Woods & Grimm, 2011). The latent moderated structural equations (LMS) method allows a full-information maximum likelihood estimation to remove the non-normality problem from the interaction (Klein & Moosbrugger, 2000). The LMS method conditions on the latent variable and transforms the covariate to a mixture of continuous conditional distributions (Barendse et al., 2012). Similar to equation 7, if group membership significantly predicts the item responses (i.e., $\omega_i \neq 0$), while controlling for group mean differences, then nonuniform DIF is present; however, if $\omega_i = 0$ and $\beta_i \neq 0$, then uniform DIF is present.

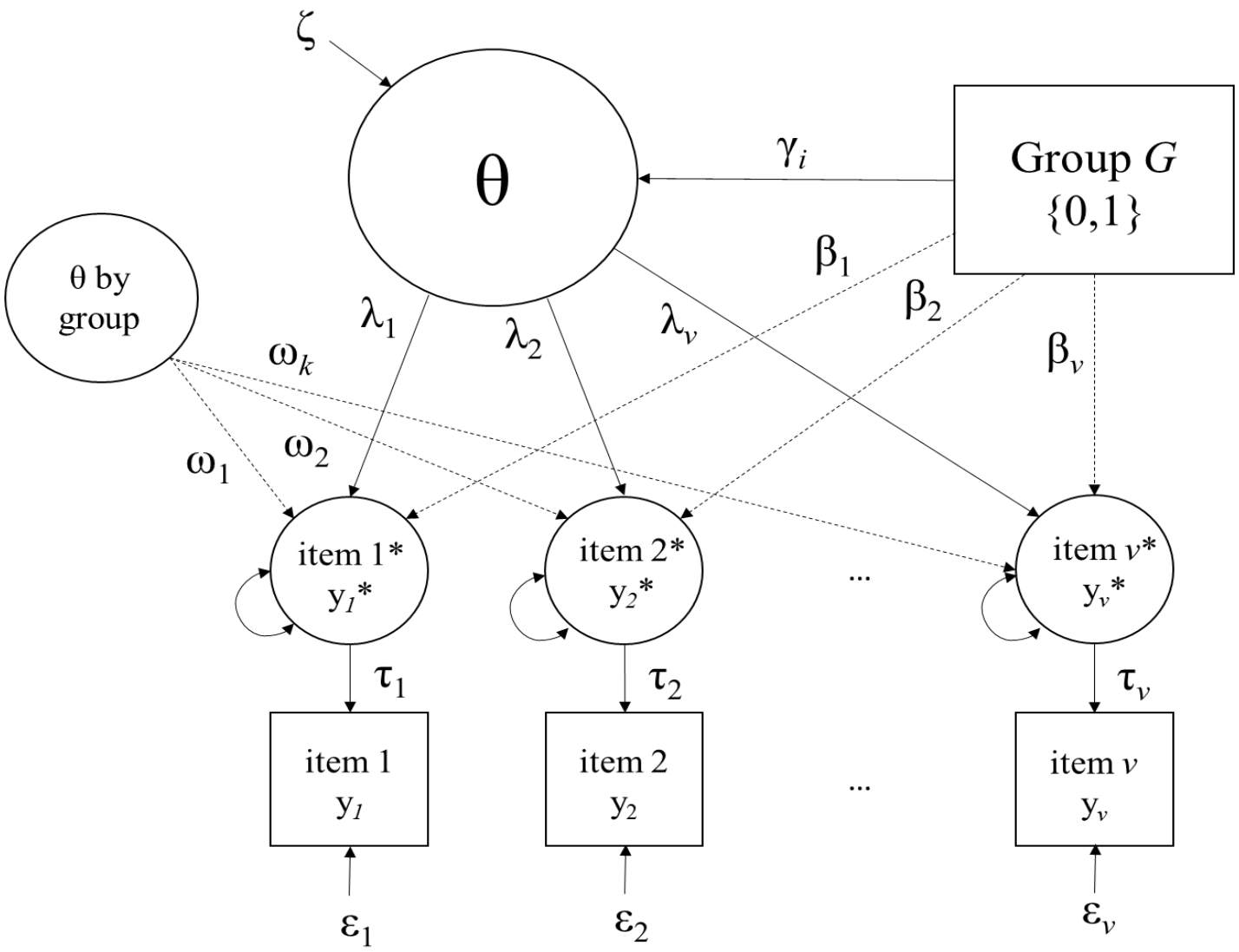

Figure 1.1. *Multiple indicator multiple causes (MIMIC) interaction DIF model*

IRT-LR test. The IRT-LR test is an omnibus test that compares two nested IRT models with varying constraints on the parameters (Thissen et al., 1993). Typically, one item is investigated at a time; this item is referred to as the studied item (Woods, 2009b). The other items serve as anchors in which all parameters are constrained to be equal between groups (i.e., measurement invariant or DIF-free; Woods, 2009b). The compact (*C*) model (i.e., the simpler model) is estimated by constraining the parameters of an item to be equal for the reference and focal group. The augmented (*A*) model (i.e., the most complex model) is estimated by allowing the parameters of an item to vary between the focal and reference groups. The difference between the natural logarithm of the likelihood functions for each model are calculated, such that:

$$G^2 = -2\ln L_C - (-2\ln L_A) \tag{9}$$

where L_C is the deviance of the compact model and L_A is the deviance of the augmented model. The IRT-LR test approximates the chi-square distribution via the statistical deviance (Maydeu-Olivares & Cai, 2006; Thissen et al., 1993). The degrees of freedom are equal to the difference in the number of parameters allowed to vary between the two models (Cohen et al., 1996). The function G^2 provides a DIF effect size; if the test statistic is significant, then additional IRT-LR tests are conducted to determine if the DIF is uniform or nonuniform (Woods, 2009b).

First, nonuniform DIF can be tested by constraining the discrimination parameter of the item, but allowing the difficulty parameter to vary (Bulut & Suh, 2017). The new model becomes the compact model and is compared to the augmented model from the original omnibus test using the IRT-LR test. A significant test statistic concludes that nonuniform DIF exists; however, if the models display invariance, the second compact model from the nonuniform IRT-LR test can be

compared to the original compact model using the IRT-LR test to determine the presents of uniform DIF (Bulut & Suh, 2017). A significant test indicates that uniform DIF is present.

Logistic regression. The logistic regression models the probability of a binary outcome that is predicted by some continuous or categorical independent variable(s) (Darlington & Hayes, 2017). The model, in its simplest form, is mathematically defined as:

$$y = \frac{1}{1 + e^{-z}} \tag{10}$$

where $y \in (0, 1)$ is some real, bounded outcome and $z \in \mathbb{R}$ is a continuous independent variable. The logistic regression can be expressed as a function of the probability of endorsing a response in an item, such that $z = \pi_0 + \pi_1\theta + \pi_2 G + \pi_3(\theta G)$, where θ is the observed ability parameter, π_0 is the intercept, π_1 is the slope, π_2 is the group effect on performance, π_3 is the interaction between the group and ability trait θ, and *G* is the group membership (Swaminathan & Rogers, 1990).

In the context of item analysis, the parameter π_0 depicts the item difficulty and the parameter π_1 denotes the item discrimination (Berger & Tutz, 2016). Uniform DIF exists when the group-specific parameter π_2 differs between the two groups such that $\pi_2 \neq 0$. Nonuniform DIF can be modeled as an extension of uniform DIF when the $\pi_3(\theta G)$ term is added to the expression, where π_3 is the group-specific parameters for nonuniform DIF such that $\pi_{1R} - \pi_{1F}$ (Berger & Tutz, 2016; Swaminathan & Rogers, 1990). Nonuniform is evident when the group-specific parameters for nonuniform DIF π_3 differs between the two groups such that $\pi_3 \neq 0$ regardless of whether $\pi_2 \neq 0$ or $\pi_2 = 0$; however, uniform DIF is present if and only if $\pi_3 = 0$ and $\pi_2 \neq 0$.

The null hypothesis that the two groups do not contain DIF can be tested using a likelihood ratio test (Penfield & Camilli, 2007). Penfield and Camilli (2007) explain that the test is conducted using a sequence of three nested models that differ by the number of free parameters, similar to the IRT-LR test. The extent to which one nested model fits the broader model elucidates the type of DIF that exists. The simplest model contains no DIF such that $z = \pi_0 + \pi_1\theta_p$, the next most complex model exhibits uniform DIF such that $z = \pi_0 + \pi_1\theta_p + \pi_2 G$, and the most complex model shows uniform and nonuniform DIF where $z = \pi_0 + \pi_1\theta_p + \pi_2 G + \pi_3(\theta G)$. A test for the presences of different forms of DIF is illustrated using a likelihood ratio test:

$$\chi^2 = -2\ln(z_{nested}) - (-2\ln(z_{complex})). \tag{11}$$

First, the nonuniform DIF model and the uniform DIF model can be compared as the complex and nested model, respectively. The ensuing statistic is tested on the chi-squared distribution with one degree of freedom. A nonsignificant result indicates that nonuniform is not present where $\pi_3 = 0$. Next, the uniform model DIF model, as the complex model, can be compared to the no DIF model, as the nested model. If the statistic is nonsignificant, then no DIF exists such that $\pi_2 = \pi_3 = 0$.

Polytomous IRT Models

Dichotomously scored items receive widespread attention from the testing industry, but use of numerous item formats scored polytomously, such as Likert scales, rubric scores, and rating scales, have created interest from the educational, medical, and psychological fields (Ostini & Nering, 2006). An item is considered polytomous when the item response options can be scored according to three or more categories, as opposed to a dichotomous item which only has two

possible outcomes (e.g. correct/incorrect). Polytomous items often include multiple score categories typically found in evaluation forms, questionnaires, and other similar instruments. They can also be modeled under the IRT framework much the same way dichotomous items are modeled; however, the IRT model grows in complexity when multiple outcomes are introduced. The purpose of polytomous IRT models is to estimate the probability of endorsing each category respective to the ability level and the item parameters (Nering & Ostini, 2010).

Several polytomous IRT models have been developed to illustrate the behaviors of polytomous data under various parameter and modeling constraints (Andrich, 1978; Bock, 1972; Masters, 1982; Muraki, 1993; Samejima, 1969). Two models, the generalized partial credit and the graded response models, were developed as polytomous extension of the 2-PL IRT model with the sole intention of extending the application of the IRT framework to ordered data (Embertson & Reise, 2000; Muraki, 1993; Samejima, 2010). The generalized partial credit and graded response models differ by the structure of their respective IRFs. The generalized partial credit model (GPCM) estimates the probability of endorsing a category *j* with respect to the latent ability θ by use of an adjacent category logistic model (Muraki, 1993; Penfield & Camilli, 2007). The GCPM compares each category individually with the adjacent categories, both directly above and below the ordered response.

The graded response model (GRM) estimates the probability of a category *j* with respect to the latent ability by use of the cumulative logistic model (Samejima, 2010; Penfield & Camilli, 2007). The GRM estimates several cumulative probability functions by condensing all categories into two groups; one group contains the endorsed category and categories with greater value, while the other group contains non-endorsed categories with less value than the endorsed one (Kang et al., 2009). The GRM and GPCM are both widely accepted models to compute

parameter estimates for polytomous items (Kang et al., 2009; Ostini et al., 2014; Penfield & Camilli, 2007); however, the models do not yield directly comparable parameters (Kang et al., 2009; Nering & Ostini, 2010). The GRM also carries greater flexibility with the aforementioned DIF detection procedures under the 2-PL IRT model due in part to their use of the cumulative logistic model.

The GRM is computed using a two-step process to estimate the probability of an endorsed response category $P_{ij}(\theta)$ for an item i, known as the category response function (Embertson & Reise, 2000; Samejima, 2010). First, the model estimates the probability at or above an endorsed ordered category $j \in \{0, 1, \dots, J-1\}$ with respect to the latent ability θ, which is referred to as the boundary response function (BRF) and denoted as $P_{ij}^{*}(\theta)$. After each BRF is computed, the difference between the adjacent probabilities $P_{ij}^{*}(\theta)$ and $P_{i(j+1)}^{*}(\theta)$ forms the category response function (CRF) of a category j for item i. The probability of endorsing above the lowest category is defined as $P_{i0}^{*}(\theta) = 1$, while the probability of endorsing above the highest category is defined as $P_{iJ}^{*}(\theta) = 0$ (Kang et al., 2009; Ostini et al., 2014). Thus, the GRM can be expressed as the combined probability of three types of conditional cumulative probability functions (i.e., CRFs) such that:

$$P_i(Y = j, \theta) = \begin{cases} 1 - \dfrac{1}{1 + e^{-Da_i(\theta - b_{i(j+1)})}}, & j = 0 \\ \dfrac{1}{1 + e^{-Da_i(\theta - b_{ij})}} - \dfrac{1}{1 + e^{-Da_i(\theta - b_{i(j+1)})}}, & 0 < j < J - 1 \\ \dfrac{1}{1 + e^{-Da_i(\theta - b_{ij})}}, & j = J - 1 \end{cases} \tag{12}$$

(Samejima, 2010).

The parameters for the GRM are analogous to the 2-PL IRT model with some subtle differences. For instance, the b_j parameter, referred to as the threshold parameter, represents the ability level needed for a 50% chance of endorsing at or above a category *j* (Samejima, 2010). The *D* parameter, similar to the 2-PL IRT model, refers to a scaling constant equal to 1.7 that conforms to the logistic distribution. The *a* parameter still denotes the discrimination of the item.

The GRM possesses many important characteristics worthy of mention: (1) the CRFs are ordered from smallest to largest, corresponding to the threshold parameters (i.e., $b_1 < \cdots < b_{J-1}$); (2) the threshold parameters indicate the location of the conditional cumulative probability functions for each category; (3) the discrimination parameter is the same across all categories of the item, but free to vary across different items; (4) the sum of the probability of all categories in an item at any ability level θ is equal to 1 (Nering & Ostini, 2010). Figure 1.2 illustrates several of these important characteristics graphically.

DIF under the GRM

For polytomous IRT models, an item is considered DIF when corresponding CRFs in the reference and focal groups are unequal (Nering & Ostini, 2010). Under the GRM, the discrimination parameter a_i remains constant while a set of threshold parameters $\boldsymbol{b_i} = \{b_{i1}, b_{i2}, \ldots, b_{i(J-1)}\}$ varies between BRFs under some item *i*. Since the GRM is just a polytomous extension of the 2-PL IRT model, then it follows that an item exhibits uniform DIF when $a_{iF} = a_{iR}$ and $\boldsymbol{b}_{iF} \neq \boldsymbol{b}_{iR}$. To that extent, nonuniform DIF is present when $a_{iF} \neq a_{iR}$ and $\boldsymbol{b}_{iF} = \boldsymbol{b}_{iR}$, or $a_{iF} \neq a_{iR}$ and $\boldsymbol{b}_{iF} \neq \boldsymbol{b}_{iR}$under the GRM.

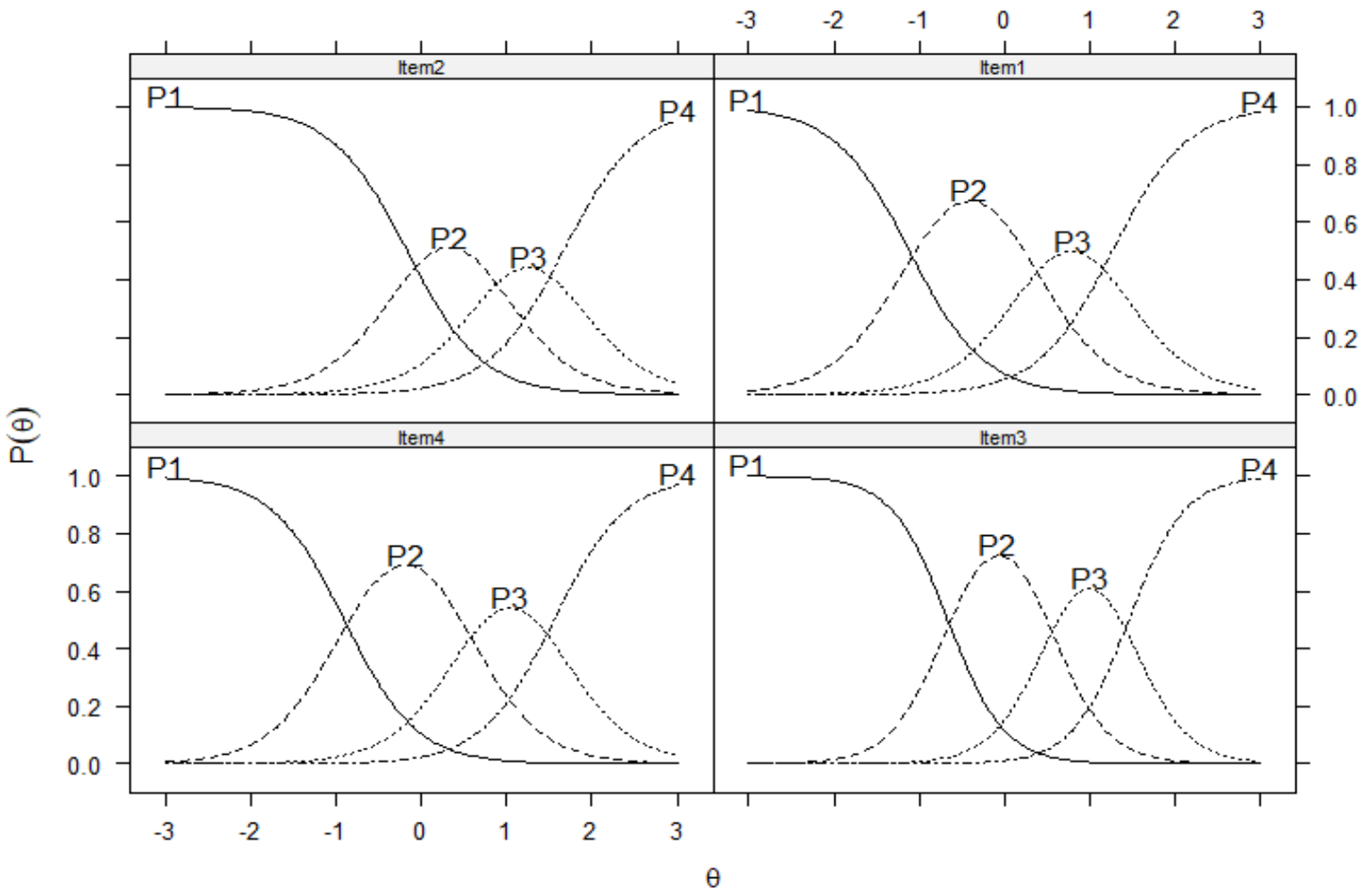

Figure 1.2. *Category response functions for four graded response model items*

Note. P1 = probability of endorsing category 1; P2 = probability of endorsing category 2; P3 = probability of endorsing category 3; P4 = probability of endorsing category 4. a_{item1} = 2.28; a_{item2} = 2.62; a_{item3} = 3.06; a_{item4} = 2.36

Uniform and nonuniform DIF effect size for an item under the GRM can be estimated in a similar fashion to the uniform and nonuniform DIF effect size under the 2-PL IRT model, respectively (Cohen et al., 1993). Analogous to equation 3, uniform DIF under the GRM, *SA*, is expressed as the sum of the difference between the reference and focal group thresholds for all categories *j*, where DIF is present when $SA \neq 0$ (Golia, 2012). Similar to *UA* under the 2-PL IRT framework, nonuniform DIF is detected when each corresponding CRF between the focal and reference group intersects along the latent continuum (Penfield, 2010b). The absolute difference between the corresponding CRFs will mitigate the effects of interaction by examining the Euclidean distance between the CRFs, such that:

$$UA = \sum_{j=1}^{J-1} (u_{i(j+1)} - u_{ij}) \left| \frac{2(a_{iF} - a_{iR})}{D a_{iF} a_{iR}} \ln\left(1 + e^{\frac{D a_{iF} a_{iR}(b_{ijF} - b_{ijR})}{a_{iF} - a_{iR}}}\right) - (b_{ijF} - b_{ijR}) \right|. \quad (13)$$

The category weights u_{ij} are typically fixed to be the same as the category values, where each successive categorical value is the next consecutive natural number (i.e., $u_{i1} = j = 1, u_{i2} = j = 2, \ldots, u_{iJ} = J$; a_{iG} is the discrimination parameter with respects to the item *i* and group $G \in \{R, F\}$; and b_{ijG} is the threshold parameter with respects to the item *i*, category *j*, and group $G \in \{R, F\}$. Comparable to the 2-PL IRT model, the function *UA* in equation 13 can be simplified to *SA* when the discriminate parameter *a* is the same for the focal and reference groups; thus, the uniform DIF function is said to be nested in the nonuniform DIF function (Cohen et al., 1993).

DIF Procedures under the GRM

The DIF detection procedures use a similar underlying framework with polytomous items as they do with dichotomous items, because the GRM is just a polytomous extension of the 2-PL IRT model. The MIMIC-interaction model, used to examine the existence of DIF and estimate

DIF effect size in the dichotomous items, stays the same under the GRM framework with the exception that the continuous latent response variable underlying the discrete manifest response variable contains multiple threshold parameters (i.e., one less than the number of categorical response options) instead of one threshold (Wang & Shih, 2010). The IRT-LR test changes only by the use of the GRM, in place of the 2-PL IRT model, to estimate the augmented and compact models (Kim & Cohen, 1998). The logistic regression test is only suitable for dichotomous items; therefore, polytomous extensions of these procedures are needed to determine the existence and effect size of DIF.

Logistic discriminant function analysis. The logistic regression analysis can be used to identify DIF in polytomous models through one of three coding schemes: the continuous ratio scheme, the cumulative scheme, and the adjacent category scheme (Agresti, 2013). The continuous ratio scheme dichotomizes polytomous scales into $Y = j$ and $Y > j$; the cumulative scheme dichotomizes polytomous scales into $Y \leq j$ and $Y > j$; the adjacent category scheme dichotomizes polytomous scales into $Y = j$ and $Y = j + 1$, where $j = 0, 1, \ldots, J - 1$ for all three schemes (Penfield & Camilli, 2007). French and Miller (1996) indicated that the three schemes adequately detect nonuniform and uniform DIF for data fitted to the GRM, but voluminous data manipulation and difficulty in interpretation of results made the process insufficient and cumbersome.

Alternatively, the variables within the logistic regression model, as described in equation 10, can be rearranged to detect DIF such that the grouping variable *G* is the dependent variable and the item response category *Y* is an independent variable (Miller & Spray, 1993). As a result, this approach, known as the logistic discriminant function analysis, describes the probability of belonging to a group $G \in (R, F)$ as a function of the item response category and the ability

parameter. A likelihood ratio test is then used to assess the presence of uniform and nonuniform DIF, similar to the logistic regression approach for dichotomous items. The logistic discriminant function analysis is advantageous in that it allows for detection of DIF without aggregating the categorical variable into multiple binary variables (i.e., dummy coding; Penfield & Camilli, 2007).

Dimensionality

A unidimensional IRT (UIRT) model characterizes the relationship between a person and measured items on a single latent trait (Reckase, 2009). That is, a set of items in a UIRT model only measures a single unobservable characteristic, such as basic mathematics skills or patient-reported experience at a medical facility. UIRT models vary by the number of observable parameters that explain the characteristics of the measured item (i.e., difficulty, discrimination, guessing parameters), but still contain a single latent trait of a person (Reckase, 2009). For instance, equation 2 represents the 2-PL UIRT model where latent ability parameter θ serves as the only unobservable measured parameter and the difficulty and discrimination parameters explain the characteristics of the measured item.

The mathematical simplicity and overall ease of interpretation makes UIRT models advantageous and useful. Yet, the likelihood of unidimensional interactions between a person and item is less common than non-unidimensional interactions (Ackerman, 1994; Embertson & Reise, 2000; Reckase, 2009). That is, items are more likely to measure more than one latent trait (Embertson & Reise, 2000). For example, a word problem may measure a student's ability to perform an arithmetic operation, but also may measure the student's reading comprehension, or ability to interpret a graph. A survey item, such as "I have people in my life that make me feel confident", may measure a respondent's social capital and self-esteem and are common

examples of non-unidimensional models. Hence, more complex IRT models need to be constructed to better articulate the multiple traits that associate with items from a measurement instrument.

Multidimensional IRT. Multidimensional IRT (MIRT) models allow for the ability to measure and interpret the effects of multiple traits present within a measurement instrument (Embertson & Reise, 2000; Hartig & Höhler, 2009; Reckase, 2009; Wu et al., 2016). MIRT models are extensions of both factor analysis and UIRT models, while the interpretation of the results are more akin to the UIRT models (Embertson & Reise, 2000; Reckase, 2009). In factor analysis, models that carry multidimensional traits are separated into two forms of dimensionality, between-item multidimensionality and within-item multidimensionality. Between-item multidimensionality is characterized as items that are only correlated to one of the multiple dimensional abilities in the model (Wu et al., 2016). For example, a set of items load on mathematical ability, while another set of items are highly correlated with reading ability. Within-item multidimensionality is present when items are loaded onto multiple traits, such as a critical thinking item highly correlating with both mathematical and reading abilities (Wu et al., 2016). MIRT models distinguish within-item multidimensionality and between-in multidimensionality in a similar manner (Ackerman, 1994).

Within-item multidimensional MIRT models are further divided into compensatory and non-compensatory models (Reckase, 2009). Respondents with low ability in one dimension can be compensated by higher abilities on the other dimensions in a compensatory model (Hartig & Höhler, 2009). Compensatory models contain disjointed abilities and produce a linear relationship between abilities (Hartig & Höhler, 2009). These models remain the most popular and simplest form of MIRT models (Bolt & Lall, 2003; Reckase, 2009). Conversely, respondents

must demonstrate high ability in all dimensions to indicate an endorsement of an item in a non-compensatory model (Hartig & Höhler, 2009; Wang & Nydick, 2015). Non-compensatory models are common among items where multiple abilities rely on one another to endorse or answer an item correctly, such as an examinee's ability to comprehend a reading passage and then use mathematical abilities to solve a word problem (Wang & Nydick, 2015).

Multidimensional GRM. The multidimensional graded response model (MGRM) is a generalization and extension of the unidimensional graded response model as described in equation 12 (Reckase, 2009). The MGRM examines the relationship between the multiple latent abilities of a person and the probability of endorsing responses from two or more score categories (Muraki & Carlson, 1995). The MGRM is computed in a comparable two-step approach as the unidimensional GRM. First, the model estimates the probability $P_{ij}^{*}(\boldsymbol{\theta})$ for an item *i* at or above each endorsed ordered category $j \in \{0, 1, \ldots, J-1\}$ with respect to the linear combination of latent abilities $\boldsymbol{\theta}$. Next, the probability $P_{ij}(\boldsymbol{\theta})$ of a category *j* for item *i* is estimated by calculating the difference between adjacent probabilities $P_{ij}^{*}(\boldsymbol{\theta})$ and $P_{i(j+1)}^{*}(\boldsymbol{\theta})$. In similar fashion to the unidimensional GRM, the probability of endorsing above the lowest category is defined as $P_{i0}^{*}(\boldsymbol{\theta}) = 1$, while the probability of endorsing above the highest category is defined as $P_{iJ}^{*}(\boldsymbol{\theta}) = 0$ (Reckase, 2009).

The probability $P_{ij}^{*}(\boldsymbol{\theta})$ is modeled by a 2-PL IRT model where the linear combination of the latent traits in the vector $\boldsymbol{\theta}$ are weighted by corresponding discrimination parameters, expressed as the vector $\boldsymbol{a}$ (Reckase, 2009; Samejima, 2010). The threshold parameter, or *d*, is the product of the discrimination parameter *a* and the difficulty parameter *b* such that $d = -ab$ and has an inverse relationship with the item categories (Reckase, 2009). Thus, the CRFs are ordered from

largest to smallest corresponding to the threshold parameters (i.e., $d_{J-1} < \cdots < d_1$). The MGRM is expressed as the combined probability of three types of conditional cumulative probability functions (i.e., CRFs) for each category *j* such that:

$$P_i(Y = j, \boldsymbol{\theta}) = \begin{cases} 1 - \dfrac{1}{1 + e^{-D\boldsymbol{a}_i\boldsymbol{\theta}' + d_{i(j+1)}}}, & j = 0 \\ \dfrac{1}{1 + e^{-D\boldsymbol{a}_i\boldsymbol{\theta}' + d_{ij}}} - \dfrac{1}{1 + e^{-D\boldsymbol{a}_i\boldsymbol{\theta}' + d_{i(j+1)}}}, & 0 < j < J - 1 \\ \dfrac{1}{1 + e^{-D\boldsymbol{a}_i\boldsymbol{\theta}' + d_{ij}}}, & j = J - 1 \end{cases} \quad (14)$$

where *D* is the scaling constant that conforms the model to the logistic distribution. Figure 1.3 graphically illustrates a two-dimensional GRM model with four categories to help visualize the complexity of the model.

As evident in equation 14, the model assumes a linear relationship between the latent abilities rather than a product of unidimensional GRMs; hence, the MGRM is based on the compensatory model (Reckase, 2009), which, in turn, focuses on the presence of within-item multidimensionality in the model. The MGRM also assumes that a correlational effect exists between latent abilities, which is not expected in the unidimensional GRM (Jiang et al., 2016). Thus, undesirable effects are produced when a unidimensional GRM is used to estimate multidimensional data. Vital parameter information is lost when a univariate IRT model is applied to instruments that have an underlying multivariate structure, producing results that are less precise than if an appropriate multivariate model had been used (Jiang et al., 2016). The effect of estimating multidimensional data in the unidimensional GRM yields improperly estimated discrimination parameters and model misfit (De Ayala, 1994). Thus, the use of the unidimensional GRM needs to be reconsidered when working with multidimensional polytomous data in exchange for the MGRM.

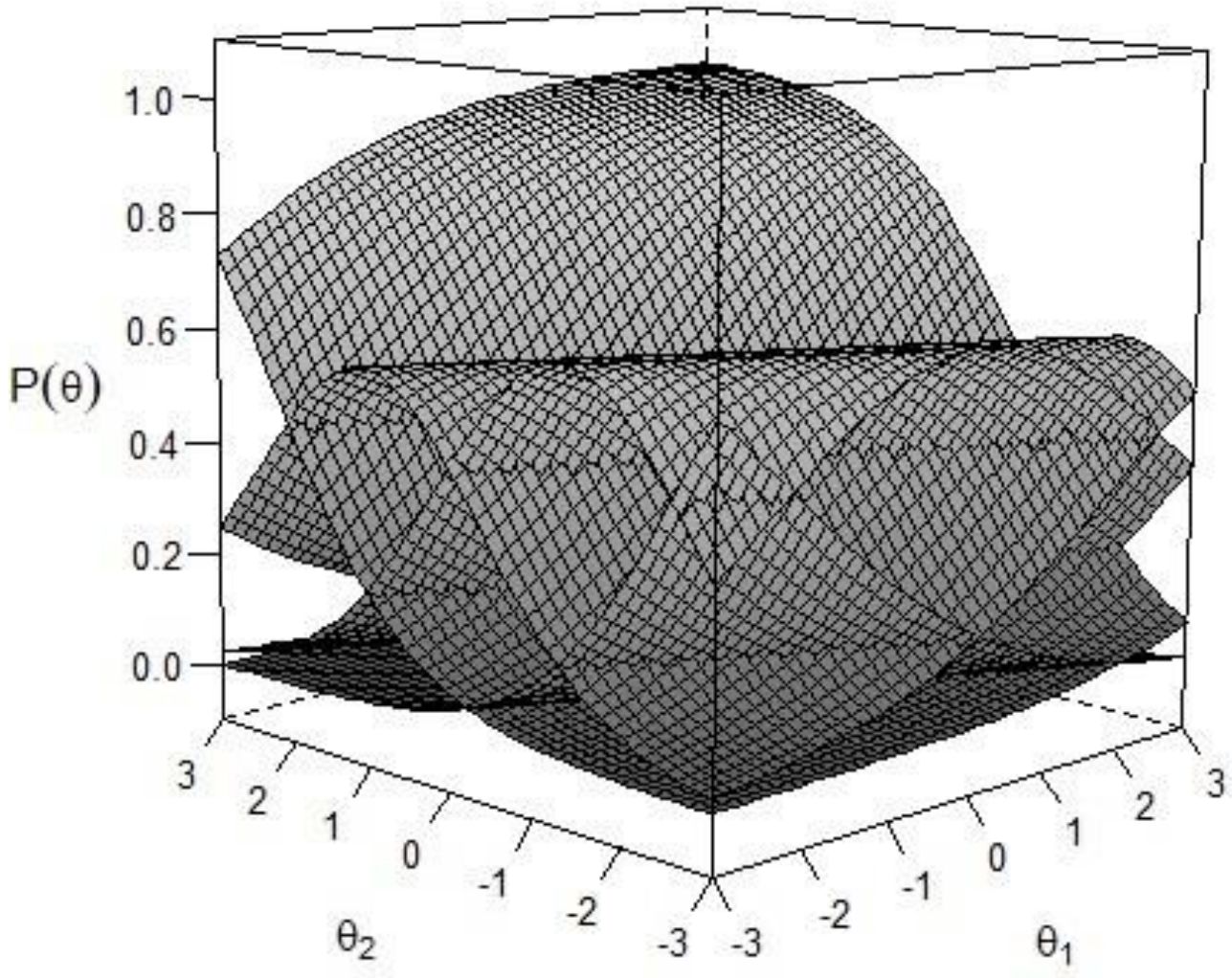

Figure 1.3. *Category response functions for a multidimensional graded response model item*

DIF Procedures under Multidimensional IRT Models

Multidimensionality in a set of items has a strong influence on DIF (Embertson & Reise, 2000). That is, items assumed to underlie a unidimensional model are likely to exhibit measurement bias because subsets of items in the model can be distinguished from one another by other undetected dimensions (Oshima & Miller, 1992). Undetectable dimensions can occur when items are highly correlated with one another, but subsets of items are subtly different enough to warrant a multidimensional model. Additionally, items can also be highly correlated with multiple latent abilities; therefore, group mean differences can lead to identification of DIF of an item, but only because other intervening latent abilities are influencing the detection of DIF (Embertson & Reise, 2000). Procedures for detection of MIRT models needs to be carefully considered to mitigate the misidentification of DIF in a model. Fortunately, multidimensional extensions of the MIMIC model, IRT-LR test, and logistic regression have been developed to detect DIF in multidimensional datasets.

Multidimensional MIMIC-interaction model. Lee et al. (2017) developed a multidimensional MIMIC-interaction model to address multiple latent abilities that may exist. The multidimensional extension follows many of the same conventions of its unidimensional counterpart; in fact, the multidimensional extension can be viewed as a generalization of the MIMIC-interaction DIF model. Thus, under the assumption that *m* number of latent abilities occur, the MIMIC-interaction model in equation 8 can be extended to:

$$y_i^* = \boldsymbol{\lambda}_i \boldsymbol{\theta} + \beta_i G + \boldsymbol{\omega}_i \boldsymbol{\theta} G + \varepsilon_i \tag{15}$$

where $\boldsymbol{\lambda}_i$ is a *m* x *m* symmetric diagonal matrix of factor loadings for item *i* on a *m* x 1 vector of latent abilities $\boldsymbol{\theta}$, such that $\boldsymbol{\lambda}_i = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{bmatrix}$, β_i is the uniform DIF effect of the grouping variable *G* on item *i* (when $\beta_i \neq 0$), $\boldsymbol{\omega}_i$ is a *m* x *m* symmetric diagonal matrix of the nonuniform DIF effect on item *i*, such that $\boldsymbol{\omega}_i = \begin{bmatrix} \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_m \end{bmatrix}$ and $\boldsymbol{\omega}_i \neq 0$, based on the interaction between the grouping variable and the latent abilities, and ε_i is the error term. Figure 1.4 provides a diagram to help illustrate the relationship between all variables across all items *i* in the model. The latent variables $\boldsymbol{\theta}$ are assumed to have a multivariate normal distribution with means of 0 and variances of 1 to establish model identification in the multidimensional MIMIC-interaction model; however the correlations between latent abilities ρ are permitted to vary freely (Lee et al., 2017). A full-information maximum likelihood estimation, using the LMS method, estimates the multidimensional MIMIC-interaction DIF model and addresses the interaction effects between the continuous latent variables $\boldsymbol{\theta}$ and the discrete covariate *G* (Barendse et al., 2012; Klein & Moosbrugger, 2000).

Multidimensional extension of the IRT-LR test. Similar to the IRT-LR test for unidimensional IRT models, the IRT-LR test for MIRT models compares nested models to evaluate the model fit as a means to determine DIF (Suh & Cho, 2014); however, the IRT-LR test for MIRT models warrants additional assumptions. First, Suh and Cho (2014) recommend a robust weighted least square with adjusted mean and variance (WLSMV) estimator to effectively and accurately assess multivariate datasets. Next, the means and variances of the latent abilities are constrained to 0 and 1s, respectively, in both groups to guarantee metric indeterminacy across multiple latent abilities. The correlation between the latent abilities is constrained to 0 in the

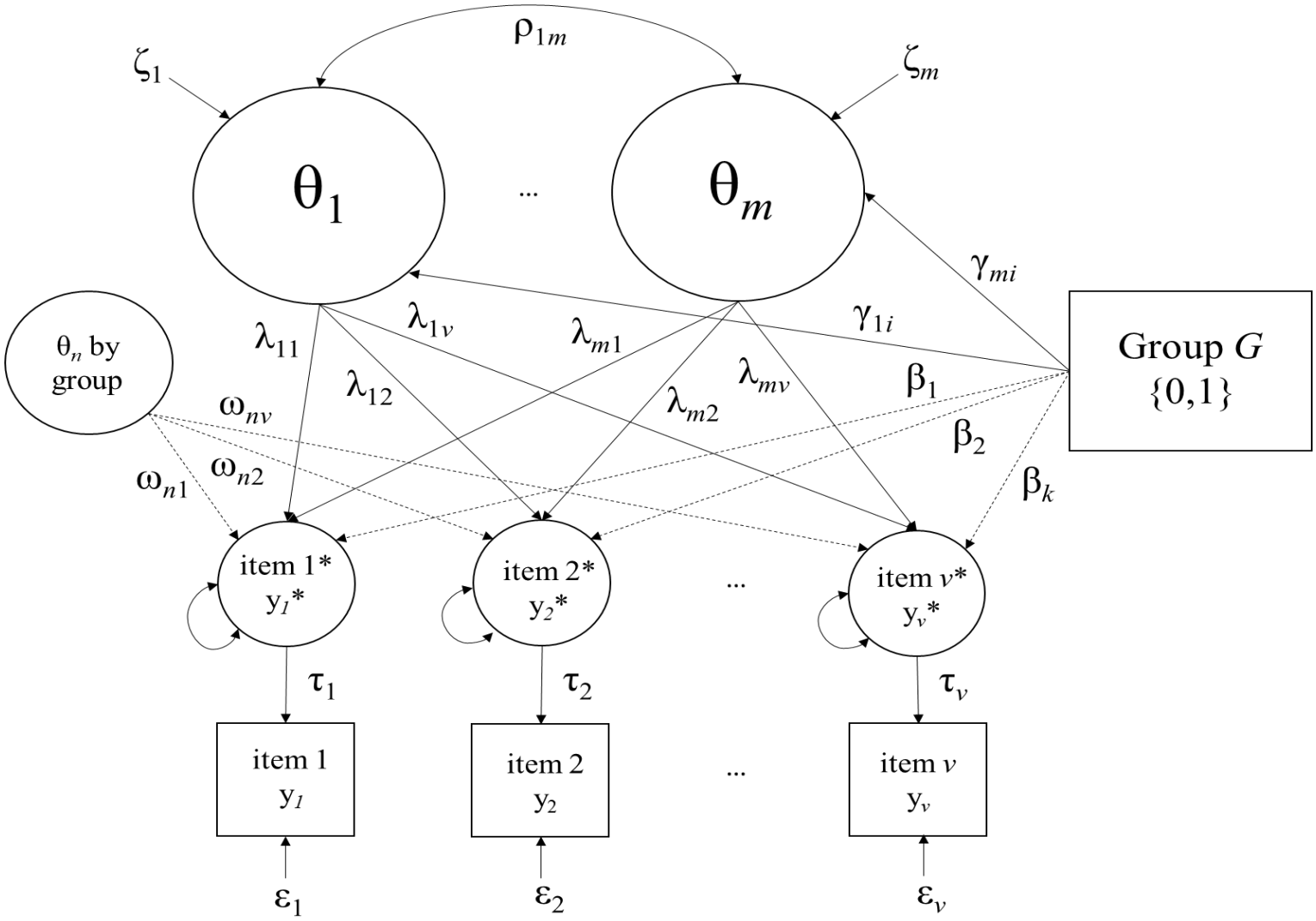

Figure 1.4. *Multiple indicator multiple causes (MIMIC) interaction DIF model for multidimensional data*

augmented model with the other item parameters estimated freely across groups. Additionally, the discrimination parameter for the studied item is constrained to 0 on at least one of the latent trait for both groups to satisfy the model identification, provided the model has a non-simple structure.

A modified IRT-LR test from the Thessian et al. (1993) method is used to accommodate the multivariate data. A flowchart, adapted from Suh and Cho (2014), is displayed in Figure 1.5 to help visualize the process under a two-dimensional model. The first test, Test 1, is an omnibus test for concurrently identifying DIF in the $\boldsymbol{a}$ and d parameters. Test 1 compares a compact model, where all the studied item parameters are constrained to be equal between groups, and the augmented model, where all the item parameters for the studied item are free to vary across groups. If the test statistic is nonsignificant, then DIF is not present; however, if Test 1 is significant, then DIF can be tested for two discrimination parameters (i.e., Test 2).

In Test 2, the compact model constrains only the studied item discrimination parameters $\boldsymbol{a}$ to be equal across groups. The augmented model allows all item discrimination parameters to freely vary between groups. In both models, the d parameter is permitted to vary freely between groups. If the test results are nonsignificant, then the Test 5 is conducted to determine the existence of DIF in the d parameter (i.e., uniform DIF). The compact model constrains all the parameters for the studied item equal across groups. The augmented model constrains only the discrimination parameter across groups, but allows the d parameter to vary. A significant test indicates the presence of uniform DIF. A significant test for Test 2 indicates the presence of DIF in one or both $\boldsymbol{a}$ parameters. Tests 3 and 4 can further ascertain the degree to which DIF is present in the discrimination parameters.

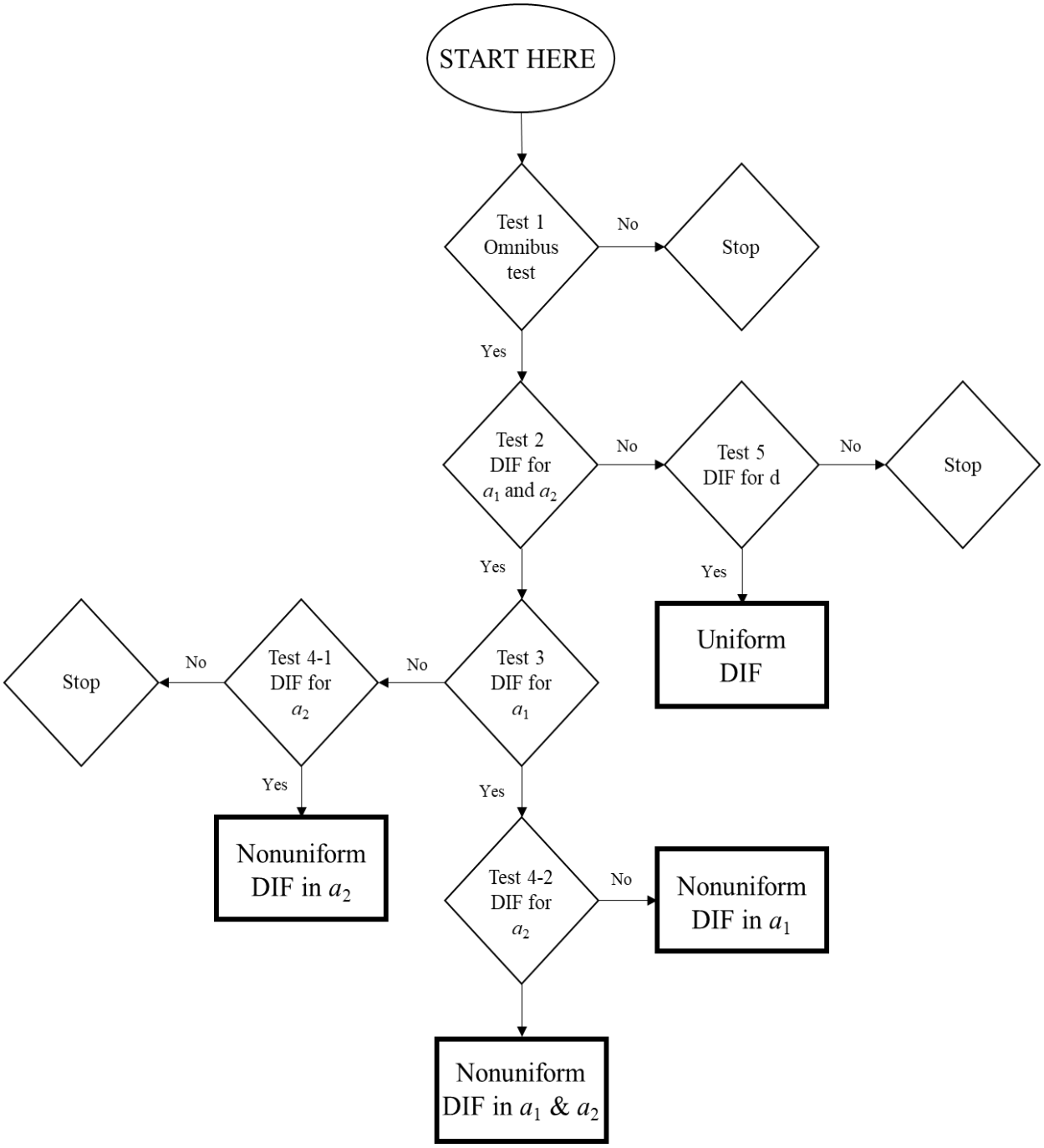

Figure 1.5. *Flowchart for assessing DIF using the multidimensional extension of IRT-LR test*

Note. This flowchart illustrates one approach to evaluating the subsequent DIF tests following the omnibus test. This particular model assumes only two dimensions exist, but extensions into larger dimensions can be applied using a similar process to determine the presence of DIF. Yes = significant test; No = nonsignificant test. Source: Suh and Cho (2014).

Under Test 3, we test for DIF in one of the discrimination parameters to determine whether the DIF in Test 2 occurs due to a_1. The compact model constrains only one discrimination parameter a_1 to be equal across groups, while the other discrimination parameter and d parameter are free to vary. The augmented model, as with the previous two tests, allows all parameters to vary freely across groups. A fourth test examines the existence of DIF caused by a_2 under the same constraints used in Test 3. If both tests are significant, then nonuniform DIF is present in both discrimination parameters. If one test is significant (e.g. Test 3), but the nonsignificant (e.g. Test 4), then nonuniform DIF would only be present in the significant test (i.e., a_1). All comparable compact and augmented models are assessed for statistical significance using equation 9. The IRT-LR test statistic, G^2, approximately follows a chi-squared distribution with *df* equal to the difference in the number of free parameters between the two compared models.

Multivariate logistic regression. The logistic regression model, as defined in equation 10, can accommodate multiple traits (Mazor et al., 1998). Mazor et al. (1998) extended the univariate logistic regression model to a multivariate framework to examine uniform DIF with multidimensional data. Bulut and Suh (2017) further expanded the logistic regression to detect nonuniform DIF in the presence of multiple abilities. Extension of the logistic regression model leads to the evaluation of both uniform and nonuniform DIF for an item associated with a set of *m* latent abilities θ_{np} such that:

$$z = \pi_0 + \sum_{n=1}^{m}(\pi_n\theta_{np}) + \pi_{m+1}G + \sum_{n=1}^{m}\pi_{m+n+1}(\theta_{np}G) \quad (16)$$

Uniform and nonuniform DIF can be identified using the same null hypotheses from the univariate logistic regression model (Penfield & Camilli, 2007). Similar to the univariate logistic regression procedure for detecting uniform and nonuniform DIF, a chi-squared test is conducted

using a sequence of three nested models, as shown in equation 11. The extent to which one nested model fits the broader model explicates the type of DIF present (Bulut & Suh, 2017; Penfield & Camilli, 2007). The simplest model contains no DIF and is expressed as the first two terms in equation 16, the uniform DIF includes the third term of equation 16, and all of equation 16 indicates the presence of nonuniform DIF (Bulut & Suh, 2017). First, the nonuniform DIF model and the uniform DIF model can be compared on the chi-squared distribution with one degree of freedom as the complex and nested model, respectively. If there is a significant group and latent ability interaction (i.e., $\pi_{m+n+1} \neq 0$), then the item exhibits nonuniform DIF. If no interaction exists, then the uniform DIF model, as the complex model, is compared with the no DIF model, as the nested model. If there is a significant group effect (i.e., $\pi_{m+1} \neq 0$), then the item exhibits uniform DIF; if the statistic is nonsignificant, then no DIF exists (i.e., $\pi_{m+1} = \pi_{m+n+1} = 0$).

Factors that Affect DIF

The accuracy of DIF detection procedures is largely dependent on their ability to minimize measurement error; that is, limit the amount of type I and II error. Type I error (i.e., α) is the rejection of a true null hypothesis and sometimes referred to as a false positive, while type II error (i.e., β) is the non-rejection of a false null hypothesis, sometimes referred to as a false negative (Coladarci et al., 2008). The inverse of type II error is known as statistical power (i.e., 1 – β). Power is the rejection of a false null hypothesis, and sometimes referred to as the rejection rate (Coladarci et al., 2008; Darlington & Hayes, 2017). Several factors have a strong effect on a statistical procedure's ability to detect DIF with accuracy. These factors (i.e., sample size, type of DIF, proportion of DIF, the correlation between the latent traits, and latent mean differences across groups) can largely affect the type I error rates and power of DIF detection procedures.

The accuracy of DIF detection procedures becomes more convoluted when these factors are examined simultaneously. Each factor, or combination of factors, can distort, improve, or otherwise influence the ability to detect DIF.

Sample size. The sample size of the focal and reference groups warrants strong consideration to ensure that DIF is accurately measured. Although the population size of groups could be known *a priori*, issues can arise when the sample size of one or more groups is too small to accurately reflect the population; that is, an insufficient sample size could elicit type II error. It has been observed that variations in sample sizes have a great effect on DIF detection in items (Acar, 2011). An early examination of the effects of sample size with detecting DIF show that small sample sizes (i.e., $n < 250$) decrease statistical power when type I error is effectively controlled (Spray, 1989).

Scott et al. (2009) found that a sample size of 300 per group is needed to detect uniform DIF for small scale unidimensional instruments (less than 5 items) with adequate power (greater than 80%). The sample size requirements decrease to 200 per group when unidimensional instruments of 5 or more items are used. Sample size requirements balloon to over 500 per group in order to detect nonuniform DIF in both small (i.e., $n = 2$) and large scaled instruments (i.e., $n = 20$). Lai et al. (2005) recommended sample sizes of 200 for data that resemble a Rasch model[4], but sample sizes of 500 or more for data estimated under the 2-PL IRT model. This figure is dependent on the number of items in the instrument and the skewness of the data. Similarly, Scott et al. (2009) advise that sample sizes of more than 500 per group be used for DIF detection studies. A simulation study and empirical example of adolescent substance use provide evidence

[4] The Rasch model is a logit function that represents the probability of a response to an item Y as a function of the latent ability of a person θ and the difficulty of an item b. The Rasch model is computationally equivalent to a one-parameter IRT model.

that using parsimonious models (i.e., Rasch, one parameter IRT models) in small samples leads to more powerful tests of DIF while adequately controlling for type I error (Belzak, 2019). Furthermore, increases in the number of DIF items in an instrument vastly improve statistical power when sample sizes are small (Belzak, 2019).

The effects of different sample sizes on type I error rates appear to be inconsistent across different DIF detection approaches for data fitted to a 2-PL MIRT model; however, large sample sizes (i.e., $n = 2000$) appear to positively influence rejection rates (Bulut & Suh, 2017). Statistical power was also increased when a more balanced sample size scheme was used to compare focal and reference groups, regardless of what DIF detection procedure was used (Bulut & Suh, 2017; Lee et al., 2017). Sample size recommendations for 2-PL MIRT models becomes complex because the percentage of DIF, type of DIF (i.e., uniform and nonuniform), and correlations between latent traits need to be considered in conjunction with sample size (Bulut & Suh, 2017; Lee et al., 2017; Oshima et al., 1997). Each of these factors can influence the recommended sample size for multidimensional models.

Similar to dichotomous models, several studies found that statistical power was improved when larger samples were used in a polytomous IRT model (Elosua & Wells, 2013; French & Miller, 1996; Hildalgo & Gómez, 2006). Data that fitted GRM were able to distinguish DIF and improve power when sample sizes were large enough. More specifically, sample sizes that were too small (e.g., 300 for each group) did not provide adequate parameter estimates to properly detect DIF, while type I error rates were consistent across different sample sizes (Kim & Cohen, 1998). Sample sizes of 500 to 1000 per group were reported as adequate sample sizes depending on the DIF procedure used (French & Miller, 1996; Kim & Cohen, 1998). Although DIF detection procedures have not been examined under multidimensional polytomous conditions,

Jiang et al. (2016) found that a total sample size of 500 provided accurate MGRM parameter estimates, except for tests with large numbers of items. Larger instruments needed more than a 1000 examinees to obtain accurate parameter estimates. An increase in sample size beyond 1000 did not improve the accuracy of MGRM parameter estimates (Jiang et al., 2016).

Type of DIF. Nonuniform DIF consider more parameters when detecting measurement bias than procedures only concerned with finding uniform DIF; therefore, detection of uniform DIF is likely to yield higher statistical power than nonuniform DIF when other factors are controlled (Thissen et al., 1993). For instance, MIMIC models include an interaction effect when detecting nonuniform DIF, but not when detecting uniform DIF. The addition of an interaction effect results in higher power when detecting nonuniform DIF, but the chance of incurring higher rates of type I error is likely to follow (Woods & Grimm, 2011). Statistical power is also substantially higher for the detection for uniform DIF compared to nonuniform DIF in MIRT models, regardless of sample size or the detection procedure (Bulut & Suh, 2017; Lee et al., 2017; Oshima et al., 1997).

For polytomous models, nonuniform DIF is most discernable when the discrimination parameters are largely different between groups, while uniform DIF is easily detected when threshold parameters are only slightly different between groups (French & Miller, 1996). It has been found that statistical power obtained from DIF found in the threshold parameters is higher than DIF found only in the discrimination parameters, but not the threshold parameters (Wang & Yeh, 2003). Under the GRM, DIF detection is difficult to detect when only the discrimination parameters differ between groups, even largely (Cohen et al., 1993). Thus, nonuniform DIF becomes challenging to detect under special circumstances of equal threshold parameters for data

fitted under GRM. Similar results were found for other polytomous models (Wang & Suh, 2010; Wang & Yeh, 2003).

Proportion of DIF. The ratio of items that exhibit DIF in a measurement instrument is the proportion of DIF (Finch, 2005; Hildalgo & Gómez, 2006; Su & Wang, 2005; Wang & Shih, 2010). The proportion of DIF refers to the number of DIF items in a set of *a priori* non-DIF items (i.e., anchors or DIF-free items). Under dichotomous and polytomous models, an instrument with higher rates of DIF typically invokes higher statistical power (Finch, 2005; Hildalgo & Gómez, 2006; Wang & Shih, 2010). This is because DIF items that contaminate the set of *a priori* non-DIF items leads to more conservative rejection rates of the null hypothesis; that is, the degree to which an item is DIF must be large enough to be significantly detectable despite the effects from the DIF anchor items. As a result, more stringent rejection rates for DIF leads to the reduction of type II error, and thus, increases power.

However, instruments that include a higher percentage of DIF items are also likely to yield an increase in type I error (Finch, 2005; Su & Wang, 2005); thus, items that are functioning similarly are more likely to be incorrectly identified as DIF (Magis et al., 2010). Item purification is recommended to remedy this effect within empirical studies (Holland & Thayer, 1988; Magis et al. 2010; Oshima & Miller, 1993; Su & Wang, 2005; Wang & Shih, 2010; Wang & Yeh, 2003; Woods, 2009b). Item purification seeks to eliminate problematic DIF items from a set of anchor items to better gauge each potential DIF item one-by-one. This process has been shown to better control type I error, while marginally affecting power (Su & Wang, 2005; Wang & Shih, 2010).

Similar to UIRT models, the proportion of DIF highly and positively influences the rejection rates in detecting both uniform and nonuniform DIF in 2-PL MIRT models (Bulut & Suh, 2017).

As demonstrated in the Lee et al. (2017) study, statistical power tends to be greater when the proportion of DIF in an instrument is high (i.e., 40 %) after other simulation conditions are controlled (e.g. sample size, correlation between latent traits). The pattern was more prominent under uniform DIF and large sample conditions. Likewise, type I error rates increase as the proportion of DIF in 2-PL multidimensional instruments also increases (i.e., about 30% DIF; Liaw, 2015). Liaw (2015) showed that the proportion of DIF has the largest effect on detection of type I error rates ($\omega^2 = .40$) compared to other factors that influence type I error detection (i.e., correlation between latent traits, $\omega^2 = .00$; balanced sample sizes between groups, $\omega^2 = .13$; nuance discrimination parameter, $\omega^2 = .17$).

Correlation between latent traits. Multidimensional models produce multiple latent traits that have a statistical association with one another. The relationship between latent traits have varying effects on the detection of DIF and are mostly dependent on other, more prominent factors (e.g., sample size, proportion of DIF, type of DIF). Bulut and Suh (2017) found that type I error rates increased when sample size increases and no correlation existed between latent traits but found type I error rates showed no discernable pattern when correlations were higher. However, when all other conditions are controlled (i.e., sample size, type of DIF, proportion of DIF, latent mean differences across groups), high and low correlations between latent traits proved to have little difference in the type I error rates in a 2-PL MIRT model (Mazor et al., 1998). Similarly, the correlation between latent traits did not meaningfully affect rejection rates when other conditions are controlled (Bulut & Suh, 2017). Bulut and Suh (2017) used a multivariate analysis of variance to simultaneously examine the effects of multiple factors on rejection rates. Results from the between-subject factors indicate that the correlation between latent traits was the only factor that did not significantly affect the rejection rates when other

conditions were controlled; however, there is evidence to suggest that power increases when correlations between latent traits are increased and other conditions are considered (Bulut & Suh, 2017; Lee et al., 2017).

Latent mean differences across groups. The latent mean differences between focal and reference groups are defined as the group impact (Holland & Wainer, 1993; Stark et al., 2006). No group impact is present when the latent means are estimated to be equal (e.g. $\mu_R = \mu_F = 0$); whereas, unequal latent means (e.g. $\mu_R = 0 \wedge \mu_F = -.5$) between the focal and reference groups does indicate group impact (DeMars & Lau, 2011; Jin & Chen, 2020; Stark et al., 2006). When the latent means between groups are equivalent in a multidimensional dichotomous model, the statistical power rates are typically higher or equal to the power rates when latent means between groups are different (Lee et al., 2017). Most procedures to identify uniform and nonuniform DIF are robust against latent mean differences. Latent mean differences are influential only when the magnitude of DIF is low and the DIF type is uniform (Bulut & Suh, 2017). The latent means can cause a distributional difference between the focal and reference group in the context of multidimensional DIF studies (Oshima et al., 1997); thus, investigation of the effects of the latent means on a multidimensional polytomous model is advisable.

Studies on Comparative DIF Procedures

Examination of the literature does not point to one optimal DIF detection procedure, but instead suggestions DIF detection procedures on a conditional basis. Optimal DIF detection procedures are predicated on the method's ability to minimize measurement error and produce high levels of certainty and power. In general, the best DIF detection procedure varies based on sample size, type of DIF, proportion of DIF, the correlation between the latent traits, latent mean differences across groups, and combinations of these factors. Additionally, the framework in

which the data underlies (e.g. 2-PL IRT, GRM, 2-PL MIRT) appears to play a prominent role in the selection of the DIF detection procedure (Bulut & Suh, 2017; Elosua & Wells, 2013; Finch, 2005; Kim & Cohen, 1998; Woods & Grimm, 2011).

DIF detection procedures for unidimensional dichotomous models. Several different DIF detection procedures can examine DIF in unidimensional dichotomous models (e.g. Bifactor model, IRT-LR test, logistic regression, Mantel-Haenszel method, MIMIC-interaction model, SIBTEST). Many of these procedures are much less labor intensive to compute than others, such as the IRT-LR test (Penfield & Camilli, 2007); however, with the advancement of computers, the IRT-LR test has proven to be a more reliable source to detect general DIF and nonuniform DIF than less complex methods (e.g., Bifactor model, Mantel-Haenszel method; Woods, 2009b). Cohen et al. (1996) found that when the 2-PL IRT model underlies the data, type I error rates were within an acceptable range for the IRT-LR test. Statistical power was also shown to be high under a number of conditions (i.e., sample size, proportion of DIF, and type of DIF) for the IRT-LR test (Woods, 2009b).

Other detection procedures also showed promising results. For instance, the MIMIC model showed higher rates of power for DIF identification when instrument has 50 items or the 2-PL IRT model underlies the data than the IRT-LR test (Finch, 2005). Tests of uniform DIF with dichotomous and polytomous responses were more accurate with MIMIC models than IRT-LR test when sample size was small. Type I error was well below the nominal alpha level and power was greater for the MIMIC approach than for IRT-LR test across various sample sizes (Woods et al., 2009). However, this study did not examine uniform and nonuniform DIF, specifically. At the time of the study, the MIMIC approach was not developed to account for nonuniform DIF. Woods and Grimm (2011) added an interaction effect to the MIMIC model to detect nonuniform

DIF. They found that accuracy was higher from IRT-LR test than the MIMIC model with an interaction effect. Although the MIMIC model with an interaction effect is a powerful DIF detection procedure, it produces severely inflated type I error (Woods & Grimm, 2011).

DIF detection procedures for unidimensional polytomous models. Swaminathan and Rogers (1990) introduced the logistic regression model as a means to detect DIF in unidimensional dichotomous models. Results from the French and Miller (1996) study indicate that the logistic regression model is powerful in detecting most forms of DIF; however, it requires large amounts of data manipulation and interpretation of the results can be difficult under polytomous conditions. Instead, the logistic discriminant function procedure is better suited for DIF identification on polytomously scored items. The procedure is simpler and more practical than polytomous extensions of the logistic regression DIF procedure and appears to be more powerful than the contingency table methods (Miller & Spray, 1993). The discriminant logistic analysis appears to be a better balance between statistical power and type I error rate than does a polytomous extension of the logistic regression (Hildalgo & Gómez, 2006). The discriminant logistic analysis also yields higher power and lower type I error rates when compared to other polytomous DIF detection procedures (Su & Wang, 2005). While polytomous extensions of the logistic regression and the discriminant logistic analysis function had good control of type I error, as well as strong statistical power for detecting uniform DIF, the logistic discrimination function analysis performed poorly in identifying nonuniform DIF, principally when item discrimination was high (Kristjansson et al., 2005).

The IRT-LR test is a practical alternative to the logistic regression approach. The test has greater statistical power than nonparametric methods (e.g., Liu-Agresti estimator; Bolt, 2002), and also controls type I error well under different sample sizes when data fits the GRM, but the

proportion of DIF can influence type I error, and subsequently, the power of the IRT-LR test (Kim & Cohen, 1998). A more recent study from Elosua and Wells (2013) found that the IRT-LR test exhibited more power compared to polytomous extensions of the logistic regression when detecting DIF under the GRM, but type I error rates were inflated for both the logistic regression approach and IRT-LR test. The polytomous extensions of the logistic regression appear to display marginally higher power than the IRT-LR test for smaller sample sizes when detecting uniform DIF; however, the IRT-LR approach exhibited much more power than polytomous extensions of the logistic regression for nonuniform DIF (Elosua & Wells, 2013).

DIF detection procedures for multidimensional dichotomous models. Studies have only recently begun to develop and compare DIF procedures for multidimensional data. The consequences of using unidimensional procedures to identify DIF when multidimensionality underlies the data have been documented to show increases in type I error and limited statistical power (Mazor et al., 1998). At least one recent studies has compared different DIF detection procedures specifically for multidimensional data fitted to the 2-PL MIRT model. Bulut and Suh (2017) compared the multidimensional MIMIC-interaction model (Lee et al., 2017), MIRT-LR test (Suh & Cho, 2014), and the multiple logistic regression model (Mazor et al., 1998). This study examined the performance of each DIF detection method to determine the optimal method under various factor constraints (i.e., type of DIF, magnitude of DIF, test length, sample size, correlation between latent traits, and latent mean differences between the reference and focal groups) with respects to type I error and rejection rates.

The analysis revealed that the MIRT-LR appears to be preferable over the other three methods in detecting uniform and nonuniform DIF based on type I error and rejection rates. However, the MIMIC-interaction and multiple logistic regression models performed better under

short tests (i.e., 10 items or less) compared to the MIRT-LR test when detecting uniform DIF. When test lengths and the number of anchor items increased (i.e., lower proportion of DIF), all three methods performed about the same in detecting uniform DIF. Additionally, the multiple logistic regression model consistently inflated type I error rates under most factor constraints. The MIMIC-interaction model demonstrated high rejection rates when sample sizes were high, but also showed inflated type I errors under similar conditions (cf. Lee et al., 2017). The MIRT-LR outperformed the other two methods considerably when detecting nonuniform DIF, regardless of the factor constraints. In general, the MIRT-LR test is a more balanced and powerful approach than the MIMIC-interaction model and the multiple logistic regression in uniform and nonuniform DIF detection for multidimensional, dichotomous data with a non-simple structure.

Statement of the Problem

DIF procedures have been thoroughly scrutinized under dichotomous UIRT models (Cohen et al., 1996; Finch, 2005; Magis, 2010; Mazor et al., 1998; Montoya & Jeon, 2020; Raju, 1988; Stark et al., 2006; Woods & Grimm, 2011; Woods et al., 2009), polytomous UIRT models (Cohen et al., 1993; Elosua & Wells, 2013; French & Miller, 1996; Golia, 2012; Kim & Cohen, 1998; Kristjansson et al., 2005; Miller & Spray, 1993; Potenza & Dorans, 1995; Su & Wang, 2005; Wang & Shih, 2010), and dichotomous MIRT models (Bulut & Suh, 2017; Lee et al., 2017; Mazor et al., 1998; Oshima & Miller, 1993; Oshima et al., 1997; Reise et al., 2014; Suh & Cho, 2014). In addition, studies have used MGRM and other polytomous MIRT models with empirical data to study health surveys (Depaoli et al., 2018), educational instruments (Immekus & Imbrie, 2008; Immekus et al., 2019), and psychological tests (De Ayala, 1994; Jin & Chen, 2020); however, optimal methods to examine DIF under polytomous MIRT models have not

been explored. This study aims to produce substantial evidence of the optimal DIF methods for data that fits a polytomous MIRT model (i.e., multidimensional graded response model).

Polytomous IRT-based models provide important details about the individual items found in a polytomous measurement instrument that cannot be replicated under classical statistical techniques (e.g., reliability and factor analysis; Depaoli et al., 2018; Nering & Ostini, 2010). For instance, both factor analysis and IRT-based models provide information about dimensionality, covariance structure, and model fit, but only IRT models provide details on response patterns across items (Depaoli et al., 2018). DIF can be used to support the existence of construct validity in polytomous measurement instruments, such as a rubric or evaluation form, but more commonly in surveys (Walker, 2011). DIF methods have been used widely in applied settings to argue the presence of construct validity in polytomous instruments, such as to evaluate the translations of questionnaire items (Kwakkenbos et al., 2014) or assess differences in diagnostic interviews between racial or gender groups (Uebelacker et al., 2009). However, many of these items are evaluated under the assumption of a unidimensional framework, despite the multiple traits that inherently exist in these types of instruments.

Significance of the Problem

Most measurement instruments assume a unidimensional framework, but in truth, these instruments are rarely unidimensional; instead, most measurement data are intrinsically multidimensional (Edwards, 2001; Embertson & Reise, 2000; Polites et al., 2012). The fundamental design of most measurement instruments implicitly and unwittingly captures multiple latent traits. For example, an item may ask a respondent to rate their attitude toward a specific sports team, but that attitude could also capture the respondent's attitude toward a player, coach, or region the team represents. Although unidimensional DIF detection procedures

are computationally less arduous and yield straightforward interpretations of data, their application with real-world data fails to account for the pragmatic and complex structure of multidimensional data (Embertson & Reise, 2000; Reckase, 2009).

Accurate DIF methods are imperative to ascertain the true presence and degree of measurement bias in a measurement instrument. Until now, DIF detection procedures have not been examined for data that fits polytomous MIRT-based models. Thus, this study explores DIF procedures for data that fits a polytomous MIRT-based model (i.e., multidimensional graded response model) to determine the best methods to compare groups in commonly used measurement instruments (e.g., surveys, evaluation forms, rubrics). In particular, assessment practitioners, survey researchers, psychometricians, and other measurement experts could find the results to be valuable in demonstrating construct validity in multidimensional measurement instruments. The optimal DIF method can vary depending on the nature of the data and selecting the correct DIF method can mean the difference of finding true DIF and reporting deficient results. Therefore, this study helps to minimize measurement error and assist practitioners by correctly identifying areas to improve or support construct validity.

Research Objectives

This study (a) compares the rejection rates of three DIF detection procedures (i.e., the multidimensional MIMIC-interaction model, MIRT-LR test, and the multidimensional logistic discrimination function analysis) to identify uniform and nonuniform DIF on data that fits the multidimensional graded response model (MGRM) framework; (b) compares the type I error rates of the three DIF detection procedures to estimate DIF on data that fits the MGRM framework; and (c) examines the effect of different factors (type of DIF, proportion of DIF,

sample size, correlation between latent traits, and latent mean differences between the focal and reference groups) on the behaviors of the three DIF detection procedures.

Chapter II

Methods

This study incorporated a Monte Carlo simulation to generate data fitted to the multidimensional graded response model. Specific constraints was applied to examine how different factors (i.e., sample size, correlation between latent traits, type of DIF, proportion of DIF, and latent mean differences) affected DIF detection. Furthermore, three DIF detection procedures (i.e., MIRT-LR test, multidimensional MIMIC-interaction model, and the multidimensional extension of the logistic discriminant function analysis) were studied to determine the optimal procedure to detect DIF under the aforementioned constraints. Type I error rates and rejection rates (i.e., power) measured the strength and accuracy of each DIF detection procedure. Mixed-design analyses of variance (ANOVA) examined the effects of different factor constraints across different DIF method treatments on the rejection and type I error rates. Results from the type I error rates, rejection rates, and the mixed-design ANOVA addressed the research objectives. In addition, an empirical example provided practical support for the simulation findings.

Data Generation

A Monte Carlo simulation generated polytomous item responses according to the multidimensional graded response model (MGRM) in the R package `SimDesign` (Chalmers & Adkins, 2020). Monte Carlo simulations generate random data from specified conditions through multiple iterations of the data generation procedure to fit a certain model (Kelley, 2010; Sigel & Chalmers, 2016). The data generation procedure pulls random values from a probability density function (e.g., normal distribution, 2-PL IRT, GRM) and utilizes the properties of the central

limit theorem to produce a plausible data frame (Sigel & Chalmers, 2016). Hallgren (2014) identified the steps to generate data from Monte Carlo simulations. First, simulation studies need to establish a set of assumptions (i.e., constraints) regarding the nature and parameters of the generated data sets. Next, numerous datasets are constructed (e.g., 1,000 datasets) according to the established assumptions to obtain an empirical distribution of parameter estimates. Statistical analyses are performed on each generated dataset and parameter estimates are retained. Afterwards, assumptions are modified to address the other constraints, new replicated datasets are generated, and the statistical analyses are methodically repeated under each new assumption.

In this study, specified factor constraints (i.e., four sample sizes, three correlations between latent traits, two types of DIF, four proportions of DIF, and two latent mean differences) generated multiple simulated data sets using full information maximum likelihood estimates (FIML) to fit the MGRM. A single value for each factor constraint was assumed in combination with the other constraints to generate a dataset. The process was systematically replicated using the same constraints to obtain a distribution of parameter estimates. The number of replications can vary based on the enormity and complexity of the simulation; most DIF studies conduct 100 or fewer replications for simulated data (Bolt, 2002; Bulut & Suh, 2017; Hildalgo & Gómez, 2006; Lee et al., 2017; Suh & Cho, 2014). Thus, 100 replications were generated for each combination of factor constraints. Next, a single constraint (e.g. sample size) changed from the combination of other constraints and the process will repeat itself. Data generation continued until all combinations of factor constraints, 120 in total, were exhausted.

Data framework. The data fit in accordance with a compensatory, within-item MGRM. The data sets imitated a 20-item survey with a 5-level scale for all items and two latent abilities. Data from a 5-level scale resembled Likert scale responses, a commonly used scale format in

polytomous item surveys, and 20 items allowed for the instrument to be easily partitioned into multiple combination of anchor and DIF items. A within-item model assumed a non-simple structure, such that any given item can be regressed on multiple latent traits. A compensatory model assumed disjointed latent traits (i.e., latent traits are independent of one another for endorsement of an item) and a linear relationship between all latent traits (Hartig & Höhler, 2009). Other similar MIRT studies have used two latent dimensions to investigate the nature of multiple dimensions and simplified some of the complexities associated with multidimensional structures (Bolt & Lall, 2003; Cohen et al., 1993; Cohen et al., 1996; Su & Wang, 2005). Therefore, this study also assumed only two latent traits exist in each data set. The two latent traits were assumed to be normally distributed $\theta \sim N(0, 1)$, similar to other DIF and MIRT studies (Cohen et al., 1996; Lee et al., 2017; Suh & Cho, 2014). Furthermore, the data sets were divided randomly into two groups, the reference and focal groups, of equal ability-levels based on the sample size constraints.

Factor constraints. The ability of a statistical procedure to detect DIF was largely dependent on several factors, namely the sample size, the type of DIF (i.e., uniform and nonuniform), the proportion of DIF, the correlation between the latent traits, and latent mean differences across groups. These factors can largely affect the type I error rates and power of DIF detection procedures and lead to inaccurate detection or missed detection of DIF. Each factor, or combination of factors, can distort, improve, or otherwise effect the ability to detect DIF. Therefore, this simulation study constrained these factors across different values to understand the influence they may have on DIF detection under the MGRM framework.

Kim and Cohen (1998) specified that sample sizes large enough to adequately estimate model parameters are needed for accurate DIF detection (i.e., at least 300 responses). Jiang et al.

(2016) noted that sample sizes less than 200 did not adequately estimate MGRM parameters, while sample sizes over 1000 did not demonstrate improved accuracy of MGRM parameters. Thus, the range of sample sizes (250 to 1000 responses) were selected based on these recommendations. Therefore, the combination of sample sizes between the reference and focal groups were decided based on other comparable DIF studies (cf. Bulut & Suh, 2017; Lee et al., 2017; Oshima et al., 1997; Suh & Cho, 2014; Woods & Grimm, 2011; Woods, 2009a, b). The simulation study constrained on four sample sizes. The sample size conditions consist of two unbalanced and two balanced conditions between the reference *R* and focal *F* groups. The two unbalanced conditions (R1000/F500; R500/F250) were selected to reproduce conditions found in real empirical data. Since reference group is typically larger than the focal group (Camilli, 2006), the reference group was twice as large as the focal group. The balanced conditions (R1000/F1000; R500/F500) were selected to compare the sample size effect between balanced and unbalanced conditions.

Multidimensional polytomous item response data were simulated with use of anchor discrimination parameters found in the Reckase (2009, p. 204) analysis and anchor threshold parameters from the Cohen et al. (1993) study. The data were constrained to four levels of DIF proportions (0%, 5%, 10%, and 20% DIF). This was comparable to the levels of DIF used in similar studies (cf. Finch, 2005; Hildalgo & Gómez, 2006; Su & Wang, 2005; Wang & Shih, 2010). Fixed parameters on the anchor items generated anchor item responses for both the reference and focal groups (see Table 2.1). Three groups of loading patterns formed the anchor discrimination parameters for this study. For data tested with DIF proportions of 5% and 10%, the first six items mostly loaded on the first latent trait, the second six items mostly loaded on the second latent trait, and the last six or seven items loaded approximately equal on both latent

Table 2.1

Anchor item parameter generation of multidimensional graded response model

	0%, 5%, 10% DIF						20% DIF					
Item	a_1	a_2	d_1	d_2	d_3	d_4	a_1	a_2	d_1	d_2	d_3	d_4
1	.89	.14	1.14	.22	-.16	-.57	.89	.14	1.14	.22	-.16	-.57
2	1.06	.04	1.47	.24	-.33	-1.34	1.06	.04	1.47	.24	-.33	-1.34
3	1.05	.02	.54	.19	-.23	-1.12	1.05	.02	.54	.19	-.23	-1.12
4	1.18	.02	.76	.41	-.59	-1.40	1.18	.02	.76	.41	-.59	-1.40
5	1.03	.23	.44	.05	-.44	-.91	1.03	.23	.44	.05	-.44	-.91
6	.93	.08	.89	.44	.18	-1.44						
7	.18	.98	.55	.06	-.43	-1.10	.18	.98	.55	.06	-.43	-1.10
8	.20	.94	-.02	-.57	-1.33	-1.97	.20	.94	-.02	-.57	-1.33	-1.97
9	.26	1.08	.88	.18	-.83	-1.54	.26	1.08	.88	.18	-.83	-1.54
10	.15	.81	2.06	-.42	-.82	-1.10	.15	.81	2.06	-.42	-.82	-1.10
11	.10	1.03	.65	.13	-.96	-1.48	.10	1.03	.65	.13	-.96	-1.48
12	.07	1.09	1.89	-.06	-.38	-.62						
13	.74	.76	1.24	.22	-.33	-.83	.74	.76	1.24	.22	-.33	-.83
14	.66	.85	.63	.08	-.50	-1.69	.66	.85	.63	.08	-.50	-1.69
15	.81	.73	1.96	.34	-.90	-1.61	.81	.73	1.96	.34	-.90	-1.61
16	.67	.60	.83	.41	-.07	-.56	.67	.60	.83	.41	-.07	-.56
17	.70	.74	1.79	.45	.19	-.31	.70	.74	1.79	.45	.19	-.31
18	.72	.72	1.27	.99	.04	-.93	.72	.72	1.27	.99	.04	-.93
19	.55	.71	1.39	.98	.46	.26						
20	.68	.62	-.01	-.86	-1.05	-1.36						
M	.63	.61	1.02	.17	-.42	-1.08	.65	.61	1.01	.19	-.48	-1.15
SD	.35	.38	.61	.44	.46	.54	.36	.39	.59	.35	.41	.46

Note. *M* = Mean; *SD* = standard deviation; Item 20 will be dropped as an anchor when the 5% DIF condition is applied; Items 19 – 20 will be dropped as anchors when the 10% DIF condition is applied.

traits. For data tested with a DIF proportion of 20%, the first five items mostly loaded on the first latent trait, the second five items mostly loaded on the second latent trait, and the last six items loaded approximately equal on both latent traits. The threshold parameters decreased in value from the lowest threshold level d_1 to the highest threshold level d_4.

Table 2.2 showed item parameters used to generate the DIF items. Selective DIF item parameters were designated to simulate the uniform and nonuniform DIF conditions. One DIF item simulated data with 5% DIF, two DIF items simulated data with 10% DIF, and five DIF items simulated data with 20% DIF. Discrimination parameters were the equal across groups under uniform DIF, but differed between one or both latent traits under DIF. Threshold parameters were consistent between items, but differed by .5 across groups. The differences in the *a* and *d* parameters between the focal and the reference groups were controlled to replicate similar MIRT DIF studies (e.g., Bulut & Suh, 2017; French & Miller, 1996; Oshima et al., 1997; Suh & Cho, 2014; Wang & Shih, 2010; Wang & Yeh, 2003). Thus, non-DIF, uniform DIF, and DIF conditions were generated through use of the anchor and DIF items.

In the context of multidimensional studies, a distributional difference can occur from differences in the covariance structure (i.e., correlation) of the latent traits, the location of latent means, or both (Oshima et al., 1997). Thus, two conditions, balanced and unbalanced, were considered to investigate the effect of having a mean difference between the two groups. First, the unbalanced condition specified equal means and variances on both latent traits for the focal and reference groups, such that a multivariate normal distribution was assumed and each dimension had a mean of 0 and a variance of 1. Second, the balanced condition specified unequal means, but equal variance between the reference and focal groups. For the focal group in the balanced condition, both latent traits followed a bivariate normal distribution where each

Table 2.2

Item parameter generation of multidimensional graded response model for DIF conditions

Type of DIF	% of DIF	Item	Reference Group						Focal Group					
			a_1	a_2	d_1	d_2	d_3	d_4	a_1	a_2	d_1	d_2	d_3	d_4
Uniform	5%	20	1.00	1.00	2.00	1.00	-1.00	-2.00	1.00	1.00	2.50	1.50	-.50	-1.50
	10%	19	.70	.70	2.00	1.00	-1.00	-2.00	.70	.70	2.50	1.50	-.50	-1.50
		20	1.00	1.00	2.00	1.00	-1.00	-2.00	1.00	1.00	2.50	1.50	-.50	-1.50
	20%	6	1.00	.10	2.00	1.00	-1.00	-2.00	1.00	.10	2.50	1.50	-.50	-1.50
		12	.10	1.00	2.00	1.00	-1.00	-2.00	.10	1.00	2.50	1.50	-.50	-1.50
		19	.70	.70	2.00	1.00	-1.00	-2.00	.70	.70	2.50	1.50	-.50	-1.50
		20	1.00	1.00	2.00	1.00	-1.00	-2.00	1.00	1.00	2.50	1.50	-.50	-1.50
Non uniform	5%	20	1.00	1.00	2.00	1.00	-1.00	-2.00	1.00	1.50	2.50	1.50	-.50	-1.50
	10%	19	.70	.70	2.00	1.00	-1.00	-2.00	1.00	1.00	2.50	1.50	-.50	-1.50
		20	1.00	1.00	2.00	1.00	-1.00	-2.00	1.00	1.50	2.50	1.50	-.50	-1.50
	20%	6	1.00	.10	2.00	1.00	-1.00	-2.00	.70	.10	2.50	1.50	-.50	-1.50
		12	.10	1.00	2.00	1.00	-1.00	-2.00	.10	1.50	2.50	1.50	-.50	-1.50
		19	.70	.70	2.00	1.00	-1.00	-2.00	1.00	1.00	2.50	1.50	-.50	-1.50
		20	1.00	1.00	2.00	1.00	-1.00	-2.00	1.50	1.50	2.50	1.50	-.50	-1.50

Note. a_1 represents the item discrimination parameter for the first latent trait; a_2 represents the item discrimination parameter for the second latent trait; d_1 represents the first threshold parameter; d_2 represents the second threshold parameter; d_3 represents the third threshold parameter; d_4 represents the fourth threshold parameter

dimension had a mean of -.5 and a variance of 1. For the reference group, two latent traits were generated from a bivariate normal distribution and each dimension has a mean of .5 and a variance of 1. The mean and variance values and distributional choice were selected from similar DIF studies (DeMars & Lau, 2011; Jin & Chen, 2020; Stark et al., 2006). In addition, the correlations between the two latent traits under each distributional condition were tested at three levels ($\rho = 0$, $\rho = .3$, $\rho = .6$). The levels were selected based on similar correlational constraints found in the Bolt and Lall (2003) study, which took values from Cohen's (1988, 1992) correlational benchmark of no correlation (0), medium correlation (.3), and large correlation (.6).

Analytical Procedures

Several analytical procedures have been developed and tested to examine DIF in polytomous, dichotomous, and multidimensional dichotomous data, but few, if any, studies have tested DIF procedures in multidimensional, polytomous data (e.g. MGRM). This study tested the accuracy of three statistical procedures (i.e., MIRT-LR test, multidimensional MIMIC-interaction model, and the multidimensional extension of the logistic discriminant function analysis) to detect uniform and DIF under various factor constraints. This section outlines the rationale, the assumptions, estimation methods, and identification procedures of each DIF method.

MIRT likelihood ratio test (MIRT-LR). The MIRT likelihood ratio (MIRT-LR) chi-squared statistic, G^2, examined DIF under the MGRM framework using the procedures established by Suh and Cho (2014). A compensatory, within-item MGRM first estimated the model parameters. Since only two dimensions were present in this simulation study, the MIRT-LR test compared nested models through a pair of tests to evaluate the model fit, and thus, determine the presence and extent of DIF, as outlined in Figure 1.5. However, this study did not conduct the omnibus DIF test or tests for nonuniform DIF on individual latent traits, so that the

three DIF detection procedures were as comparable as possible; instead, the process began with test 2. In this test, and all subsequent tests, the augmented model constrained the parameters for all anchor items (see Table 2.1), but not the studied items (see Table 2.2), equal across the reference and focal groups. The augmented model were compared to a sequence of compacted models in which the item threshold parameters and item discrimination parameters (a_1 and a_2) for each studied item were fixed equal between the two groups, successively. The difference between likelihood tests for each hierarchical model were distributed as chi-squared variables with degrees of freedom equal to the difference in the number of parameters estimated between models. Each test examined the statistical significance of improved model fit with the addition of a new term. The MIRT-LR test tested the null hypothesis that DIF was present between the focal and reference groups.

Model identification and the estimation methods also needed to be established to estimate the augmented and compact models of each test correctly. The association between the parameters of IRT models, MIRT models included, and factor analysis models has been widely accepted in the measurement community (e.g., Kamata & Bauer, 2008; Stark et al., 2006). Thus, DIF can be studied by applying commonly used factor analysis estimation methods to estimate IRT item parameters (e.g., Kim & Yoon, 2011; Suh & Cho, 2014). In this study, a full-information maximum likelihood estimator (FIML) [5] and a χ^2 test statistic will compare models in the MIRT-LR test. Although Suh and Cho (2014) recommended a WLSMV estimation method to more accurately assess ordinal data, the decision to alter the estimation method for the MIRT-LR test was to ensure estimation methods were consistent between other DIF procedures used in this

[5] FIML is used for this simulation study because the dataset is assumed complete. Empirical studies with missing data should consider the maximum likelihood estimation method with Huber-White robust standard errors (MLR) and the Yuan-Bentler test statistic for better parameter approximations, provided the number of scale levels is greater than 4. If not, consider an alternative estimation method (e.g. MLR-CAT; Lei & Shiverdecker, 2020).

study (i.e., multidimensional extension of the MIMIC model). Research also showed that a maximum likelihood estimation method was shown to perform as well as the WLSMV estimation method for ordered categorical data when sample sizes were over 500 (Bandalos, 2014; Li, 2016). The MIRT-LR test was obtained from the maximum likelihood estimation (FIML) of the MGRM using the `mirt` package (Chalmers, 2012) in R v. 4.0.3 (R Core Team, 2020).

Model identification resolved issues with metric indeterminacy across the latent ability parameters; thus, the means and variances of both latent abilities was normally distributed and fixed to 0 and 1s, respectively, in both the focal and reference groups. In test 2, the item discrimination parameters was constrained equal across groups for the studied items in the compact model, while all other parameters were unconstrained to test for nonuniform DIF. In the augmented model, all item parameters were free to vary between groups. In test 5, all item parameters (i.e, both discrimination parameters and threshold parameters) were set equal across both groups in the compact model to test for uniform DIF. In the augmented model, only the item discrimination parameters were set equal across groups for the studied item.

Multidimensional MIMIC-interaction model. A multidimensional MIMIC-interaction model examined the presence of DIF in multidimensional, polytomous data that fits to a MGRM framework. A latent factor regressed on a covariate examines mean differences between the values of the covariate (i.e., group membership) in a standard MIMIC model (i.e., equations 5 and 6; Kline, 2012). Comparatively, the regression of an endogenous manifest variable on a covariate allows one to examine the difference between the values of the covariate (i.e., group membership) on a response option, while controlling for the latent factor (Brown & Moore,

2012; Wang & Wang, 2012). In other words, an additional coefficient between an endogenous indicator and covariate in a standard MIMIC model reveals the presence of DIF.

Additional extensions to this model helped to specify a new model that fits multidimensional data and detects nonuniform DIF. First, an interaction term between the latent ability factor and group membership (i.e., the covariate) determined the presence of nonuniform DIF. Second, a non-simple structure, such that continuous latent response variables were allowed to regress on multiple latent factors, was assumed. Thus, a multidimensional MIMIC-interaction model for some item *i* was expressed as:

$$y_i^* = \boldsymbol{\lambda}_i \boldsymbol{\theta} + \beta_i G + \boldsymbol{\omega}_i \boldsymbol{\theta} G + \varepsilon_i \quad (17)$$

where $\boldsymbol{\lambda}_i$ was a *m* x *m* symmetric diagonal matrix of factor loadings for item *i* on a *m* x 1 vector of latent abilities $\boldsymbol{\theta}$, β_i was the uniform DIF effect of the grouping variable *G* on item *i* (when $\beta_i \neq 0$), $\boldsymbol{\omega}_i$ was a *m* x *m* symmetric diagonal matrix of the nonuniform DIF effect on item *i* (when $\boldsymbol{\omega}_i \neq 0$), and ε_i was the normally distributed residual. The grouping variable *G* was dummy coded such that the reference group *R* was 0 and the focal group *F* was 1. Figure 2.1 displayed the complete model for illustrative purposes. The multidimensional MIMIC-interaction model tested the null hypothesis that DIF was present between the focal and reference groups.

Group-invariant anchor items, continuous latent item responses, and fixed latent variance between groups was assumed for model identification, in addition to the common SEM assumptions (Kelloway, 2015; Kline, 2012; Ullman, 2018). As such, the latent variables $\boldsymbol{\theta}$ assumed to have a multivariate normal distribution with means of 0 and variances of 1. Uniform and nonuniform DIF effects was also fixed to 0 for all anchor items. Factor loading $\boldsymbol{\lambda}_i$, residuals,

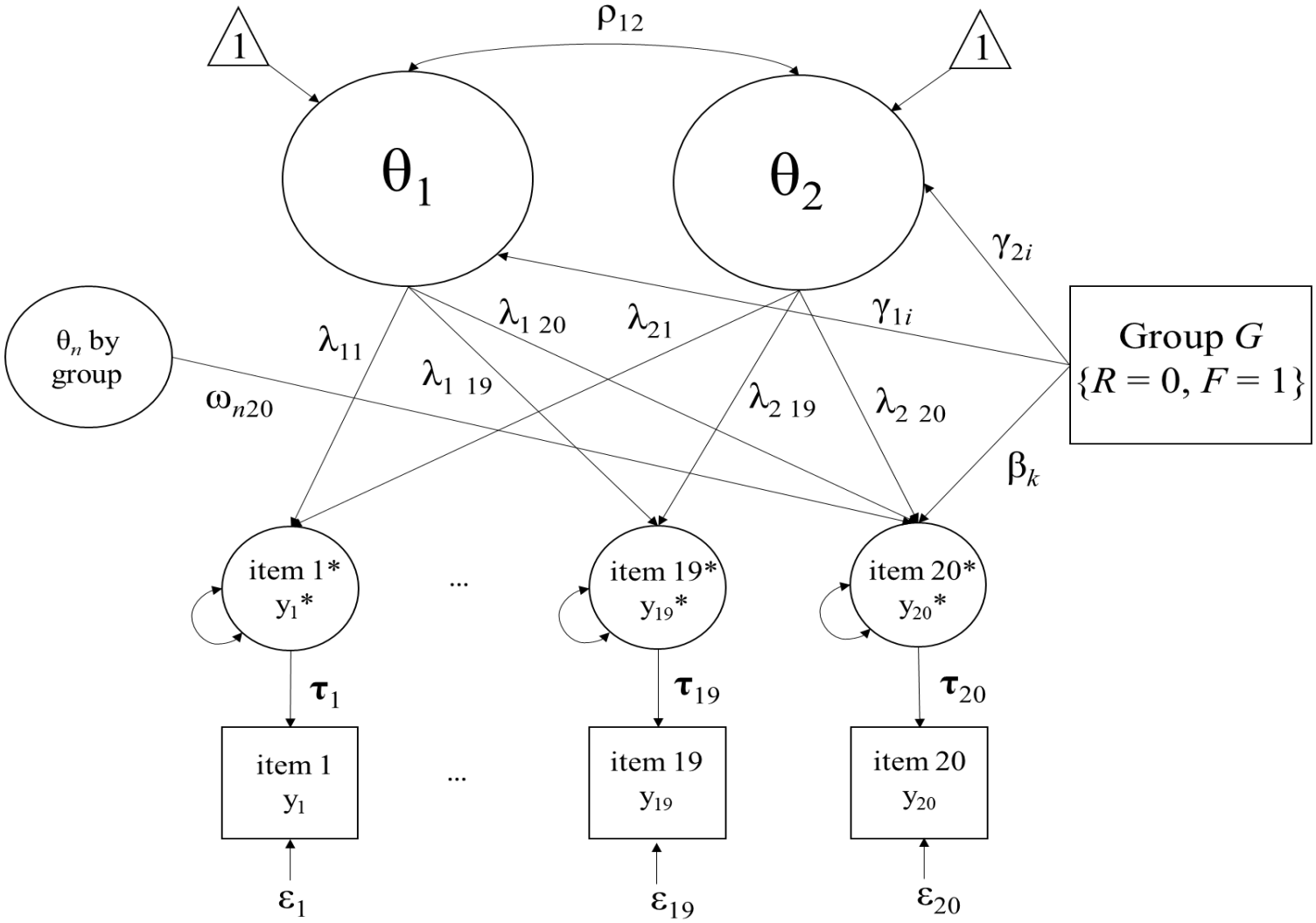

Figure 2.1. *Two-dimensional MIMIC-interaction model with 5% DIF and fixed variance*

the correlation between the latent abilities ρ, the uniform β_i and nonuniform $\boldsymbol{\omega}_i$ DIF effects for the studied item(s), and group mean differences γ_i were estimated. A continuous latent response y_i* underlied each categorical item response y_i, and the thresholds $\boldsymbol{\tau}_i$ represented the discrete responses for each categorical option in y_i.

The LMS method with a MLR estimator assisted in properly estimating the interaction term to examine nonuniform DIF. The LMS method assumed all variables in an interaction term and all associated error terms were normally distributed (Marsh et al., 2012; Maslowsky et al., 2015). Woods and Grimm (2011) acknowledged that the LMS method was not appropriate for interactions with categorical variables because they violate the normality assumption. However, Muthén and Muthén (1998-2017, p. 759) recommended a multiple group approach or the "XWITH" command in Mplus for interactions involving a categorical manifest variable (i.e., group membership) and continuous latent variable. Woods and Grimm (2011) noted that the "XWITH" command elicited a maximum likelihood approach with robust standard errors that was analogous to the LMS method; hence, this analysis used the "XWITH" command to address the violation of normality. A robust maximum likelihood estimation (i.e., MLR) and the LMS method estimated nonuniform DIF in the multidimensional MIMIC-interaction model using Mplus v. 8.5 (Muthén & Muthén, 1998-2017), while uniform DIF was tested with a MLR estimator using the `lavaan` package (Rosseel, 2012) in R v. 4.0.3 (R Core Team, 2020).

Multidimensional extension of the logistic discriminant function analysis. This study also examined a DIF procedure that accounts for the polytomous structure of the model using the logistic discriminant function analysis and the multidimensional structure of the model using the Mazor et al. (1998), and Bulut and Suh (2017) extension. The logistic discriminant function analysis rearranged the logistic regression model, such that the grouping variable *G* was the

dependent variable and the item response category Y was an independent variable (Miller & Spray, 1993). The logistic discriminant function analysis had a distinct advantage over the linear discriminant function analysis in situations where explanatory variables violate the multivariate normality assumption, such as the response category Y and the interaction effects $\theta_{np}Y$. Mazor et al. (1998), and Bulut and Suh (2017) extended the univariate logistic regression model to a multivariate framework to examine uniform and nonuniform DIF in the presence of multiple abilities, respectively, as described in equation 16. Application of both these principles to the logistic discriminant function analysis produced a multidimensional extension of the logistic discriminant function analysis and account for the multiple traits associated with the MGRM model. The new DIF method allowed for the evaluation of both uniform and nonuniform DIF for an item associated with a set of m latent abilities θ_{np} such that:

$$z = \pi_0 + \sum_{n=1}^{m}(\pi_n\theta_{np}) + \pi_{m+1}Y + \sum_{n=1}^{m}\pi_{m+n+1}(\theta_{np}Y) \quad . \tag{18}$$

When substituted in equation 10, this approach described the probability of belonging to a group G as a function of the item response category Y and the multiple ability parameters θ_{np}.

For the simulation study, the model contained two dimensions (m = 2), five response categories, and two groups, such that the probability that an item favors the reference group R was:

$$z_R = \pi_0 + \pi_1\theta_{1p} + \pi_2\theta_{2p} + \pi_3 Y + \pi_4(\theta_{1p}Y) + \pi_5(\theta_{2p}Y) \tag{19}$$

for some normally distributed, continuous latent traits θ_{np} and an ordered response category Y. The results from the multidimensional extension of the logistic discriminant function analysis indicated the probability that a category favored the reference group with respects to the latent

traits. Intrinsically, the analysis was a dichotomous multiple logistic regression model, where the criterion variable was group membership (i.e, focal and reference group). Therefore, the `glm` function in the R `stats` package estimated the logistic discriminant analysis parameters (R Core Team, 2020), where the observed score of each dimension for each person provided an estimate for the latent abilities θ_{np}.

A likelihood ratio chi-squared test (see equation 11) examined the null hypothesis that DIF was present with at least one item category. The likelihood test statistic, χ^2, approximately follows a chi-squared distribution with *df* equal to the difference in the number of terms between the two compared models. A two-step process, similar to the aforementioned MIRT-LR test procedure, tested for the significant difference at the category level of the item-scale using the `anova` function in the R `stats` package (R Core Team, 2020). First, a likelihood ratio chi-squared test with two degrees of freedom examined the presence of nonuniform DIF. The complex model was the full model found in equation 19. The nested model contained the first four terms of the model (i.e., π_0 through π_3). If the test results was statistically nonsignificant, then a test for uniform DIF was conducted. Under this circumstance, the complex model contained the first four terms, $z_R = \pi_0 + \pi_1\theta_{1p} + \pi_2\theta_{2p} + \pi_3 Y$, while the null model (i.e., the first three terms of equation 19) was the nested model. A significant test indicated the presence of uniform DIF, while a nonsignificant test meant that no DIF was present.

Evaluation Criteria

Error reduction is essential to improve the reliability of any statistical procedure. Typically, a significance level, the probability of rejecting a true null hypothesis, is declared, *a priori*, to establish a benchmark of acceptable error. Lower significant levels reduce the occurrence of type I error, but limit the probability of finding statistically significant results and likely increase the

probability of type II error (Martin & Roberts, 2010; Litière, 2010a). In contrast, type II error, the non-rejection of a false null hypothesis, detrimentally increases the probability of statistically significant results in cases where it should not (Coladarci et al., 2008). Therefore, rejection of a false null hypothesis also strengthens the reliability of results. In an effort to verify the strength and reliability of the three DIF detection methods, assessment of type I error rates from non-DIF conditions and rejection rates from uniform and DIF conditions was compared as the evaluation criteria.

Type I error rates. Type I error is the rejection of a true null hypothesis (Coladarci et al., 2008). Large type I error increases the probability of falsely accepting an incorrect outcome; thus, mitigating type I error was paramount to verify the reliability and validity of the DIF methods. Type I error rates identify the degree to which a procedure contains type I error (Litière, 2010b). In the context of this study, type I error rates represented the probability of DIF detection when no DIF was actually present in the studied item. Multiple hypothesis testing for each DIF method was conducted to measure the occurrence of false identification of DIF.

Type I error rates for the MIRT-LR test was computed as the proportion of significant likelihood ratio test statistics out of 100 replications in the non-DIF conditions. Statistically significant test results inferred false identification of DIF. The proportion of significant DIF parameter estimates represented the type I error rates for the MIMIC-interaction model. Items that exhibited either uniform or nonuniform DIF were considered falsely identified as DIF. The proportion of significant likelihood ratio tests represented the type I error rates for the multidimensional extension of the logistic discriminant function analysis. Datasets with 0% DIF and 24 factor constraint combinations (four sample sizes, three correlations between latent traits, and two mean differences between groups) were used to identify type I error rates for both

uniform and nonuniform DIF. Items 6, 12, 19, and 20 were tested for DIF using the values found in Table 2.1 with all other items serving as anchors. Items falsely identified as either uniform or nonuniform DIF represented type I errors. A significance level of $\alpha = 0.05$ was used to evaluate the three DIF detection methods.

Issues with multiple hypothesis testing frequently result in type I error inflation (Higdon, 2013). That is, large amounts of replicated hypothesis testing increases the likelihood of erroneous type I error detection. Inflated type I error abates the decision to reject the null hypothesis and weakens the results of a study (Litière, 2010b). Therefore, inflation of type I error needs to be considered and managed when multiple hypothesis testing is conducted. Similar DIF detection studies have observed inflated type I error rates when examining false identification of DIF (Bulut, & Suh, 2017; Lee et al., 2017; Magis & De Boeck, 2014; Woods & Grimm, 2011). Several adjustment techniques were available (e.g., Benjamini-Yekutieli, Bonferroni, Hochberg, Holm, and Hommel methods), but the procedure developed by Benjamini and Hochberg (1995) provided the least conservative approach. Thus, the Benjamini-Hochberg (BH) correction was used to control incorrect DIF detection and provided better estimates of type I error rates.

Rejection rates. Type II error, in contrast from type I error, is the non-rejection of a false null hypothesis (Coladarci et al., 2008). This form of error leads to acceptance of an incorrect statistical statement and result in weakened reliability. Inversely, statistical power is the rejection of a false null hypothesis (Darlington & Hayes, 2017), and as such, type II error decreases when power increases (Vo & James, 2010); thus, evincing the presence of high rejection rates of a false null hypothesis (i.e., power) indicates strong reliability of a given statistical method. In the context of this study, rejection rates (i.e., power) indicated the probability of DIF detection when DIF was actually present in the studied item.

Rejection rates for DIF conditions were calculated similarly to type I error rates, using the data sets generated with uniform and nonuniform DIF conditions. Rejection rates for the MIRT-LR test were computed as the proportion of significant likelihood ratio test statistics out of 100 replications in DIF conditions. Statistically significant test results inferred rejection of the false null hypothesis for DIF items. The proportion of significant DIF parameter estimated under DIF conditions represented the rejection rates for the MIMIC-interaction model. Items that exhibit either uniform or nonuniform DIF will be considered correctly identified as DIF. The proportion of statistically significant likelihood ratio tested for items under DIF conditions represented the rejection rates for the multidimensional extension of the logistic discriminant function analysis. Items correctly identified as either uniform or nonuniform DIF represented the statistical power.

Effect of factor constraints. Nine mixed-design analyses of variance (ANOVA) were used to examine the effects of factor constraints across DIF method treatments on the rejection and type I error rates. The dependent variable for the first set of four mixed-design ANOVAs was the type I error rates, while the dependent variable for the second set of five analyses was rejection rates. The different levels within the four sample sizes, three correlations between latent traits, two types of DIF, and two latent mean differences served as the between-subjects variable when type I error rate was the dependent variable. The different levels within the four sample sizes, three correlations between latent traits, two types of DIF, three proportions of DIF, and two latent mean differences served as the between-subjects variable when rejection rate was the dependent variable. The DIF detection methods and its interactions with each simulated factor functioned as the within-subjects variable. Interactions with more than one simulated factor were not important in addressing the research objectives and difficult to interpret; thus, only one between-subjects factor was tested at a time for nine total mixed-design ANOVAs. Omega-

squared ω^2 were computed as a measure of effect size for each ANOVA factor (Olejnik, & Algina, 2003). The `anova_test` function in the R `rstatix` package was used to conduct the mixed-design ANOVAs (Kassambara, 2020). Follow-up tests interpreted any statistically significant interaction effects or differences between groups in the between-subjects and within-subjects variables.

The mixed-design ANOVA required several assumption tests to reduce the likelihood of type I and II errors. That is, conditions of normality, no univariate outliers, homogeneity of variance, and sphericity of the covariance matrix must be satisfied before further analysis was conducted (Field, 2009; Murrar & Brauer, 2018). First, Q-Q plots determined the presence of univariate normality (Field, 2009). Violation of normality resulted in the use of a transformation of the data (e.g., logit transformation). Outliers were defined as data that exceed three median absolute deviation from the median (Leys et al., 2013). Outliers were neither be removed, nor altered, because the nature of the outliers helped to illustrate the behaviors of factor constraints on DIF. Homogeneity of variance for the between-subjects variables was assumed, because sample sizes between groups were equal (Tabachnick & Fidell, 2018). Mauchly's test assessed the sphericity of the covariance matrix assumption. The Greenhouse-Geisser correction was be applied if the estimated value of sphericity is less than .75. If the estimated value of sphericity was .75 or more, then the Huynh-Felt correction was applied in cases where the sphericity assumption was violated (Verma, 2016).

In the follow-up test, statistically significant interaction effects were interpreted first. If interaction effects were significant, then simple effects with a Bonferroni correction provided an interpretation of the effect of the between-subjects variable on each within-subjects group, as well as the effect of each within-in subjects on each between-subjects group, in the significant

interaction (Armstrong, 2014; Lee & Lee, 2018). Results from the simple effects analyses provided an interpretation of the interaction effect, while main effects were only reported and not interpreted. In addition, boxplots provided a graphical illustration of the significant interactions. Main effects for the between-subjects and within-subjects variables were interpreted in cases where the interactions were not statistically significant. The main effects interpreted differences between groups in each variable, while controlling for all other variables.

Empirical Example

An empirical example illustrated the findings in the simulated study with real-world data. The three DIF procedures used real-world survey data, assumed to fit a MGRM, and compared results with findings from the simulated study. The survey data were produced from responses to the Experience Learning post-experience student survey (Walker & Rocconi, 2021). The survey provided supportive evidence for the institutional continuous improvement initiative, Experience Learning. The initiative incorporated experiential learning into college-level courses in order to improve student performance and learning across several desirable attributes.

The Experience Learning initiative sought to enrich student learning in four specific areas: lifelong learning, application of knowledge and skills, collaboration, and structured reflection. The four interconnected Experience Learning student learning outcomes (SLOs) were stemmed from the Experience Learning mission statement, which called for "enhancing opportunities for students to learn through actual involvement with problems and needs in the larger community", and Kolb's (1984) experiential learning cycle. SLO 2 ("students will apply knowledge, values, and skills in solving real-world problems") and SLO 4 ("students will engage in structured reflection as part of the inquiry process") were aligned with the stages in Kolb's learning cycle. SLO 1, ("students will value the importance of engaged scholarship and lifelong learning") and

SLO 3 (“students will work collaboratively with others”) addressed missing components of Kolb’s theory, collaboration and lifelong learning (Bergsteiner & Avery, 2014; Bergsteiner et al., 2010).

Instrument. The Experience Learning post-experience survey measures a self-assessment of student achievement of the Experience Learning SLOs. The survey items, modified from the AAC&U VALUE rubrics (Rhodes, 2010), are aligned with and operationalize each SLO. The survey serves as an indirect measure of these learning outcomes such that a set of items are theoretically aligned with a one or more SLOs, as shown in Table 2.3. For example, the first survey item, related to lifelong learning, was constructed from language found in the “Foundations and Skills for Lifelong Learning” rubric and aligned with SLO 1. AAC&U rubric language was reconstructed into survey items with simpler language that was more suitable for undergraduate students to comprehend and answer. Survey items aligned with SLOs 2, 3, and 4 also used a similar method to reconstruct rubric language for survey use. In some cases, survey items were aligned with multiple SLOs based on exploratory analyses and alignment of the item language with other items under the concerning SLOs (e.g., items 4 and 16).

A panel of 25 experiential learning campus experts inspected the survey for content validity. Panelists suggested dropping or consolidating three items because they were considered to be redundant and to limit survey fatigue. Furthermore, the survey was pilot tested with a cohort of 80 students from five experiential learning courses. Results from an exploratory factor analysis bared strong evidence that survey items factored onto the anticipated latent trait (i.e., SLOs). Two changes occurred as a result of the pilot test. First, the rating scale was expanded from a 5-point Likert scale to a 7-point Likert scale (i.e., (1) *strongly disagree*, (2) *disagree*, (3) *somewhat disagree*, (4) *neither agree nor disagree*, (5) *somewhat agree*, (6) *agree*, (7) *strongly agree*) to

Table 2.3

Alignment of survey items with Experience Learning SLOs and AAC&U VALUE rubrics

Item No.	Survey Item Description	SLO	VALUE Rubric Item	VALUE Rubric
1	I am interested in continuing to explore the problems of society (i.e. the needs of others)	1	Show evidence of interest in the problems of society (needs of others)	Foundations and Skills for Lifelong Learning
2	I think it is important for the university to use its resources for the benefit of society	1	Value (i.e., offer a positive attitude toward) the use of engaged scholarship to address societal problems	Foundations and Skills for Lifelong Learning
3	I am interested in using the skills and knowledge that I have acquired from this course to contribute to the public good	1	Demonstrate a desire to utilize engaged scholarship	Civic Engagement
4	I want to continue to develop relevant skills that are related to this experience	1, 4	Demonstrate a commitment to lifelong learning	Foundations and Skills for Lifelong Learning
5	I can clearly describe a real-world problem related to this course to someone that knows little about the problem	2	Clearly describe a real-world problem amenable to engaged scholarship	Critical Thinking
6	I have been introduced to more than one way to address real world problem(s) that my faculty member/professor brought up in this course.	2	Analyze literature (content/research methods) related to the problem	Critical Thinking
7	I feel confident in my ability to develop a logical, consistent approach to address a real world problem related to this course	2	Formulate an inquiry approach driven by questions relevant to the problem	Critical Thinking
8	I can list many potential ethical issues for real world problems related to this course	2	Recognize potential ethical issues related to addressing the problem	Ethical Reasoning

Table 2.3 (continued).

Item No.	Pre-Experience Survey Items	SLO	VALUE Rubric Items	VALUE Rubric
9	I can draw conclusions from data collected through this experience	2	Employ the selected inquiry approach • Collect and analyze data • Draw conclusions/ inferences (interpret)	Inquiry and Analysis
10	I am able to identify and apply information from this course to address and potentially improve real world problem(s)	2	Apply findings toward addressing the problem	Global Learning
11	My classmates would say that I often listened to and respected the ideas of others	3	Participate in collaborative interactions; Support group processes; Be attentive to the ideas of others	Teamwork
12	My classmates would say that I was able to offer relevant questions and comments within a group setting	3	Participate in collaborative interactions; Support group processes; Offer relevant questions and comment	Teamwork and Civic Engagement
13	I met obligations for group assignments on a timely basis	3	Support group processes; Meet obligations for group assignments on a timely basis	Teamwork
14	I purposefully reflected on what I learned from problems I encountered during this experience	4	Use structured reflection in assessing an engaged inquiry experience; Use reflection on the inquiry process to guide lifelong learning	Integrative Learning
15	During this experience, I reflected on what I learned about myself from this experience	4	Assess what they have learned about themselves as an individual (self-awareness) from experiences; Use reflection on the inquiry process to guide lifelong learning	Integrative Learning
16	During this experience, I thought about what it means to be a member of the broader community	1, 4	Assess what they have learned about themselves as members of the broader community	Integrative Learning

better examine variability between responses and to mitigate a ceiling effect. Second, items 4 and 16 were assumed to be align with SLO 1 and 4 because of factor loading results and similarity between language in the items with other items in SLO 1 and 4.

Model fit procedures. The survey data were assumed to fit to a multidimensional graded response model; however, the data could be fit to various plausible models (i.e., the generalized partial credit model (GPCM), the one-parameter multidimensional graded response model (1-PL MGRM), GRM, and MGRM). A comparison between each plausible polytomous IRT model confirmed the relationship between the survey items and the Experience Learning SLOs. First, the data set was fitted to each polytomous IRT model using the R `mirt` package (Chalmers, 2012). The quasi-Monte Carlo EM method estimated item parameters for all IRT models, because the estimator handled high dimensions well and converged the model quickly (Asmussen & Glynn, 2007; Sloan & Woźniakowski, 1998). Next, fit indices, M_2^* statistic[6], comparative fit index (CFI), Tucker-Lewis index (TLI), and the root mean square error of approximation based on the M_2 statistic ($RMSEA_2$), evaluated and compared the MGRM, GPCM, and 1-PL MGRM using the `M2*` function in the R `mirt` package (Cai & Hansen, 2013; Chalmers, 2012). More specifically, the following guidelines assessed model fit: CFI and TLI $\geq$ .90 for adequate fit and $\geq$ .95 for good fit; and $RMSEA_2$ $\leq$ .08 for adequate fit and $\leq$ .05 for good fit (Brown, 2015; Gana & Broc, 2019; Maydeu-Olivares, 2013). Lastly, the `anova` function in the R `mirt` package compared the GRM and MGRM (Chalmers, 2012). The MGRM was a generalization of the GRM (Reckase, 2009); therefore, comparison of the deviances between the two models sufficed in order to determine the most parsimonious model.

[6] The M_2 statistic follows a chi-square distribution, but provides better type I error rate control than full-information tests (e.g., Pearson's χ^2, likelihood ratio test) when data are sparse (Maydeu-Olivares & Joe, 2005). The M_2^* statistic has the added benefit of better calibration than the M_2 statistic when tests are long and items are polytomous (Cai & Hansen, 2013).

Analytical procedures. The empirical example used the same DIF detection procedures from the simulation study. For all DIF detection procedures, female respondents served as the reference group, because they were the majority in the experiential learning courses, while male respondents were classified as the focal group. Anchor items were needed to properly identify the MIMIC-interaction, MLDFA, and MIRT-LR models. An item purification procedure identified anchor items in the empirical data (Woods, 2009b). Item purification, as a means to identify anchor items, reduces type I error rates (Navas-Ara & Gómez-Benito, 2002). Woods (2009b) suggested an empirical, brute-force selection of anchor items by testing each item, sequentially, for DIF with the omnibus MIRT-LR test, treating all other items as anchors. Each item were ranked-ordered based on the ratio of the MIRT-LR test statistic and degrees of freedom, χ^2/df, from smallest to largest. The three items with the smallest χ^2/df value were selected as anchors and rest of the items were individually tested for DIF using the designated anchors.

The empirical example demonstrated how selection of DIF statistical procedures can influence the conclusions made on DIF detection in multidimensional items. Items in which DIF detection procedures detected DIF were examined and compared to the simulated study. As such, the *p*-values for each method were reported for both uniform and nonuniform DIF detection. The findings assisted in the evaluation and comparison between DIF detection procedures. In addition, the discrimination and threshold parameter estimates for both groups were provided as supplemental information to help with the assessment. Results from the simulation study were compared to the parameter estimates in the empirical study to illustrate how DIF statistical procedures can detect DIF differently in multidimensional items.

Chapter III

Results

A Monte Carlo simulation produced data fitted to the multidimensional graded response model to study the optimal DIF detection method under various constraints and examined how different constraints affect DIF detection. The ratio of identified false positives from the replicated data sets under different factor constraints measured type I error, while the ratio of identified true positives from the replicated data sets measured the rejection rates. Type I error rate is defined as the proportion of significant test statistics out of 100 replications in the non-DIF conditions. Rejection rate is the proportion of significant test statistics out of 100 replications in DIF conditions. The alpha criterion for statistical significance for type I error and rejection rates was .05. Mixed-design analyses of variance (ANOVA) examined the effects of different factor constraints across different DIF detection method treatments on the rejection and type I error rates. The results from the type I error rates, rejection rates, and the mixed-design ANOVAs specifically address the research objectives. In addition, an empirical example from an experiential learning survey (Walker & Rocconi, 2021), assumed to fit the MGRM framework, compared the DIF detection procedures to reinforce the simulation findings.

Type I Error Rates

Different combinations of factor constraints were measured on their type I error rates over three DIF detection methods. The factor constraints included DIF type, latent mean differences across groups, correlation between latent traits, and sample size. There were two groups of DIF type (uniform and nonuniform DIF), two levels of latent mean differences (balanced [$\mu_R = 0$, $\mu_F = 0$] and unbalanced [$\mu_R = .5$, $\mu_F = -.5$] latent means), three correlation levels ($\rho = 0$, $\rho = .3$, and ρ

= .6), and four groups of sample sizes (R1000/F1000, R1000/F500, R500/F500, R500/F250). Factor constraints were measured across three different DIF detection methods: multidimensional IRT likelihood ratio test (MIRT-LR), multidimensional MIMIC-interaction model (MIMIC), and the multidimensional extension of the logistic discriminant function analysis (MLDFA).

Overall, the results from the MIRT-LR (*M* = .13, *SD* = .133), MIMIC (*M* = .06, *SD* = 0.028), and MLDFA (*M* = .10, *SD* = .023) tests showed that average type I error rates between uniform and nonuniform DIF differed across the three different methods. The MIRT-LR test wildly differed across different factor constraints. In particular, the test produced over 40% significant tests when no uniform DIF was present under unbalanced latent mean differences and large, balanced sample size (R1000/F1000), regardless of correlation size. Conversely, the test produced only one significant test when no nonuniform DIF was present under balanced latent mean differences, a medium correlation (ρ = .3), and large, unbalanced sample size (R1000/F500). The MLDFA test produced consistent type I error rates between the uniform and nonuniform DIF detection and the various factor constrain combinations. The type I error rate for the MLDFA method ranged from .06 to .14 across uniform and nonuniform DIF detection. The MIMIC model varied more than the MLDFA method (.01 to .16), but produced the lowest type I error rate in two-thirds of all factor constraint combinations.

Uniform DIF. The results from Figure 3.1 indicated that type I error rates for the MIRT-LR (*M* = .182, *SD* = .170), MIMIC (*M* = .046, *SD* = .018), and MLDFA (*M* = .101, *SD* = .022) tests vary across the three different methods when detecting uniform DIF. Other patterns were less obvious and pronounced when detecting uniform DIF. In general, the proportion of statistically significant replications remained below .15 when detecting uniform DIF, except for the MIRT-

LR test. Figure 3.1 indicated that type I error rates were generally higher for larger sample sizes when mean differences across groups were unbalanced (i.e., $\mu_R = .5$, $\mu_F = -.5$) and using the MIRT-LR test to assess uniform DIF. The MIRT-LR test for uniform DIF produced consistently higher type I error rates than the other two methods when latent mean differences between the reference and focal groups were unbalanced, but generally performed better than the MLDFA method under balanced latent mean differences.

In general, the multidimensional MIMIC-interaction model produced relatively small type I error rates when attempting to detect uniform DIF. The type I error rates of the MIMIC model did not exceed .08, nor did the upper limit of the 95% confidence interval surpass .12 under any condition. No consistent pattern of type I error rates was evident across sample sizes, correlation between latent means, or latent mean differences for the MIMIC model when detecting uniform DIF. In fact, the method produced lower type I error rates than the other two methods under all but four factor constraint combinations.

The multidimensional extension of the logistic discriminant function analysis (MLDFA) produced the highest type I error rates for uniform DIF detection and balanced latent mean differences, except when correlations between latent means and sample sizes were large, $\rho = .6$ and R1000/F1000, respectively. Type I error rates remained around .10 under all factor constraint combinations for the MLDFA method when detecting uniform DIF, ranging from .06 to .15. Overall, the MLDFA kept consistent type I error rates across all factor constraints when testing the non-detection of uniform DIF in non-DIF items.

Nonuniform DIF. Results for nonuniform DIF detection, as shown in Figure 3.2, also indicated that type I error rates also vary between the three DIF detection methods, MIRT-LR (*M* = .068, *SD* = .024), MIMIC (*M* = .080, *SD* = .027), and MLDFA (*M* = .106, *SD* = .024). Figure

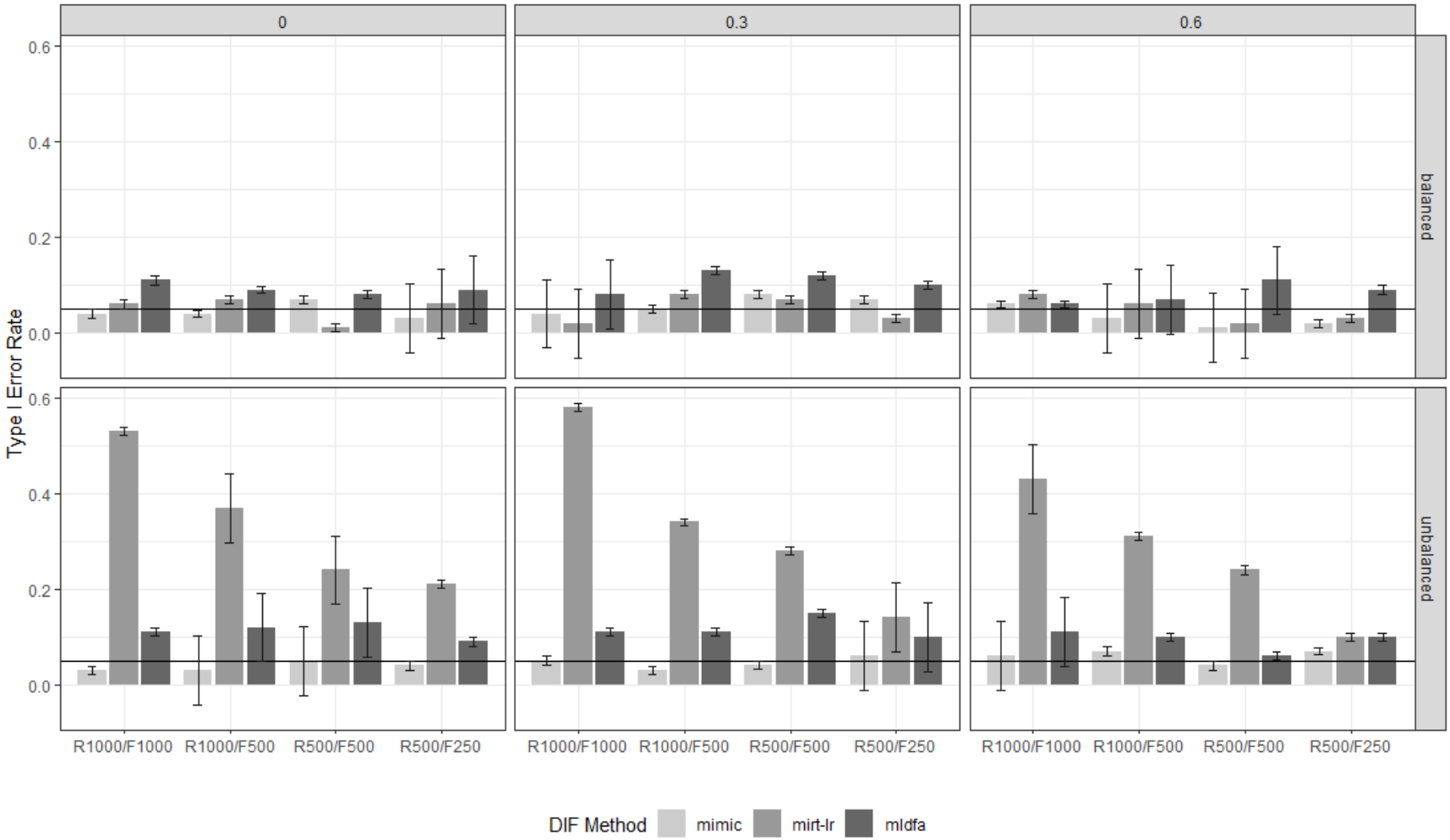

Figure 3.1. *Type I error rates for uniform DIF results with 95% confidence intervals and α = .05 reference line*

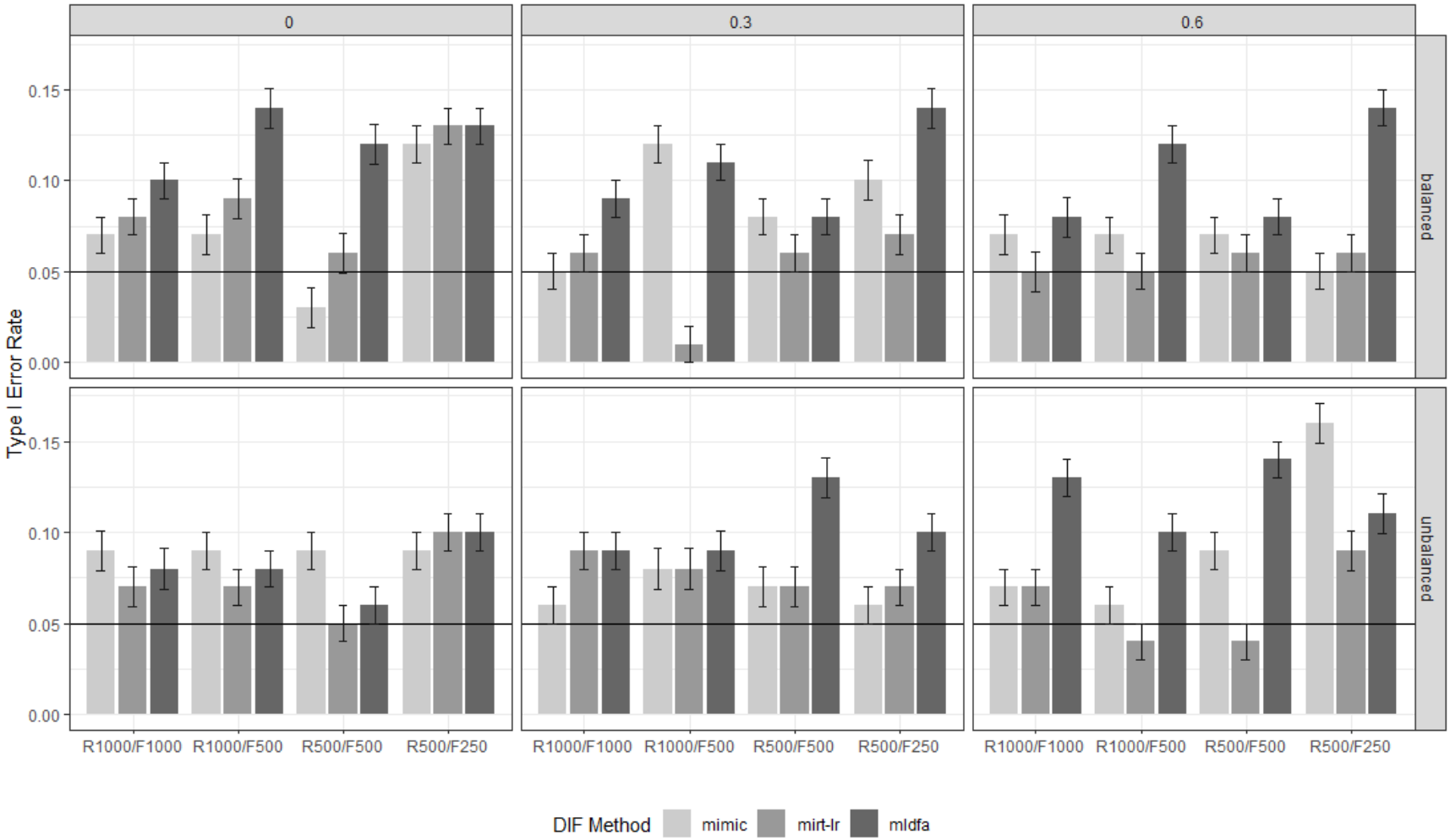

Figure 3.2. *Type I error rates for nonuniform DIF results with 95% confidence intervals and α = .05 reference line*

3.2 indicated that type I error rates for nonuniform DIF detection with the MIRT-LR test were comparatively lower than the MLDFA method, with a few exceptions, under the same factor constraints. Additionally, the MIRT-LR test also produced lower or equivalent type I error rates when compared with the MIMIC model in 62.5% of the constraint combinations. However, there was no distinct pattern across sample sizes, latent correlations, or latent mean difference that suggested the MIRT-LR produced lower type I error rates under specific factor constraint combinations. Interestingly, the MIRT-LR test produced a substantially lower type I error rate, .01, with balanced latent mean differences, medium correlation ($\rho = .3$), and large, unbalanced sample sizes (R1000/F500) compared to the other two methods and across other factor constraint combinations.

The multidimensional MIMIC-interaction model produced lower type I error rates than the other two methods when attempting to detect nonuniform DIF under no correlation and balanced latent mean differences. No consistent pattern of type I error rates was evident across sample sizes, correlation between latent means, or latent mean differences for the MIMIC model when detecting nonuniform DIF. The model, however, did produce a high type I error rate, .16, when correlations were high ($\rho = .6$), latent mean differences were unbalanced, and sample size was small and unbalanced (R500/F250).

The MLDFA method produced several high type I error rates when detecting nonuniform DIF. In particular, type I error rates exceed .10 in 62.5% of constraint combinations and ranged from .06 to .14. The MLDFA method produced higher type I error rates than the other two methods in all but five constraint combinations. In fact, the MLDFA method never produced the lowest type I error rate for any constraint combination when detecting nonuniform DIF. However, the MLDFA did produce lower type I error rates than the MIMIC model when

detecting nonuniform DIF under no latent correlation and unbalanced latent mean differences, with the exception of small, unbalanced sample sizes (R500/F250).

Effect of factor constraints on type I error. The effect of the factor constraints and DIF detection methods on type I error rates were assessed with multiple mixed-design analyses of variance (ANOVA). More specifically, four two-way mixed-design ANOVAs allowed for examination of one factor constraint at a time. This approach allowed one to examine each factor constraint's effect on type I error rates separately without adding the complexity of interpreting high-level interaction effects. Therefore, this study looked at the interaction effects between one factor constraint and the DIF detection methods to determine the effect on type I error rates. While this method simplified the process, multiple hypothesis testing on the same dependent variable increased the likelihood of making a type I error (Darlington & Hayes, 2017; Higdon, 2013). A Bonferroni correction was used to protect against the chance of a type I error occurring when testing the interaction effects (Frane, 2015); thus, the alpha criterion for significance testing was lowered to .0125.

The mixed-design ANOVAs assumed multivariate normality. To assess this assumption, Q-Q plots of the type I error rates indicated extreme deviations in the tails and 10 observations were at least three median absolute deviations from the median, indicating issues with univariate normality. A scatterplot of the data by DIF detection method, coupled with the kernel density estimates of each DIF detection method, showed that data were clustered near and bounded by 0 (Figure B.1 in Appendix B). Furthermore, the logistic probability density function mapped on the kernel density distribution of the data indicated that the data are non-normal (Figure B.2 in Appendix B).

Consequently, a logit transformation was determined to be an appropriate adjustment of the type I error rates. A logit transformation was favored, because all factor constraints were discrete and logit transformations on proportional data were practical to interpret through odds and odds ratios (Osbourne, 2017). The logit function also produced a near linear transformation of the data by elongating the scaled difference between the high bounded values and between the low bounded values. Application of the logit transformation resulted in better-fitted Q-Q plots and fewer troublesome outliers.

Outliers were neither removed, nor altered, before analysis, because the values are of substantive interest to the research objectives. More specifically, all nine outliers were produced from three specific factor constraint combinations (i.e., uniform DIF, unbalanced latent mean differences, R1000/F1000; uniform DIF, unbalanced latent mean differences, R1000/F500; uniform DIF, unbalanced latent mean differences, R500/F500). Adjustments or removal of the outliers could produce inauthentic conclusions on the nature of the DIF detection methods; therefore, it was determined that these values remain unaltered to investigate the true effect of the factor constraints and DIF detection methods.

Effect of DIF type. This study first explored the effect of the DIF type (uniform, nonuniform) and DIF detection method (MIRT-LR, MLDFA, and MIMIC-interaction model) on the log-odds of type I error rates. The interaction between the DIF type and method was examined using a 3 (DIF detection method) x 2 (DIF type) mixed design ANOVA, where the DIF type was the between-subjects variable and DIF detection method was the within-subjects variable. A test of sphericity on repeated measures was performed before conducting the mixed-design ANOVA to reduce the likelihood of type I errors. Mauchly's test on the within-subjects ANOVA interaction was found to violate the sphericity assumption, $\chi^2(2) = 37.00$, $\varepsilon = .641$, $p <$

.001. A Greenhouse-Geisser correction was used to adjust the degrees of freedom to account for the lack of sphericity and limit the possibility of an overly inflated test statistic in the ANOVA test (Field, 2009). A Greenhouse-Geisser correction is the preferred method when the epsilon statistic is below .75 (Verma, 2016).

The ANOVA interaction was found to be significant, $F(1.28, 58.95) = 6.93$, $p < .001$, $\omega^2 = .273$, such that log-odds of type I error rates depends on the combination of the DIF type and the DIF detection method (see Figure 3.3A). In addition, the main effect for DIF detection method was significant, $F(1.28, 58.95) = 12.41$, $p < .001$, $\omega^2 = .234$, but the main effect for DIF type was not statistically significant, $F(1, 46) = .49$, $p = .486$, $\omega^2 = -.009$. A simple effects analysis on the between-subjects effect for the MIRT-LR test was found to be significant, $F(1, 46) = 8.00$, $p_{adj} = .021$, $\omega^2 = .127$. The results indicate significant logit-adjusted mean differences for the MIRT-LR test between uniform DIF ($M = -1.71$, $SD = 1.070$) and nonuniform DIF ($M = -2.36$, $SD = .304$). In terms of odds, the mean odds of committing a type I error for uniform DIF was roughly 1 for every 5.53 correct non-rejection of the null hypothesis, while the mean odds of making a type I error for nonuniform DIF was about 1 for every 10.59 correct non-rejection of the null hypothesis.

Similarly, a simple effects analysis on the between-subjects effect for the MIMIC-interaction model was statistically significant, $F(1, 46) = 26.90$, $p_{adj} < .001$, $\omega^2 = .350$, indicating differences of logit-adjusted means existed between the uniform DIF ($M = -2.64$, $SD = .281$) and nonuniform DIF ($M = -2.22$, $SD = .274$) for the MIMIC-interaction model. The mean odds for type I error to occur with uniform DIF was roughly 1 for every 14.01 correct non-rejection of the null hypothesis, while the mean odds of a type I error occurring for nonuniform DIF was about 1 for every 9.21 correct non-rejection of the null hypothesis.

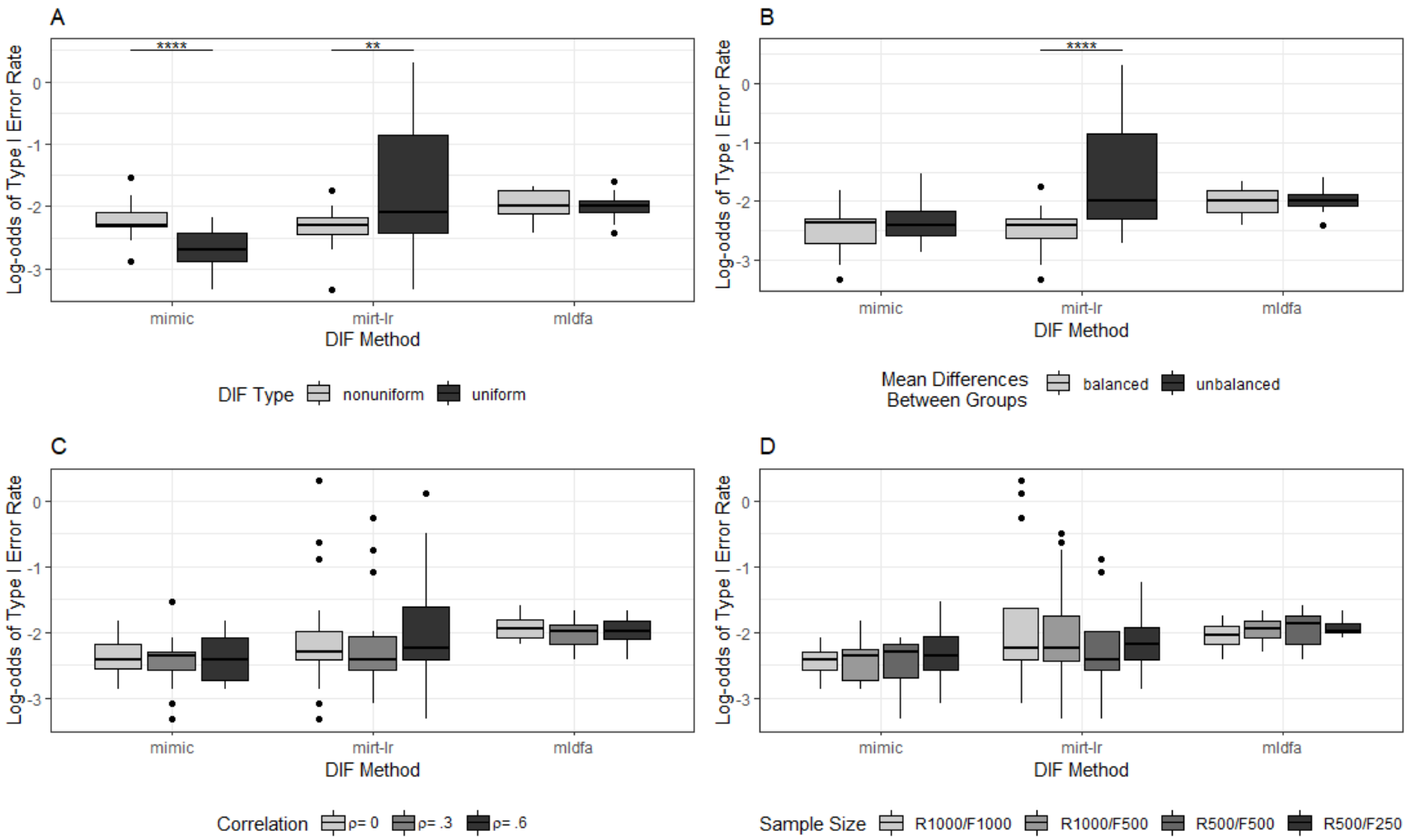

Figure 3.3. *Boxplot of interactions and significant simple main effects of factor constraints and DIF detection method for log-odds of type I error rate*

****$p_{adj} < .0001$, **$p_{adj} < .01$

Finally, a simple effects analysis on the between-subjects effect for the MLDFA method was found to be not statistically significant, $F(1, 46) = .51$, $p_{adj} = .999$, $\omega^2 = -.010$. Thus, the logit-adjusted mean differences between the uniform DIF ($M = -2.00$, $SD = .201$) and nonuniform DIF ($M = -1.96$, $SD = .211$) for the MLDFA method is not significantly different. The mean odds of committing a type I error for uniform DIF was roughly 1 for every 7.39 correct non-rejection of the null hypothesis, while the mean odds of making a type I error for nonuniform DIF was about 1 for every 7.10 correct non-rejection of the null hypothesis.

A simple effects analysis on the within-subjects effects for uniform DIF was significant, $F(1.15, 26.53) = 13.53$, $p_{adj} = .003$, $\omega^2 = .324$. The results indicated that the logit-adjusted means for uniform DIF was significantly lower for the MIMIC model ($p_{adj} < .001$) than the MIRT-LR test. The logit-adjusted means for uniform DIF was also significantly lower for the MIMIC model ($p_{adj} = .004$) than the MLDFA method, but no significant logit-adjusted means difference was present between the MIRT-LR test ($p_{adj} = .401$) and the MLDFA method. A simple effects analysis on the within-subjects effects for nonuniform DIF was significant, $F(2, 46) = 14.86$, $p_{adj} < .001$, $\omega^2 = .273$. The results indicated that the logit-adjusted means for nonuniform DIF was significantly lower for the MIMC-interaction model ($p_{adj} = .003$) than the MLDFA method. The logit-adjusted means for nonuniform DIF was also significantly lower for the MIRTL-LR test ($p_{adj} < .001$) than the MLDFA method, but no there was no significant logit-adjusted means difference between the MIMIC-interaction model ($p_{adj} = .223$) and the MIRT-LR test.

Effect of mean differences between groups. This study next explored the effect of the mean differences between groups (balanced [$\mu_R = 0$, $\mu_F = 0$], unbalanced [$\mu_R = .5$, $\mu_F = -.5$]) and DIF detection method on the log-odds of type I error rates. A 3 (DIF detection method) x 2 (mean differences between groups) mixed design ANOVA examined the interaction between the mean

differences and DIF detection method, where the mean differences between groups was the between-subjects variable and DIF detection method was the within-subjects variable. Before conducting the mixed-design ANOVA, sphericity on repeated measures was tested to reduce the likelihood of type I error. Mauchly's test on the within-subjects ANOVA interaction was found to be in violation of sphericity, $\chi^2(2) = 39.45$, $\varepsilon = .631$, $p < .001$, and a Greenhouse-Geisser adjustment was applied to the ANOVA test.

The interaction effect was statistically significant, $F(1.26, 58.09) = 13.80$, $p < .001$, $\omega^2 = .252$, such that log-odds of type I error rates depends on the combination of the mean differences between groups and the DIF detection method (see Figure 3.3B). In addition, the main effect for DIF detection method was significant, $F(1.26, 58.09) = 12.16$, $p < .001$, $\omega^2 = .227$. The main effect for the logit-adjusted mean differences between groups was also found to be significant, $F(1, 46) = 23.61$, $p < .001$, $\omega^2 = .233$. A simple effects analysis on the between-subjects effect for the MIRT-LR test was found to be significant, $F(1, 46) = 21.90$, $p_{adj} < .001$, $\omega^2 = .303$. The results indicate significant differences of logit-adjusted means between balanced latent means ($M = -2.51$, $SD = .393$) and unbalanced latent means ($M = -1.56$, $SD = .914$) for the MIRT-LR test. In connection with odds, the mean odds for incorrectly rejecting the null hypothesis with balanced latent means was roughly 1 for every 12.30 correct non-rejection of the null hypothesis, while the mean odds of incorrectly rejecting the null hypothesis for unbalanced latent means was about 1 for every 4.76 correct non-rejection of the null hypothesis.

Conversely, a simple effects analysis on the between-subjects effect for the MIMIC-interaction model was found to be nonsignificant, $F(1, 46) = .073$, $p_{adj} = .999$, $\omega^2 = -.006$. Therefore, there was not a significant logit-adjusted mean difference between the balanced latent means ($M = -2.47$, $SD = .374$) and unbalanced latent means ($M = -2.38$, $SD = .317$) for the

MIMIC-interaction model. The mean odds for type I error to occur with balanced latent means was roughly 1 for every 11.82 correct non-rejection of the null hypothesis, while the mean odds of a type I error occurring for unbalanced latent means was about 1 for every 10.80 correct non-rejection of the null hypothesis.

Lastly, a simple effects analysis on the between-subjects effect for the MLDFA method was also found to be nonsignificant, $F(1, 46) = .07$, $p_{adj} = .999$, $\omega^2 = -.020$; thus, no significant differences of logit-adjusted means exists between the balanced latent means ($M = -1.99$, $SD = .211$) and unbalanced latent means ($M = -1.97$, $SD = .203$) for the MLDFA method. The mean odds for incorrectly rejecting the null hypothesis with balanced latent means was roughly 1 for every 7.32 correct non-rejection of the null hypothesis, while the mean odds of incorrectly rejecting the null hypothesis for unbalanced latent means was about 1 for every 7.17 correct non-rejection of the null hypothesis.

A simple effects analysis on the within-subjects effects for balanced latent mean differences was significant, $F(2, 46) = 22.84$, $p_{adj} < .001$, $\omega^2 = .330$. The results indicated that the logit-adjusted means for balanced latent mean differences was significantly lower for the MIRT-LR test ($p_{adj} < .001$) than the MLDFA method. The logit-adjusted means for balanced latent mean differences was also significantly lower for the MIMIC-interaction model ($p_{adj} < .001$) than the MLDFA method, but no significant logit-adjusted means difference was present between the MIMIC-interaction model ($p_{adj} = .999$) and the MIRT-LR test. A simple effects analysis on the within-subjects effects for unbalanced latent mean differences was significant, $F(1.12, 25.87) = 10.67$, $p_{adj} = .008$, $\omega^2 = .251$. The results indicated that the logit-adjusted means for unbalanced latent mean differences was significantly lower for the MIMC-interaction model ($p_{adj} < .001$) than the MIRT-LR test. The logit-adjusted means for the unbalanced latent mean differences was

also significantly lower for the MLDFA method (p_{adj} = .047) than the MIRT-LR test, as was the logit-adjusted means difference for the MIMIC-interaction model (p_{adj} = .043) compared to the MLDFA method.

Effect of correlation between latent traits. The effect of the correlation between latent traits ($\rho = 0$, $\rho = .3$, $\rho = .6$) and DIF detection method on the log-odds of type I error rates were examined using a 3 (DIF detection method) x 3 (correlation between latent traits) mixed design ANOVA. The correlation between latent traits is the between-subjects variable and DIF detection method is the within-subjects variable. Mauchly's test of sphericity on the within-subjects ANOVA interaction was found to be in violation, $\chi^2(2) = 41.87$, $\varepsilon = .620$, $p < .001$, and a Greenhouse-Geisser correction was added to adjust the degrees of freedom.

The results from the ANOVA interaction effect was found to be nonsignificant, $F(2.48, 55.77) = .233$, $p = .837$, $\omega^2 = -.041$, such that log-odds of type I error rates did not depend on the combination of the correlation between latent traits and the DIF detection method (see Figure 3.3C). The main effect for DIF detection method was significant, $F(1.24, 55.77) = 9.24$, $p = .002$, $\omega^2 = .176$. The results of a Bonferroni corrected pairwise *t*-test found that significant differences of logit-adjusted means existed between the MIRT-LR test ($M = -2.04$, $SD = .845$, $p_{adj} = .024$) and MIMIC-interaction model ($M = -2.43$, $SD = .346$) when controlling for the correlation between latent traits. These post-hoc results were also applicable to the DIF detection method main effects post-hoc results for the effects of sample size on the type I error rates. The results from another Bonferroni corrected pairwise *t*-test also found that significant differences of logit-adjusted means existed between MIMIC-interaction model ($p_{adj} < .001$) and the MLDFA method ($M = -1.98$, $SD = .205$), but no significant differences of logit-adjusted means exists

between the MIRT-LR test (p_{adj} = .999) and the MLDFA method. Figure 3.4 illustrated the post-hoc results for the DIF detection method main effects on the type I error rates.

In regards to odds, the mean odds for type I error to occur with the MIRT-LR test was roughly 1 for every 7.69 correct non-rejection of the null hypothesis, regardless of correlation size. The mean odds of a type I error occurring for MIMIC-interaction model was about 1 for every 11.36 correct non-rejection of the null hypothesis, while the mean odds of a type I error occurring for MLDFA method was about 1 for every 7.24 correct non-rejection of the null hypothesis when controlling for the correlation between latent traits.

The main effect for the correlation between latent traits was found not to be significant, $F(2, 45) = .391$, $p = .679$, $\omega^2 = -.018$. Therefore, the logit-adjusted means for no correlation between latent traits ($M = -2.12$, $SD = .586$), the correlation of .3 between latent traits ($M = -2.12$, $SD = .590$), and the correlation of .6 between latent traits ($M = -2.20$, $SD = .549$) were likely to produce similar type I error rates when controlling for DIF detection method. Thus, the mean odds for incorrectly rejecting the null hypothesis when there was no correlation between the latent traits was roughly 1 for every 8.33 correct non-rejection of the null hypothesis regardless of DIF detection method. The mean odds of incorrectly rejecting the null hypothesis for a correlation between latent traits of .3 was about 1 for every 8.33 correct non-rejection of the null hypothesis. The mean odds of incorrectly rejecting the null hypothesis for a correlation of .6 was about 1 for every 9.03 correct non-rejection of the null hypothesis when controlling for the DIF detection method.

Effect of sample size. Finally, this study explored the effect of the sample size (R1000/F1000, R1000/F500, R500/F500, R500/F250) and DIF detection method on the log-odds of type I error rates. The interaction between sample size and method was examined using a 3

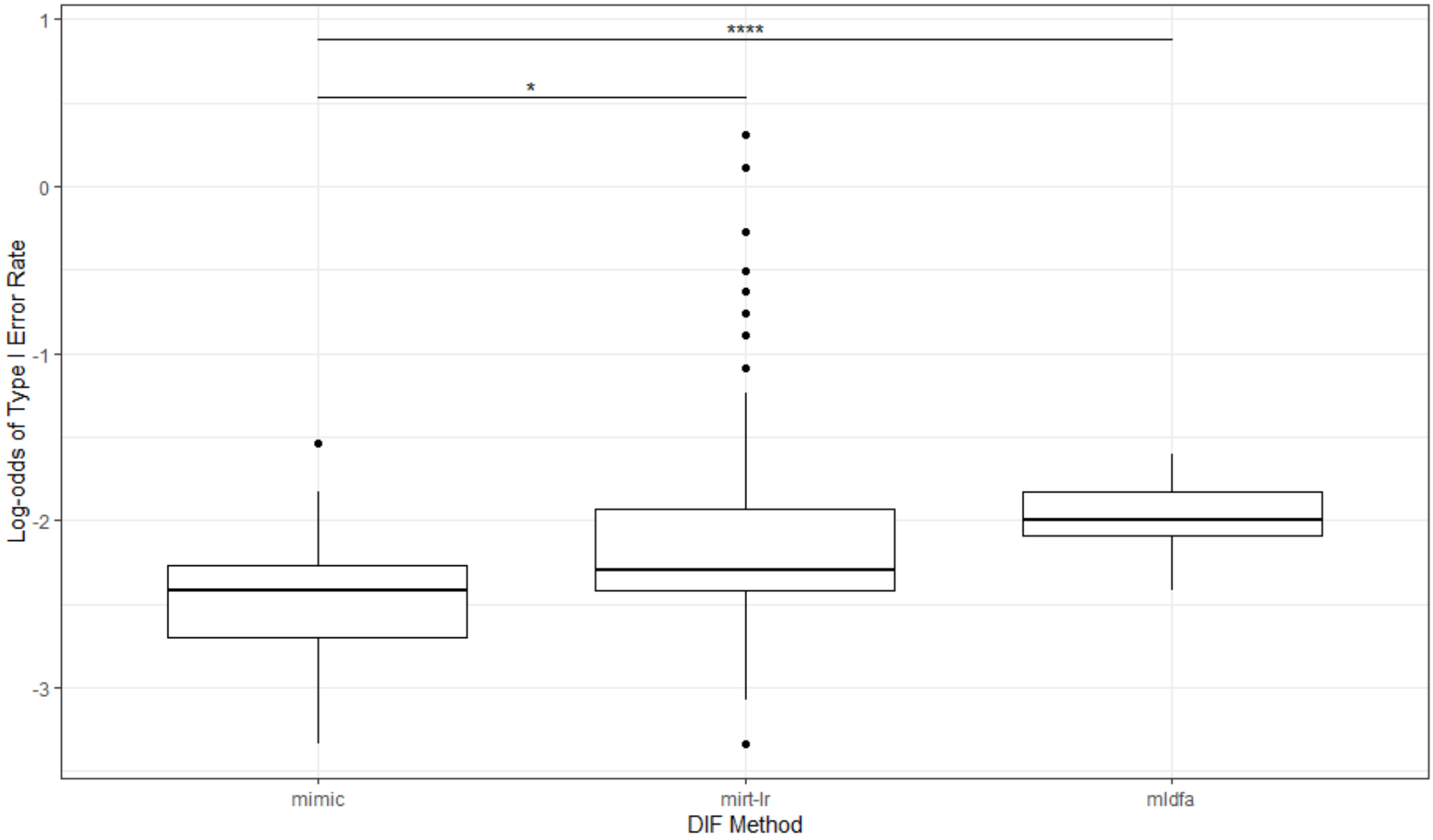

Figure 3.4. *Boxplot of DIF detection method main effects for log-odds of type I error rate*

****$p_{adj} < .0001$, *$p_{adj} < .05$

(DIF detection method) x 4 (sample size) mixed design ANOVA, where the sample size was the between-subjects variable and DIF detection method was the within-subjects variable. Mauchly's test on the within-subjects ANOVA interaction was found to be in violation of sphericity, $\chi^2(2) = 40.47$, $\varepsilon = .621$, $p < .001$; thus, a Greenhouse-Geisser correction was added to the analysis.

The ANOVA interaction effect was found to be nonsignificant, $F(3.73, 54.66) = .702$, $p = .584$, $\omega^2 = -.024$, such that log-odds of type I error rates did not depend on the combination of the sample size and the DIF detection method (see Figure 3.3D). The main effect for DIF detection method was significant, $F(1.24, 54.66) = 9.37$, $p = .002$, $\omega^2 = .178$ (see Figure 3.4), but the main effect for sample size was found not to be significant, $F(3, 44) = .328$, $p = .805$, $\omega^2 = -.031$. Therefore, sample size constraints, R1000/F1000 ($M = -2.10$, $SD = .718$), R1000/F500 ($M = -2.13$, $SD = .586$), R500/F500 ($M = -2.22$, $SD = .553$), and R500/F250 ($M = -2.14$, $SD = .409$), are likely to produce similar type I error rates when controlling for DIF detection method.

In terms of odds, the mean odds for type I error to occur with a large, balanced sample size of R1000/F1000 was roughly 1 for every 8.17 correct non-rejection of the null hypothesis regardless of DIF detection method. The mean odds of a type I error occurring for a large, unbalanced sample size of R1000/F500 was about 1 for every 8.41 correct non-rejection of the null hypothesis. The mean odds of a type I error occurring for a small, balanced sample size of R500/F500 was about 1 for every 9.21 correct non-rejection of the null hypothesis. The mean odds of a type I error occurring for a small, unbalanced sample size of R500/F250 was about 1 for every 8.50 correct non-rejection of the null hypothesis, when controlling for DIF detection method.

Rejection Rates

Different combinations of factor constraints were measured on their rejection rates (i.e., power) over three DIF detection methods. The factor constraints included the percentage of DIF, DIF type, latent mean differences across groups, correlation between latent traits, and sample size. There were three levels of DIF percentage (5%, 10%, 20%), two groups of DIF type (uniform and nonuniform DIF), two levels of latent mean differences (balanced [$\mu_R = 0$, $\mu_F = 0$] and unbalanced [$\mu_R = .5$, $\mu_F = -.5$] latent means), three correlation levels ($\rho = 0$, $\rho = .3$, and $\rho = .6$), and four groups of sample sizes (R1000/F1000, R1000/F500, R500/F500, R500/F250). Similar to the analyses on type I error rates, factor constraints were measured across three different DIF detection methods: multidimensional IRT likelihood ratio test (MIRT-LR), multidimensional MIMIC-interaction model (MIMIC), and the multidimensional extension of the logistic discriminant function analysis (MLDFA).

Overall, the results from the MIRT-LR (*M* = .542, *SD* = .427), MIMIC (*M* = .920, *SD* = .114), and MLDFA (*M* = .961, *SD* = .095) tests showed that average rejection rates between uniform and nonuniform DIF varied across the three different methods. For the MIRT-LR test, the rejection rates when detecting uniform DIF radically differed from nonuniform DIF detection. The MIMIC model and MLDFA method produced more consistently high rejection rates between detection of uniform and nonuniform DIF than the MIRT-LR test. Data that contained 5% DIF and balanced latent mean differences tended to produce the lowest rejection rates with lower sample sizes exacerbating the effect. In general, the proportion of statistically significant replications predominately increased or stayed roughly equal as the sample size increased for all methods when detecting uniform and nonuniform DIF.

Uniform DIF. Figures 3.5 – 3.7 indicated that rejection rates for the MIRT-LR (*M* = .945, *SD* = .123), MIMIC (*M* = .882, *SD* = .149), and MLDFA (*M* = .979, *SD* = .059) tests varied across the three different methods when detecting uniform DIF. For instance, Figure 3.5 showed that MIRT-LR test consistently had the highest rejection rates and MIMIC consistently had the lowest for uniform DIF when data contain 5% DIF and unbalanced latent mean differences between groups. Inversely, the MIRT-LR test had the lowest rejections rates for uniform DIF, while the MIMIC model typically had the highest rejection rates when data contain 5% DIF and balanced latent mean differences between groups.

Figure 3.6 also showed similar patterns for uniform DIF when data contained 10% DIF; however, Figure 3.7 indicated that rejection rates remained consistent across all factor constraints when data contain 20% DIF. In fact, the rejection rates were roughly equal between data that contained balanced and unbalanced latent mean differences between groups. The MIRT-LR test correctly rejected every null hypothesis for uniform DIF in roughly 79% of factor constraint combinations when data contained 20% DIF.

Sample size did not appear to influence the rejection rates when data contained 10% or 20% uniform DIF, but a directly proportional relationship between sample size and rejection rate when data contain 5% DIF was apparent. The pattern of rejection rates for the MIRT-LR test across correlations, however, remained the same for uniform DIF across all percentages of DIF. The MIRT-LR test appeared to exhibit the lowest rejection rates when sample size was small with balanced latent mean differences and high correlation between latent traits.

In general, the multidimensional MIMIC-interaction model produced relatively low rejection rates when attempting to detect uniform DIF with low sample sizes (R500/F250). That is, the MIMIC model correctly rejected the null hypothesis every time when the sample size was large

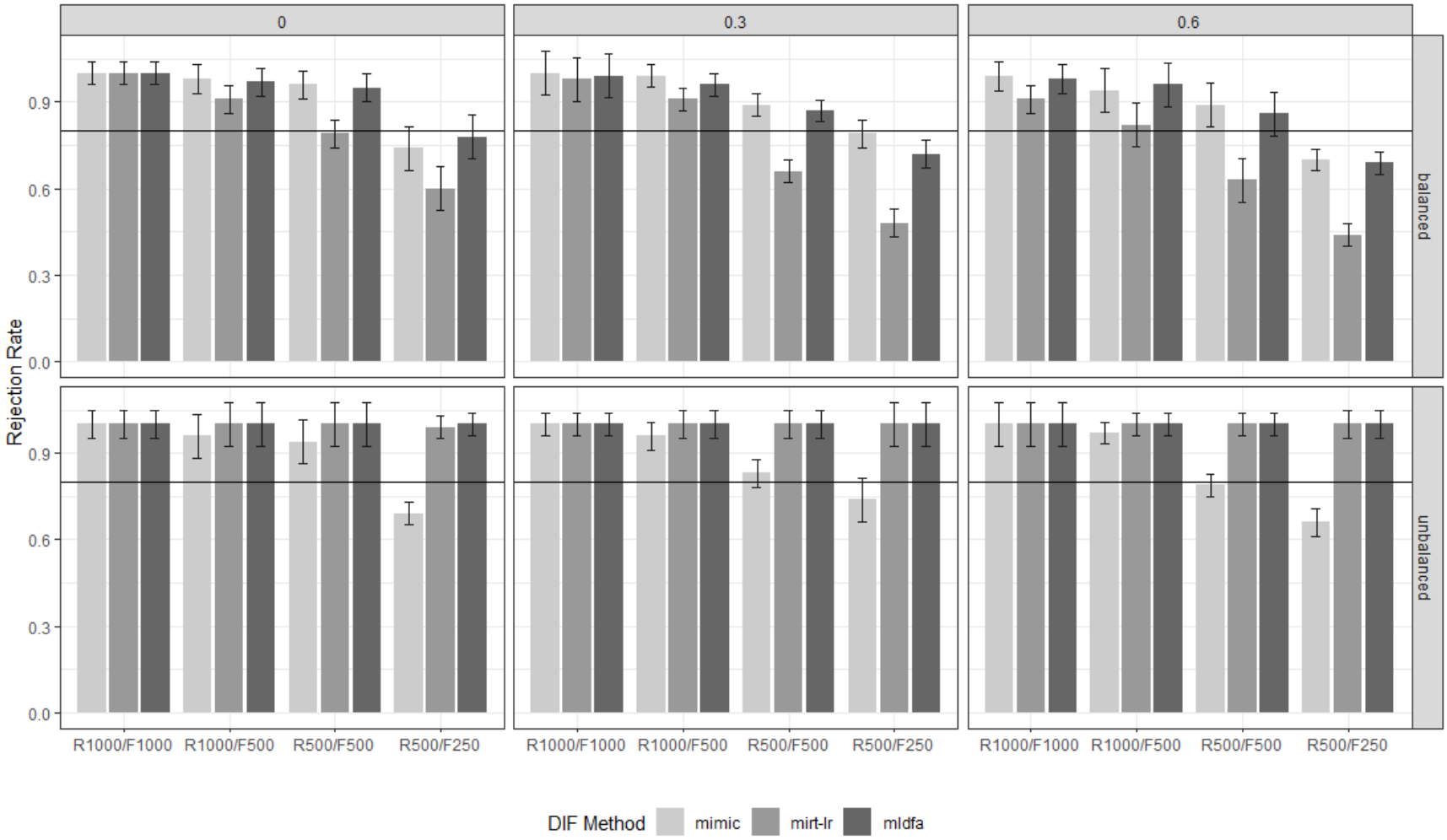

Figure 3.5. *Rejection rates for 5% uniform DIF results with 95% confidence intervals and 1 – β = .80 reference line*

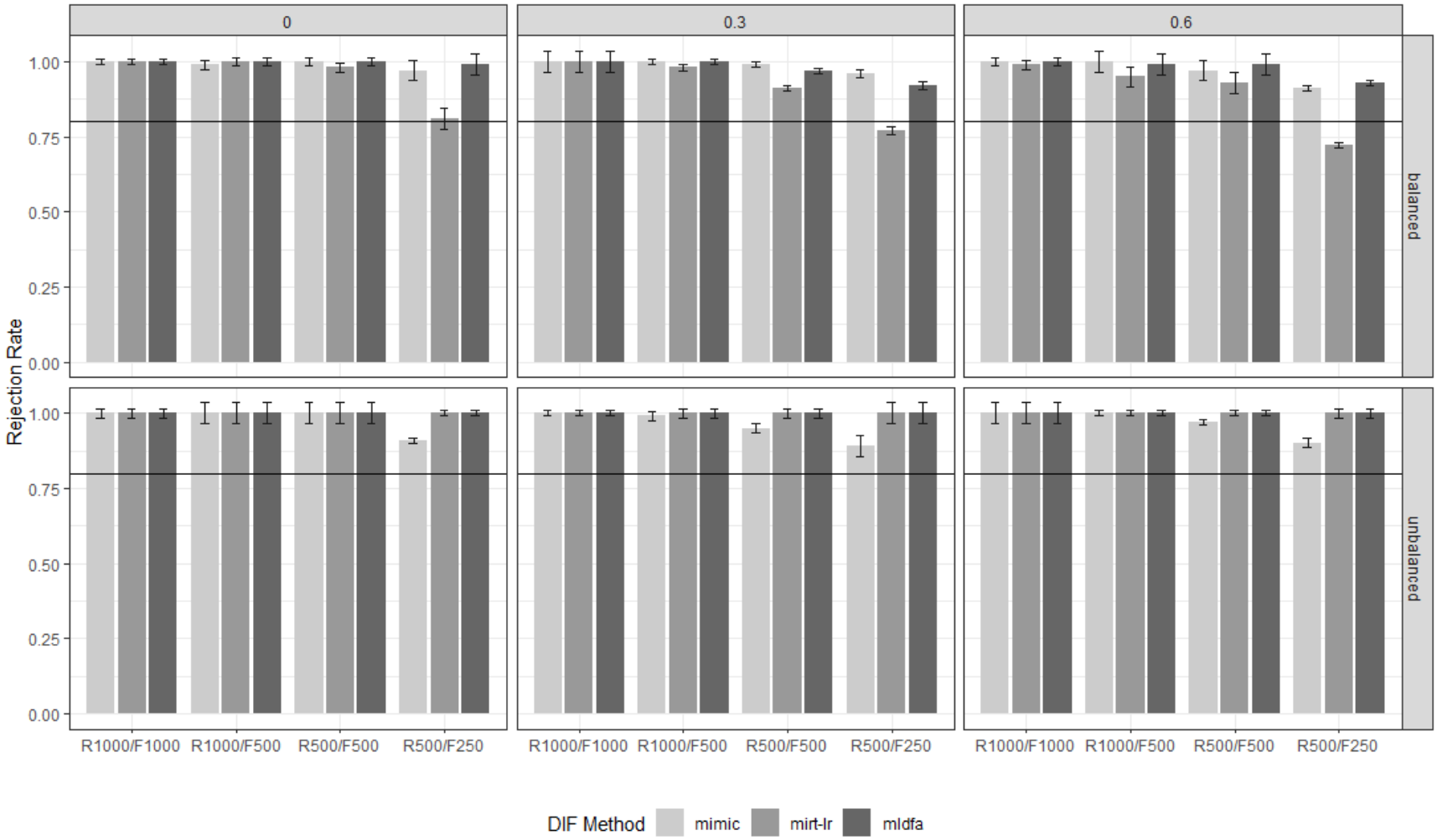

Figure 3.6. *Rejection rates for 10% uniform DIF results with 95% confidence intervals and 1 – β = .80 reference line*

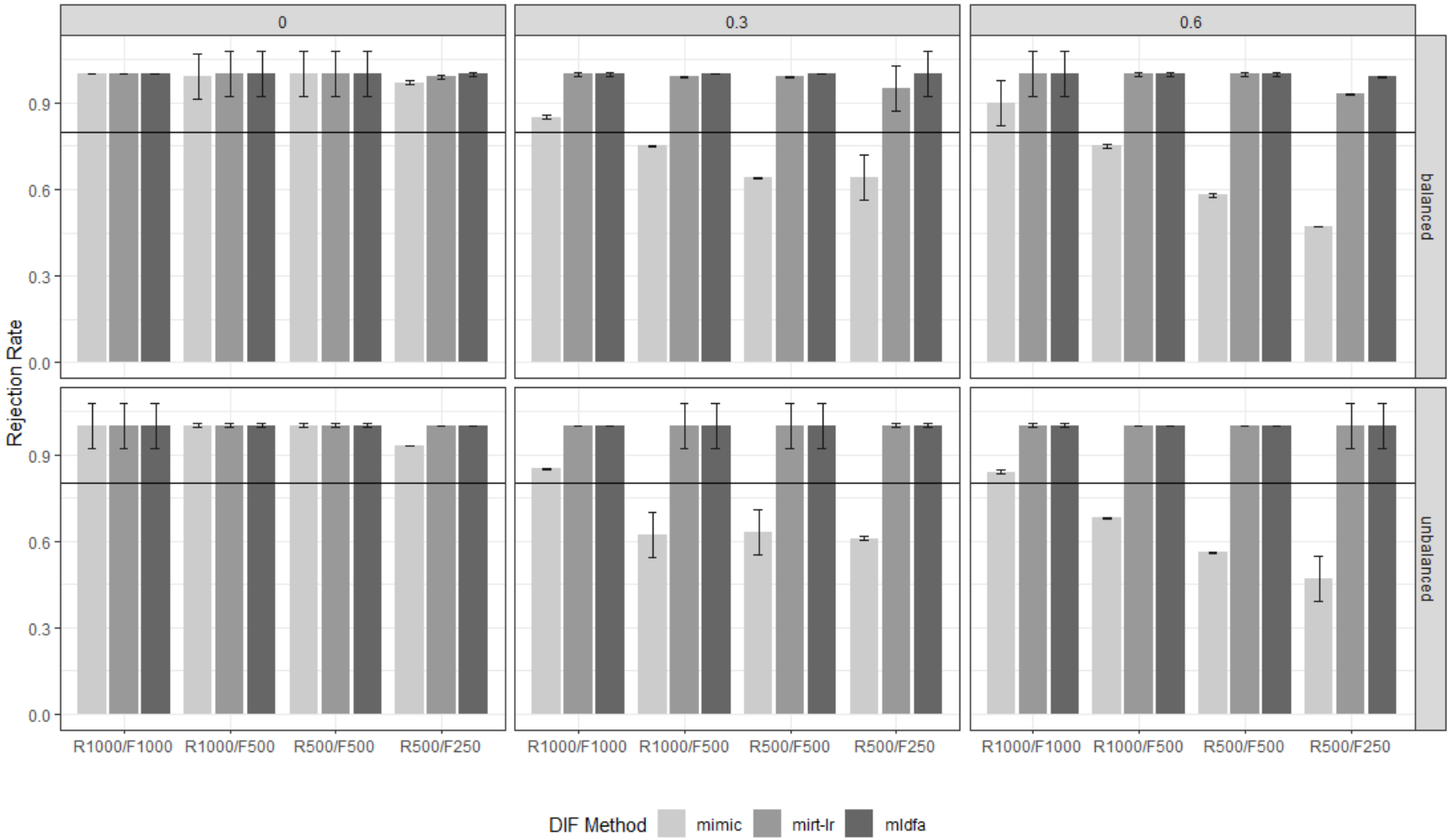

Figure 3.7. *Rejection rates for 20% uniform DIF results with 95% confidence intervals and 1 – β = .80 reference line*

and unbalanced (R1000/F1000), but rejection rates gradually decreased or stayed the same as the sample size decreased across all factor constraint combinations. The highest rejection rates for uniform DIF detection appeared to be when the data contained 10% DIF, with no fewer than 91% of replications correctly rejecting the null hypothesis. The rejection rates appeared to be similar whether the data contained balanced or unbalanced latent mean differences between groups for 5%, 10%, and 20% DIF. The results were also consistent across different correlation sizes for 5%, 10%, and 20% DIF.

The multidimensional extension of the logistic discriminant function analysis produced similar rejection rates as the MIRT-LR test for uniform DIF with a few exceptions. The MLDFA method produced noticeably higher rejection rates when the data contained balanced latent mean differences for low sample size on 10% DIF compared to the MIRT-LR test. In general, the MLDFA method (M = .979, SD = .059) had the highest rejection rates in comparison with the other two methods. Rejection rates remained above .85 for all factor constraint combinations when detecting uniform DIF, except when sample size was low (R500/F250) and latent mean differences were balanced ($\mu_R = 0$, $\mu_F = 0$) on data that contained 5% DIF.

Nonuniform DIF. Nonuniform DIF detection results, as displayed in Figures 3.8 – 3.10, showed that rejection rates severely differ between the three DIF detection methods, MIRT-LR (M = .139, SD = .150), MIMIC (M = .958, SD = .033), and MLDFA (M = .943, SD = .119). In general, the MIRT-LR appeared to perform considerably worse than the other two methods at correctly detecting nonuniform DIF. Comparable bar graphs indicated the rejection error rates for nonuniform DIF detection with the MIRT-LR test were strikingly lower than the rejection rates of the MLDFA method and MIMIC model across all factor constraint combinations. In fact, the rejection rates did not exceed .25, except when data contained 20% DIF and the correlation

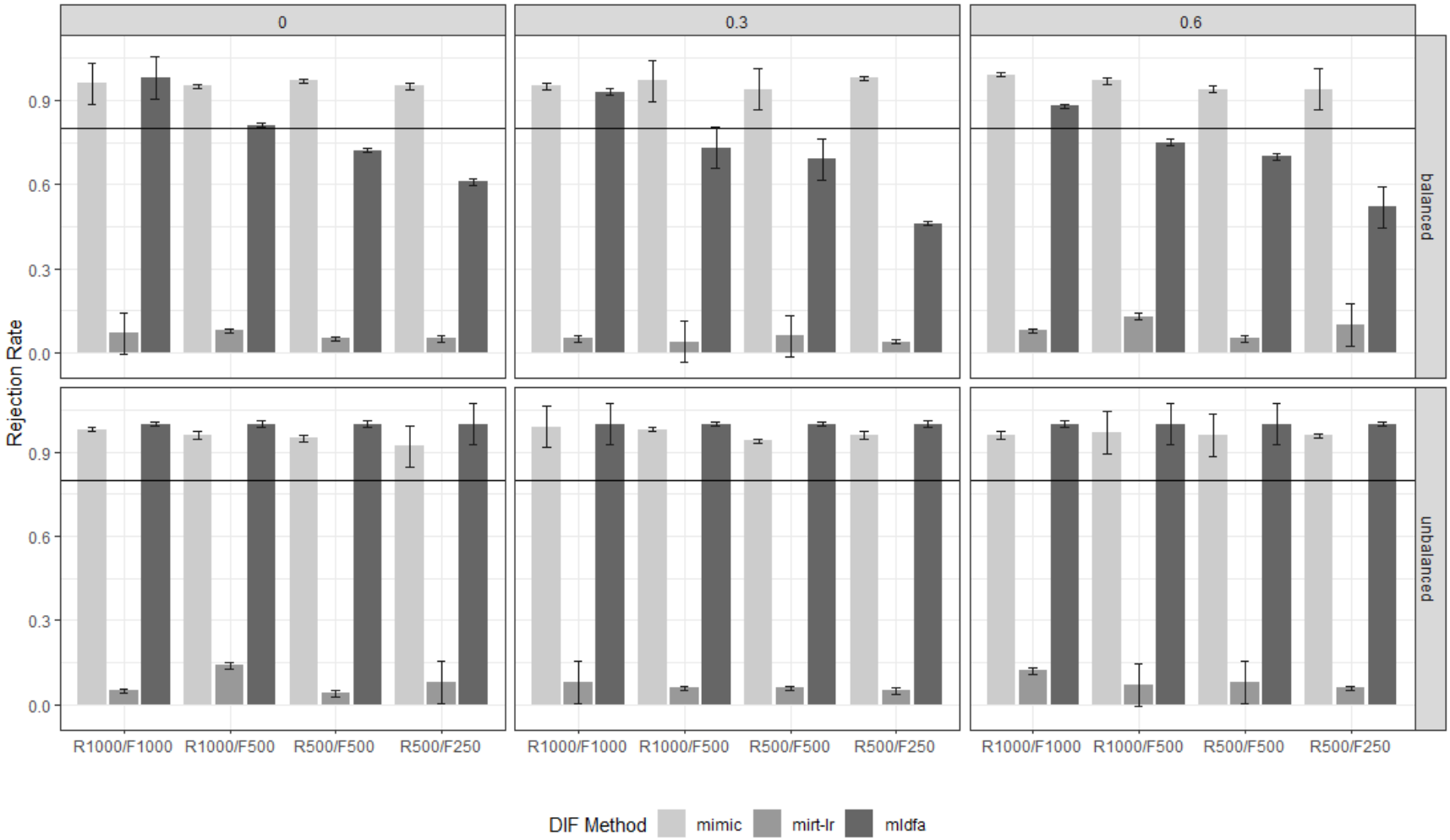

Figure 3.8. *Rejection rates for 5% nonuniform DIF results with 95% confidence intervals and 1 – β = .80 reference line*

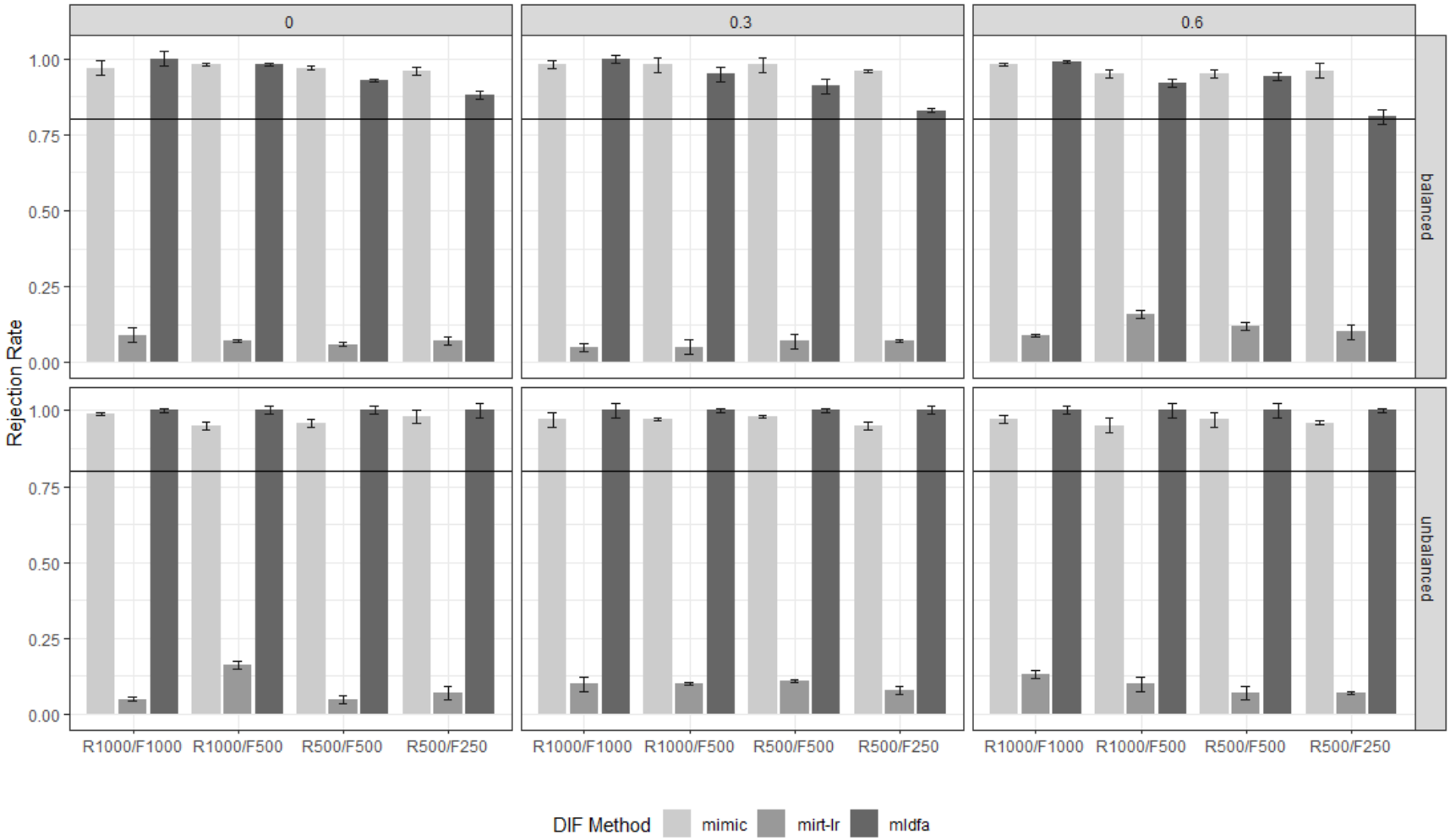

Figure 3.9. *Rejection rates for 10% nonuniform DIF results with 95% confidence intervals and 1 – β = .80 reference line*

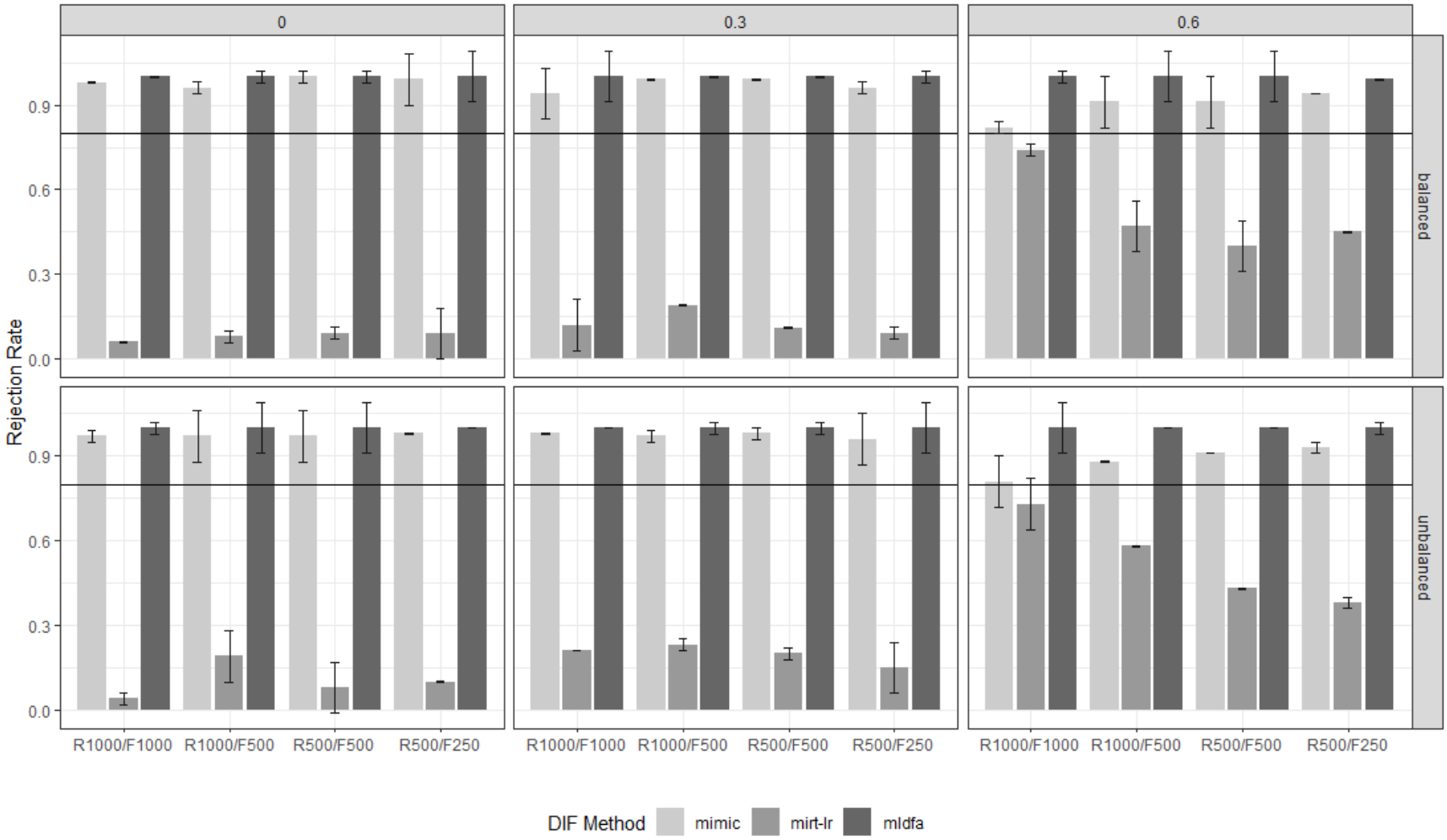

Figure 3.10. *Rejection rates for 20% nonuniform DIF results with 95% confidence intervals and 1 – β = .80 reference line*

between latent traits was strong ($\rho = .6$). The MIRT-LR test produced comparable rejection rates to the MIMIC model when sample sizes are high (R1000/F1000) under these factor conditions, but rejection rates quickly fell when smaller sample sizes were used.

The multidimensional MIMIC-interaction model produced consistently high rejection rates when attempting to detect nonuniform DIF across all factor constraint combinations. Rejection rates remained above .90, except when data contained 20% DIF, the correlation between latent traits was strong ($\rho = .6$), and sample size was high (R1000/F1000; R1000/F500). Even so, rejection rates never fell below .80 for any factor constraint combination. The MIMIC model produced rejection rates that were comparable to the MLDFA method under most factor constraint combinations, but appeared to produce noticeably higher rejection rates when data contained 5% DIF and balanced latent mean difference ($\mu_R = 0$, $\mu_F = 0$).

The MLDFA method produced several high rejection rates when detecting nonuniform DIF. In particular, the MLDFA method correctly detected nonuniform DIF in every replication when data contained 20% DIF, except in one instance when 99% of replications correctly detected nonuniform DIF. Overall, rejection rates exceed .90 in 81.9% of constraint combinations. The method also correctly detected nonuniform DIF in every replication when the data contained unbalanced latent mean differences ($\mu_R = .5$, $\mu_F = -.5$). When the data contained balanced latent mean differences, rejection rates generally increased as the percentage of DIF in the data and sample size increased. The pattern of increased rejection rates from small to large sample size was particularly apparent when data contained 5% DIF.

Effect of factor constraints on rejection rates. The effect of the factor constraints and DIF detection methods on rejection rates was studied with multiple mixed-design analyses of variance (ANOVA). More specifically, five two-way mixed-design ANOVAs examined one

factor constraint at a time to understand each factor's effect on rejection rates, separately. Therefore, this study looked at the interaction effects between one factor constraint and the DIF detection methods to determine the effect on rejection rates. Although this approach simplified the analysis, multiple hypothesis testing on the same dependent variable increased the likelihood of making a type I error (Darlington & Hayes, 2017; Higdon, 2013). A Bonferroni correction, such that the alpha criterion significance testing was decreased to .01, reduced the occurrence of type I error when testing each interaction effect independently (Frane, 2015).

Multivariate normality was assumed in the mixed designed ANOVAs. An observation of the Q-Q plots of the rejection rates indicated severe departures from univariate normality. In addition, 130 observations were at least three median absolute deviations from the median. A scatterplot of the data by DIF detection method, coupled with the kernel density estimates of each DIF detection method, showed that data were clustered near and bounded by 1 (Figure B.3 in Appendix B). Furthermore, the logistic probability density function mapped on the kernel density distribution of the data indicates that the data was non-normal (Figure B.4 in Appendix B). Although the mixed ANOVAs were robust to violations of normality (Tabachnick & Fidell, 2018), a logit transformation was applied to the rejection rates to aid in the interpretation of results and protect against over-inflated standard errors (Osbourne, 2017). Rejection rate values of 1.00 were adjusted to the R program default of .975 to avoid undefined odds-ratios and infinite log transformation values. Application of the logit transformation resulted in better-fitted Q-Q plots and fewer troublesome outliers.

The data still contained 82 outliers that were more than three median absolute deviations from the median after the logit transformation. Similar to the analyses on type I error rates, all outliers were of substantive interest to the research objectives. More explicitly, 70 outliers

represented all results from the MIRT-LR test when testing for the presence of nonuniform DIF. The other 12 outliers contained results from data with a small, unbalanced sample size (R500/F250). If alterations were made on the outliers, then any conclusions on the nature of the DIF detection methods could be met with skepticism; therefore, it was determined that these values would neither be removed, nor altered.

Effect of percentage of DIF. This portion of the study first explored the effect of the percentage of DIF (5%, 10%, 20%) and DIF detection method (MIRT-LR, MLDFA, and MIMIC-interaction model) on the log-odds of rejection rates. The interaction between the percentage of DIF and method was examined using a 3 (DIF detection method) x 3 (DIF percentage) mixed design ANOVA, where the DIF type was the between-subjects variable and DIF detection method was the within-subjects variable. Before conducting the mixed-design ANOVA, sphericity on repeated measures was tested to reduce the likelihood of type I errors. Mauchly's test on the within-subjects ANOVA interaction was found to be in violation of sphericity, $\chi^2(2) = 144.00$, $\varepsilon = .609$, $p < .001$. A Greenhouse-Geisser correction was added to adjust the degrees of freedom and limit the possibility of an overly inflated test statistic in the ANOVA test (Field, 2009).

A nonsignificant interaction was found, $F(2.44, 171.69) = 4.05$, $p = .013$, $\omega^2 = .049$, such that log-odds of rejection rates did not depend on the combination of the percentage of DIF and the DIF detection method (see Figure 3.11A). Recall that a Bonferroni correction to adjust for multiple hypothesis testing amended the alpha criterion for the ANOVA test to $\alpha = .01$; therefore, a *p*-value of .013 was nonsignificant. In addition, the main effect for DIF detection method was significant, $F(1.22, 171.69) = 100.78$, $p < .001$, $\omega^2 = .458$, but the main effect for percentage of DIF was nonsignificant, $F(2, 141) = 4.27$, $p = .016$, $\omega^2 = .035$.

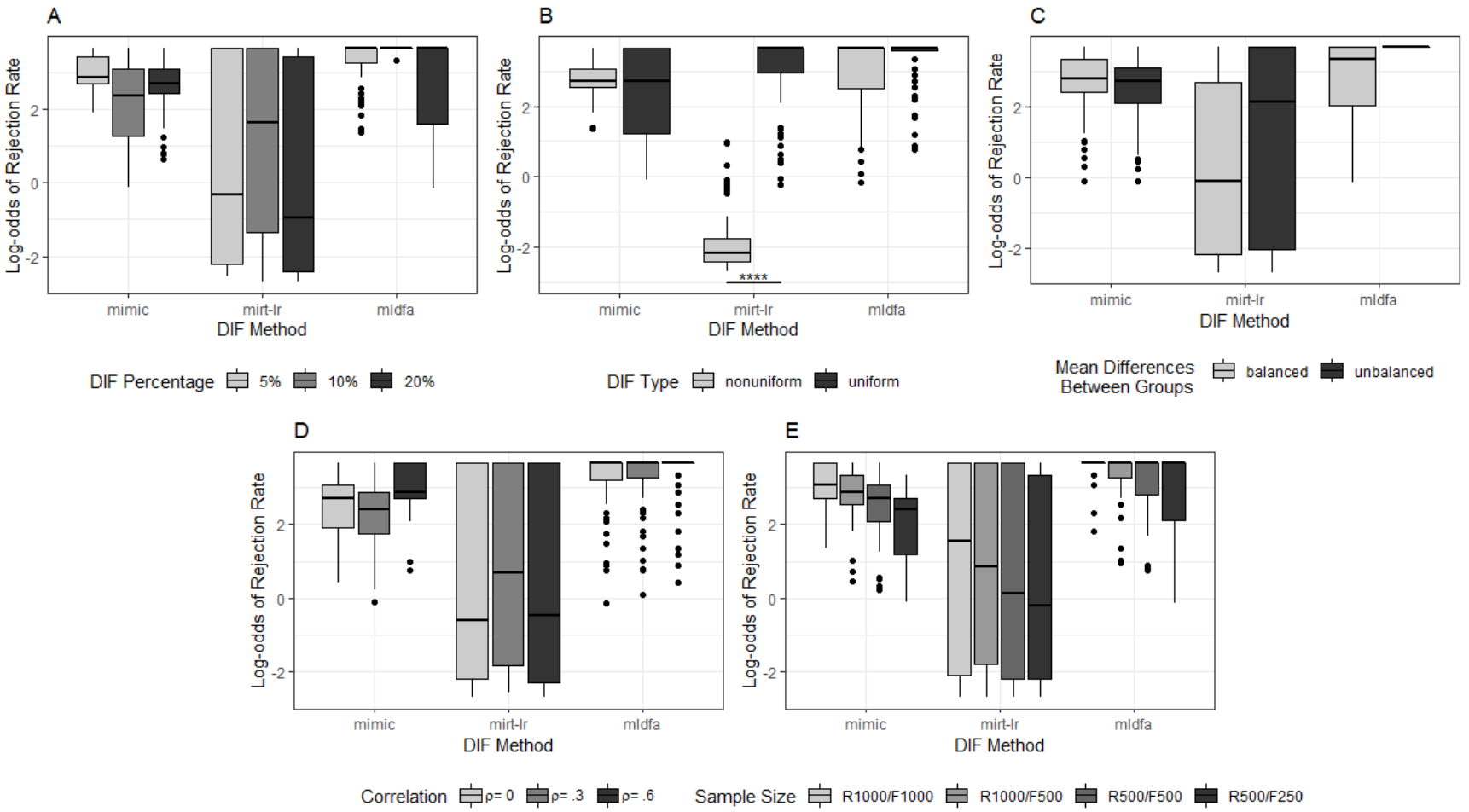

Figure 3.11. *Boxplot of interactions and significant simple main effects of factor constraints and DIF detection method for log-odds of rejection rate*

****$p_{adj} < .0001$

The results of a Bonferroni corrected pairwise *t*-test found that significant differences of logit-adjusted means existed between the MIRT-LR test (M = .581, SD = 2.65, $p_{adj} < .001$) and MIMIC-interaction model (M = 2.56, SD = .935) when controlling for the percentage of DIF. The results from another Bonferroni corrected pairwise *t*-test also found that significant differences of logit-adjusted means existed between MIMIC-interaction model ($p_{adj} < .001$) and the MLDFA method (M = 3.20, SD = .913), and significant differences of logit-adjusted means exists between the MIRT-LR test ($p_{adj} < .001$) and the MLDFA method. These post-hoc results were also applicable to the DIF detection method main effects post-hoc results for the effects of mean differences between groups, correlation between latent traits, and sample size on the rejection rates. Figure 3.12 details the relationship between the DIF detection methods when all factor constraints were controlled. In connection with odds, the mean odds of correctly rejecting the null hypothesis with the MIRT-LR test was roughly 1.78 to every occurrence of type II error regardless of percentage of DIF. The mean odds of correctly rejecting the null hypothesis with the MIMIC-interaction model was about 12.94 for every failed rejection of the null hypothesis, while the mean odds of correctly rejecting the null hypothesis with the MLDFA method was about 24.53 to every occurrence of type II error, when controlling for the percentage of DIF.

The logit-adjusted means for 5% DIF (M = 1.77, SD = 2.11), 10% DIF (M = 2.25, SD = 2.07), and 20% DIF (M = 2.32, SD = 1.89), when controlling for DIF detection method, were statistically likely to yield similar rejection rates. In terms of odds, the mean odds of correctly rejecting the null hypothesis with 5% DIF was roughly 5.87 to every occurrence of type II error regardless of DIF detection method. The mean odds of correctly rejecting the null hypothesis with the 10% DIF was about 9.49 for every failed rejection of the null hypothesis, while the

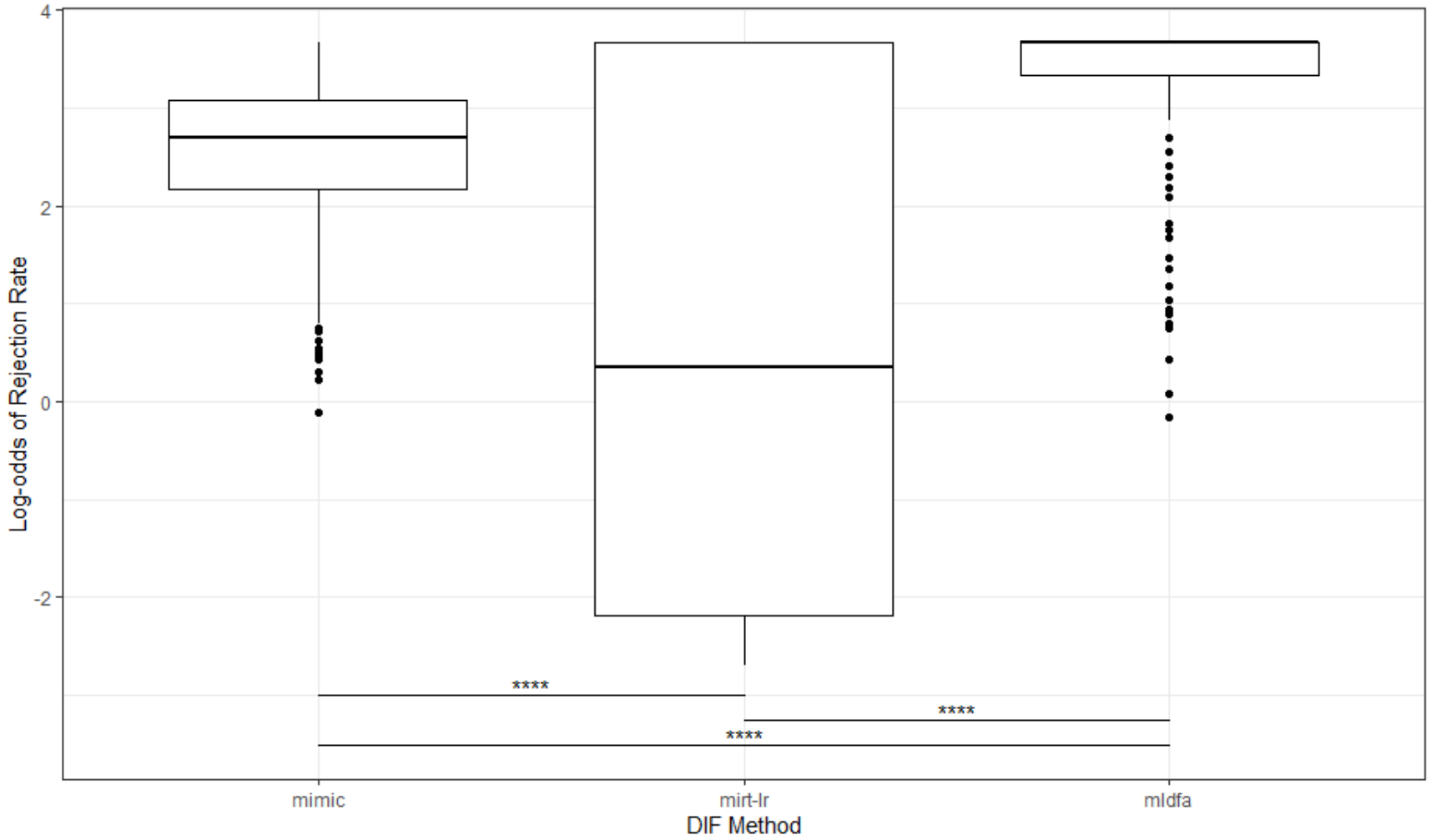

Figure 3.12. *Boxplot of DIF detection method main effects for log-odds of rejection rate*

****p_{adj} < .0001

mean odds of correctly rejecting the null hypothesis with 20% DIF was about 10.18 to every occurrence of type II error.

Effect of DIF type. This study next explored the effect of the DIF type (uniform, nonuniform) and DIF detection method (MIRT-LR, MLDFA, and MIMIC-interaction model) on the log-odds of rejection rates. A 3 (DIF detection method) x 2 (DIF type) mixed design ANOVA examined the interaction effect between DIF type, the between-subjects variable, and DIF detection method, the within- subjects variable. Mauchly's test on the within-subjects ANOVA interaction was found to be in violation of sphericity, $\chi^2(2) = 30.98$, $\varepsilon = .844$, $p < .001$. A Huynd-Feldt correction adjusted the degrees of freedom to account for the lack of sphericity and limit the possibility of an overly inflated test statistic in the ANOVA test (Field, 2009). A Huynd-Feldt correction is the preferred method when the epsilon test statistic is above .75 (Verma, 2016).

Interaction effect was significant, $F(1.69, 239.69) = 440.39$, $p < .001$, $\omega^2 = .836$, such that log-odds of rejection rates depends on the combination of the DIF type and the DIF detection method (see Figure 3.11B). In addition, the main effect for DIF detection method, $F(1.69, 239.69) = 393.63$, $p < .001$, $\omega^2 = .820$ and the main effect for DIF type were significant, $F(1, 142) = 240.28$, $p < .001$, $\omega^2 = .675$. A simple effects analysis on the between-subjects effect for the MIRT-LR test was significant, $F(1, 142) = 977.00$, $p_{adj} < .001$, $\eta_p^2 = .873$, $\omega^2 = .871$. The results indicated that significant differences of logit-adjusted means for the MIRT-LR test existed between uniform DIF ($M = 3.05$, $SD = 1.060$) and nonuniform DIF ($M = -1.89$, $SD = .818$). In terms of odds, the mean odds of correctly rejecting the null hypothesis with uniform DIF was roughly 21.12 to every occurrence of type II error. The mean odds of failing to reject

the null hypothesis with nonuniform DIF was roughly 6.62 to every occurrence of a correct rejection of the null hypothesis.

A simple effects analysis on the between-subjects effect for the MIMIC-interaction model was found not to be significant, $F(1, 142) = 5.87$, $p_{adj} = .051$, $\omega^2 = .033$. The results indicated that no significant differences of logit-adjusted means for the MIMIC-interaction model existed between the uniform DIF ($M = 2.37$, $SD = 1.23$) and nonuniform DIF ($M = 2.74$, $SD = .413$). The mean odds of correctly rejecting the null hypothesis with uniform DIF was roughly 10.70 to every failed rejection of the null hypothesis. The mean odds of correctly rejecting the null hypothesis with nonuniform DIF was roughly 15.49 to every failed rejection of the null hypothesis.

Finally, a simple effects analysis on the between-subjects effect for the MLDFA method was also nonsignificant, $F(1, 142) = 4.62$, $p_{adj} = .099$, $\omega^2 = .024$. Thus, no significant logit-adjusted mean difference between the uniform DIF ($M = 3.36$, $SD = .678$) and nonuniform DIF ($M = 3.04$, $SD = 1.08$) was present for the MLDFA method. The mean odds of correctly rejecting the null hypothesis with uniform DIF was roughly 28.79 to every failed rejection of the null hypothesis. The mean odds of correctly rejecting the null hypothesis with nonuniform DIF was roughly 20.91 to every failed rejection of the null hypothesis.

A simple effects analysis on the within-subjects effects for uniform DIF was significant, $F(1.14, 80.80) = 25.95$, $p_{adj} < .001$, $\omega^2 = .032$. The results indicated that the logit-adjusted means for uniform DIF was significantly higher for the MIRT-LR test ($p_{adj} < .001$) than the MIMIC-interaction model. The logit-adjusted means for uniform DIF was also significantly higher for the MLDFA method ($p_{adj} < .001$) than the MIMIC model, but no significant logit-adjusted means difference was present between the MIRT-LR test ($p_{adj} = .069$) and the MLDFA method. A

simple effects analysis on the within-subjects effects for nonuniform DIF was significant, $F(2, 142) = 842.35$, $p_{adj} < .001$, $\omega^2 = .509$. The results indicate that the logit-adjusted means for nonuniform DIF was significantly higher for the MIMC-interaction model ($p_{adj} < .001$) than MIRT-LR test. The logit-adjusted means for nonuniform DIF was also significantly higher for the MLDFA method ($p_{adj} < .001$) than the MIRT-LR test, but no there was no significant logit-adjusted means difference between the MIMIC-interaction model ($p_{adj} = .069$) and the MLDFA method.

Effect of mean differences between groups. A 3 (DIF detection method) x 2 (mean differences between groups) mixed design ANOVA explored the effect of the latent mean differences between groups (balanced [$\mu_R = 0$, $\mu_F = 0$], unbalanced [$\mu_R = .5$, $\mu_F = -.5$]) and DIF detection method on the log-odds of rejection rates. The mean differences between groups was the between-subjects variable and DIF detection method was the within-subjects variable. Mauchly's test on the within-subjects ANOVA interaction was found to be in violation of sphericity, $\chi^2(2) = 143.20$, $\varepsilon = .611$, $p < .001$, and a Greenhouse-Geisser adjustment was applied to the ANOVA test.

The results from the ANOVA interaction was nonsignificant, $F(1.22, 173.39) = 3.87$, $p = .043$, $\omega^2 = .024$, such that log-odds of rejection rates did not depend on the combination of the logit-adjusted mean differences between groups and the DIF detection method (see Figure 3.11C). The main effect for DIF detection method was significant, $F(1.22, 173.39) = 98.59$, $p < .001$, $\omega^2 = .453$. The main effect for the mean differences between groups was also found to be significant, $F(1, 142) = 8.70$, $p = .004$, $\omega^2 = .041$. The results of a Bonferroni corrected pairwise *t*-test verified that significant differences of logit-adjusted means existed between the balanced latent means ($M = 1.87$, $SD = 1.99$, $p_{adj} = .012$) and unbalanced latent means ($M = 2.36$, $SD =$

2.05) when controlling for DIF detection method. The mean odds of correctly rejecting the null hypothesis with balanced latent means was about 6.49 to every failed rejection of the null hypothesis, while the mean odds of a correct null hypothesis for unbalanced latent means was about 10.59 to every occurrence of type II error, regardless of DIF detection method.

Effect of correlation between latent traits. The effect of the correlation between latent traits ($\rho = 0$, $\rho = .3$, $\rho = .6$) and DIF detection method on the log-odds of rejection rates was explored using a 3 (DIF detection method) x 3 (correlation between latent traits) mixed design ANOVA. The correlation between latent traits was the between-subjects variable and DIF detection method was the within-subjects variable. Mauchly's test on the within-subjects ANOVA interaction was found to be in violation of sphericity, $\chi^2(2) = 122.00$, $\varepsilon = .632$, $p < .001$, and a Greenhouse-Geisser correction was added to adjust the degrees of freedom and limit the possibility of an overly inflated test statistic.

The results from the within-subjects ANOVA interaction was found to be nonsignificant, $F(2.53, 178.28) = 1.22$, $p = .301$, $\omega^2 = .004$, such that log-odds of rejection rates depended on the combination of the correlation between latent traits and the DIF detection method (see Figure 3.11D). The main effect for DIF detection method was significant, $F(1.26, 178.28) = 96.95$, $p < .001$, $\omega^2 = .457$, but the main effect for the correlation between latent traits was found not to be significant, $F(2, 141) = .635$, $p = .532$, $\omega^2 = -.004$. Therefore, the logit-adjusted means for no correlation between latent traits ($M = 2.25$, $SD = 2.14$), the correlation of .3 between latent traits ($M = 2.04$, $SD = 2.09$), and the correlation of .6 between latent traits ($M = 2.05$, $SD = 1.88$), when controlling for DIF detection method, were likely to produce similar rejection rates. In terms of odds, the mean odds of correctly rejecting the null hypothesis with no correlation between the latent traits was roughly 9.49 to every occurrence of type II error, regardless of DIF detection

method. The mean odds of correctly rejecting the null hypothesis with a medium correlation of .3 was 7.69 to every failed rejection of the null hypothesis, while the mean odds of correctly rejecting the null hypothesis with a large correlation of .6 was about 7.77 to every occurrence of type II error, when controlling for the DIF detection method.

Effect of sample size. Lastly, this study explored the effect of the sample size (R1000/F1000, R1000/F500, R500/F500, R500/F250) and DIF detection method on the log-odds of rejection rates. The interaction between sample size and method was examined using a 3 (DIF detection method) x 4 (sample size) mixed design ANOVA, where the sample size was the between-subjects variable and DIF detection method was the within-subjects variable. Mauchly's test on the within-subjects ANOVA interaction was found to be in violation of sphericity, $\chi^2(2) = 121.30$, $\varepsilon = .632$, $p < .001$, and a Greenhouse-Geisser correction was added to the ANOVA test.

The results from the ANOVA interaction effect was nonsignificant, $F(3.79, 176.96) = .086$, $p = .984$, $\omega^2 = -.025$, such that log-odds of rejection rates depended on the combination of the sample size and the DIF detection method (see Figure 3.11E). The main effect for DIF detection method was significant, $F(1.26, 176.96) = 94.80$, $p < .002$, $\omega^2 = .452$, but the main effect for sample size was nonsignificant, $F(3, 140) = 3.66$, $p = .014$, $\omega^2 = .042$. Therefore, the logit-adjusted means for R1000/F1000 ($M = 2.46$, $SD = 2.05$), R1000/F500 ($M = 2.26$, $SD = 1.96$), R500/F500 ($M = 2.03$, $SD = 2.06$), R500/F250 ($M = 1.71$, $SD = 2.03$) sample sizes, when controlling for DIF detection method, were statistically likely to produce similar rejection rates.

In terms of odds, the mean odds of correctly rejecting the null hypothesis with a large, balanced sample size (R1000/F1000) was roughly 11.70 to every occurrence of type II error, regardless of DIF detection method. The mean odds of correctly rejecting the null hypothesis with a large, unbalanced sample size (R1000/F500) was 9.58 to every failed rejection of the null

hypothesis, while the mean odds of correctly rejecting the null hypothesis with a small, balanced sample size (R500/F500) was about 7.61 to every occurrence of type II error, when controlling for the DIF detection method. The mean odds of correctly rejecting the null hypothesis with a large, unbalanced sample size (R500/F250) was 5.53 to every failed rejection of the null hypothesis.

Empirical Example

An empirical example provided context for the application of the three DIF detection methods. In this example, data were collected with the Experience Learning Student Survey, which was intended to capture student perception of student learning outcome (SLO) attainment after an experiential learning occurrence (Walker & Rocconi, 2021). The sample included students who enrolled in an experiential learning course at a large, southeastern U.S. university. An experiential learning course included pedagogies focused on internships, service learning, simulation/gaming/role-playing, study abroad, undergraduate research, or some combination of the five types. The survey was disseminated to students through Qualtrics at the end of each semester from fall 2017 to spring 2020. Overall, the survey had a 56.8% response rate with 711 completers. The students were from 35 experiential learning courses across 23 disciplines. From the sample, 686 (96.5%) were undergraduate students and 25 (3.5%) were graduate students. The gender[7] demographic comprised of 290 (40.8%) males and 421 (59.2%) females.

For this example, gender served as the grouping variable, since gender dichotomized the data in a similar fashion to one of the simulation's sample size groups (R500/F250). Therefore, Females served as the reference group, because they represent the largest group, and males

[7] Although the author acknowledges that gender is fluid and non-binary, the demographic was partitioned into two categories for the purpose of simplicity and demonstration. Students who identify as non-binary or other will be excluded from the study.

served as the focal group. Sizable differences across the four statistical moments were present between the two genders, as shown in Table 3.1. In particular, mean differences between the groups provided indirect evidence of group favorability across items, such that some DIF items may have favor one group and others favored the other group. Additionally, all items were positively skewed and platykurtic across both groups; thus, both male and female responses favored agreement over disagreement and both groups selected across a small distribution of categorical response options on all items.

Model fit. The survey was assumed to fit the multidimensional graded response model (MGRM) and contained four factors; however, other plausible models could fit the data, as well. Thus, a comparison between four models (i.e., MGPCM, 1-PL MGRM, GRM, and MGRM) confirmed the association between the survey items and the four Experience Learning SLOs. First, a model fit comparison between the four factor 2-PL MGRM, MGPCM, and 1-PL MGRM models indicated that the 2-PL MGRM, $M_2^*(23) = 45.91$, $p < .05$, TLI = .960, CFI = .978, was a better fit for the data than the other models, as shown in Table 3.2. Next, a log-likelihood comparison between nested models indicated that the multidimensional GRM with four factors, $\chi^2(8) = 1449.59$, $p < .001$, deviance = 19504.01, AIC = 19742.01, BIC = 20285.44, was a better fit than the unidimensional GRM, deviance = 20953.59, AIC = 21175.59, BIC = 21682.49, for the survey data; therefore, the 2-PL MGRM was confirmed to be the best fit for the survey.

The four factors were labeled after their associated SLO: (1) lifelong learning (LL), (2) application of knowledge and skills (AKS), (3) collaboration (COL), and (4) structured reflection (SR). A summary of the factor statistics indicated that factors LL and AKS ($\rho = .616$), AKS and COL ($\rho = .645$), ASK and SR ($\rho = .632$), and COL and SR ($\rho = .717$) were strongly correlated, while LL and COL ($\rho = .371$), and LL and SR ($\rho = .371$) were moderately correlated. The latent

Table 3.1
Item descriptive statistics by gender

	Female (N = 421)				Male (N = 290)			
	Mean	SD	Kurtosis	Skew	Mean	SD	Kurtosis	Skew
Item 1	6.19	1.08	-1.79	3.53	5.97	1.19	-1.80	3.97
Item 2	6.55	.75	-2.36	8.91	6.29	1.01	-2.26	7.20
Item 3	6.30	1.00	-1.85	4.10	6.10	1.17	-1.97	4.84
Item 4	6.27	1.07	-2.07	5.29	6.07	1.17	-1.83	4.29
Item 5	6.20	.96	-1.54	3.02	6.17	.97	-2.06	7.24
Item 6	6.10	1.11	-1.49	2.32	6.16	1.01	-1.83	5.25
Item 7	6.14	1.03	-1.57	2.92	6.25	0.88	-1.76	5.71
Item 8	6.12	1.08	-1.66	3.36	6.10	1.06	-1.82	5.01
Item 9	6.15	1.06	-1.55	2.48	6.28	.90	-1.91	6.85
Item 10	6.22	1.02	-1.94	5.27	6.24	.95	-2.01	6.73
Item 11	6.59	.65	-1.67	2.94	6.39	.80	-2.29	9.79
Item 12	6.36	.86	-1.72	4.39	6.27	.98	-1.90	5.20
Item 13	6.59	.70	-1.88	3.36	6.38	.98	-1.96	4.58
Item 14	6.19	.98	-1.54	2.83	6.06	1.05	-1.61	3.83
Item 15	6.06	1.15	-1.51	2.42	5.91	1.28	-1.61	2.91
Item 16	6.09	1.16	-1.55	2.41	6.00	1.19	-1.73	3.64

Note. Items 1 – 4, 16 are loaded on the lifelong learning factor (LL), items 5-10 are loaded on the application of knowledge and skills factor (AKS), items 11-13 are loaded on the collaboration factor (COL), and items 4, 14-16 are loaded on the structured reflection factor (SR). All factors are assumed to have a latent mean of 0 and variance of 1.

Table 3.2
Model fit statistics for comparison of 2-PL MGRM, MGPCM, and 1-PL MGRM models

Model	M_2*	*df*	$RMSEA_2$	$RMSEA_2$ CI(5, 95)	TLI	CFI
2-PL MGRM	45.91*	23	.037	(.021, .053)	.960	.978
MGPCM	56.42***	23	.045	(.030, .060)	.942	.968
1-PL MGRM	98.05***	35	.050	(.039, .062)	.928	.939

***$p < .001$, *$p < .05$

mean for each factor was assumed to be 0 and the variance was equal to 1. Appendix A contained a supplemental table (Table A.4) that outlined the item discrimination and threshold parameters for the reader's interest.

Comparison of DIF detection methods. An item purification method was utilized to empirically select anchor items to assist in DIF detection. The Woods (2009b) method identified the best items to select as anchors through a sequential and ranked-ordered based approach on χ^2/df values from the MIRT-LR test. The three items with the smallest χ^2/df value were selected anchors. The Metropolis-Hastings Robbins-Monro (MHRM) algorithm was used to estimate the augmented and compact models. The MHRM algorithm allowed the model to estimate the higher dimensional structure of the models, produced faster results, and converged more likelihood ratio tests than the quasi-Monte Carlo EM method; however, four items (2, 5, 9, and 12) still had issues with convergence and were dropped from the item purification analysis. Items 3, 10, and 13 produced the lowest χ^2/df values and were designated as anchors for this study. Table A.5 in Appendix A provided details on the χ^2/df values for all items.

All non-anchor items were tested for uniform and nonuniform DIF using the MIRT-LR test, MLDFA method, and MIMIC-interaction model under the MGRM framework. Parameters for anchor items were fixed equal across groups and used to calculate discrimination and threshold parameters for the male and female groups. Tables A.6 and A.7 in Appendix A showed the discrimination and threshold parameters for both groups. Correlations between latent traits appeared to be similar across groups, except between the LL and AKS (ρ_{female} = .627; ρ_{male} = .206) factors. Factors LL and COL (ρ_{female} = .262; ρ_{male} = .296), LL and SR (ρ_{female} = .252; ρ_{male} = .332), ASK and COL (ρ_{female} = .425; ρ_{male} = .313), ASK and SR (ρ_{female} = .359; ρ_{male} = .277), and

COL and SR (ρ_{female} = .332; ρ_{male} = .289) were moderately correlated across both groups. The latent mean for each factor was assumed to be 0 and the variance was equal to 1.

A comparison of the DIF detection methods, as shown in Table 3.3, provided empirical confirmation of their ability to detect uniform and nonuniform DIF. The MIRT-LR test detected uniform DIF in 12.5% of items. More specifically, items 9 and 11 exhibited uniform DIF under the MIRT-LR test. The MLDFA method detected uniform DIF in 25% of items, where items 2, 7, 9, and 11 demonstrated uniform DIF. The MIMIC-interaction model detected uniform DIF in 18.75% of items. Items 2, 9, and 12 exhibited uniform DIF under the MIMIC-interaction model. Neither the MIRT-LR test nor the MLDFA method detected nonuniform DIF in any item, while the MIMIC-interaction model only detected nonuniform DIF in item 2.

Table 3.3
DIF detection method p-values for uniform and nonuniform DIF

	Uniform DIF *p*-values			Nonuniform DIF *p*-values		
	MIRT-LR	MLDFA	MIMIC	MIRT-LR	MLDFA	MIMIC
Item 1	.085	.780	.517	.998	.699	.479
Item 2	.139	.013	.011	.998	.406	.022
Item 3†						
Item 4	.493	.519	.941	.999	.134	.520
Item 5	.999	.439	.303	.998	.219	.660
Item 6	.194	.168	.495	.998	.614	.325
Item 7	.122	.007	.087	.998	.771	.053
Item 8	.591	.798	.616	.999	.426	.768
Item 9	.002	.021	.032	.998	.168	.196
Item 10†						
Item 11	.006	.002	.608	.998	.220	.218
Item 12	.186	.313	.009	.998	.168	.441
Item 13†						
Item 14	.455	.526	.999	.998	.864	.265
Item 15	.787	.427	.999	.998	.172	.364
Item 16	.614	.347	.999	.999	.408	.701

†Anchor item

Chapter IV

Discussion, Limitations, and Conclusion

The purpose of this study is to determine the optimal DIF detection method for data fitted to the multidimensional graded response model (MGRM). This study compares the type I error and rejection rates of three DIF detection procedures, MIRT-LR test, MIMIC-interaction model, and the MLDFA method, to identify uniform and nonuniform DIF on data that fits the MGRM framework. In addition, this study examines other factors (e.g., type of DIF, proportion of DIF, sample size) to understand the conditions that can influence the rate of type I errors and rejection rates. A series of mixed-design ANOVA tests are conducted to analyze the effect of these factors on the performance of the three DIF detection procedures.

Previous DIF simulation studies have compared unidimensional and multidimensional extensions of the MIMIC, IRT-LR tests, and multiple regression methods (Bulut & Suh, 2017; Cohen et al., 1996; Finch, 2005; Lee et al., 2017; Woods & Grimm, 2011; Woods et al., 2009). Other studies have extended these methods for polytomous, unidimensional data in detecting uniform and nonuniform DIF (Elosua & Wells, 2013; French & Miller, 1996; Miller & Spray, 1993; Wang & Shih, 2010). These studies, however, are limited to data that fit the 2-PL IRT, GRM, or 2-PL MIRT models. Similar DIF detection methods need to be considered for polytomous items that are estimated to measure multiple latent traits. Many measurement instruments, such as surveys, rubrics, and evaluation forms, produce data fitted to polytomous, multidimensional models and correct selection of a DIF detection method to identify DIF is imperative to establishing construct validity and minimizing measurement error. This chapter

discusses the DIF detection method performance in terms of type I error and rejection rates, the practical implications of the simulation results, and the connection between DIF detection methods for data fitted to the MGRM framework and other similar DIF simulation studies.

Type I Error Rates

Type I error rates indicate the probability of rejecting a true null hypothesis (Coladarei et al., 2018). Ideally, a researcher wants to mitigate the occurrence of type I error to ensure that instances of non-DIF are not identified as DIF. The repercussions for type I error results in failed claims of construct validity, recommendations to omit or alter a properly functioning item, or incorrect statements of DIF. Results from the simulation study provide information on the occurrence of type I error for the MIRT-LR test, MIMIC-interaction model, and MLDFA method.

MIRT-LR test. The MIRT-LR test produces reasonable, but slightly inflated (i.e., .06 - .08), type I errors for uniform DIF detection when the latent mean differences between groups are balanced, but unacceptably high type I error rates when the latent mean differences are unbalanced. These results appear to be consistent across different correlation effects and sample size. In many DIF studies, latent means are fixed to 0 for both groups to help identify the model (Steinmetz, 2010); therefore, it is unlikely the MIRT-LR test produce excessively high type I error rates in practice, if the aforementioned approach is used to identify the model and estimate the intercepts (i.e., thresholds).

The simulation results indicate that the MIRT-LR test may be a reasonable detection method for uniform DIF when latent mean differences between the reference and focal groups are balanced. This conclusion coincides with the Elosua and Wells (2013) study that found type I error rates are well controlled in non-DIF conditions (i.e., 0% DIF) for polytomous data fitted to

the GRM. In addition, Cohen et al. (1996) found the use of the IRT-LR test on data that underlies the 2-PL IRT model, a special case of the GRM, also produced acceptable type I error rates. Although the results indicate the MIRT-LR test may not induce problematic type I error rates in practice, extreme caution is advised when the MIRT-LR test is to detect uniform DIF under the MGRM framework, particularly when group mean differences are unbalanced.

In general, the MIRT-LR appears to perform better than the MDLFA method at reducing the likelihood of incorrectly detecting nonuniform DIF. This is consistent with the findings from Bulut and Suh (2017) study that also found the MIRT-LR test produced better type I error control than multiple logistic regression method. However, that study examined type I error rates as any type of DIF detection, uniform and nonuniform, on non-DIF detection. The MIRT-LR test mostly produces acceptable type I error rates for nonuniform DIF detection regardless of the sample size, which is similar to findings investigating DIF detection in unidimensional GRM (Kim & Cohen, 1998).

MIMIC-interaction model. The MIMC-interaction model produces adequate levels of type I error across all conditions when detecting uniform DIF and only slightly increased when attempting to detect nonuniform DIF. The model frequently exhibits lower type I error rates than the other two DIF detection methods when detecting uniform DIF. In particular, an inferential test shows that the MIMIC-interaction model produces substantially lower type I error rates than the MIRT-LR test and MLDFA method. These findings are in congruence with Woods et al. (2009) who found type I error rates for the MIMIC approach to be less than IRT-LR test across different sample sizes. Other studies also found that the MIMIC-interaction model produced inflated type I error rates to detect DIF in 2PL-MIRT models (Grimm & Woods, 2011; Lee et al., 2017), similar to the findings in this study. Although there is no distinctive pattern across the

factor constraints on type I error rates, the MIMIC-interaction model does appear to over inflate type I error under some extreme conditions. In particular, the combination of unbalanced latent mean differences between groups, small sample sizes, and high correlations between latent traits has a large type I error rate (i.e., .16). Increased sample sizes and an assumption of balanced latent traits radically improve the use of the model to detect no DIF in these situations.

MLDFA method. The average occurrence for type I error is consistent when the MLDFA method tested uniform and nonuniform DIF detection on non-DIF items, but the method regularly produces inflated type I error rates for all combinations of factor constraints. In fact, the type I error rates surpass the nominal alpha level of .05 under all factor constraint conditions. That is, the MLDFA method incorrectly rejects the null hypothesis in more than 5% of random trials across every factor constraint combination. In the Bulut and Suh (2017) study, the multiple logistic regression model, a similar approach to the MLDFA, was used to detect DIF for dichotomous MIRT data. Their findings indicate the multiple logistic regression model consistently inflated type I error rates under most factor constraints, as well. Thus, a more stringent alpha level (e.g., .01) is advisable when using the MLDFA method; however, further research is necessary to offer a better rule of thumb for setting the alpha level.

Su and Wang (2005) noted the logistic discriminant function analysis, the unidimensional variant of the MLDFA, produced lower type I error rates when the magnitude of the DIF was small. That is, type I error rates were better controlled when the discrimination and threshold parameters of the reference and focal groups were roughly equal. In this study, discrimination and threshold parameters in non-DIF items is estimated to be equal through the data generation process. Despite this constraint, type I error is abundant across non-DIF items; however, over selection of anchor items, such as the case in this study, can increase the likelihood that freely

estimated parameters are unequal between groups and lead to inflated type I error rates (Stark et al., 2006; Wang & Yeh, 2003; Woods, 2009b). More intentional research directed toward the effect of the magnitude of DIF could provide better insight into the nature of type I error rates from the MLDFA method.

Type of DIF. Type I error rates significantly differ between uniform and nonuniform detection of non-DIF items for the MIMIC-interaction model and MIRT-LR test. Woods and Grimm (2011) initially introduced the use of the interaction effect in the MIMIC model to detect nonuniform DIF. Their study on unidimensional data indicated that type I error rates were inflated when the interaction effect was applied to the model, but reasonable when an interaction effect was not used to detect uniform DIF in non-DIF items. Similarly, this study also finds the interaction effect produces overinflated type I error rates (i.e., .03 - .16) to detect nonuniform DIF in non-DIF items. The MIMIC model also identifies few instances of uniform DIF in non-DIF items.

False identification of uniform DIF from the MIRT-LR test occurs more often than nonuniform DIF detection in non-DIF items. In addition, the variability of uniform DIF detection in non-DIF items is larger than nonuniform DIF detection in non-DIF items. These inferential findings suggest that the MIRT-LR test may not be suitable to reliability reduce the occurrence of type I error for uniform DIF detection. This runs counter to other studies that have suggested the use of the MIRT-LR test on multidimensional data based on acceptable type I error rates for both uniform and nonuniform DIF detection (Bulut & Suh, 2017; Lee et al., 2017; Oshima et al., 1997). Cohen et al. (1996), however, indicated that type I error rates for nonuniform DIF detection with data fitted to the 2PL-IRT model, a special case of the GRM, were approximately

in the expected theoretical range, similar to the findings in this study. Thus, the MIRT-LR test could be a viable option for nonuniform DIF detection, if limiting type I error takes priority.

Latent mean differences across groups. The MIRT-LR test controls type I error better when the latent mean differences between the reference and focal groups are balanced (or equal). This result runs in contrast with the findings from data fitted to 2PL-IRT model and GRM, such that latent mean differences between groups had no substantive effect on the type I error rates when using the MIRT-LR test (Stark et al, 2006). However, the latent mean difference between groups does not largely affect type I error rates for the MIMIC-interaction model, nor the MLDFA method. These results are similar to other findings from studies with multidimensional data (Bulut & Suh, 2017; Lee et al., 2017).

Latent mean differences between groups are impossible to control and estimate without influencing the discrimination and threshold parameters (Müller & Schäfer, 2017; Steinmetz, 2010). Researchers have the option to fix the corresponding latent means equal between the groups, or fix the latent means of one group to a value (typically 0) and allow the latent means to freely estimate for the other group (Steinmetz, 2010). The former choice results in freely estimated discrimination and threshold parameters. In addition, the selection of a DIF detection method is inconsequential to the degree of type I error present. The latter decision to fix only one group latent means will likely create unbalanced latent mean differences. If the decision is to allow a group to freely estimate its latent means, then either the MIMIC-interaction model or MDLFA method is advisable to control type I error.

Correlation between latent traits. The correlation between latent traits does not meaningfully affect the likelihood of a type I error occurrence for any of the DIF detection methods. In other words, different correlation sizes between the latent traits produce similar type

I error rates for all three DIF detection methods. Further examination of the main effects shows that the different correlations sizes does not affect the type I error occurrence when controlling for the DIF detection method. In addition, the effect size of the correlation between latent traits is zero ($\omega^2 = .00$), indicating the factor has virtually no effect on the type I error rates. Similarly, correlations between latent traits, when all other factors are controlled, demonstrated no significant effect on type I error rates in a 2-PL MIRT model (Bulut & Suh, 2017; Mazor et al., 1998). Liaw (2015) also found that the correlation between latent traits for data fitted to the 2PL-MIRT model had no effect on type I error rates ($\omega^2 = .00$). Therefore, consideration of the correlation between latent traits is not necessary to reduce the possibility of type I error.

Sample Size. Sample size has a minimal effect on type I error rates across all three DIF detection methods. In fact, type I error rates mainly stay consistent across all four sample sizes when the DIF detection method is controlled. This conclusion appears to align with previous studies that examined the effect on sample size on DIF detection in non-DIF items (Bulut & Suh, 2017; Kim & Cohen, 1998). Despite no difference in type I error occurrence between sample sizes, the probability of detecting DIF in non-DIF items is large across all sample sizes. The mean odds for rejecting the true null hypothesis range from 1 for every 8.17 to 9.21 non-rejections of the true null hypothesis. That is, DIF detection of non-DIF items occurs in 10.9% to 12.2% of cases; thus, supplementing further cause for lowering the nominal alpha level when detecting DIF.

Rejection Rates

Rejection rates refer to the probability of rejecting a false null hypothesis (Coladarei et al., 2018). In other words, the rejection rates indicate the statistical power of a DIF detection method. Statistical power is essential to claim an instrument contains construct validity. High rejection

rates indicate the absence or near absence of type II error, which results from the non-rejection of a false null hypothesis (Coladarci et al., 2008). In this study, type II error occurs when instances of DIF are reported as non-DIF. Type II error can result in false conclusions of construct validity, failure to omit or alter an improperly functioning item, or incorrect claims of measurement invariance. Results from the simulation study provide information on the proportion of correct rejections of the null hypothesis (i.e., statistical power) for the MIRT-LR test, MIMIC-interaction model, and MLDFA method.

MIRT-LR test. The MIRT-LR test demonstrated adequate power across different percentages of DIF when detecting uniform DIF. In particular, the test performs best when the latent mean difference between groups are unbalanced, but still appear to display adequate statistical power for balanced latent mean differences, provided the sample size is large (i.e., R1000/F500) and percentage of DIF items is greater than 5%. The MIRT-LR test performs increasingly better as the percentage of DIF items also increases. This is similar to the findings in the Lee et al. (2017) study that found instruments with greater percentage of DIF items, between 20% and 40% DIF, produced high rates of statistical power under uniform DIF for the MIRT-LR test. Despite increased DIF detection performance with high percentage of DIF, the MIRT-LR test does not exhibit as high rates of statistical power as the MLDFA method when detecting uniform DIF with low sample sizes. Similarly, the logistic discriminate function analysis displayed higher statistical power than the IRT-LR test for data fitted to the unidimensional GRM when detecting uniform DIF with low sample sizes (Elosua & Wells, 2013).

The MIRT-LR test fails when attempting to detect nonuniform DIF. Overall, the MIRT-LR test rejects the false null hypothesis in about 64% of cases, indicating that the occurrence of a type II error is nearly 40%. These results are indicative of the poor performance of the MIRT-LR

test to detect nonuniform DIF. As a result, the MIRT-LR test is an unreliable DIF detection method, specifically in detection of nonuniform DIF. In contrast, the MIRT-LR test performed well under the multidimensional 2-PL IRT model (Bulut & Suh, 2017), unidimensional 2-PL IRT model (Woods & Grimm, 2011), and the unidimensional GRM (Elosua & Wells, 2013). In fact, the MIRT-LR test produced noticeably higher rejection rates than the MIMIC model, multiple logistic regression model, and logistic discriminate function analysis when detecting nonuniform DIF under other IRT frameworks, regardless of the factor constraints (Bulut & Suh, 2017; Elosua & Wells, 2013).

MIMIC-interaction model. The MIMIC-interaction model produces adequate levels of statistical power when controlling all factor constraints. The model correctly detects DIF in approximately 93% of cases. In general, the MIMIC-interaction model shows higher rejection rates than MIRT-LR test. A similar finding was also discovered when the 2-PL IRT model underlies the data (Finch, 2005). The MIMIC-interaction model produces higher average rejection rates under nonuniform DIF detection than uniform DIF detection, except when data is 20% DIF. A similar conclusion was presented in an earlier study that also examined the multidimensional MIMIC-interaction model under the 2PL-MIRT model (Lee et al., 2017). However, the MIMIC-interaction model still yields high rates of statistical power when detecting nonuniform DIF in DIF items. Rejection rates remain above .80 across all factor constraint combinations when detecting nonuniform DIF. Statistical power above 80% indicates that the MIMIC-interaction model produces acceptable levels of statistical power when detecting DIF across a multitude of factor combinations (Cohen, 1988).

MLDFA method. Overall, the MLDFA method produces higher statistical power than the MIMIC-interaction model and MIRT-LR test when controlling all factor constraints. The

MLDFA demonstrates slightly higher rejection rates than the MIRT-LR test when detecting uniform DIF, and much higher statistical power when detecting nonuniform DIF and data fitted to the MGRM. Elosua and Wells (2013) found that the univariate logistic discriminant function analysis also displayed marginally higher rejection rates in uniform DIF detection than the IRT-LR test under the unidimensional GRM, but, in direct contrast, they showed that the IRT-LR test had higher statistical power in nonuniform DIF detection than the logistic discriminant function analysis. The difference in the Elousa and Wells (2013) study and this study is best reflected in the poor detection capability of the MIRT-LR test rather than the strength of the MLDFA method.

The simulation indicates the MLDFA method rejects the false null hypothesis in roughly 96% of cases. This shows the MLDFA method is a powerful method to detect DIF. Su and Wang (2005) found the univariate logistic discriminant function analysis also produced higher rejection rates than other comparable polytomous DIF detection procedures. Furthermore, the MLDFA method proves to reliably measure uniform and nonuniform DIF in DIF items, provided the percentage of DIF is not small (i.e., 5%) and latent mean differences are unbalanced. The univariate logistic discriminate function analysis also exhibited strong statistical power for uniform DIF detection in a similar simulation study, but found the method did not produce high rejection rates when detecting nonuniform DIF (Kristjansson et al., 2005).

Percentage of DIF. The percentage of DIF in an instrument has a minimal effect on rejection rates across the three DIF detection methods. Although rejection rates do not meaningfully differentiate between different percentages of DIF in the instrument, the statistical power does appear to be within an acceptable range for 5%, 10%, and 20% DIF. The correct detection of DIF occurs in 85.4% to 91.1% of cases, indicating the likelihood of correctly rejecting the null

hypothesis steadily, but nonsignficantly, increases as the percentage of DIF increased. Most similar DIF studies under the unidimensional framework and multidimensional, dichotomous framework showed that statistical power proportionally increased with the percentage of DIF, but all of these studies indicated significant changes between different percentages of DIF (Belzak, 2019; Bulut & Suh, 2017; Finch, 2005; Hildalgo & Gómez, 2006; Wang & Shih, 2010).

Lee et al. (2017) demonstrated that rejection rates tend to be greater when the percentage of DIF was large (i.e., 40%). In contrast, Bulut and Suh (2017) found that the MIMIC-interaction model, MIRT-LR test, and multiple regression all exhibited high rejection rates for uniform DIF detection when percentage of DIF was low. This Monte Carlo simulation finds high rejection rates for uniform DIF detection across all three DIF detection methods, but results indicate that rejection rates stay relatively high for uniform DIF detection, regardless of the percentage of DIF items in the instruments. Moreover, rejection rates are consistent across different percentages of DIF when detecting nonuniform DIF for all three DIF detection methods, with the exception that the MIRT-LR test slightly increases statistical power under a large percentage of DIF (i.e., 20%).

Type of DIF. Rejection rates are considerably different between uniform and nonuniform DIF for the MIRT-LR test, but roughly equal for the MLDFA method and MIMIC-interaction model. A comparable early study on the IRT-LR test with data fitted to the 2PL-IRT model indicated that uniform DIF detection is likely to produce higher rejection rates than nonuniform DIF when controlling for other factors (Thissen et al., 1993). In this study, the MIRT-LR test detects uniform DIF items in roughly 95.5% of cases, but only 13.1% of tests detect DIF in nonuniform DIF items. The discrepancy between the two types of DIF detection is stark, but examination of the item parameters could explain this palpable difference.

In polytomous models, nonuniform DIF was most distinguishable when the discrimination parameters were considerably different between groups, while only small group differences in the threshold parameters were needed to detect uniform DIF (French & Miller, 1996). Suh and Cho (2014) showed that discrimination parameters that differed widely between groups (e.g., a_{1F} = 1.0, a_{1R} = .7), which they referred to as high magnitude DIF, produced much higher rejection rates than discrimination parameters that only differed slightly between groups (e.g., a_{1F} = 1.3, a_{1R} = .7), known as low magnitude DIF. Tests on uniform DIF detection indicate little discrepancies between low and high magnitude DIF parameters. The threshold differences between the reference and focal groups ($\Delta\tau_n$ = .5) are similar to threshold and difficulty parameters in other DIF simulation studies (Bulut & Suh, 2017; Lee et al., 2017; Oshima et al., 1997; Suh & Cho, 2014). Rejection rates for the MIRT-LR test on uniform DIF detection in this simulation study are similar to MIRT-LR test rejection rates in other multidimensional DIF studies (Bulut & Suh, 2017; Suh & Cho, 2014). The discrimination parameters, however, are similar to parameters found in low magnitude DIF and rejection rates are much lower for the MIRT-LR test with these parameters. Therefore, it may be worthwhile to examine the effect of the magnitude of DIF in future studies with the MGRM.

Latent mean differences across groups. Latent mean differences between the focal and reference groups are important to investigate, because large enough distributional differences between the groups can influence variations in the ability to detect DIF accurately in the multidimensional context (Oshima et al., 1997). Rejection rates between balanced and unbalanced latent mean differences across groups are relatively similar across DIF detection methods. Moreover, rejection rates do not differ much between the two factor groups regardless of the DIF detection method. This conclusion appears to align with a previous study that

concluded balanced and unbalanced latent mean differences had similar results in a multidimensional dichotomous model (Lee et al., 2017). However, DIF tests with balanced latent mean differences tend to have slightly higher rejection rates than tests with unbalanced latent mean differences between groups. Correct detection of DIF occurs in 86.6% of tests with balanced latent mean differences and 91.4% of tests with unbalanced latent mean differences. Thus, sufficient statistical power is highly probable to occur, regardless of the latent mean differences between groups.

Correlation between latent traits. The correlation between latent traits has a negligible effect on rejection rates across the DIF detection methods. That is, the rejection rates are similar across the MIMIC-interaction model, MLDFA method, and the MIRT-LR test when there is no correlation between latent traits. The same findings are apparent when the correlations are of medium ($\rho = .3$) and large ($\rho = .6$) sizes. The Bulut and Suh (2017) study also found that correlations between latent traits did not different between DIF detection methods for data fitted to the 2PL-MIRT model. Rejection rates stay consistent at different correlational sizes when controlling the DIF detection method, as well. The correct detection of DIF occurs in 90.5% of tests with no correlation ($\rho = 0$), 88.5% of tests with medium correlation ($\rho = .3$), and 88.6% of tests with large correlations ($\rho = .6$). No distinctive pattern of rejection rates is present with correlational size and DIF detection is statistical powerful (i.e., above .80) across all correlation sizes. Other studies also found that the correlation between latent traits did not produce significantly different rejection rates between different correlational sizes (Bulut & Suh, 2017; Lee et al., 2017).

Sample Size. The sample size does not meaningfully affect the correct rejection of the null hypothesis for any of the DIF detection methods, except 20% DIF for the MIMIC model. That is,

different sample sizes produce similar statistical power for all three DIF detection methods. An obvious pattern in the data is apparent; increased probability of correctly rejecting the null hypothesis appears proportional to the increase in sample size, although further examination of the main effect shows the statistical differences in probability between the different sample sizes is negligible. This finding contradicts other findings that suggested sample size might influence DIF detection (Acar, 2011; Elosua & Wells, 2013; French & Miller, 1996; Hildalgo & Gómez, 2006).

Despite differences between sample sizes having little effect on DIF detection, rejections rates are still within an acceptable range for every sample size. Correct DIF detection occurrs in roughly 84.7% to 92.1% of tests. Under these pretenses, sample sizes as low as 750 (i.e., R500/F250), without the influence of other factors, may be sufficient to detect DIF with acceptable levels of statistical power. This conclusion slightly differs from other studies. An early DIF detection study showed that small sample sizes (i.e., $n < 250$) decrease statistical power (Spray, 1989). Scott et al. (2009) and Lai et al. (2005) recommended sample sizes of at least 500 per group to detect DIF under the 2-PL IRT model. Kim and Cohen (1998) lowered that requirement to 300 per group for data fitted to the GRM.

As mentioned in other multidimensional studies, additional factor constraints and DIF detection methods can affect sample size requirements for multidimensional models (Bulut & Suh, 2017; Lee et al., 2017; Oshima et al., 1997). This appears to be true in this study, as well. For instance, high percentage of DIF could serve to detect uniform and nonuniform DIF, but sample size requirements vary based on the correlational size between latent traits. Sample sizes as low as 750 (i.e., R500/F250) provide adequate statistical power for the MIMIC-interaction method and no correlation, but much larger sample sizes (R1000/F1000) are needed to

adequately assess DIF when correlations are large (i.e., $\rho = .6$) and data contain 20% DIF. Future research is required to investigate the influence of other factors on sample size needs for DIF detection of data fitted to the MGRM.

Empirical and Practical Implications

The empirical example provides real-world context to examine the results of the simulation study. The data clearly fit the MGRM framework better than other possible IRT models. In addition, the data are assumed to have balanced latent means between groups such that $\mu_R = 0$, $\mu_F = 0$ for all latent means in each group. After model estimation, the correlation between latent traits are roughly .6 and the sample size best followed R500/F250 from the simulation study. Results reveal comparative and contrasting relationships between the DIF detection methods with empirical data.

In the simulation study, the MLDFA model demonstrates very high rejection rates (i.e., .99) in detecting uniform DIF when correlations between traits are .6, sample size is 500 for the reference group and 250 for the focal group, latent means are balanced between groups, and the percentage of DIF are 20%. This indicates that the MLDFA method correctly identifies DIF nearly every time; however, the MLDFA method also as a higher tendency to incorrectly detect uniform DIF compared to the MIMC model and MIRT-LR test.

In the empirical example, the MLDFA method flags items 2, 7, 9, and 11 as uniform DIF. All three DIF detection methods observe uniform DIF in item 9, which indicates that this item is likely functioning differently between groups. The MLDFA method and MIMIC model both report uniform DIF is present in items 2 and 9, but differentiate on items 7, 11, and 12 for uniform DIF detection. In addition, the MLDFA method and MIRT-LR test both detect uniform DIF in item 11. These results appear to match with the simulation findings. The MLDFA is more

likely to indicate uniform DIF detection, because of high rejection and type I error rates, and the other two DIF detection methods have considerably lower rejection and type I error rates under the given factor constraints.

For nonuniform DIF detection, the MIMIC-interaction model indicates nonuniform DIF in item 2, but the other two methods do not observe nonuniform DIF in any item. The simulation findings suggest that the MIMIC-interaction model correctly detects nonuniform DIF reliably in an instrument with 5% DIF (.94), while the MIRT-LR test (.10) and MLDFA method (.52) correctly detect nonuniform DIF in fewer than 80% of cases. Therefore, the simulation results likely provide reliable benchmarks for empirical data.

This example provides an illustration of what practitioners should carefully examine before deciding on a DIF detection procedure. For instance, the MLDFA method detects four items functioning differently between males and females. The simulation indicates that this method correctly identifies uniform DIF with great precision, but does have a tendency to overestimate the number of uniform DIF items when the alpha level is not low (i.e, $\alpha = .05$). On the other hand, the MIMIC model detects fewer items with uniform DIF than the MLDFA method, but identifies nonuniform DIF detection in one item, whereas the other two methods do not detect nonuniform DIF. The shortcomings of the MIRT-LR test are profound in the empirical example. Not only does the MIRT-LR test fail to detect nonuniform DIF on any item, the test indicates that the probability of nonuniform DIF detection is exceedingly low across all items. Thus, practitioners should consider the rejection and type I error rates of each DIF detection procedure based on the contributing factors before selection of a method. In some cases, one method may be better suited for uniform DIF detection testing and another for nonuniform DIF detection testing.

Limitations and Future Research

This study is limited in scope based on the fixed constraints selected for this study. First, the results are only generalizable for a 20 item, 5-option scaled instrument with two groups. One needs to exercise caution when justifying the use of a DIF detection method based on the results of this study. For instance, some studies may want to test multiple groups for DIF simultaneously, such as marital status and gender, or contain a non-dichotomized categorical variable to test DIF, such as race or class standing. In these instances, some DIF detection methods may not be possible to conduct, such as the MLDFA method to test DIF among different racial groups, or the addition of another covariate group to the MIMIC-interaction model may influence the probability of DIF detection for the other tested covariate.

Other factors using UIRT and MIRT models can influence the statistical power and type I error rates of some DIF detection methods. For example, discrimination and threshold parameters that vary in magnitude between groups have shown to significantly differ in type I error and rejection rates for data fitted to the 2PL-MIRT and GRM models (Bulut & Suh, 2017; Lee et al., 2017; Oshima et al., 1997; Suh & Cho, 2014). In this study, group differences between threshold parameters are fixed to .5 and group differences between discrimination parameters only vary between .3 and .5, but these values are not considered in the study. Finch (2005) examined different instrument sizes and found that instruments that contained more items produced higher statistical power. More specifically, instruments with more anchor items correctly identified DIF more often than instruments with fewer anchors. Stark et al. (2005) discovered that increasing the number of response options also increases the statistical power of the DIF detection method. Future DIF studies should consider the effects of varying magnitudes

of item parameters, instrument size, and the number of response options to deepen understanding on the reliability of these DIF detection methods under the MGRM framework.

Next, the simulation results are restricted to data fitted to the MGRM. Other polytomous MIRT frameworks, such as the multidimensional generalized partial credit model (MGPCM), need to be considered as viable frameworks for DIF analysis. Similar to the MGRM, the MGPCM is a polytomous extension of the 2PL-IRT model (Embertson & Reise, 2000; Muraki, 1993; Samejima, 2010); thus, the same procedures could test for DIF for data fitted to the MGPCM. The MGRM and MGPCM, however, do not estimate comparable item parameters, since the mathematical structure differs between the models, but Bolt (2002) indicated that data fitted to the GPCM could fit to the GRM, which was later supported by Kang et al. (2009). DeMars (2007) found that that type I error rates of the same DIF detection method were equally inflated for data fitted to the GRM and GPCM. Similar findings could occur in a future comparative study of DIF detection procedures with MGRM and MGPCM data, which could extend support for the use of these procedures to data fitted to the MGPCM. Future DIF research should concentrate on the implications of DIF detection methods when data are fitted to the MGPCM and how it compares to data fitted under the MGRM.

Finally, the simulated data assumes a uniform distribution across response options; however, survey results, evaluation forms, and rubric scores have a tendency to be skewed. Gelin and Zumbo (2007) examined the effects of the skewness on DIF detection in a unidimensional DIF study and found that type I error rates increased with skewed data. Kristjansson et al. (2005) found little difference in type I error and rejection rates between skewed and symmetric ability distributions in polytomous data. Since these studies have conflicting conclusions on the role of

skewness in DIF detection, skewness should be examined in future DIF studies to determine its effect on the DIF detection procedures in MGRM data.

Recommendations and Conclusion

This study provides a comparative analysis of the MIRT-LR test, the MIMIC-interaction model, and the MLDFA method for data fitted to the MGRM. In particular, type I error rates inflation appear to occur across all of the DIF detection methods. The MIMIC model exhibits significantly lower mean type I error rates than the other two DIF detection procedures. The MLDFA method and MIRT-LR test display negligibly different type I error rates. The MIMIC model shows lower type I error rates when detecting uniform DIF than nonuniform DIF, but the MIRT-LR test displays lower type I error rates when detecting nonuniform compared to uniform DIF. Rejection rates indicate acceptable statistical power for the MIMIC-interaction model and MLDFA under most factor conditions, but the MIRT-LR test has inadequate statistical power when detecting nonuniform DIF. Comparatively, the MLDFA method shows significantly higher mean rejection rates than the other two DIF detection procedures. The MIRT-LR test has the lowest mean rejection rates and large variability across the different factor constraint combinations.

Additionally, inferential analyses on the effects of common factor constraints provide context on strength and reliability of each DIF detection procedure. Sample size, correlation between latent traits, and latent mean differences do not appear to show significant differences in mean rejection rates between different factor condition groups. That is, the rejection rate for a sample size of 2000 (R1000/F1000) is similar to a sample size of 750 (R500/F250), a large correlation ($\rho = .6$) is similar to no correlation ($\rho = 0$), and a balanced latent mean difference between groups is similar to an unbalanced latent mean difference. Type I error rates indicate a similar pattern for

sample size and correlation between latent traits. The balanced latent mean differences between groups, however, display lower type I error rates than the unbalanced latent mean differences for the MIRT-LR test.

In general, the MIMIC-interaction model provides the most dependable measure of DIF in MGRM data. Although the MLDFA method indicates strong statistical power, it has inflated type I error rates. The MIRT-LR test exhibits egregious discrepancy between uniform and nonuniform DIF detection that limit the utility of the method. The MIMIC-interaction provides consistent power across all factor constraints that only slightly differs from the MDLFA method, and provides reasonably low type I error rates that can be easily controlled through lower alpha levels. Therefore, the following recommendations are advised for practitioners who are examining DIF with data fitted to the MGRM:

- MIMIC-interaction model is the recommended DIF detection procedure for data fitted to MGRM when lower type I error takes precedent; the MLDFA method is suggested when strong statistical power is paramount.
- Inflated type I error rates are likely to incur false rejections of the null hypothesis. Alpha level reduction is recommended to alleviate the chance of false rejection. Although a sufficient rule of thumb cannot be provided in this study, the alpha level should be lower than .05, but not so low as to limit the statistical power of the analysis.
- A sample size of 750 (R500/F250) is likely to provide adequate power to detect DIF for all three methods; however, a sample of 750 appears to be near the threshold for unacceptable power. A larger, more balanced sample size, such as 1000 (R500/F500), is a more preferable minimum sample size requirement, especially if the alpha level is lowered to account for inflated type I error rates.

It is important to state that these recommendations are generalized guidelines for DIF detection. The decisions to use a particular DIF detection procedure are conditional and predicated on the factor constraints and availability of software to conduct the necessary analyses (e.g., currently, only Mplus can run a MIMIC-interaction model). The findings from this study are useful for practitioners to reference in order to effectively measure DIF and improve construct validity, but future research is still needed to improve guidelines for factor constraints not explored in this study and to control inflated type I error rates.

References

Acar, T. (2011). Sample size in differential item functioning: An application of hierarchical linear modeling. *Educational Science: Theory and Practice*, *11*(1), 284-288.

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255-278.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

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

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Lawrence Erlbaum Associates.

Armstrong, R.A. (2014). When to use the Bonferroni correction. *Ophthalmic & Physiology Optics*, *34*(5), 502-508. https://doi.org/10.1111/opo.12131

Asmussen, S., & Glynn, P.W. (2007). *Stochastic simulation: Algorithms and analysis*. Springer.

Bandalos, D.L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 102-116. https://doi.org/10.1080/10705511.2014.859510

Barendse, M.T., Oort, F.J., Werner, R.L., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(4), 561-579. https://doi.org/10.1080/10705511.2012.713261

Belzak, W.C.M. (2019). Testing differential item functioning in small samples. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2019.1671162

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289-300.

Berger, M., & Tutz, G. (2016). Detection of uniform and non-uniform differential item functioning by item focused trees. *Psychometrika*, *81*(3), 727-750. https://doi.org/10.1007/s11336-015-9488-3

Bergsteiner, H., & Avery, G.C. (2014). The twin-cycle experiential learning model: Reconceptualising Kolb's theory. *Studies in Continuing Education, 36*(3), 257-274. https://doi.org/10.1080/0158037X.2014.904782

Bergsteiner, H., Avery, G.C., & Neumann, R. (2010). Kolb's experiential learning model: critique from a modelling perspective. *Studies in Continuing Education*, *32*(1), 29-46. https://doi.org/10.1080/01580370903534355

Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Brown, T.A., & Moore, M.T. (2012). Confirmatory factor analysis. In R.H. Hoyle (Ed.), *Handbook of structural equation modeling*, (pp. 361-379). Guilford Press.

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

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113-141. https://doi.org/10.1207/S15324818AME1502_01

Bolt, D.M., & Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395-414. https://doi.org/10.1177/0146621603258350

Bulut, O., & Suh, Y.-S. (2017). Detecting multidimensional differential item functioning with multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*(51). https://doi.org/10.3389/feduc.2017.00051

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*, 245-276.

Camilli, G. (2006). Test fairness. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 220-256). American Council on Education.

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for R environment. *Journal of Statistics Software*, *48*(6), 1-29. https://doi.org/0.18637/jss.v048.i06

Chalmers, R.P., & Adkins, M.C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248-280. https://doi.org/10.20982/tqmp.16.4.p248

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaption of the SIBTEST procedure. *Journal of Educational Measurement*, *33*(3), 333-353.

Cohen, A.S., Kim, S.-H., & Baker, F.B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*(4), 335-350. https://doi.org/10.1177/014662169301700402

Cohen, A.S., Kim, S.-H., & Wollack, J.A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26. https://doi.org/10.1177/014662169602000102

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Coladarci, T., Cobb, C.D., Minium, E.W., & Clarke, R.B. (2008). *Fundamentals of statistical reasoning in education* (2nd ed.). John Wiley & Sons.

Darlington, R.B., & Hayes, A.F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford Press.

De Ayala, R.J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*(2), 155-170, https://doi.org/10.1177/014662169401800205

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

DeMars, C.E. (2008). Polytomous differential item functioning and violations of ordering of the expected latent trait by the raw score. *Educational and Psychological Measurement, 68*(3), 379-396. https://doi.org/10.1177/0013164407308511

DeMars, C.E. (2010). *Item response theory*. Oxford University Press.

DeMars, C.E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement*, *71*(4), 597-616. https://doi.org/10.1177/0013164411404221

Depaoli, S., Tiemensma, J., & Felt, J.M. (2018). Assessment of health surveys: Fitting a multidimensional graded response model. *Psychology, Health, and Medicine*, *23*(1), 1299-1317. https://doi.org/10.1080/13548506.2018.1447136

Edwards, J.R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods*, *4*(2), 144-192.

Elosua, P., & Wells, C.S. (2013). Detecting DIF in polytomous items using MACS, IRT, and ordinal logistic regression. *Psicológica*, *34*, 327-342.

Embertson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Psychology Press.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Sage Publications.

Finch, 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.

Frane, A.V. (2015). Are per-family type I error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods*, *14*(1), 12-23. https://doi.org/10.22237/jmasm/1430453040

French, A.W., & Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315-332.

Furr, R.M. (2018). Item response theory and Rasch models. In *Psychometrics: An introduction* (3rd ed., pp. 455-487). Sage Publications.

Gana, K., & Broc, G. (2019). *Structural equation modeling with lavaan*. Wiley.

Gelin, M.N., & Zumbo, B.D. (2007). Operating characteristics of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation for short scales. *Journal of Modern Applied Statistical Methods*, *6*(2), 573-588. https://doi.org/10.22237/jmasm/1193890860

Golia, S. (2012). Differential item functioning classification for polytomously scored items. *Electronic Journal of Applied Statistical Analysis*, *5*(3), 367-373.

Hallgren, K.A. (2014). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 43-60.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer.

Hanson, B.A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics*, *23*(3), 244-253.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Higher Education*, *35*(2), 57-63. https://doi.org/10.1016/j.stueduc.2009.10.002

Hidalgo, M.D., & Gómez, J. (2006). Non-uniform DIF detection using discriminant logistic analysis and multinominal logistic regression: A comparison for polytomous items. *Quality and Quantity*, *40*, 805-823. https://doi.org/10.1007/s11135-005-3964-2

Higdon, R. (2013). Multiple hypothesis testing. In W. Dubitzky, O. Wolkenhauer, K.H. Cho, & H. Yokota (Eds.). *Encyclopedia of Systems Biology*. Springer. https://doi.org/10.1007/978-1-4419-9863-7_1211

Holland, P.W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 129-145). Lawrence Erlbaum Associates.

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

Immekus, J.C., & Imbrie, P.K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: An illustration with the state metacognitive inventory. *Educational and Psychological Measurement*, *68*(4), 695-709. https://doi.org/10.1177/0013164407313366

Immekus, J.C., Snyder, K.E., & Ralston, P.A. (2019). Multidimensional item response theory for factor structure assessment in educational psychological research. *Frontiers in Education*, *4*(45). https://doi.org/10.3389/feduc.2019.00045

Jiang, S., Wang, C., & Weiss, D.J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*(109). https://doi.org/10.3389/fpsyg.2016.00109

Jin, K., & Chen, H. (2020). MIMIC approach to assessing differential item functioning with control of extreme response style. *Behavioral Research*, *52*, 23-35. https://doi.org/2010.3758/s13428-019-01198-1

Kamata, A., & Bauer, D.J. (2008). A note on the relation between factor analytic and item response models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136-153. https://doi.org/10.1080/10705510701758406

Kassambara, A. (2020). *Rstatix: Pipe-friendly framework for basic statistical tests*. R package v. 0.6.0. https://CRAN.R-project.org/package=rstatix

Kang, T., Cohen, A.S., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*(7), 499-518. https://doi.org/10.1177/0146621608327800

Kelley. K. (2010). Monte Carlo simulation. In N.J. Salkind (Ed.), *Encyclopedia of research design*, (pp. 1065-1067). Sage Publications.

Kelloway, E.K. (2015). *Using Mplus for structural equation modeling: A researcher's guide* (2nd ed.). Sage Publications.

Kim, S.-H., & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345-355.

Kim, S.-H., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 212-228. https://doi.org/10.1080/10705511.2011.557337

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457-474.

Kline, R.B. (2012). Assumptions in structural equation modeling. In R.H. Hoyle (Ed.), *Handbook of structural equation modeling*, (pp. 111-125). Guilford Press.

Kolb, D.A. (1984). *Experiential learning: Experience as a course of learning and development*. Prentice-Hall.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B.D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Kwakkenbos, L., Willems, L.M., Baron, M., Hudson, M., Cella, D., van den Ende, C.H.M., Thombs, B.D., & Canadian Scleroderma Research Group. (2014). The comparability of English, French, and Dutch scores on the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F): An assessment of differential item functioning in patients with systemic sclerosis. *PLoS One*, *9*(3), e91979, https://doi.org/10.1371/journal.pone.0091979

Lai, J.-S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation and Health Professions*, *28*(3), 283-294. https://doi.org/10.1177/0163278705278276

Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of the multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*, *77*(4), 545-569. https://doi.org/10.1177/0013164416651116

Lee, S., & Lee, D.K. (2018). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*,*71*(5), 353-360. https://doi.org/10.4097/kja.d.18.00242

Lei, P.-W., & Shiverdecker, L.K. (2020). Performance of estimators for confirmatory factor analysis of ordinal variables with missing data. *Structural Equation Modeling*, *27*(4), 584-601, https://doi.org/10.1080/10705511.2019.1680292

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of*

Experimental Social Psychology, *49*(1), 764-766. http://doi.org/10.1016/j.jesp.2013.03.013

Liaw, Y.-L. (2015). *When can multidimensional item response theory (MIRT) models be a solution for differential item functioning (DIF)?: A Monte Carlo simulation study*. (Doctoral dissertation). Available from University of Washington Research Works Archive (ID No. 0250E_15172).

Litière, S. (2010a). Type II error. In N.J. Salkind (Ed.), *Encyclopedia of research design*, (pp. 1577-1579). Sage Publications.

Litière, S. (2010b). Type I error. In N.J. Salkind (Ed.), *Encyclopedia of research design*, (pp. 1574-1577). Sage Publications.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936-949. https://doi.org/10.3758/s13428-015-0619-7

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847-862.

Magis, D., & De Boeck, P. (2014). Type I error inflation of DIF identification with Mantel-Haenszel: An explanation and a solution. *Educational and Psychological Measurement*, *74*(4), 713-728. https://doi.org/10.1177/0013164413516855

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*, (pp. 436-458). Guilford Press.

Martin, M.A., & Roberts, S. (2010). Beta. In N.J. Salkind (Ed.), *Encyclopedia of Research Design*, (pp. 81-82). Sage Publications.

Maslowsky, J., Jager, J., & Hemken, D. (2015). Estimating and interpreting latent variable interactions: A tutorial for applying the latent moderated structural equations method. *International Journal of Behavioral Development*, *39*(1), 87-96. https://doi.org/10.1177/0165025414552301

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

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary research and perspectives*, *11*(3), 71-101. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55-64.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009-1020.

Mazor, K.M., Hambleton, R.K., & Clauser, B.E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement*, *22*(4), 357-367. https://doi.org/10.1177/014662169802200404

McDonald, R.P. (1997). Normal ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). Springer.

Miller, T.M., & Spray, J.A. (1993). Logistic discriminant function analysis for DIF identification of polytomously score items. *Journal of Educational Measurement*, *30*, 107-122.

Montoya, A.K., & Jeon, M. (2020). MIMIC models for uniform and non-uniform DIF as moderated mediation models. *Applied Psychological Measurement*, *44*(2), 118-136. https://doi.org/10.1177/0146621619835496

Müller, P., & Schäfer, S. (2017). Latent mean (comparison). In J. Matthes, C.S. Davis, & R.F. Potter (Eds.), *The international encyclopedia of communication research methods*. John Wiley & Sons.

Muraki, E. (1993). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E., & Carlson, J.E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*(1), 73-90.

Murrar, S., & Brauer, M. (2018). Mixed model analysis of variance. In B.B. Frey (Ed.), *The Sage encyclopedia of educational research, measurement, and evaluation* (pp. 1075-1078). Sage Publications.

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (8th ed.). Author.

Navas-Ara, M.J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment*, *18*, 9-15.

Nering, M.L. & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models: Developments and applications*. Taylor and Francis Group.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434-447. https://doi.org/10.1037/1082-989X.8.4.434

Osbourne, J.W. (2017). *Regression and linear modeling: Best practices and modern methods*. Sage Publications.

Oshima, T.C., & Miller, M.D. (1993). Multidimensionality and item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14*, 163-173.

Oshima, T.C., Raju, N.S., & Flowers, C.P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement*, *34*(3), 253-272.

Osterlind, S.J., & Everson, H.T. (2009). *Differential item functioning* (2nd ed.). Sage Publications. https://doi.org/10.4135/9781412993913.d3

Ostini, R., Finkelman, M., & Nering, M. (2014). Selecting among polytomous IRT models. In S.P. Reise & D.A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.

Ostini, R., & Nering, M. (2006). *Polytomous item response theory models*. Sage Publications.

Penfield, R.D. (2010a). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47*, 129-149.

Penfield, R.D. (2010b). Explaining crossing DIF in polytomous items using differential step functioning effects. *Applied Psychological Measurement, 34*(8), 563-579.

Penfield, R.D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353-370.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics vol. 26: Psychometrics* (pp. 125-167). Elsevier.

Polites, G., Roberts, N., & Thatcher, J.B. (2012). Conceptualizing models using multidimensional constructs: A review and guidelines for their use. *European Journal of Information Science Systems*, *21*(1), 22-48. https://doi.org/10.1057/eijis.2011.10

Potenza, M.T., & Dorans, N.J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*(1), 517-529.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502.

R Core Team (2020). *R: A language and environment for statistical computing*. Retrieved from https://www.r-project.org/

Reckase, M.D. (2009). *Multidimensional item response theory*. Springer.

Reise, S.P., Cook, K.F., & Moore, T.M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S.P. Reise & D.A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.

Rhodes, T. (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Association of American Colleges and Universities.

Rogers, H. J., & Swaminathan, H. (2016). Concepts and methods in research on differential item functioning of test items: Past, present, and future. In C. Wells & M. Faulkner-Bond

(Eds.), *Educational measurement: From foundations to future* (pp. 126-142). Guilford Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *4*8(2), 1-36.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph 17*(34).

Samejima, F. (2010). The general graded response model. In M.L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 77-107). Taylor and Francis Group.

Scott, N.W., Fayers, P.M., Aaronson, N.K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Peterson, M.A., Sprangers, M.A., EORTC Quality of Life Group, & Quality of Life Cross-Cultural Meta-Analysis Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*(3), 288-295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Sigel, M.J., & Chalmers, R.P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education*, *24*(3), 136-156. https://doi.org/10.1080/10691898.2016.1246853

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Sloan, I.H., & Woźniakowski, H. (1998). When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? *Journal of Complexity*, *14*(1), 1-33. https://doi.org/10.1006/jcom.1997.0463

Spray, J.A. (1989). *Performance of three conditional DIF statistics in detecting differential item functioning on simulated tests* (Research Report No. 89-7). ACT, Inc.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unifying strategy. *Journal of Applied Psychology*, *91*, 1291-1306.

Steinmetz, H. (2010). Estimation and comparison of latent means across cultures. In P. Schmidt, J. Billiet, & E. Davidov (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 85-116). Routledge.

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4), 313-350. https://doi.org/10.1207/s15324818ame1804_1

Suh, Y., & Cho, S.-J. (2014). Chi-square difference tests for detecting functioning in a multidimensional IRT model: A Monte Carlo study. *Applied Psychological Measurement*, *38*(5), 359-375. https://doi.org/10.1177/0146621614523116

Swaminathan, H., & Rogers, H.J. (1990). Detecting item bias using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tabachnick, B.G., & Fidell, L.S. (2018). *Using multivariate statistics* (6th ed.). Allyn & Bacon.

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

Uebelacker, L.A., Strong, D., Weinstock, L.M., & Miller, I.W. (2009). Use of item response theory to understand differential item functioning of DSM-IV major depression symptoms by race, ethnicity, and gender. *Psychological Medicine*, *39*(4), 591-601. https://doi.org/10.1017/S0033291708003875

Ullman, J.B. (2018). Structural equation modeling. In B.G. Tabachnick & L.S. Fidell (Eds.), *Using multivariate statistics* (6th ed., pp. 731-836). Allyn & Bacon.

Verma, J.P. (2016). *Repeated measures design for empirical researchers*. John Wiley & Sons.

Vo, H.T., & James, L.M. (2010). Power. In N.J. Salkind (Ed.), *Encyclopedia of research design*, (pp. 1065-1067). Sage Publications.

Walker, C.M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376. https://doi.org/10.1177/0734282911406666

Walker, C.M., Beretvas, S.N., & Ackerman, T. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. *Applied Measurement in Education*, *14*(1), 3-16. https://doi.org/10.1207/S15324818AME1401_02

Walker, C.M., & Sahin, S.G. (2017). Using a multidimensional IRT framework to better understand differential item functioning (DIF): A tale of three DIF detection procedures.

Educational and Psychological Measurement, *77*(6), 945-970. https://doi.org/10.1177/0013164416657137

Walker, J.P., & Rocconi, L.M. (2021). Experiential learning student surveys: Indirect measures of student growth. *Research and Practice in Assessment*, *16*(1), 21-35.

Wang, C., & Nydick, S.W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*(2), 119-134. https://doi.org/10.1177/0146621614545983

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons.

Wang, W.-C., & Shih, C.-S. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*. *34*(3), 166-180.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498.

Woods, C.M. (2009a). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1-27.

Woods, C.M. (2009b). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42-57. https://doi.org/0.1177/0146621607314044

Woods, C.M., & Grimm, K.J. (2011). Testing for non-uniform differential item functioning with multiple indicator multiple cause models. *Psychological Measurement*. *35*(5), 339-361.

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, 320. https://doi.org/10.1007/s10862-008-9118-9

Wu, M., Tam, H.P., & Jen, T.-S. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 337-347). Lawrence Erlbaum Associates.

Appendices

Appendix A

Supplemental Tables

Table A.1

Rejection and type I error rates for multidimensional item response theory likelihood ratio test (MIRT-LR)

			Uniform DIF				Nonuniform DIF			
ρ	Mean Diff.	N	0%	5%	10%	20%	0%	5%	10%	20%
0	Balanced	R1000/F1000	.06	1.00	1.00	1.00	.08	.07	.09	.06
		R1000/F500	.07	.91	1.00	1.00	.09	.08	.07	.08
		R500/F500	.01	.79	.98	1.00	.06	.05	.06	.09
		R500/F250	.06	.60	.81	.99	.13	.05	.07	.09
	Unbalanced	R1000/F1000	.53	1.00	1.00	1.00	.07	.05	.05	.04
		R1000/F500	.37	1.00	1.00	1.00	.07	.14	.16	.19
		R500/F500	.24	1.00	1.00	1.00	.05	.04	.05	.08
		R500/F250	.21	.99	1.00	1.00	.10	.08	.07	.10
0.3	Balanced	R1000/F1000	.02	.98	1.00	1.00	.06	.05	.05	.12
		R1000/F500	.08	.91	.98	.99	.01	.04	.05	.19
		R500/F500	.07	.66	.91	.99	.06	.06	.07	.11
		R500/F250	.03	.48	.77	.95	.07	.04	.07	.09
	Unbalanced	R1000/F1000	.58	1.00	1.00	1.00	.09	.08	.10	.21
		R1000/F500	.34	1.00	1.00	1.00	.08	.06	.10	.23
		R500/F500	.28	1.00	1.00	1.00	.07	.06	.11	.20
		R500/F250	.14	1.00	1.00	1.00	.07	.05	.08	.15
0.6	Balanced	R1000/F1000	.08	.91	.99	1.00	.05	.08	.09	.74
		R1000/F500	.06	.82	.95	1.00	.05	.13	.16	.47
		R500/F500	.02	.63	.93	1.00	.06	.05	.12	.40
		R500/F250	.03	.44	.72	.93	.06	.10	.10	.45
	Unbalanced	R1000/F1000	.43	1.00	1.00	1.00	.07	.12	.13	.73
		R1000/F500	.31	1.00	1.00	1.00	.04	.07	.10	.58
		R500/F500	.24	1.00	1.00	1.00	.04	.08	.07	.43
		R500/F250	.10	1.00	1.00	1.00	.09	.06	.07	.38

Table A.2

Rejection and type I error rates for multiple indicator multiple causes interaction structural equation model (MIMIC)

			Uniform DIF				Nonuniform DIF			
ρ	Mean Diff.	N	0%	5%	10%	20%	0%	5%	10%	20%
0	Balanced	R1000/F1000	.04	1.00	1.00	1.00	.07	.96	.97	.98
		R1000/F500	.04	.98	.99	.99	.07	.95	.98	.96
		R500/F500	.07	.96	1.00	1.00	.03	.97	.97	1.00
		R500/F250	.03	.74	.97	.97	.12	.95	.96	.99
	Unbalanced	R1000/F1000	.03	1.00	1.00	1.00	.09	.98	.99	.97
		R1000/F500	.03	.96	1.00	1.00	.09	.96	.95	.97
		R500/F500	.05	.94	1.00	1.00	.09	.95	.96	.97
		R500/F250	.04	.69	.91	.93	.09	.92	.98	.98
0.3	Balanced	R1000/F1000	.04	1.00	1.00	.85	.05	.95	.98	.94
		R1000/F500	.05	.99	1.00	.75	.12	.97	.98	.99
		R500/F500	.08	.89	.99	.64	.08	.94	.98	.99
		R500/F250	.07	.79	.96	.64	.10	.98	.96	.96
	Unbalanced	R1000/F1000	.05	1.00	1.00	.85	.06	.99	.97	.98
		R1000/F500	.03	.96	.99	.62	.08	.98	.97	.97
		R500/F500	.04	.83	.95	.63	.07	.94	.98	.98
		R500/F250	.06	.74	.89	.61	.06	.96	.95	.96
0.6	Balanced	R1000/F1000	.06	.99	1.00	.90	.07	.99	.98	.82
		R1000/F500	.03	.94	1.00	.75	.07	.97	.95	.91
		R500/F500	.01	.89	.97	.58	.07	.94	.95	.91
		R500/F250	.02	.70	.91	.47	.05	.94	.96	.94
	Unbalanced	R1000/F1000	.06	1.00	1.00	.84	.07	.96	.97	.81
		R1000/F500	.07	.97	1.00	.68	.06	.97	.95	.88
		R500/F500	.04	.79	.97	.56	.09	.96	.97	.91
		R500/F250	.07	.66	.90	.47	.16	.96	.96	.93

Table A.3

Rejection and type I error rates for multidimensional extension of the logistic discriminant function analysis (MLDFA)

			Uniform DIF				Nonuniform DIF			
ρ	Mean Diff.	N	0%	5%	10%	20%	0%	5%	10%	20%
0	Balanced	R1000/F1000	.11	1.00	1.00	1.00	.10	.98	1.00	1.00
		R1000/F500	.09	.97	1.00	1.00	.14	.81	.98	1.00
		R500/F500	.08	.95	1.00	1.00	.12	.72	.93	1.00
		R500/F250	.09	.78	.99	1.00	.13	.61	.88	1.00
	Unbalanced	R1000/F1000	.11	1.00	1.00	1.00	.08	1.00	1.00	1.00
		R1000/F500	.12	1.00	1.00	1.00	.08	1.00	1.00	1.00
		R500/F500	.13	1.00	1.00	1.00	.06	1.00	1.00	1.00
		R500/F250	.09	1.00	1.00	1.00	.10	1.00	1.00	1.00
0.3	Balanced	R1000/F1000	.08	.99	1.00	1.00	.09	.93	1.00	1.00
		R1000/F500	.13	.96	1.00	1.00	.11	.73	.95	1.00
		R500/F500	.12	.87	.97	1.00	.08	.69	.91	1.00
		R500/F250	.10	.72	.92	1.00	.14	.46	.83	1.00
	Unbalanced	R1000/F1000	.11	1.00	1.00	1.00	.09	1.00	1.00	1.00
		R1000/F500	.11	1.00	1.00	1.00	.09	1.00	1.00	1.00
		R500/F500	.15	1.00	1.00	1.00	.13	1.00	1.00	1.00
		R500/F250	.10	1.00	1.00	1.00	.10	1.00	1.00	1.00
0.6	Balanced	R1000/F1000	.06	.98	1.00	1.00	.08	.88	.99	1.00
		R1000/F500	.07	.96	.99	1.00	.12	.75	.92	1.00
		R500/F500	.11	.86	.99	1.00	.08	.70	.94	1.00
		R500/F250	.09	.69	.93	.99	.14	.52	.81	.99
	Unbalanced	R1000/F1000	.11	1.00	1.00	1.00	.13	1.00	1.00	1.00
		R1000/F500	.10	1.00	1.00	1.00	.10	1.00	1.00	1.00
		R500/F500	.06	1.00	1.00	1.00	.14	1.00	1.00	1.00
		R500/F250	.10	1.00	1.00	1.00	.11	1.00	1.00	1.00

Table A.4

Item discrimination and threshold parameters for empirical data set

	a_1	a_2	a_3	a_4	d_1	d_2	d_3	d_4	d_5	d_6
Item 1	2.25				8.51	6.42	5.81	4.67	2.90	-.42
Item 2	1.83				7.53	6.80	6.50	5.29	3.66	.84
Item 3	4.56				14.14	12.19	11.27	9.10	5.85	.49
Item 4	3.43			.47	11.06	9.64	8.66	7.07	4.46	.28
Item 5		3.31			10.58	9.44	7.76	6.21	3.63	-.45
Item 6		3.56			11.62	8.58	7.78	5.88	3.30	-.30
Item 7		3.86			13.43	10.07	8.58	6.81	4.04	-.46
Item 8		2.49			8.16	6.81	5.82	4.63	2.61	-.41
Item 9		2.90			10.01	8.35	6.63	5.45	3.19	-.07
Item 10		5.03			14.12	12.00	10.79	8.96	5.27	.06
Item 11			5.36		13.35	12.32	9.16	6.90	1.25	
Item 12			4.03		9.99	8.88	8.47	6.19	4.02	.37
Item 13			2.69		8.65	7.88	7.44	4.68	3.55	1.16
Item 14				6.29	11.70	10.22	8.33	6.77	3.73	-.65
Item 15				7.56	12.07	10.34	8.65	6.48	3.52	-.84
Item 16	.53			4.63	9.69	7.80	6.42	5.09	2.81	-.47

Note. Blank discrimination parameters are equal to zero. No respondent selected a response of 1 (the lowest response value) on Item 11; thus, d_6 is left blank.

Table A.5
Item purification test for anchor and studied item selection

Item	χ^2	*df*	χ^2/df
Item 1	999.41	8	124.93
Item 3†	902.36	8	112.80
Item 4	929.58	8	116.20
Item 6	981.05	8	122.63
Item 7	951.46	8	118.93
Item 8	996.27	8	124.53
Item 10†	919.39	8	114.92
Item 11	989.24	7	141.32
Item 13†	868.81	8	108.60
Item 14	960.35	8	120.04
Item 15	1003.70	8	125.46
Item 16	968.86	8	121.11

Note. The analysis excludes items 2, 5, 9, and 12 because of convergence issues.
†Anchor item

Table A.6
MGRM discrimination parameters for empirical data by gender

	Female				Male			
	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
Item 1	1.23				1.20			
Item 2	1.13				1.24			
Item 3†	1.24				1.24			
Item 4	1.22			.63	1.19			.61
Item 5		1.27				1.29		
Item 6		1.31				1.27		
Item 7		1.38				1.22		
Item 8		1.20				1.17		
Item 9		1.27				1.15		
Item 10†		1.33				1.33		
Item 11			1.31				1.29	
Item 12			1.28				1.26	
Item 13†			1.19				1.19	
Item 14				1.44				1.41
Item 15				1.47				1.38
Item 16	.74			1.20	.63			1.10

Note. Blank discrimination values are equal to zero.
†Anchor item

Table A.7
MGRM threshold parameters for empirical data by gender

	Female						Male					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Item 1	5.26	3.97	3.45	2.87	1.88	-.38	5.18	3.81	3.65	2.90	1.70	-.60
Item 2	5.16	4.79	4.53	3.68	2.76	.59	5.24	4.65	4.50	3.77	2.50	.28
Item 3†	5.08	4.23	3.94	3.27	2.13	-.14	5.08	4.23	3.94	3.27	2.13	-.14
Item 4	5.83	4.92	4.50	3.64	2.35	-.04	5.72	5.05	4.47	3.62	2.14	-.16
Item 5	5.29	4.72	4.04	3.42	2.09	-.47	5.11	4.75	4.10	3.37	2.05	-.57
Item 6	5.53	4.27	3.87	3.03	1.71	-.44	5.42	4.18	3.92	3.06	1.79	-.47
Item 7	5.93	4.68	4.02	3.31	2.06	-.62	5.86	4.58	4.05	3.32	2.07	-.42
Item 8	5.09	4.28	3.71	3.10	1.82	-.51	4.94	4.30	3.79	3.01	1.71	-.51
Item 9	5.53	4.76	3.80	3.24	2.00	-.39	5.40	4.70	3.98	3.28	1.94	-.19
Item 10†	5.07	4.50	4.18	3.69	2.36	-.37	5.07	4.50	4.18	3.69	2.36	-.37
Item 11	5.87	5.50	4.06	3.09	.67		5.89	5.43	4.05	3.09	.16	
Item 12	5.42	4.96	4.57	3.36	2.26	.02	5.49	4.69	4.69	3.41	2.13	.13
Item 13†	5.89	5.46	5.23	3.47	2.66	.80	5.89	5.46	5.23	3.47	2.66	.80
Item 14	5.50	4.70	3.81	3.13	1.85	-.41	5.42	4.75	3.81	3.13	1.66	-.50
Item 15	4.75	4.07	3.37	2.53	1.44	-.47	4.73	3.89	3.36	2.61	1.31	-.53
Item 16	6.15	4.97	4.04	3.24	1.83	-.47	6.09	4.86	4.13	3.32	1.83	-.42

Note. No respondent selected a response of 1 (the lowest response value) on Item 11; thus, d_6 is left blank.
†Anchor item

Appendix B

Supplemental Figures

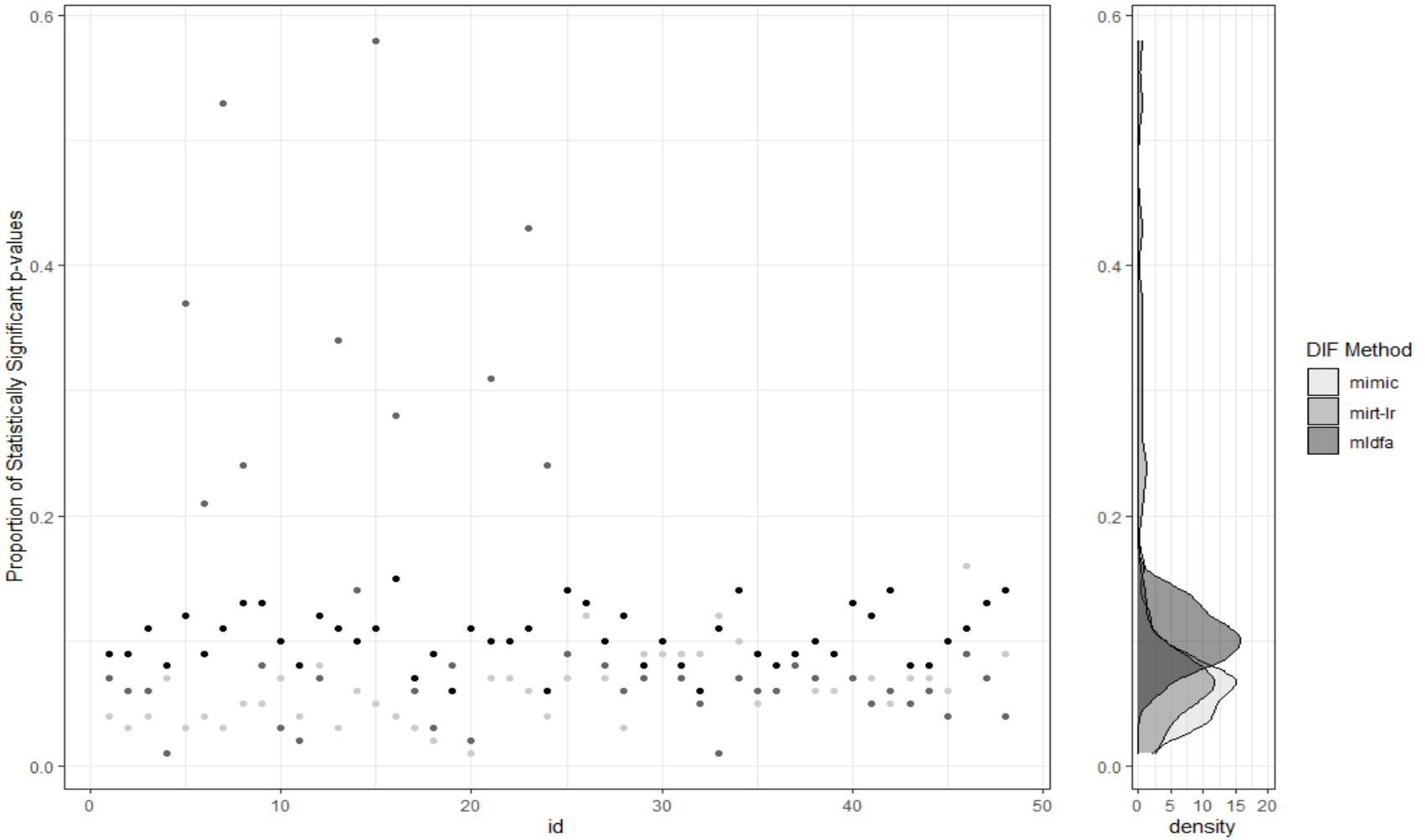

Figure B.1. *Kernel density distributions for type I error rates by DIF detection method*

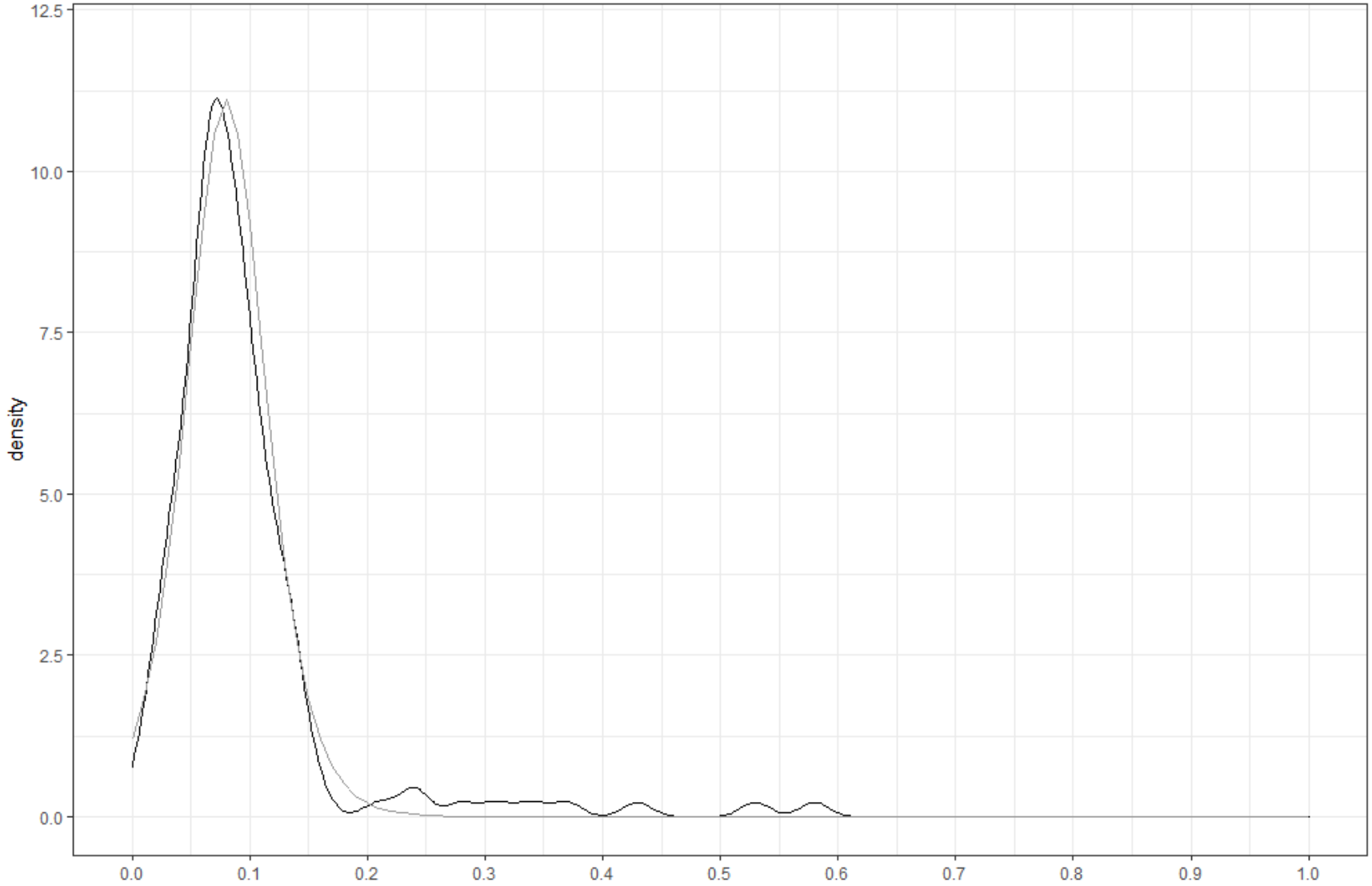

Figure B.2. *Kernel density estimate for type I error rates and logistic probability density function with scaling and mean adjustments*

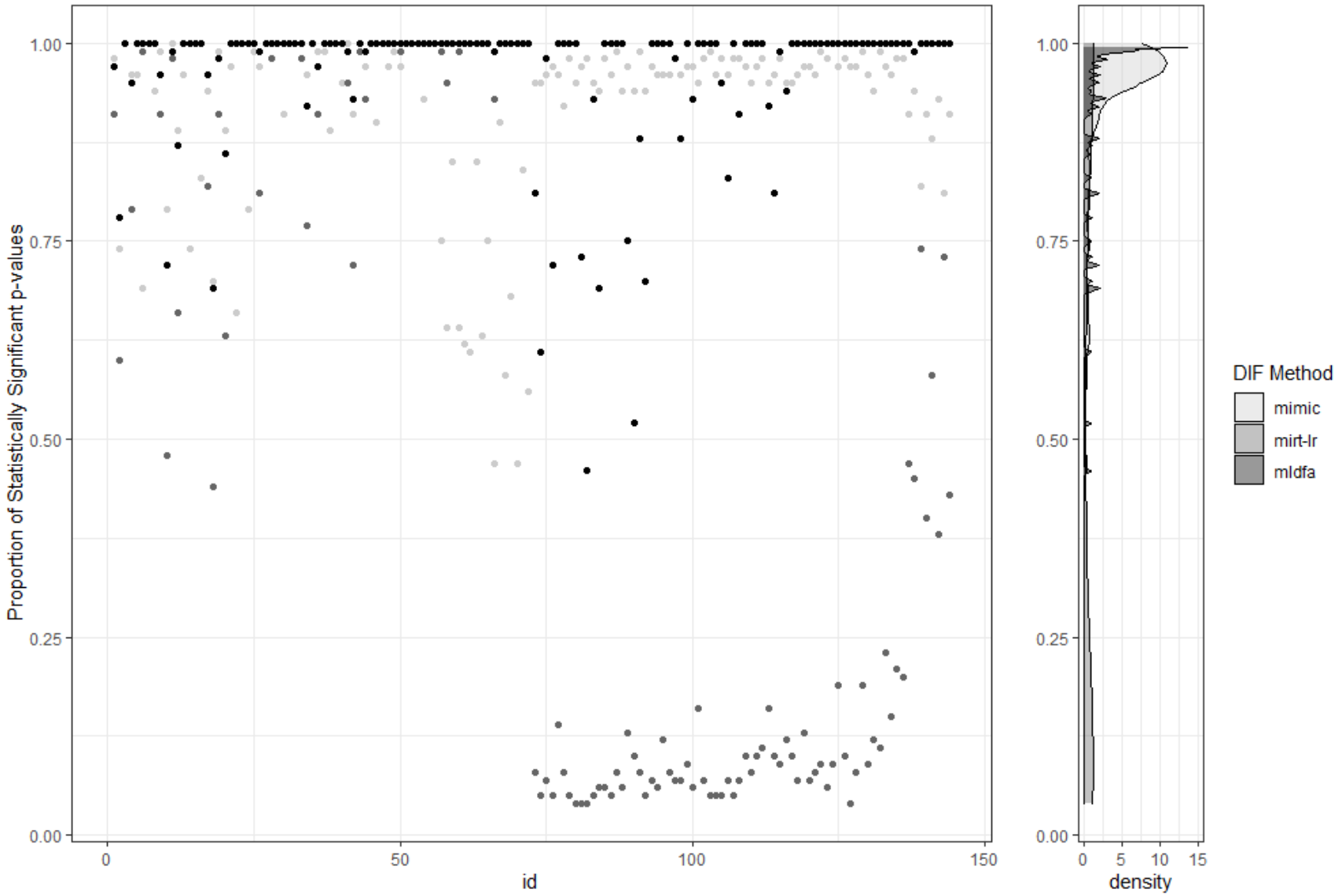

Figure B.3. *Kernel density distributions for rejection rates by DIF detection method*

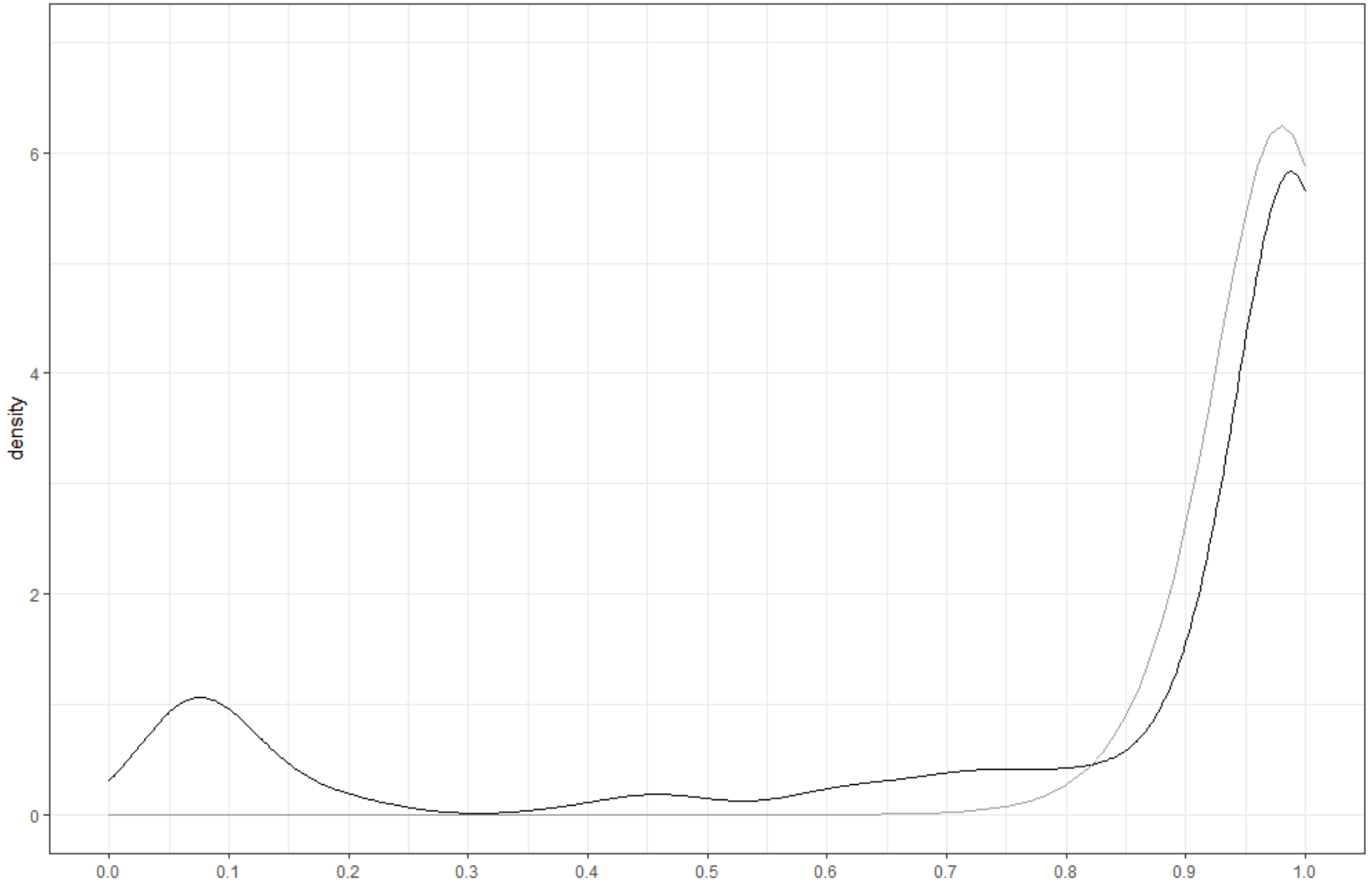

Figure B.4. *Kernel density estimate for rejection rates and logistic probability density function with scaling and mean adjustments*

Vita

John Walker was born in Lompoc, California and raised in Farmington Hills, Michigan. He earned his Master's in Policy Studies from the University of Sydney in Sydney, Australia and a BS in mathematics from Grand Valley State University. He currently serves as the Research and Assessment Coordinator in Teaching and Learning Innovation at the University of Tennessee, Knoxville. Prior to his arrival in Knoxville, John served as the Assessment Specialist in the Office of Operations, Curriculum, and Assessment at Schoolcraft College in Livonia, Michigan. He also taught English as a Foreign Language in the public school system in Incheon, South Korea and completed an internship as a survey researcher with Fyusion Asia-Pacific in Sydney while pursuing his Master's degree. During this PhD years, he authored and co-authored articles in *Research and Practice in Assessment*, *Journal for Research and Practice in College Teaching*, *International Journal of Teaching and Learning in Higher Education*, and *Nursing Education Perspectives* and presented at numerous conferences across the country. In July 2021, John accepted a Psychometrician position with Fastbridge Assessment, a division of Illuminate Education. John's research interests are focused on psychometrics, measurement and assessment, quantitative methods, and measurement bias.