for factor invariance Mplus syntax, see p. 194

RELAXED METHODS FOR EVALUATING MEASUREMENT INVARIANCE WITHIN A MULTIPLE-GROUP CONFIRMATORY FACTOR ANALYTIC FRAMEWORK

by

JORDAN CAMPBELL BRACE

B.Sc., Memorial University of Newfoundland, 2013

M.A., The University of British Columbia, 2015

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Relaxed methods for evaluating measurement invariance within a multiple-group confirmatory factor analytic framework

submitted by Jordan Campbell Brace in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Psychology

Examining Committee:

Victoria Savalei, Psychology, UBC
Supervisor

Jeremy Biesanz, Psychology, UBC
Supervisory Committee Member

Rachel Fouladi, Psychology, Simon Fraser University
Supervisory Committee Member

Amery Wu, ECPS, UBC
University Examiner

Sterett Mercer, ECPS, UBC
University Examiner

Abstract

Measurement Invariance (MI) refers to the equivalent functioning of psychometric instruments when applied across different groups. Violations of MI can lead to spurious between-group differences, or obscure true differences, on observed scores, means, and covariances. Chapter 1 introduces the multiple-group confirmatory factor analysis (MGCFA) approach to evaluating MI. The present research seeks to identify overly restrictive assumptions of the MGCFA approach, and to provide alternative recommendations. Chapter 2 notes that typical MGCFA MI models assume equivalent functioning of each item, while in practice, applied researchers are often primarily interested in equivalent functioning of composite scores. Chapter 2 introduces an approach to assessing MI of composite scores that does not assume MI of all items, by placing between-group equality constraints on measurement parameter totals. Invariance of parameter totals is referred to as "scale-level MI", while the invariance of parameters is referred to as "item-level MI". Analyses of tests of scale-level and item-level MI illustrate that, despite item-level MI models being nested within scale-level MI models, tests of scale-level MI are often more sensitive to violations of MI that affect the between-group comparability of composite scores. Chapter 3 introduces an approach to quantifying between-group differences in classification accuracy when critical composite scores are used for selection and a minimum of partial scalar MI – MI of some, but not all, loadings and intercepts – is retained. Chapter 3 illustrates that different patterns of violations of MI differentially affect classification accuracy ratios for different measures of classification accuracy. Between-group differences on multiple sets of measurement parameters can have compensatory or additive effects on classification accuracy ratios. Finite sample variability of classification accuracy ratios is discussed, and a Bollen-Stine bootstrapping approach for estimating confidence intervals around classification accuracy ratios is recommended. Chapter 4 addresses limitations of popular methods of assessing fit of nested MI models. Chapter 4 introduces a modified RMSEA, RMSEA_{D}, for comparing fit of nested MI models, which avoids the sensitivity to minor misspecifications of chi-square tests, as well as the differential interpretation of Δ*GFIs* depending on model degrees of freedom. Recommendations, limitations, and future research are discussed in Chapter 5.

Lay Summary

The present research is concerned with evaluating whether measures of psychological constructs, e.g., intelligence or personality tests, are unbiased across different groups. When questionnaires are biased, responses for participants from different groups with equal standing on the psychological construct of interest are not expected to be the same, making it difficult to compare scores across groups, as given scores reflect different unobservable ‘ each group. The present research proposes new methods of evaluating whether measures show bias of this nature for specified groups. Chapter 2 introduces a method of testing whether total scores are unbiased across groups. Chapter 3 introduces a method of quantifying how accurately measures classify participants above or below a given threshold for selection when measures are biased. Chapter 4 introduces a method for quantifying the degree of incorrectness of the assumption of unbiased measures across specified groups.

Preface

In consultation with my advisor, Dr. Victoria Savalei, I was responsible for the identification and design of the research program, as well as the evaluation of all proposed methods. With respect to Chapter 2, I was responsible for the design, analysis, and presentation of this research, with Dr. Savalei providing guidance and feedback throughout its development. An earlier version of the research discussed in Chapter 3 was presented as a poster at the 2016 meeting of the *Society for Multivariate Experimental Psychology*. An abstract of this research was also published as Brace, J. C. (2017). Sensitivity and specificity ratios as indices of measurement invariance. *Multivariate Behavioral Research, 52*(1), 107-108. doi:10.1080/00273171.2016.1263792. I was responsible for the design, analysis, and presentation of this research. An earlier version of the research discussed in Chapter 4 was presented as a poster at the 2018 *Modern Modelling Methods* conference. I was responsible for the design, analysis, and presentation of this research. Dr. Savalei provided feedback at all stages of the development of the dissertation document, but I was ultimately responsible for all writing herein.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
F	Focal group
GFI	Goodness-of-Fit Index
MGCFA	Multiple-Group Confirmatory Factor Analysis
MI	Measurement Invariance
NPV	Negative Predictive Value
PPV	Positive Predictive Value
R	Reference group
RMSEA	Root-Mean-Square Error of Approximation
SEM	Structural Equation Modelling
Sens	Sensitivity
Spec	Specificity
SRMR	Standardized Root-Mean-Square Residual

List of Symbols

G	# of groups
g	The index of a group, such that g Î $\{1,...,G\}$
P	# of modeled manifest variables
Q	# of modeled latent variables
S	Sample covariance matrix
$S(q)$	Model-implied covariance matrix
S	Population covariance matrix
Y	$P\ {}^{3}Q$ matrix of observed scores
y_{comp}	Observed composite score
x	Factor score
DGFI	Change on goodness-of-fit index
F	$Q\ {}^{3}Q$ matrix of latent variances
k	$Q\ {}^{3}1$ vector of latent means
L	$P\ {}^{3}Q$ matrix of factor loadings
l^{*}	Weighted sum of factor loadings
t	$P\ {}^{3}1$ vector of indicator intercepts
t^{*}	Weighted sum of indicator intercepts
Y	$P\ {}^{3}P$ diagonal matrix of residual variances
y^{*}	Weighted sum of residual variances
n	Classification accuracy
ʻ	Classification accuracy ratio such that $i= {}_{R}n/ \ {}_{F}$

Acknowledgements

I would like to extend a sincere thank you to everyone who has helped me throughout this process. First and foremost, I would like to thank my advisor, Dr. Victoria Savalei, for all of her education, feedback, advice, recommendations, encouragement, support, sympathy, and patience over the past seven years. I doubt that I would have gotten through this without her. I also owe a debt of gratitude to my supervisory committee, Dr. Jeremy Biesanz and Dr. Rachel Fouladi, for their expertise and advice which helped shape this document into its current form. I would also like to thank the Social Sciences and Humanities Research Council of Canada (SSHRC) for funding this research.

I would like to thank my friends and colleagues who have offered advice, support, and feedback in the development of this document: Cathy Xijuan Zhang, Bill Chen, Dr. Ellen Stephenson, Dr. Jennifer Lay, and Victoria Michaelowski. I would also like to thank friends who contributed indirectly to the production of this document with their love and support: Dr. Jennifer Campbell, Brianne Glazier, Siba Ghrear, Spencer Murch, Madison Elliott, Ryan Downey, and Kirsten Morry. I would like to thank my parents for their continued love and support from 7,000 kilometers away. Finally, I would like to thank my partner, Eri
offering more love and support than anyone has ever deserved.

To Erin

Chapter 1: Introduction

1.1. What is Measurement Invariance?

Measurement invariance (MI) refers to the equivalent functioning of psychometric instruments when applied across different groups (Millsap, 2011). Measurement invariance is necessary to justify comparing observed scores across different groups, either in terms of means and structural relationships (Steenkamp & Baumgartner, 1998), or individual scores (Millsap & Kwok, 2004). When MI cannot be secured for a measure, observed between-group differences may be spurious, while true differences may be obscured, depending on the nature of violations of MI (Steinmetz, 2013).

Measurement invariance can be defined in terms of the following conditional independence (Mellenbergh, 1989; Meredith & Millsap, 1992):

$$P(Y \mid \chi, g) = P(Y \mid \chi), \tag{1.1}$$

where *Y* is a vector of observed measures of some unobservable latent construct of interest, χ is a vector of unobservable true scores on that latent construct, and *g* is group membership. Thus, the sampling distribution of *Y* is dependent only on the true state on the construct of interest, χ, and is independent of group membership, *g*. The definition of MI given in Equation (1.1) is also referred to as *strict invariance* (Meredith, 1993).

A less restrictive definition of MI says that the predictive relationship between *Y* and χ is not moderated by *g* (Mellenbergh, 1989):

$$E(Y \mid \chi, g) = E(Y \mid \chi). \tag{1.2}$$

The expected value of *Y* is dependent only on the true state on the construct of interest, ξ, and is independent of group membership, *g*. The definition of MI given in Equation (1.2) is also referred to as *strong invariance*. Note that Equation (1.2) necessarily holds whenever Equation (1.1) does.

Assuming a linear common factor model (Spearman, 1904; Thurstone, 1947), the relationship between ξ and *Y* is given by:

$$Y_{ig} = \Lambda_g \xi_{ig} + \tau_g + \varepsilon_{ig}, \tag{1.3}$$

where Y_{ig} is a $P \times 1$ vector of observed scores on *P* items for individual *i* in group *g*, τ_g is a $P \times 1$ vector of item intercepts in group *g*, ξ_{ig} is a $Q \times 1$ vector of latent variable scores for individual *i* in group *g*, Λ_g is a $P \times Q$ matrix of factor loadings, and e_{ig} is a $P \times 1$ vector of residual scores. An implication of Equation (1.2) for the linear common factor model is that $\tau_g =$ and $\Lambda_g =$ for all *g*: the predicted set of observed scores, $\hat{Y}_{ig}$, for an individual, *i*, in group *g*, with true scores ξ_{ig} is independent of *g*. Conversely, the predicted set of true scores, $\hat{\xi}_{ig}$, for an individual with observed scores Y_{ig} is also independent of *g*. Under Equation (1.2), two test respondents with identical true scores have identical expected observed scores, and two test respondents with identical observed scores are expected to have identical true scores, independent of group membership. Thus, when Equation (1.2) is violated, comparing scores across groups is difficult, as any given value of Y_{ig} reflects different values of ξ_{ig} across groups, and any given value of ξ_{ig} reflects different values of Y_{ig} across groups (Millsap & Kwok, 2004; Steinmetz, 2013). For the purposes of many applied behavioral researchers, satisfaction of strong invariance, as defined in Equation (1.2), is often sufficient. Consequently, strict invariance is often ignored in practice (Vandenberg & Lance, 2000).

Golembiewski, Billingsley, and Yeagar (1975) describe three types of change that can occur in behavioral research: Beta change, gamma change, and alpha change. *Beta change* refers to a difference in subjective calibration of a measurement scale, rather than true differences on the construct being measured (Millsap & Hartog, 1988). Differences in subjective understanding of what it means to be a 3 or a 4 on a Likert scale, resulting in differences in responding that are independent of actual differences on what is being measured, is an example of beta change. *Gamma change* refers to subjective redefinition of the construct being measured (Millsap & Hartog, 1988). A group of employees who receive workplace diversity training may have higher scores on a measure of attitudes towards diversity in the workplace than employees who do not receive diversity training. While this may suggest that diversity training was effective, it is also plausible that diversity training has clarified the meaning of “workplace that divers received it. Thus, employees in the two groups may respond to the measure based on different definitions of workplace diversity, and may not actually differ in their attitudes towards either conceptualization of diversity (Meade, Lautenschlager & Hecht, 2005). This is an example of gamma change. *Alpha change*, thus, refers to the sort of differences behavioral researchers are usually interested in. A between-group difference in depression scores for participant groups receiving different treatments being motivated by a true difference in levels of depression is an example of alpha change. The pursuit of MI can be thought of as the verification that groups being compared are not subject to beta change or gamma change, thus allowing score differences to justifiably be interpreted as true differences on the measured construct.

To illustrate with an example, cultural differences in individualism and collectivism are believed to play a role in cultural differentiation within a number of psychological domains (Hui & Triandis, 1986; Markus & Kitayama, 1991). However, it has also been documented that cross-cultural differences in response-style to psychometric instruments exist between members of individualist and collectivist cultures, such that members of individualist cultures are more likely to endorse extreme responses to Likert items, while individuals from collectivist cultures are more inclined towards the midpoint (Chen, Lee, & Stevenson, 1995). Thus, when evaluating measured differences between individualist and collectivist cultures, one cannot be certain if observed differences are due to differences on the construct of interest (i.e., alpha change) or a fundamental difference in how respondents interact with the instrument across groups (i.e., beta or gamma change), unless one evaluates MI (Chen & West, 2008; Heine, Lehman, Peng, & Greenholtz, 2002).

In the behavioral sciences, the presence of MI is most frequently evaluated using linear *Confirmatory Factor Analysis* (CFA) ‒a formal specification of the linear common factor model ‒which is introduced in section 1.2. I begin by introducing the general CFA framework, before transitioning to a discussion of multiple-group CFA (MGCFA), the specific factor analytic model used in the analysis of MI.

1.2. Confirmatory Factor Analysis

Confirmatory factor analysis (CFA; Joreskog, 1971) is a statistical technique which allows researchers to define unobservable latent variables in terms of observable manifest variables hypothesized to correlate with unobservable psychological constructs of interest. Confirmatory factor analyses consist of a measurement model which defines latent variables in terms of observed indicator variables, and a structural model which identifies hypothesized correlations (or lack thereof) between these latent variables. Confirmatory factor analysis assumes that error terms are uncorrelated with latent variables and each other, and that error terms and have a mean of 0 (Kline, 2016).

In a typical CFA of P indicators on Q latent variables, the following general covariance structure is fit to data (Brown, 2006):

$$\Sigma(\theta) = \Lambda \Phi \Lambda \ , \tag{1.4}$$

where Φ is a $Q \times Q$ covariance matrix among latent variables, Ψ is a $P \times P$ diagonal matrix of residual variances, and $\Sigma(\theta)$ is the $P \times P$ model-implied covariance matrix, an approximation to the population covariance matrix, Σ, given the proposed model, and the sample covariance matrix, S.

Additionally, the following general mean structure is fit to data (Brown, 2006):

$$\mu(\theta) = \Lambda \kappa +, \tag{1.5}$$

where κ is a $Q \times 1$ vector of latent means, and $\mu(\theta)$ is a $P \times 1$ model-implied mean vector, an approximation to the population mean vector, μ, given the proposed model, and sample mean vector, $\bar{Y}$.

CFA seeks to find a solution for θ, a vector containing all estimated model parameters used to define each element of $\Sigma(\theta)$ and $\mu(\theta)$. The solution to θ is the set of estimates which produces the $\Sigma(\theta)$ and $\mu(\theta)$ for which observed data are most likely given the constraints on covariance and mean estimates imposed by Equation (1.4) and Equation (1.5). This is done by iteratively solving the systems of linear equations $\Sigma(\theta) = S$ and $\mu(\theta) = \bar{Y}$ to minimize deviations between observed and estimated covariances and means. The output of a CFA includes parameter estimates, standard errors for those estimates, and an overall chi-square test of model fit, *T*. Assuming normal distribution of data and a correctly specified model, *T* follows a central chi-square distribution (Steiger, Shapiro, & Browne, 1985) with degrees of freedom $df = \frac{P(P+1)}{2} + P - m$, where *m* is the length of θ. A non-significant chi-square test of fit allows researchers to retain the null hypothesis, $H_0: \Sigma = \Sigma(\theta), \ \mu = \mu(\theta)$, and thus, infer that their data are not unlikely given the proposed model.

1.2.1. Model Identification in CFA

Prior to estimating model parameters, CFA models must be identified. CFA models are identified when they meet two criteria: all latent variables are given a scale, and model degrees of freedom are greater than 0 (Kline, 2016). Because latent variables are not actually observable, their variances are essentially arbitrary, and must be manually specified. The most common methods of identifying latent variables is by either fixing latent variances to 1, or by fixing the loading associated with an indicator of that latent variable to 1. Generally, which method of identification is used has no bearing on model fit.

When model degrees of freedom are less than 0, the model is said to be *under-identified.* When models are under-identified, the number of estimated model parameters is greater than the number of unique elements of *S* and $\bar{Y}$, resulting in infinitely many possible solutions to θ which allow $\Sigma(\theta)$ and $\mu(\theta)$ to perfectly reproduce *S* and $\bar{Y}$. When model degrees of freedom are exactly 0, the number of estimated model parameters is equal to the number of unique variances, covariances, and means in *S* and $\bar{Y}$. Such a model is said to be *just-identified* or *saturated*. While saturated models produce valid parameter estimates, they also necessarily perfectly reproduce *S* and $\bar{Y}$. Thus, no hypotheses about the fit of covariance and mean structures are evaluated when saturated models are fit to data. When model degrees of freedom are greater than 0, the model is said to be *over-identified*. Over-identified models have fewer estimated model parameters than the number of unique elements of *S* and $\bar{Y}$. When models are over-identified, it becomes theoretically possible for $\Sigma(\theta)$ and $\mu(\theta)$ to not perfectly reproduce *S* and $\bar{Y}$, and thus, the null hypothesis of model fit, $H_0 : \Sigma = \Sigma(\theta),\ \mu = \mu(\theta)$ becomes testable. When models are saturated or under-identified, it is impossible to obtain evidence that leads to rejection of a null hypothesis of perfect model fit.

In practice, it is often the case that the covariance structure is over-identified, while the mean structure is saturated. This is because the general mean structure given in Equation (1.5) introduces *P* observed means, and $P + Q$ estimated parameters: *P* estimated intercepts, and *Q* estimated latent means, producing an under-identified mean structure. By convention, latent means are generally fixed to zero when the mean structure would otherwise be under-identified, resulting in a just-identified mean structure of the form $\mu(\theta) =$ with *P* estimated intercepts. Thus, this general mean structure is able to perfectly reproduce observed means by fixing τ to $\bar{Y}$, and by extension, $\mu(\theta) = \bar{Y}$. For researchers who are primarily interested in testing hypotheses about the covariance structure, the general mean structure given in Equation (1.5) has no impact on model fit, model degrees of freedom, or covariance structure parameter estimates, and can safely be ignored.

If *Q* constraints are placed on intercept estimates, the *Q* latent means no longer need to be fixed to 0 for identification, and can be freely estimated, producing an alternate saturated mean structure with freely estimated latent means. If *Q+1* or more constraints are placed on latent means and intercepts, the mean structure becomes over-identified, and contributes to the overall model degrees of freedom and model fit. Therefore, at least *Q+1* constraints need to be placed on mean structure parameters before hypotheses about the mean structure become testable.

1.2.2 Estimating CFA Models

Confirmatory factor analytic models are typically estimated using structural equation modelling (SEM) software such as lavaan (Rosseel, 2012), Mplus (Muthen & Muthen, 1998-2017), or EQS (Bentler, 2006). The most popular estimation method in SEM is normal theory maximum likelihood (ML) (Anderson & Gerbing, 1984; Joreskog & Sorbom, 1996; Steiger, 1990). When structure is only imposed on the covariance matrix, the ML fit function is given by

$$F_{ML}(S,\hat{\Sigma}) = tr\{S\hat{\Sigma}^{-1}\} - \ln|S\hat{\Sigma}^{-1}| - P. \tag{1.6}$$

When the mean structure is over-identified, (1.6) becomes

$$F_{ML}(S,\hat{\Sigma},\bar{Y},\hat{\mu}) = tr\{S\hat{\Sigma}^{-1}\} - \ln|S\hat{\Sigma}^{-1}| + (\bar{Y}-\hat{\mu})'\hat{\Sigma}^{-1}(\bar{Y}-\hat{\mu}) - P. \tag{1.7}$$

ML estimation is used to identify a solution to q by first proposing a set of starting values, computing F_{ML} as defined in Equation (1.6) or Equation (1.7), and subsequently updating candidate values of q to decrease F_{ML}. This process continues iteratively until a specified threshold is reached such that the difference in successive values of F_{ML} is sufficiently small. At this point, the model is said to have converged, and the final value of F_{ML} is denoted $\hat{F}_{ML}$, the minimization of the fit function. Larger values of $\hat{F}_{ML}$ indicate greater discrepancies between sample and model-implied covariances and means (Amemiya & Anderson, 1990, Joreskog & Sorbom, 1996). The magnitude of this discrepancy reflects the likelihood of observing S and $\bar{Y}$ if the null hypothesis of perfect model fit is true.

The overall test of model fit is given by $T = (N - 1)\hat{F}_{ML}$, where N is the total sample size studied. When the null hypothesis is true in the population, and data are normally distributed, T asymptotically follows a central chi-square distribution with $df = \frac{P(P+1)}{2} - m$ degrees of freedom, or $df = \frac{P(P+1)}{2} + P - m$ degrees of freedom when mean structures are identified. If one's hypothesized model is incorrect, and da T asymptotically follows a non-central chi-square distribution with non-centrality parameter λ_{nonc}, which can be estimated as $\hat{\lambda}_{nonc} = T - df$. If one's observed T is found statistically significant when tested against the appropriate reference distributio $\hat{F}_{ML}$ is sufficiently large to permit the inference that S and $\bar{Y}$ is unlikely if $\Sigma = \Sigma(\theta)$ and $\mu = \mu(\theta)$, and thus, that the specified model should be rejected. While Maximum Likelihood estimation assumes multivariate normality of data, there exists a large body of research on assessing fit of SEMs when data are not multivariate normal (e.g., Chou, Bentler, & Satorra, 1991; Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992; Satorra & Bentler, 1988; Satorra & Bentler, 1994; Savalei, 2014).

1.2.2.1 Goodness-of-Fit Indices

Traditional chi-square approaches to SEM fit test the point null hypothesis that the specified model is exactly correct in the population. In practice, however, model complexity and sample size can lead to high rejection rates for models that competently explain means and covariances among a set of observed variables, even when misspecifications are negligibly small (Edwards, 2013). *Goodness-of-fit indices* (GFIs) are alternative measures of model fit which, rather than assessing model fit in terms of the probability of observing one specified model and making a binary decision based on some convention, assess fit on a continuum in terms of degree of departure of $\Sigma(\theta)$ and $\mu(\theta)$ from S and $\bar{Y}$, with adjustments that account for the number of modelled variables, model complexity, and sample size. Hu and Bentler (1999) recommend reporting overall chi-square tests of model fit, and Standardized Root Mean Residuals (SRMR) alongside GFIs, and not reporting GFIs as the only measure of model fit. The most popular GFIs used in the behavioral sciences include the Comparative Fit Index (CFI; Bentler, 1990) and the Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980).

The CFI is estimated as

$$CFI = \frac{(T_{baseline} - df_{baseline}) - (T_{model} - df_{model})}{(T_{baseline} - df_{baseline})} = \frac{\hat{\lambda}_{nonc,baseline} - \hat{\lambda}_{nonc,model}}{\hat{\lambda}_{nonc,baseline}} = \frac{(\hat{F}_{ML,baseline} - \frac{df_{baseline}}{N-1}) - (\hat{F}_{ML,model} - \frac{df_{model}}{N-1})}{(\hat{F}_{ML,baseline} - \frac{df_{baseline}}{N-1})},$$

where terms subscripted "model" are associated with the fitted model of interest, and those subscripted "baseline" are associated with the fitted independence model, a baseline model that constrains all modeled manifest variables to orthogonality (Bentler, 1990). Because SEM typically allows covariances among exogenous variables to exist by default, some argue that variables exogenous to the model of interest should be allowed to covary under the baseline model (Muthen & Muthen, 1998-2017). The CFI can be thought of as the proportion of non-centrality introduced by the worst possible model that can be recovered by the model of interest. An important limitation of the CFI, which is often neglected in practice, is that fit depends on the fit of the baseline model. When baseline model fit is not sufficiently poor, the CFI appears to indicate poor fit for the model of interest. Kenny (2015) recommends against computing the CFI if the RMSEA of the baseline is less than 0.158. Hu and Bentler (1999) recommend a critical CFI of .95 or greater alongside a standardized root mean square residual (SRMR; Bentler, 2006) of .08 or less for retaining model fit.

The RMSEA (Steiger & Lind, 1980) is defined as

$RMSEA = \frac{\sqrt{T_{model} - df_{model}}}{\sqrt{df_{model} * (N-1)}} = \frac{\sqrt{\hat{\lambda}_{nonc,model}}}{\sqrt{df_{model} * (N-1)}} = \sqrt{\frac{\hat{F}_{ML,model}}{df_{model}} \frac{1}{N-1}}$, and is arguably the most popular GFI used in behavioral research. The RMSEA can be thought of as the amount of misfit per model degree of freedom, and controls for sample size while penalizing model complexity (or alternatively, rewarding parsimony). One of the greatest benefits of the RMSEA relative to other goodness-of-fit indices is that it is a transformation of an asymptotically non-central chi-square distribution, allowing for the computation of a confidence interval around its estimate in addition to providing a quantitative index of model fit (Browne & Cudeck, 1993).

MacCallum, Browne, and Sugawara (1996) noted that, due to its transformation of a noncentral chi-square distribution, the RMSEA can be used to construct hypotheses relative to a specified degree of model misspecification. Specifically, one can compute the non-centrality parameter of the sampling distribution associated a specified critical RMSEA value, *c*, as $\lambda_{nonc} = c^2 * df_{model} * (N-1)$. Thus, one can test the directional null hypotheses $H_0: RMSEA < c$, often referred to as a test of close fit, or $H_0: RMSEA > c$, the test of not close fit, by testing one's observed T against the appropriate non-central chi-square distribution $\chi^2_{df}(\lambda_{nonc})$ (Browne & Cudeck, 1993). Rejection of the null of close fit in the upper tail of $\chi^2_{df}(\lambda_{nonc})$ implies that a population RMSEA less than the specified *c* is unlikely, while rejection of the null of not close fit in the lower tail of $\chi^2_{df}(\lambda_{nonc})$ implies that a population RMSEA greater than the specified *c* is unlikely. Kenny, Kaniskan, and McCoach (2015) argued that the RMSEA performs poorly with, and should not be used for, models with small degrees of freedom, where it is even more sensitive than the chi-square test of model fit. Hu and Bentler (1999) recommended an RMSEA of .06 or less alongside a SRMR of .08 or less for retaining model fit.

Use of goodness-of-fit indices relaxes requirements on statistical information while still allowing researchers to make reasonable inferences with respect to their substantive hypotheses, and have, thus, become popular within the SEM framework. Unfortunately, these methods have minimal implications for analysis of MI, as MI is typically evaluated by comparing the fit of a series of nested structural equation models, as opposed to examining overall fit of a single model. Nested model comparison within the SEM framework will now be discussed.

1.2.3 Nested Model Comparison With Structural Equation Modelling

It is often the case that behavioral researchers wish to not simply evaluate a single model, but compare a set of nested models. Sets of nested models are typically obtained by introducing additional constraints on parameter estimates to an existing model, such as by fixing parameters to specific values, or by introducing equality constraints across multiple parameters. Following ML estimation, the difference in fit between nested models can be evaluated using the ML chi-square difference test:

$$D = T_A - T_B \tag{1.8}$$

where T_A is the chi-square test of fit associated with the more restrictive model, M_A, and T_B is the chi-square test associated with less restrictive model, M_B. When data are normally distributed and both models are correct, D asymptotically follows a central chi-square distribution with degrees of freedom $df_D = df_A - df_B$, where df_A is the number of degrees of freedom associated with M_A, and df_B is the number of degrees of freedom associated with M_B. When M_B is correct but M_A is not, D follows a non-central chi-square distribution whose non-centrality parameter can be estimated as $\hat{\lambda}_{nonc} = D - df_D$ (Steiger et al., 1985; Yuan & Bentler, 2004; Yuan & Chan, 2016). If D is found statistically significant when tested against a chi-square distribution with df_D degrees of freedom, researchers reject the null hypothesis that constraints introduced by M_A above and beyond those already present for M_B are true in the population.

1.3 Evaluating Measurement Invariance via Multiple-Group Confirmatory Factor Analysis

Within the factor analytic context, measurement invariance across distinct groups is most frequently evaluated using a model called *Multiple-Group Confirmatory Factor Analysis* (MGCFA). It is worth noting that MGCFA cannot be used to evaluate MI across repeated measures, as assessing MI with respect to time-points involves model constraints that cannot be implement within the MGCFA context (see Widaman, Ferrer & Conger, 2010). As is the case with single-group CFA, MGCFA assumes that error terms are uncorrelated with latent variables and each other, and that error terms and have a mean of 0.

The mean and covariance structure of a 1-factor MGCFA model with P observed variables is given by

$$\Sigma(\theta_g) = \Lambda_g \Phi_g \Lambda_g', \; \mu(\theta_g) = \tau_g + \Lambda_g . \tag{1.9}$$

Measurement invariance is typically tested in a sequential manner wherein between-group equality constraints are progressively added to a baseline model (Byrne, Shavelson, & Muthén, 1989; Millsap & Everson, 1991), and change in fit is assessed via chi-square difference tests (Steiger, Shapiro & Browne, 1985). The baseline model is typically referred to as the *configural* invariance model (Horn, McArdle, & Mason, 1983; Horn & McArdle, 1992), under which parameter matrices have the same dimensions and pattern of fixed and freely estimated parameters in all groups, with no between-group equality constraints on parameter estimates. The configural invariance model is logically identical to fitting the same single-group CFA model separately to each group, with the overall chi-square test statistic and degrees of freedom being equal to the sum of those observed for each single-group CFA run.

As with single-group CFA, latent variances must be identified by either fixing a single loading in each group to 1, or by fixing the latent variances themselves to 1. By fixing a parameter to 1 in multiple groups, one is, in effect, also assuming that that parameter is invariant across groups. It is important to select a parameter that one is confident does not violate MI, as violating this assumption can lead to biased parameter estimates (Cheung & Rensvold, 1999). In practice, researchers typically constrain a loading to 1 for identification when fitting MI models, as constraining latent variances to 1 implicitly assumes invariance of latent variances, which is not necessary for MI to hold (Cheung & Rensvold, 1999; Joreskog & Sorbom, 1996).

If the null hypothesis of configural invariance holds, the test of configural invariance is typically followed by the test of *weak* or *metric* invariance (Meredith, 1993), which introduces between-group equality constraints on factor loadings, such that $\Lambda_g =$ for all *g*. When metric invariance holds, Equation (1.9) becomes $\Sigma(\theta_g) = \Lambda_g \Phi' \Lambda_g + \Psi_g, \mu(\theta_g) = \tau_g + \kappa_g \Lambda$. Thus, between-group differences on $\Sigma(\theta_g)$ are a function of between-group differences on Φ_g and Ψ_g, and between-group differences on $\mu(\theta_g)$ are a function of between-group differences on κ_g and τ_g. Metric invariance implies that the latent construct of interest has the same scale in all populations, such that a 1-unit increase in ξ_g predicts the same increase for each item across groups (Van de Schoot, Lugtig, & Hox, 2012; Millsap & Hartog, 1998). Metric invariance permits the between-group comparability of structural relations (i.e., regression coefficients) involving latent variables (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

Substantively, metric invariance has been interpreted to mean that the latent construct has the same meaning in all groups, or an absence of Gamma change (Golembiewski et al., 1975).

The test of metric invariance, if it holds, is typically followed by the test of *strong* or *scalar* invariance (Meredith, 1993), which introduces between-group equality constraints on intercepts, such that $\tau_g = \tau$ for all g. When scalar invariance holds, Equation (1.9) becomes $\Sigma(\theta)_g = \Lambda_g \Phi' \Lambda_g + \Psi(\theta)_g$, $\mu(\theta)_g = \tau + \Lambda_g \kappa_g$, and Equation (1.3) becomes $Y_{ig} = \tau + \Lambda_g \xi_{ig} + \epsilon_{ig}$. Therefore, between-group differences on $\mu(\theta_g)$ are a function only of between-group differences on κ_g, and between-group differences on $E(Y_{ig})$ are a function only of between-group differences on ξ_{ig}, satisfying Equation (1.2). Scalar invariance is sufficient to permit between-group comparison of observed scores, observed means, and latent means (Steenkamp & Baumgartner, 1998). With constraints on intercepts, latent means become estimable, with latent means in one group –typically referred to as the *reference group* –fixed to 0 by convention. Latent mean estimates in other groups –referred to as *focal groups* –can, thus, be interpreted as latent mean differences relative to the reference group. Substantively, scalar invariance has been interpreted as implying the equivalent subjective understanding of all levels of all items across populations (Van de Schoot, Lugtig, & Hox, 2012; Millsap & Hartog, 1998).

If scalar invariance holds, the next and final degree of MI testing is the test of *strict invariance*, which introduces between-group equality constraints on residual variances (Meredith, 1993), such that $\Psi_g = \Psi$ for all g. When strict invariance holds, Equation (1.9) becomes $\Sigma(\theta_g) = \Lambda_g \Phi' \Lambda_g + \Psi$, $\mu(\theta_g) = \tau + \Lambda_g \kappa_g$. Thus, between-group differences on $\Sigma(\theta_g)$ and $\mu(\theta_g)$ are only a function of between-group differences on Φ_g and κ_g. Strict invariance is sufficient to satisfy Equation (1.1), implying equal probabilities across groups of any particular set of observed scores conditional on true scores (Meredith, 1993; Steenkamp & Baumgartner, 1998).

1.3.1 Partial Measurement Invariance

If full metric invariance cannot be retained, methodologists recommend testing *partial metric invariance* (Steenkamp and Baumgartner, 1998; Vandenberg & Lance, 2000). Partial metric invariance is the invariance of some, but not all, factor loadings across groups. The most commonly used method for testing partial MI is the method of Byrne, Shavelson and Muthén (1989). To determine which loadings are invariant across groups, between-group equality constraints on the loadings associated with a single latent variable are imposed on the configural invariance model, and loss of fit is evaluated via a chi-square difference test. If these constraints are found to lead to poor fit, they are removed, and constraints are applied to a different latent variable. If the initial constraints are found to fit, they remain when testing the invariance of loadings associated with other latent variables. Thus, the retained model becomes the new baseline model for subsequent difference tests. This procedure continues until invariance of all loadings has been tested for each latent variable. Next, invariance of individual loadings not already constrained to equality across groups is iteratively tested in the same way: if loss of fit is significant, the constraint is removed before proceeding to the next loading. If loss of fit is not significant, the constraint remains when testing invariance of other loadings. This procedure seeks to maximize the number of invariant parameters when full metric invariance across all factors cannot be retained. Partial metric invariance is retained if at least 2 factor loadings (including those fixed to 1 to identify latent variables) are invariant for each factor (Byrne, Shavelson & Muthén, 1989; Steenkamp & Baumgartner, 1998). Partial metric invariance permits valid comparison of unstandardized regression coefficients involving latent variables across groups (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). If partial metric invariance cannot be retained, then no further invariance testing or between-group comparison is warranted (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

If full scalar invariance cannot be retained, methodologists recommend testing *partial scalar invariance*, the invariance of some, but not all, intercepts across groups. The logic to specifying a partial scalar invariance model is largely identical to that of specifying a partial metric invariance model, with the exception that the invariance of intercepts cannot be tested for items whose loadings are not also invariant (Byrne, Shavelson & Muthén, 1989). Note that when a single intercept for a latent variable is constrained to equality across groups, the latent mean in the focal group becomes estimable. In a 2-group model, this results in no change in degrees of freedom, as one new constraint is introduced, and one new parameter is estimated. Therefore, a minimum of two intercept constraints is necessary before partial scalar invariance becomes testable (Byrne, Shavelson, & Muthén, 1989; Little, Slegers, & Card, 2006; Meredith, 1993; Steenkamp & Baumgartner 1998). Partial scalar invariance of at least 2 items per factor is considered sufficient to permit between-group comparisons involving latent means, latent variances, and latent covariances (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Vandenberg and Lance (2000) also argue that the addition and retention of between-group equality constraints on latent covariance matrices to a partial scalar invariance model, such that $\mathsf{F}_g =$ for all *g*, permits the inference that any structural relation involving latent variances that holds in one group necessarily holds in all groups. If partial scalar invariance cannot be retained, no further invariance testing is warranted.

While partial strict invariance is mentioned by Byrne, Shavelson, & Muthen (1989), it is rarely discussed in the applied literature, as there are largely no meaningful implications of confirming the invariance of some, but not all, residual variances for applied researchers.

It should be noted that there is little consensus among methodologists as to which algorithm for specifying partial MI models is most appropriate. Cheung and Rensvold (1999) noted that the logic of Byrne, Shavelson, and Muthen (1989) can be applied in bottom-up (such that equality constraints are progressively added to the configural invariance model) or top-down manner (such that constraints are progressively removed from the strict invariance model), and produce different results. Cheung and Rensvold (1999) argue that because choosing a reference indicator loading to constrain to 1 for latent variable identification assumes that this loading is invariant across groups, invariance of other loadings is contingent on choice of reference indicator. When full metric invariance of a given factor cannot be retained in a partial metric invariance specification search, the authors recommend testing the invariance of loadings for each possible reference indicator, and identifying the largest set of items such that all items are mutually invariant for every choice of reference indicator within the set. Some methodologists recommend computing confidence intervals around between-group differences in measurement parameter estimates to verify their invariance across groups (Cheung & Lau, 2012; Meade & Bauer, 2007). Millsap and Yoon (2007) recommend letting modification indices guide partial MI model specification searches. The alignment method (Asparouhov & Muthén, 2014) and Bayesian SEM (Muthén & Asparouhov, 2013a) are also candidate techniques for guiding partial MI model specification searches.

1.4. Overview of Dissertation Body Chapters

The remainder of this introduction provides an overview of the research discussed in chapters 2, 3, and 4. For each chapter, relevant concepts are introduced, followed by a brief outline of the primary research. Chapter 2 introduces a structural equation modelling approach to evaluating measurement invariance for observed composite scores when measurement invariance cannot be retained for each individual item. Chapter 3 introduces a method for evaluating the impact of violations of measurement invariance on between-group differences in classification accuracy when observed composite scores are used for selection. Chapter 4 introduces a new approach to evaluating fit of nested measurement invariance models using goodness-of-fit indices, focusing primarily on the RMSEA.

1.4.1. Chapter 2: Measurement Invariance of Observed Composite Scores

In practice, researchers who make use of psychometric instruments frequently use observed composite scores as proxies for unobservable true scores (e.g., Head, Allison, Lucena, Hassenstab, & Morris, 2017; Levant, Alto, McKelvey, Richmond, & McDermott, 2017; McCuish, Mathesius, Lussier, & Corrado, 2017; McDermott et al., 2017; Stevens, Blanchard, Shi, & Littlefield, 2018). Observed composite scores are generally computed as

$$y_{comp,ig} = w' Y_{ig} , \qquad (1.10)$$

where Y_{ig} is a $P \; {}^{3}1$ vector of item scores for individual *i* in group *g*, and *w* is $P \; {}^{3}1$ a vector of item weights. In the case of unweighted sum scores, every item is given a weight of 1. In the case of unweighted mean scores, every item is given a weight of $\frac{1}{P}$.

The definition of strict invariance given in Equation (1.1) states that the sampling distributions of item scores, given latent variable scores, are independent of group membership. In MGCFA this definition of MI is satisfied when $\Lambda_g =$, $\tau_g =$, and $\Psi_g =$ ` for all g (Meredith, 1993). The definition of scalar invariance, given in Equation (1.2) states that expected item scores are only a function of latent variable scores, and are independent of group membership (Steenkamp & Baumgartner, 1998). In MGCFA this definition of MI is satisfied when $\Lambda_g =$ and $\tau_g =$ for all g. For applied researchers who are primarily interested in ensuring that observed composite scores can be compared across groups, these definitions of MI may be overly restrictive. To illustrate, one can think of observed composite scores as item scores for a psychometric test with a single item, i.e., let $Y = y_{comp}$. To confirm strict invariance of this single-item test, it must be verified that Equation (1.1) holds, in this case meaning

$$P(y_{comp,ig} \mid \xi_{ig}, g) = P(y_{comp,ig} \mid \xi_{ig}). \tag{1.11}$$

To confirm strong invariance of this single-item test, it must be verified that Equation (1.2) holds, in this case meaning

$$E(y_{comp,ig} \mid \xi_{ig}, g) = E(y_{comp,ig} \mid \xi_{ig}). \tag{1.12}$$

If we recall Equation (1.10), we know that Equation (1.11) and Equation (1.12) are necessarily satisfied when Equation (1.1) and Equation (1.2) are, respectively. However, the converse is not necessarily true.

In Chapter 2 of this dissertation, I demonstrate that, for researchers who are only interested in ensuring MI of observed composites, and not of each individual item, it is only necessary to confirm equality of measurement parameter totals, such that $w'\Lambda_g = w'$, $w'\tau_g = w'$, and $w'\Psi_g w = w'\Psi w$ for all g for the general MGCFA model given in Equation

(1.9). In this chapter, I propose a structural equation modelling approach to testing Equation (1.11) and Equation (1.12) which relaxes some of the constraints imposed by MGCFA when testing Equation (1.1) and Equation (1.2). A series of power analyses is conducted to evaluate the power of tests -ol fe v tehl i"s Mpl r owphœsne dhn o"dseclasl-eare fit group populations for which violations of MI are present.

1.4.2. Chapter 3: Quantifying the Impact of Partial Measurement Invariance on Selection Accuracy in Multiple Groups.

When violations of measurement invariance are detected, methodologists often recommend testing partial measurement invariance (Byrne, Shavelson & Muthen, 1989; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000), as discussed in section 1.3.1. Unfortunately, partial MI has minimal implications for applied researchers who wish to make use of observed composite scores. While partial scalar invariance is sufficient to permit between-group comparisons on latent means, full scalar invariance is necessary to permit between-group comparisons on observed composite means (Steinmetz, 2013) or individual scores (Millsap & Kwok, 2004). In general, the only implications of partial MI for observed composite use is that one can identify a measurement invariant subset of items, and discard non-invariant items. This practice is considered problematic, as it may be capitalizing on chance, and may introduce many different reduced versions of the same scale to the literature (Browne, 2006; Millsap & Kwok, 2004).

Millsap and Kwok (2004) proposed the first method of quantifying the impact of violations of MI on observed composite use when only partial scalar invariance can be retained within a factor analytic context. Specifically, their method permits the calculation of measures of classification accuracy in two groups when a specified critical percentile on observed composite scores is used for selection/diagnosis. Researchers may then decide whether between-group differences on measures of classification accuracy are acceptably small. Unfortunately, while meritorious, the method has seen little application in applied analyses of MI.

In Chapter 3 of this dissertation, I propose method which allows researchers to identify critical percentiles for selection that minimize between-group differences in classification accuracy. Measures of classification accuracy at each percentile from 1 to 99 are computed for each group using the method of Millsap and Kwok (2004), and are presented as a plot of classification accuracy ratios against all possible critical percentiles. Classification accuracy ratios are used because they give the relative probability of being selected given that one is in the reference group, rather than the focal group, a reasonably intuitive and accessible metric of the impact of violations of MI. Researchers may then identify critical values for selection based on an acceptably small amount of expected bias on their preferred measures of classification accuracy.

A series of sample classification accuracy ratio plots are presented, illustrating the impact of different patterns of violations of MI on different measures of classification accuracy. Depending on the direction of violations of MI, as well as the measures of classification accuracy of interest, violations of MI across multiple types of measurement parameter (i.e., loadings, intercepts, or residual variances) may have additive or compensatory effects on classification accuracy differences across groups, which can readily be decoded by examining the proposed plots.

1.4.3. Chapter 4: Extending Goodness-of-Fit Indices to The Analysis of Measurement Invariance.

As was discussed in section 1.2.2.1, goodness-of-fit indices (GFIs) are alternative measures of model fit used in structural equation modelling that place misfit on continuous metrics, and are popular alternatives to traditional chi-square tests of model fit, which provide binary tests of point null hypotheses of perfect fit. A notable limitation of many GFIs is that guidelines tend to only exist for assessing fit of a single model, while guidelines for how to apply GFIs in the context of comparing nested MI models are generally few.

Some methodologists have advocated for the use of differences in goodness-of-fit indices ($\Delta GFIs$) —such that $\Delta GFI = GFI_A - GFI_B$, where GFI_A is the fit of M_A on a particular fit index, and GFI_B is the fit of M_B on that same fit index —for the purposes of comparing fit of nested MI models (Chen, 2007; Cheung & Rensvold, 2002). A study by Chen (2007) proposed critical values for rejection of the null hypothesis of equal model fit by $\Delta GFIs$. Unfortunately, critical values were found to be quite variable, varying as a function of sample size, sample evenness, pattern of violations of MI, proportion of parameters that violate MI, and location of violations of MI. The variability of proposed critical values for model rejection make it difficult to make strong recommendations that generalize to designs outside of the exact populations studied, which casts some doubt on the utility of $\Delta GFIs$ when testing MI.

Chapter 4 of this dissertation proposes an alternative approach to comparing nested SEMs via goodness-of-fit indices. The proposed method allows misfit introduced by M_A above and beyond that already present under M_B to be interpreted on the same scale as overall model GFIs. This allows the critical values for model rejection that researchers are familiar with (e.g., Hu & Bentler, 1999) to be applied to nested model comparison. While this research was motivated by its application in the context of evaluating MI, its logic should apply to the comparison of any nested SEMs. Further, while Chapter 4 primarily discusses how to use the RMSEA in the evaluation of MI, its logic should extend to any GFIs that are transformations $\hat{F}_{ML}$, such as the CFI.

Chapter 2: A Structural Equation Modelling Approach to Evaluating Measurement Invariance of Observed Composite Scores

2.1. Introduction

In psychological research, the most popular method of testing for the presence of measurement invariance (MI) is Multiple Group Confirmatory Factor Analysis (MGCFA). The MGCFA approach to evaluating MI involves fitting the same measurement model separately to data collected across different groups, and subsequently introducing between-group equality constraints on individual measurement parameter estimates. Non-significant loss of model fit as a function of added constraints is typically interpreted as evidence for the invariance of the constrained parameters (although see Yuan and Chan (2016) for cautions against such inferences). If all measurement parameters are found to be invariant, researchers generally conclude that scores on that instrument have the same meaning and interpretation across groups. Within the MGCFA framework, measurement is considered fully invariant when factor loadings, indicator intercepts, and residual variances are invariant across groups, however invariance of factor loadings and indicator intercepts in the absence of invariant residual variances is generally considered sufficient to permit comparison of scores across groups (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

In this chapter, I argue that the traditional MGCFA approach of constraining individual parameter estimates to equality across groups can be overly restrictive when evaluating measurement invariance of psychometric instruments. Specifically, if one is only interested in the comparability of total scores, rather than individual item scores, as is frequently the case in psychological research (e.g., Head, Allison, Lucena, Hassenstab, & Morris, 2017; Levant, Alto, McKelvey, Richmond, & McDermott, 2017; McCuish, Mathesius, Lussier, & Corrado, 2017;

McDermott et al., 2017; Stevens, Blanchard, Shi, & Littlefield, 2018), one needs only to confirm invariance of parameter *totals* across groups.

In the present chapter, I define *scale-level measurement invariance* as the invariance of measurement parameter totals, thus permitting comparison of observed composite scores across groups. The traditional MGCFA conception of MI, the invariance of every individual measurement parameter across groups, is henceforth referred to as *item-level measurement invariance*. I note that scale-level MI necessarily holds whenever item-level MI holds; item-level MI models are nested within scale-level MI models (Bentler & Bonett, 1980). I argue that depending on one's interests, researchers may -level MI as a proxy to scale-level invariance, which is sufficient but not necessary.

I begin by demonstrating that invariance of measurement parameter totals across groups implies between-group comparability of composite scores to the same degree that invariance of individual measurement parameters implies between-group comparability of item scores. I then introduce an SEM model for evaluating scale-level MI which does not assume item-level MI. Next, I conduct a series of asymptotic power computations using the methodology of MacCallum, Browne, and Cai (2006) to compare the power of chi-square difference tests to detect violations of scale-level MI and item-level MI. We follow up with an empirical application using real data. Sample code illustrating how to implement scale-level MI constraints on loadings, intercepts, and residual variances in the R package *lavaan* (Rosseel, 2012), and the statistical package *Mplus* (Muthén & Muthén, 1998-2017) are included in Appendices 2A and 2B, respectively.

2.1.1. *Defining Measurement* Invariance of Composites or “Scale-*Level Measurement* Invariance”

In practice, behavioral researchers frequently use observed composite scores on psychometric instruments as proxies for true, unobservable scores on psychological constructs. Let us define observed composite as $y_{comp} = w'Y$, where w is a $P \times 1$ vector of composite weights, and P is the number of items involved in the composite. In most cases w is either a vector of 1's, producing unweighted sum scores, or $\frac{1}{P}$'s, producing mean scores. For the duration of this chapter, w can be assumed to be a vec

groups.

When composite scores are used in psychological research, ensuring the measurement invariance of all items –as given in Equation (1.1) –may be overly restrictive if one is only interested in the comparability of composite scores. Statistically, we can define the strict invariance of composite scores, or scale-level strict invariance, as:

$$P(y_{comp} \mid \chi, g) = P(y_{comp} \mid \chi). \tag{2.1}$$

The probability of any particular value of y_{comp}, given χ, is independent of group membership. We can also define the strong, or scalar invariance of composite scores, or scale-level scalar invariance, as:

$$E(y_{comp} \mid \chi, g) = E(y_{comp} \mid \chi). \tag{2.2}$$

The expected value of y_{comp} given χ is independent of group membership.

As was illustrated in Chapter 1, Equation (1.1) holds when $\Lambda_g = \;$, $\tau_g = \;$, and $\Psi_g = \;$` . We can show that Equation (2.1) also holds under these constraints, as $Y_{i,g} = \Lambda \xi_{i,g} + \tau e_{i,g} +$ where $e_{i,g} \sim MVN(0, \Psi)$, for individual *i* in group *g*, meaning, $y_{comp,i,g} = w'(\Lambda\xi_{i,g} + \tau e_{i,g} +)\; w'(= {}_{i,g}\xi)\Lambda\; w' + w\tau' e_{i,g}$. Let us define $\lambda^*_g = w'\Lambda_g$, $\tau^*_g = w'{}_g$`, and $e^*_{i,g} = w'e_{i,g}$, and thus, in terms of Equation (1.3),

$$y_{comp,i,g} = \lambda^*_g\, {}_{i,g}\xi + {}^*_g\, \tau e^*_{i,g} +, \qquad (2.3)$$

where $e^*_{i,g} \sim N(0, \psi^*_g)$, and $\psi^*_g = w'\Psi_g w$. If $\Lambda_g = \;$, $\tau_g = \;$, and $\Psi_g = \;$`, it is necessarily true that $\lambda^*_g = {}^*\lambda$, $\tau^*_g = {}^*\tau$, and $\psi^*_g = \psi^*$ for all *g*, and thus, $y_{comp,i,g} = \lambda^*\, {}_{i,g}\xi + {}^*\, \tau e^*_{i,g} +$, where $e^*_{i,g} \sim N(0, \psi^*)$, satisfying Equation (2.1). Similarly, we can show that Equation (2.2) is necessarily true when Equation (1.2) is, as $\Lambda_g = \;$, $\tau_g = \;$, and Equation (2.3) imply $E(y_{comp,i,g} \mid \xi_{i,g}) = {}^*\lambda_{i,g}\xi + {}^*$.

While $\Lambda_g = \;$, $\tau_g = \;$, and $\Psi_g = \;$` necessarily imply $\lambda^*_g = {}^*\lambda$, $\tau^*_g = {}^*\tau$, and $\psi^*_g = \psi^*$, the converse need not be the case. If, in the reference group, $\Lambda_R = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, and, in the focal group, $\Lambda_F = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, $\Lambda_R \neq \Lambda_F$, but $\lambda^*_R = 6 = \lambda^*_F$. Recall that in analyses of MI across two groups, the reference group is generally the larger of the two, although this designation is, essentially, arbitrary. For applied researchers making use of observed composites, it may be unimportant that item 3 is a stronger indicator in the reference group than in the focal group, and that item 1 is a stronger indicator in the focal group than in the reference group, so long as the entire instrument, in aggregate, is an equally good indicator in both groups. Thus, a conceptualization of MI which does not assume every individual item behaves identically across groups is of interest to researchers who evaluate MI primarily to decide whether observed composite scores can be compared across groups.

2.1.2. Testing Scale-Level Measurement Invariance

To formally test for violations of different degrees of scale-level measurement invariance, we propose a series of SEM models imposing between-group equality constraints on factor loading totals (l^*), intercept totals (t^*), and residual variance totals (y^*), in that order. The *scale-level metric invariance* model introduces between-group equality constraints on factor loading totals –such that $l_g^* = {}^*{\scriptstyle /}$ –to the baseline configural invariance model. The *scale-level scalar invariance* model, subsequently, introduces between-group equality constraints on intercept totals –such that $t_g^* = {}^*{\scriptstyle ,}$ –to the scale-level metric invariance model, satisfying Equation (2.2). The *scale-level strict invariance* model, further, introduces constraints on residual variance totals –such that $y_g^* = y^*$ –to the scale-level scalar invariance model, satisfying Equation (2.1). Scale-level MI models are nested within the configural invariance model, and thus, the plausibility of the constraints can be assessed jointly, or sequentially via difference testing. Note that there is no scale-level configural MI model.

A model with scale-level measurement invariance constraints is identified as long as the typical latent variance and latent mean identification constraints are present. Specifically, it is still necessary to fix a single loading to 1 or to fix the factor variance to 1 to identify the scale of the latent factor. When testing equality of intercept totals, it is also necessary to additionally constrain a single intercept to equality across groups. This is because a single between-group equality constraint on intercepts provides the information to estimate the latent mean difference across groups, resulting in a model with identical fit and degrees of freedom to that with no intercept constraints and a latent mean difference fixed to 0. Thus, a minimum of two mean structure constraints is necessary before mean structure constraints are testable (Byrne, Shavelson, & Muthén, 1989; Little, Slegers, & Card, 2006; Meredith, 1993; Steenkamp & Baumgartner 1998). See Appendices 2A and 2B for sample code fitting our proposed model in *lavaan* and *Mplus*, respectively.

2.1.3. Power Analyses

To evaluate the utility of scale-level MI models, relative to item-level MI models, a series of asymptotic power analyses is conducted evaluating power of chi-square difference tests to detect violations of scale-level and item-level MI when scale-level and item-level MI models are fit to data. Power analyses are performed using the methodology of MacCallum, Browne, and Cai (2006). This method involves computing the population RMSEA for a baseline model, M_B, and constrained model, M_A, being compared via a chi-square difference test, from which the non-centrality parameter, λ_{nonc}, for the sampling distribution of the chi-square difference test statistic, *D*, can be computed. The proportion of the density of this non-central chi-square distribution which exceeds the critical value for rejection of the null hypothesis of equal model fit is, thus, the power of the chi-square difference test to detect the studied misspecifications. This methodology is useful because it does not require data simulation, only the specification of population and fitted models. This technique is described in greater detail in the methodology section.

In the present chapter, I use the method of MacCallum, Browne and Cai (2006) to evaluate the power of chi-square difference tests to detect a variety of violations of scale-level and item-level MI when scale-level and item-level MI models are fit to data. Power is plotted over a range of sample sizes, from $N_R = N_F = 50$ to $N_R = N_F = 1{,}000$, where N_R and N_F are sample sizes in the reference group and focal group, respectively.

2.2. Methodology

2.2.1. Simulation Design

A series of four simulation studies, each consisting of multiple power analyses, is conducted to compare the behavior of tests of scale-level and item-level measurement invariance. Power analyses are conducted in *R* (R Core Team, 2016) using the methodology of MacCallum, Browne, and Cai (2006), an extension of MacCallum, (1996) SEM power analysis procedure to difference testing. MacCallum, Browne, and Sugawara's (1996) method noted that one could model, in a given population, by calculating the non-centrality parameter of the alternative distribution as $\lambda_{nonc} = (N-1) * df * RMSEA^2$. The population RMSEA for a given model can be obtained by fitting that model to population covariance matrices and population mean vectors. Power to detect misfit can then be estimated by computing the density of the chi-square distribution with degrees of freedom *df* and non-centrality parameter λ_{nonc} that exceeds the 95th percentile of the central chi-square distribution with *df* degrees of freedom. This method can be extended to chi-square difference testing by computing $\lambda_{nonc} = (N-1)(df_A * RMSEA_A^2 - df_B * RMSEA_B^2)$ (MacCallum, Browne & Cai, 2006), or $\lambda_{nonc} = (\frac{N-G}{G})(df_A * RMSEA_A^2 - df_B * RMSEA_B^2)$, in the case of multiple-group models (Steiger, 1998), where $RMSEA_A$ and df_A are the RMSEA and degrees of freedom associated with M_A, the more restrictive model in the comparison, and $RMSEA_B$ and df_B are the RMSEA and degrees of freedom associated with M_B, the less restrictive model in the comparison. Power of the chi-square difference test can then be computed as the density of the chi-square distribution with $df_D = df_A - df_B$ degrees of freedom and non-centrality parameter λ_{nonc} that exceeds the 95th percentile of the central chi-square distribution with df_D degrees of freedom.

In each of the four simulation studies, scale-level and item-level measurement invariance models are fit to a variety of multiple-group population models for which violations of scale-level or item-level MI are present, in order to compute the population RMSEAs associated with those models in that multiple-group population. All model fitting is done in the SEM package *lavaan* (version 0.5-23) (Rosseel, 2012). In any given power analysis, violations of MI are present in the population on only one group of measurement parameters: loadings, intercepts, or residual variances. In any given power analysis, a single scale-level and item-level MI model is fit to the multiple-group population model: scale-level and item-level metric invariance when loadings violate MI, scale-level and item-level scalar invariance when intercepts violation MI, and scale-level and item-level strict invariance when residual variances violate MI. Only one MI model needs to be fit to each multiple-group population because, given that violations of MI are only present on one group of measurement parameters, one can assume that less restrictive MI models fit perfectly. To clarify, when violations of metric MI are present in the population, one can assume that the RMSEA value associated with the configural MI model is 0. When violations of scalar invariance are present in the population, one can assume that the RMSEA value associated with the metric invariance model is 0, and so on. Thus, one does not need to compute $RMSEA_B$ to compute the λ_{nonc} for the sampling distribution of D.

Calculated RMSEA values are subsequently converted to non-centrality parameters, and used to compute power to detect violations of measurement invariance relative to correctly specified baseline models. In all power analyses, the baseline model is one degree of MI lower than the constrained model, such that, for example, fit of the scale-level strict MI model is compared to the fit of the scale-level scalar MI model. Because the baseline model is properly specified in all power analyses, its population RMSEA can be assumed to be 0, and the non-centrality parameter for the difference test can be computed as $\lambda_{nonc} = \frac{(N - G)}{G} * df_A * RMSEA_A^2$. The results of power analyses are presented as plots of sample size, ranging from $N_R = N_F = 50$ to $N_R = N_F = 1{,}000$, against power of the chi-square difference test.

It is worth mentioning that there exists some skepticism of the validity of asymptotic power computations produced by the methods of MacCallum, Browne, and Sugawara (1996) and MacCallum, Browne, and Cai (2006). To verify that power computations are accurate, power analyses were also conducted using Monte Carlo simulation for a subset of sample sizes in a subset of conditions. All power computations using the method of MacCallum, Browne, and Cai (2006) were consistent with the results of the Monte Carlo simulation. It should be noted that this observation does not dismiss skepticism of the method, but rather highlights that the method appears to behave as intended for unidimensional multiple-group models when sample sizes are equal.

2.2.2. Study Conditions

A series of four simulation studies consisting of multiple power analyses is conducted. In the first study, power of tests of scale-level and item-level MI is evaluated when scale-level and item-level MI models are fit to multiple-group populations for which violations of item-level MI that do not violate scale-level MI are present. Thus, in Study 1, type I error rates are evaluated for tests of scale-level MI, while power is evaluated for tests of item-level MI. In the subsequent three simulation studies, power of tests of scale-level and item level MI is evaluated when scale-level and item-level MI models are fit to multiple-group populations for which violations of scale-level MI –and by extension, violations of item-level MI –are present. Thus, in studies 2, 3 and 4, power is evaluated for tests of both scale-level and item-level MI.

In simulation studies 1 through 3, $P = 8$, and population parameter values in the reference group are always $\Lambda'_R = [.7,.7,.7,.7,.7,.7,.7,.7]$, $\tau'_R = [0,0,0,0,0,0,0,0]$, $\text{diag}(\Psi_R) = [.51,.51,.5]$, $\Phi_R = 1$, and $\kappa_R = 0$, in all conditions. By extension, $\lambda^*_R = 5.6$, $\tau^*_R = 0$ and $\psi^*_R = 4.08$. Thus, in studies 1 through 3, $y_{comp,i,R} = 5.6\xi_{i,R} + \varepsilon^*_{i,R}$, with $\varepsilon^*_{i,R} \sim N(0,4.08)$. Population parameters in the focal group are described within each study below, and summarized in tables 1 through 3. Population parameters used in Study 4 are described below.

In each power analysis within each study, power of tests of either metric, scalar, or strict invariance is evaluated. In studies 1 through 3, 40 degrees of freedom are associated with the configural invariance model. When MI is evaluated using scale-level MI models, 41, 42, and 43 degrees of freedom are associated with the metric, scalar, and strict invariance models, respectively. When MI is evaluated using item-level MI models, 47, 54, and 62 degrees of freedom are associated with the metric, scalar, and strict invariance models, respectively. Chi-square difference tests evaluating metric, scalar, and strict invariance are, thus, based on 1, 1, and 1 degree of freedom, respectively, when scale-level constraints are used, and 7, 7, and 8 degrees of freedom, respectively, when item-level constraints are used. All four simulation studies will now be described in greater detail.

Study 1: To illustrate that tests of scale-level invariance are only sensitive to violations of scale-level MI, while tests of item-level MI are sensitive to all violations of MI, a series of six power analyses is conducted using multiple-group populations where item-level violations of MI do not violate scale-level MI. Two power analyses compute power to detect violations of metric invariance, two compute power to detect violations of scalar invariance, and two compute power to detect violations of strict invariance. In all power analyses population parameters are invariant across groups, with the exception of the violations of MI present on one of either loadings, intercepts, or residual variances. Multiple-group population parameters are given in *Table 2.1*.

In conditions where L_R ¸ $_F\mathrm{L}$, either $\Lambda_F' = [. 7 , . 6 ,$, or $\Lambda_F' = [. 7 , . 5 ,$. Note that in both cases, $l_F^* = 5.6$, and thus, $l_R^* = {}_F^*l = ^*$. In conditions where t_R ¸ $_Ft$, either $\tau_F' = [0 , - . 1 ,$, or $\tau_F' = [0 , - . 2 ,$. Note that in both cases $t_F^* = 0$, and thus $t_R^* = {}_F^*t = ^*$. In conditions where Y_R ¸ $_F\mathrm{Y}$, either $\mathrm{diag}(\Psi_F) = [. 5 1 , . 4 1 , . 6]$, or $\mathrm{diag}(\Psi_F) = [. 5 1 , . 3 1 , . 7]$. Note that in both cases, $y_F^* = 4.08$, and thus $y_R^* = y_F^* = ^*y$. Power of chi-square difference tests to detect violations of metric, scalar, or strict invariance when scale-level and item-level MI constraints are fit to data is evaluated for conditions where L_R ¸ $_F\mathrm{L}$, t_R ¸ $_Ft$, or Y_R ¸ $_F\mathrm{Y}$, respectively.

Study 2: A series of 18 power analyses is conducted to explore the power of tests of scale-level and item-level MI to detect violations of scale-level and item-level MI. Six power analyses compute power to detect violations of scale-level and item-level metric invariance, six compute power to detect violations of scale-level and item-level scalar invariance, and six compute power to detect violations of scale-level and item-level strict invariance. Study 2 varies both the number of individual non-invariant measurement parameters, and between-group differences on measurement parameter totals ($l_F^* - {}_R^* l$, $t_F^* - {}_R^* \iota$, or $y_F^* - y_R^*$). In each power analysis, between 2 and 7 loadings, intercepts or residual variances are .1 higher in the focal group. By extension, measurement parameter totals are between .2 and .7 higher in the focal group across power analyses. Population parameters in the focal group are otherwise identical to those in the reference group. Population parameters for each power analysis in Study 2 are summarized in *Table 2.2.*

In conditions where L_R ¸ $_F L$, either $\Lambda_F{}' = [\ .\ 7\ ,\ .\ 8\ ,,$ $\Lambda_F{}' = [\ .\ 7\ ,\ .\ 7\ ,,$ $\Lambda_F{}' = [\ .\ 7\ ,\ .\ 7\ ,,$ $\Lambda_F{}' = [\ .\ 7\ ,\ .\ 7\ ,,$ $\Lambda_F{}' = [\ .\ 7\ ,\ .\ 7\ ,,$ or $\Lambda_F{}' = [\ .\ 7\ ,\ .\ 7\ ,.$ In each condition, $l_F^* = 6.3$, $l_F^* = 6.2$, $l_F^* = 6.1$, $l_F^* = 6.0$, $l_F^* = 5.9$, or $l_F^* = 5.8$, respectively. By extension, $l_F^* - {}_R^* l = .7$, $l_F^* - {}_R^* l = .6$, $l_F^* - {}_R^* l = .5$, $l_F^* - {}_R^* l = .4$, $l_F^* - {}_R^* l = .3$, or $l_F^* - {}_R^* l = .2$, respectively.

In conditions where t_R ¸ $_F \iota$, either $\tau_F{}' = [\ 0\ ,\ .\ 1\ ,\ ,$ $\tau_F{}' = [\ 0\ ,\ 0\ ,\ .\ ,$ $\tau_F{}' = [\ 0\ ,\ 0\ ,\ 0,$ $\tau_F{}' = [\ 0\ ,\ 0\ ,\ 0,$ $\tau_F{}' = [\ 0\ ,\ 0\ ,\ ($,or $\tau_F{}' = [\ 0\ ,\ 0\ ,$. In each condition, $t_F^* = .7$, $t_F^* = .6$, $t_F^* = .5$, $t_F^* = .4$, $t_F^* = .3$, or $t_F^* = .2$, respectively. By extension, $t_F^* - {}_R^* t = .7$, $t_F^* - {}_R^* t = .6$, $t_F^* - {}_R^* t = .5$, $t_F^* - {}_R^* t = .4$, $t_F^* - {}_R^* t = .3$, or $t_F^* - {}_R^* t = .2$.

In conditions where Y_1 ¸ Y_2, either $\text{diag}(\Psi_F) = [\,.5\,1\,,\,.6\,1\,,\,.6]$, $\text{diag}(\Psi_F) = [\,.5\,1\,,\,.5\,1\,,\,.6]$, $\text{diag}(\Psi_F) = [\,.5\,1\,,\,.5\,1\,,\,.5]$, $\text{diag}(\Psi_F) = [\,.5\,1\,,\,.5\,1\,,\,.5]$, $\text{diag}(\Psi_F) = [\,.5\,1\,,\,.5\,1\,,\,.5]$, or $\text{diag}(\Psi_F) = [\,.5\,1\,,\,.5\,1\,,\,.5]$. In each of these conditions, $y^*_F = 4.78$, $y^*_F = 4.68$, $y^*_F = 4.58$, $y^*_F = 4.48$, $y^*_F = 4.38$, and $y^*_F = 4.28$, respectively. By extension, $y^*_F - y^*_R = .7$, $y^*_F - y^*_R = .6$, $y^*_F - y^*_R = .5$, $y^*_F - y^*_R = .4$, $y^*_F - y^*_R = .3$, and $y^*_F - y^*_R = .2$, respectively. Power of chi-square difference tests to detect violations of metric, scalar, or strict invariance when scale-level and item-level MI models are fit to data is evaluated for conditions where L_R ¸ L_F, t_R ¸ t_F, or Y_R ¸ Y_F, respectively.

Study 3: A series of 18 power analyses is conducted to explore the power of tests of scale-level and item-level MI to detect violations of scale-level and item-level MI. Six power analyses compute power to detect violations of scale-level and item-level metric invariance, six compute power to detect violations of scale-level and item-level scalar invariance, and six compute power to detect violations of scale-level and item-level strict invariance. Study 3 varies the number of non-invariant parameters in each multiple-group population, while holding between-group differences on measurement parameter totals constant. Thus, when fewer parameters violate MI, the magnitude of between-group differences on individual non-invariant parameters increases. Specifically, for $w' = [1,1,1,1,1,1,1,1]$ either $l^*_F - l^*_R = .6$, $t^*_F - t^*_R = .6$, or $y^*_F - y^*_R = .6$, in all power analyses, while between 1 and 6 parameters are non-invariant across groups, for a total of 18 power analyses. Population parameters in Study 3 are given below, and in *Table 2.3.*

Population parameters in the focal group are identical to those used in the reference group, with the exception that impactful violations of MI are present on one of either loadings, intercepts, or residual variances. In conditions where L_R ¸ $_F\mathrm{L}$, either $\Lambda_F' = [\ .7\ ,\ .7\ ,$, $\Lambda_F' = [\ .7\ ,\ .7\ ,\ .7\ ,$ $\Lambda_F' = [\ .7\ ,\ .7\ ,\ .7,$ $\Lambda_F' = [\ .7\ ,\ .7\ ,,$ $\Lambda_F' = [\ .7\ ,\ .7$, or $\Lambda_F' = [\ .7\ ,\ .7\ ,$. In conditions where t_R ¸ $_F\iota$, either $\tau_F' = [\ 0\ ,\ 0\ ,\ .\ ,$ $\tau_F' = [\ 0\ ,\ 0\ ,\ 0\ ,\ .\,$, $\tau_F' = [\ 0\ ,\ 0\ ,\ 0\ ,\ 0,$ $\tau_F' = [\ 0\ ,\ 0\ ,\ ($, $\tau_F' = [\ 0\ ,\ 0\ ,$,or $\tau_F' = [\ 0\ ,\ 0\ ,$. In conditions where Υ_R ¸ $_F\Upsilon$, either $\mathrm{diag}(\Psi_F) = [\ .51\ ,\ .51\ ,\ .6],$ $\mathrm{diag}(\Psi_F) = [\ .51\ ,\ .51\ ,\ .5],$ $\mathrm{diag}(\Psi_F) = [\ .51\ ,\ .51\ ,\ .5],$ $\mathrm{diag}(\Psi_F) = [\ .51\ ,\ .51\ ,\ .5],$ $\mathrm{diag}(\Psi_F) = [\ .51\ ,\ .51\ ,\ .5],$ or $\mathrm{diag}(\Psi_F) = [\ .51\ ,\ .51\ ,\ .511]$. Power of chi-square difference tests to detect violations of scale-level and item-level metric, scalar, or strict invariance is evaluated for conditions where L_R ¸ $_F\mathrm{L}$, t_R ¸ $_F\iota$, or Υ_R ¸ $_F\Upsilon$, respectively.

Study 4: A series of 18 power analyses is conducted to explore the power of tests of scale-level and item-level MI to detect violations of scale-level and item-level MI. Six power analyses compute power to detect violations of scale-level and item-level metric invariance, six compute power to detect violations of scale-level and item-level scalar invariance, and six compute power to detect violations of scale-level and item-level strict invariance. Study 4 holds both the number of non-invariant parameters and between-group differences on measurement parameter totals constant across power analyses, while the total number of items, *P*, varies from 4 to 14 across power analyses. In all power analyses, either L_R ¸ $_F\mathrm{L}$, t_R ¸ $_F\iota$, or Υ_R ¸ $_F\Upsilon$, with all other parameter matrices being invariant. In all conditions, $\Phi_R = \Phi_F = 1$ and $\kappa_R = \kappa_F = 0$.

In all conditions, $\Lambda'_R = [.7,.7,.7,.7,...]$, with "..." being a placeho $(P-4) \in \{0,2,4,6,8,10\}$ additional .7's. Thus, in each multiple-group population, $\lambda^*_R = 2.8$, $\lambda^*_R = 4.2$, $\lambda^*_R = 5.6$, $\lambda^*_R = 7.0$, $\lambda^*_R = 8.4$, or $\lambda^*_R = 9.8$, depending on whether *P* is 4, 6, 8, 10, 12 or 14. In all conditions, $\tau_R = [0,0,0,0,...]$, with "..." being a placeholder for $(P-4) \in \{0,2,4,6,8,10\}$ additional 0's. Thus, in each multiple-group population, $\tau^*_R = 0$. $\text{diag}(\Psi_R) = [.51,.$ with "..." being a placeholder for $(P-4) \in \{0,2,4,6,8,10\}$ additional .51's. Thu multiple-group population, $\psi^*_R = 2.04$, $\psi^*_R = 3.06$, $\psi^*_R = 4.08$, $\psi^*_R = 5.10$, $\psi^*_R = 6.12$, or $\psi^*_R = 7.14$, depending on whether *P* is 4, 6, 8, 10, 12 or 14.

In conditions where $\Lambda_R \neq \Lambda_F$, $\Lambda'_F = [.7,.9,.9,.9,...]$, with "..." being a pl $(P-4) \in \{0,2,4,6,8,10\}$ additional .7's. Thus, in each multiple-group population where $\Lambda_R \neq \Lambda_F$, $\lambda^*_F = 3.4$, $\lambda^*_F = 4.8$, $\lambda^*_F = 6.2$, $\lambda^*_F = 7.6$, $\lambda^*_F = 9.0$, or $\lambda^*_F = 10.4$, depending on whether *P* is 4, 6, 8, 10, 12 or 14. Note that across all populations where $\Lambda_R \neq \Lambda_F$, $\lambda^*_F - \lambda^*_R = .6$.
In conditions where $\tau_R \neq \tau_F$, $\tau'_F = [0,.2,.2,.2,...]$, with "..." being a placeh $(P-4) \in \{0,2,4,6,8,10\}$ additional 0's. Thus, in each multiple-group population where $\tau_R \neq \tau_F$, $\tau^*_F = .6$, and $\tau^*_F - \tau^*_R = .6$. When $\Psi_R \neq \Psi_F$, $\text{diag}(\Psi_F) = [.51,..,$ with "..." bein placeholder for $(P-4) \in \{0,2,4,6,8,10\}$ additional .51's. Thus, in each multiple-group population where $\Psi_R \neq \Psi_F$, $\psi^*_F = 2.64$, 3.66, 4.68, 5.70, 6.72, or 7.74, depending on whether *P* is 4, 6, 8, 10, 12 or 14. Note that across all populations where $\Psi_R \neq \Psi_F$, $\psi^*_F - \psi^*_R = .6$.

Power of chi-square difference tests to detect violations of scale-level and item-level metric, scalar, or strict MI is evaluated for conditions where $\Lambda_R \neq \Lambda_F$, $\tau_R \neq \tau_F$, or $\Psi_R \neq \Psi_F$, respectively. Degrees of freedom associated with the configural invariance model are $df = P(P+1) - 4*P$. Scale-level MI constraints on loadings, intercepts, and residual variances each increase model degrees of freedom by 1. Item-level MI constraints on loadings, intercepts, and residual variances each increase degrees of freedom by $p-1$, $p-1$, and p, respectively. Degrees of freedom associated with difference tests are, thus, 1 for all tests of scale-level MI, and $p-1$, $p-1$, and p, for tests of item-level metric, scalar, and strict MI, respectively.

2.3. Results

Study 1. Figure 2.1 illustrates the results of Study 1, in which violations of item-level MI that do not violate scale-level MI are present in the population. Population models used in power analysis are presented in *Table 2.1.* Consistent with expectations, chi-square difference tests associated with scale-level MI models reject at .05 in all conditions, while those associated with item-level constraints do not. When violations of item-level MI are present, power of chi-square difference tests to detect violations of item-level MI is .8 near $N_R = N_F = 300$ in all conditions where non-invariant parameters have between-group differences of .1 (see *Figure 2.1(a), Figure 2.1(c),* and *Figure 2.1(e)*). When non-invariant parameters have between-group differences of .2, power of chi-square difference tests to detect violations of item-level MI is .8 near $N_R = N_F = 100$ in all conditions (see *Figure 2.1(b)*, *Figure 2.1(d)*, and *Figure 2.1(f)*). Power to detect item-level violations of MI appears largely uninfluenced by whether non-invariant parameters are at the level of loadings, intercepts, or residual variances.

Study 2. Study 2 evaluates power of tests of scale-level and item-level measurement invariance to detect violations of scale-level MI, and by extension, item-level MI. In Study 2, both the number of non-invariant parameters and between-group differences on non-invariant parameter totals vary across power analyses. *Figure 2.2*, *Figure 2.3* and *Figure 2.4* give power analysis results when violations of measurement invariance are present on loadings, intercepts, and residual variances, respectively. Conditions are presented in order of decreasing number of non-invariant parameters, from 7 parameters .1 higher in the focal group, to 2 such parameters.

Figure 2.2 illustrates the power of chi-square difference tests to detect violations of scale-level and item-level metric MI when L_R ¸ $_F$L in the population. Population models used in power analysis are presented in *Table 2.2. Figure 2.2(a)* through *Figure 2.2(f)* give the power to detect 7, 6, 5, 4, 3, or 2 non-invariant loadings .1 higher in the focal group, meaning $/^*_F - {}^*_R/$ ∓7 , .6, .5, .4, .3, or .2 in each population, respectively. Power of tests of scale-level metric MI is greater than that of tests of item-level metric MI when 7/8 loadings are non-invariant, meaning $/^*_F - {}^*_R/$ ∓7 . Power of tests of scale-level and item-level metric MI are similar when 6/8 loadings are non-invariant. Power of tests of item-level metric MI is greater than that of tests of scale-level metric MI when 5 or fewer loadings are non-invariant. As the number of non-invariant loadings –and by extension, $/^*_F - {}^*_R/$ –decreases, so does the power of tests of scale-level metric MI. This pattern is not observed for tests of item-level metric MI. Power of tests of item-level metric MI is greatest when 4/8 loadings are non-invariant, with power decreasing as the number of non-invariant loadings increases or decreases. This diminished power of tests of item-level MI when the majority of loadings violate MI has been previously documented by Chen (2007).

Figure 2.3 illustrates the power of chi-square difference tests to detect violations of scale-level and item-level scalar MI when $\tau_R \neq \tau_F$ in the population. Population models used in power analysis are presented in *Table 2.2*. *Figure 2.3(a)* through *Figure 2.3(f)* give the power to detect 7, 6, 5, 4, 3, or 2 non-invariant intercepts .1 higher in the focal group, meaning $\tau^*_F - \tau^*_R = .7$, .6, .5, .4, .3, or .2 in each population, respectively. Power of tests of scale-level scalar MI is greater than that of tests of item-level scalar MI when 7/8 intercepts are non-invariant, meaning $\tau^*_F - \tau^*_R = .7$. Power of tests of scale-level and item-level scalar MI are similar when 6/8 intercepts are non-invariant. Power of tests of item-level scalar MI is greater than that of tests of scale-level scalar MI when 5 or fewer intercepts are non-invariant. As the number of non-invariant intercepts —and by extension, $\tau^*_F - \tau^*_R$ —decreases, so does the power of tests of scale-level scalar MI. As was the case when $\Lambda_R \neq \Lambda_F$, this pattern is not observed for tests of item-level scalar MI. Power of tests of item-level scalar MI is greatest when 4/8 loadings are non-invariant, with power decreasing as the number of non-invariant intercepts increases or decreases. This diminished power of tests of item-level MI when the majority of intercepts violate MI has been previously documented by Chen (2007).

Figure 2.4 illustrates the power of chi-square difference tests to detect violations of scale-level and item-level strict MI when $\Psi_R \neq \Psi_F$ in the population. Population models used in power analysis are presented in *Table 2.2*. *Figure 2.4(a)* through *Figure 2.4(f)* give the power to detect 7, 6, 5, 4, 3, or 2 non-invariant residual variances .1 higher in the focal group, meaning $\psi^*_F - \psi^*_R = .7$, .6, .5, .4, .3, or .2 in each population, respectively. Power of tests of scale-level strict MI is greater than that of tests of item-level strict MI when 7, 6, or 5 residual variances are non-invariant. Power of tests of scale-level and item-level strict MI are similar when 4 or 3 residual variances are non-invariant. Power of tests of item-level strict MI is greater than that of tests of scale-level strict MI when 2 residual variances are non-invariant. Unlike when L_R ¸ $_F L$ or t_R ¸ $_F t$, as the number of non-invariant residual variances –and by extension $y_F^* - y_R^*$ – decreases, the power of tests of both scale-level and item-level strict MI decreases. This exception was also noted by Chen (2007).

Study 3. Study 3 evaluates power of tests of scale-level and item-level MI to detect violations of scale-level MI, and by extension, item-level MI. In Study 3, the number of non-invariant measurement parameters varies across multiple-group populations, while between-group differences on non-invariant parameter totals are held constant across power analyses. Thus, as the number of non-invariant parameters decreases, the magnitude of between-group differences on individual non-invariant parameters increases. *Figure 2.5*, *Figure 2.6*, and *Figure 2.7* give power analysis results when violations of MI are present on loadings, intercepts, and residual variances, respectively. Populations are presented in order of decreasing number of non-invariant parameters, from 6 parameters higher in the focal group, to 1 such parameter.

Figure 2.5 illustrates the power of chi-square difference tests to detect violations of scale-level and item-level metric MI when L_R ¸ $_F L$ in the population. Population models used in power analysis are presented in *Table 2.3. Figure 2.5(a)* through *Figure 2.5(f)* give the power to detect 6, 5, 4, 3, 2, or 1 non-invariant loading, respectively. Note that $l_F^* - {}_R^* l$ =6 across all power analyses presented in *Figure 2.5.* Power of tests of scale-level and item-level metric MI are similar when 6 loadings are non-invariant. Power of tests of item-level metric MI is greater than that of tests of scale-level metric MI when 5 or fewer loadings are non-invariant. As the number of non-invariant loadings decreases, with $l_F^* - {}_R^* l$ held constant, power of tests of item-level metric invariance increases, while power of tests of scale-level metric invariance is stable across power analyses.

Figure 2.6 illustrates the power of chi-square difference tests to detect violations of scale-level and item-level scalar MI when $\tau_R \neq \tau_F$ in the population. Population models used in power analysis are presented in *Table 2.3. Figure 2.6(a)* through *Figure 2.6(f)* give the power to detect 6, 5, 4, 3, 2, or 1 non-invariant intercept, respectively. Note that $\tau_F^* - \tau_R^* = .6$ across all power analyses presented in *Figure 2.6.* Power of tests of scale-level and item-level scalar MI are similar when 6 intercepts are non-invariant. Power of tests of item-level scalar MI is greater than that of tests of scale-level scalar MI when 5 or fewer intercepts are non-invariant. As the number of non-invariant intercepts decreases, with $\tau_F^* - \tau_R^*$ held constant, power of tests of item-level scalar MI increases, while power of tests of scale-level scalar MI is stable across power analyses.

Figure 2.7 illustrates the power of chi-square difference tests to detect violations of scale-level and item-level strict MI when $\Psi_R \neq \Psi_F$ in the population. Population models used in power analysis are presented in *Table 2.3. Figure 2.7(a)* through *Figure 2.7(f)* give the power to detect 6, 5, 4, 3, 2, or 1 non-invariant residual variance, respectively. Note that $\psi_F^* - \psi_R^* = .6$ across all power analyses presented in *Figure 2.7.* Power of tests of scale-level strict MI is greater than that of tests of item-level strict MI when 6 or 5 residual variances are non-invariant. Power of tests of scale-level and item-level strict MI are similar when 4 or 3 residual variances are non-invariant. Power of tests of item-level strict MI is greater than that of tests of scale-level MI when 2 or 1 residual variances are non-invariant. As the number of non-invariant residual variances decreases, with $\psi_F^* - \psi_R^*$ held constant, power of tests of item-level strict MI increases, while power of tests of scale-level strict MI is stable across power analyses.

Study 4. Study 4 evaluates power of tests of scale-level and item-level MI to detect violations of scale-level MI, and by extension, item-level MI. In Study 4, both the number of non-invariant measurement parameters and between-group differences on non-invariant parameter totals is constant across populations, while the number of items, *P*, is varied. *Figure 2.8*, *Figure 2.9*, and *Figure 2.10* give power analysis results when violations of MI are present on loadings, intercepts, and residual variances, respectively. Populations are presented in order of increasing *P*, from *P*=4 to *P*=14.

Figure 2.8 illustrates the power of chi-square difference tests evaluating the plausibility of scale-level and item-level metric MI to detect 3 non-invariant loadings that are .2 higher in the focal group. By extension, $l_F^* - l_R^* = .6$ in all power analyses presented in *Figure 2.8*. *Figure 2.8(a)* through *Figure 2.8(f)* give the power to detect 3 non-invariant loadings on scales where $P = 4$, 6, 8, 10, 12, or 14, respectively. Power of tests of scale-level metric MI is greater than that of tests of item-level metric MI when $P = 4$. Power of tests of item-level metric MI is greater than that of tests of scale-level metric MI when $P = 6$ or more. As *P* increases, the power of tests of scale-level metric MI to detect 3 non-invariant loadings decreases, while the power of tests of item-level metric MI increases.

Figure 2.9 illustrates the power of chi-square difference tests evaluating the plausibility of scale-level and item-level scalar MI to detect 3 non-invariant intercepts that are .2 higher in the focal group. By extension, $t_F^* - t_R^* = .6$ in all power analyses presented in *Figure 2.9*. *Figure 2.9(a)* through *Figure 2.9(f)* give the power to detect 3 non-invariant intercepts on scales where $P = 4$, 6, 8, 10, 12, or 14, respectively. Power of tests of scale-level scalar MI is greater than that of tests of item-level scalar MI when $P = 4$. Power of tests of item-level scalar MI is greater than that of tests of scale-level scalar MI when $P = 6$ or more. As *P* increases, the power of tests of scale-level scalar MI to detect 3 non-invariant intercepts decreases, while the power of tests of item-level scalar invariance increases.

Figure 2.10 illustrates the power of chi-square difference tests evaluating the plausibility of scale-level and item-level strict MI to detect 3 non-invariant residual variances that are .2 higher in the focal group. By extension, $y_F^* - y_R^* = .6$ in all power analyses presented in *Figure 2.10*. *Figure 2.10(a)* through *Figure 2.10(f)* give the power to detect 3 non-invariant residual variances on scales where $P = 4$, 6, 8, 10, 12, or 14, respectively. Power of tests of scale-level strict MI is greater than that of tests of item-level strict MI when $P = 4$. Power of tests of scale-level and item-level strict MI are similar when $P = 6$ or 8. Power of tests of item-level strict MI is greater than that of tests of scale-level strict MI when $P = 10$ or more. As *P* increases, the power of tests of scale-level strict MI to detect 3 non-invariant residual variances decreases, while the power of tests of item-level scalar invariance appears stable across scale lengths.

2.3.1. Empirical Application

In this section, scale-level measurement invariance is evaluated for real responses to a real psychometric instrument. In 2008, the Dutch healthcare system introduced Diagnosis Related Groups (DRGs), a system used to streamline health services by standardizing the price of medical treatment in terms of diagnostic categories, rather than specific services provided. A 2012 Study by Tummers and Van de Walle explored the willingness of psychologists, psychiatrists, and psychotherapists to use DRGs using a 4-item measure of “willingness to change” developed by Metselaar, Walet, Cozijn

(2013) examined measurement invariance of this scale between psychologists and psychiatrists in a methodological paper illustrating a Bayesian SEM (Muthén & Asparouhov, 2012) approach to evaluating MI. We now present a replication of this analysis using tests of scale-level and item-level MI.

A total of 5199 psychologists and psychiatrists were contacted to complete the survey, with 570 psychologists and 504 psychiatrists returning valid cases. Summary statistics are presented in *Table 2.4*. See Tummers and Van de Walle (2012) for detailed sampling methodology. A series of five multiple-group 1-factor CFAs are fit to data: 1) Configural invariance, 2) item-level metric invariance, 3) item-level scalar invariance, 4) scale-level metric invariance, and 5) scale-level scalar invariance. Subsequently, nested models are compared via chi-square difference tests. Analyses are conducted in *lavaan* (Version 0.5-23). Results of these analyses are presented in *Table 2.5*.

While the configural invariance model does not fit by the chi-square ($\chi^2 = 16.727, df = 4, p = .002$), it does fit well by the CFI (.992) and adequately by the RMSEA (.077). Item-level metric invariance fits by the chi-square difference test relative to configural invariance ($\Delta\chi^2 = 5.143,\ df = 3, p = .161$), however a significant loss of fit is observed when subsequently introducing between-group equality constraints on intercepts ($\Delta\chi^2 = 14.335,\ df = 3, p = .002$). Scale-level constraints on both loadings ($\Delta\chi^2 = 2.212,\ df = 1, p = .137$) and intercepts ($\Delta\chi^2 = 1.540,\ df = 1, p = .215$), however, fit by the appropriate chi-square difference tests, suggesting that while individual items do not perform identically in both populations, the set of items in aggregate do, permitting between group comparison of composite scores.

2.4. Discussion

The present chapter introduces the concept of scale-level measurement invariance, the invariance of measurement parameter totals across groups, implying the measurement invariance and between-group comparability of observed composite scores on psychometric instruments. The traditional MGCFA approach to evaluating MI involves testing the invariance of each individual factor loading, indicator intercept, and residual variance across groups. I demonstrate that such assumptions can be relaxed if one is primarily interested in the comparability of observed composite scores, which is frequently the case in behavioral research. I propose a SEM based approach to testing scale-level MI (See *Appendix A* and *Appendix B* for sample syntax using *lavaan* and *Mplus*, respectively.)

2.4.1. Summary and Interpretation of Results

A series of power analyses is conducted illustrating the utility of tests of scale-level measurement invariance. As expected, when violations of item-level MI that do not violate scale-level MI are present in the population (Study 1), tests of scale-level MI maintain expected type I error rates, while tests of item-level MI do not. Power of tests of item-level MI to detect violations of MI increases with the magnitude of between-group differences on individual parameters. See *Table 2.1* for population and fitted models used in Study 1, and *Figure 2.1* for the results of Study 1.

Subsequent simulations evaluate the power of chi-square difference tests associated with scale-level and item-level MI models to detect violations of both scale-level and item-level measurement invariance. Study 2 evaluates the power of tests of scale-level and item-level measurement invariance to detect violations of MI while varying both the number of non-invariant parameters, and between-group differences on measurement parameter totals. See *Table 2.2* for population and fitted models used in Study 2. Power of chi-square difference tests to detect violations of scale-level and item-level metric invariance —meaning $\Lambda_R \neq \Lambda_F$ and $\lambda^*_R \neq \lambda^*_F$ in the population —are presented in *Figure 2.2.* Power of chi-square difference tests to detect violations of scale-level and item-level scalar invariance —meaning $\tau_R \neq \tau_F$ and $\tau^*_R \neq \tau^*_F$ in the population —are presented in *Figure 2.3.* Power of chi-square difference tests to detect violations of scale-level and item-level strict invariance —meaning $\Theta_R \neq \Theta_F$ and $\theta^*_R \neq \theta^*_F$ in the population —are presented in *Figure 2.4.*

In general, power of chi-square difference tests to detect violations of scale-level MI is greater than the power of chi-square difference tests to detect violations of item-level MI when a large number (7 of 8) of individual measurement parameters violate MI. Power to detect violations of item-level MI is generally greater than power to detect violations of scale-level MI when a smaller number (5 or fewer) of individual measurement parameters violate MI. In general, power of chi-square difference tests to detect violations of scale-level MI increases as a function of between-group differences on measurement parameter totals. For example, power to detect violations of scale-level metric MI is greatest when $|\lambda^*_F - \lambda^*_R| = .7$, and lowest when $|\lambda^*_F - \lambda^*_R| = .2$.

Curiously, power to detect violations of item-level metric or scalar invariance is greatest when violations of MI are present for half of loadings or intercepts, and power decreases with increasing or decreasing number of parameters that violate MI. This phenomenon has previously been documented by Chen (2007). I hypothesize that this phenomenon may be due, in part, to the fact that when between-group equality constraints are placed on loadings or intercepts, and the majority of those loadings and intercepts are greater in one group in the population, the maximum likelihood algorithm may be compensating by adjusting latent mean or latent variance estimates. This explanation is consistent with the fact that this phenomenon is only observed when all parameters that violate MI are greater in the same group, and the fact that the phenomenon is not observed for violations of strict MI (Chen, 2007), as such violations of MI cannot be compensated for with latent mean or variance adjustments.

In general, I believe that tests of scale-level MI behave in a manner that is desirable to researchers interested in MI of observed composite scores. Tests of scale-level MI are more powerful than tests of item-level MI under the most severe violations of MI studied. Further, power of tests of scale-level MI decreases with the magnitude of between-group differences on measurement parameter totals –which can be thought of as an unstandardized effect size of violations of MI –while power of tests of item-level MI does not.

Study 3 evaluates the power of tests of scale-level and item-level measurement invariance to detect violations of MI while varying the number of non-invariant parameters in each multiple-group population, but holding between-group differences on measurement parameter totals constant. See *Table 2.3* for population and fitted models used in Study 3. Power of chi-square difference tests to detect violations of scale-level and item-level metric invariance – meaning $\Lambda_R \neq \Lambda_F$ and $\lambda^*_R \neq \lambda^*_F$ in the population –are presented in *Figure 2.5.* Power of chi-square difference tests to detect violations of scale-level and item-level scalar invariance – meaning $\tau_R \neq \tau_F$ and $\tau^*_R \neq \tau^*_F$ in the population –are presented in *Figure 2.6.* Power of chi-square difference tests to detect violations of scale-level and item-level strict invariance – meaning $\Psi_R \neq \Psi_F$ and $\psi^*_R \neq \psi^*_F$ in the population –are presented in *Figure 2.7.*

In general, in Study 3, power of tests of item-level MI is greater than that of tests of scale-level MI. Power of tests of item-level MI increases as violations of MI are distributed over fewer items, and by extension, as between-group differences on non-invariant measurement parameters increases. Power of tests of scale-level MI, however, is unrelated to the number of parameters that violate MI in the population. I believe that tests of scale-level MI behave in a manner more desirable to researchers interested in MI of observed composite scores. This is because across simulations, between-group differences on measurement parameter totals are held constant, and as such, the magnitude of violations of scale-level violations of MI are also held constant. Phrased differently, in all populations studied, the relationship between $y_{comp,i,R}$ and $\chi_{i,R}$ is held constant, the relationship between $y_{comp,j,F}$ and $\chi_{j,F}$ is held constant, and thus, the difference between these relationships is unchanged. It is, therefore, appropriate that tests of scale-level MI are equally sensitive in all conditions. Applied researchers who wish to verify MI of observed composites likely have little interest in the number of parameters that contribute to between-group differences on parameter totals, and are likely more interested in the between-group differences themselves.

Study 4 evaluates the power of tests of scale-level and item-level MI to detect violations of MI while varying scale length, but holding both the number of non-invariant parameters in each population constant, and between-group differences on measurement parameter totals constant. Power of chi-square difference tests to detect violations of scale-level and item-level metric invariance –meaning $\Lambda_R \neq \Lambda_F$ and $\lambda^*_R \neq \lambda^*_F$ in the population –are presented in *Figure 2.8.* Power of chi-square difference tests to detect violations of scale-level and item-level scalar invariance –meaning $\tau_R \neq \tau_F$ and $\tau^*_R \neq \tau^*_F$ in the population –are presented in *Figure 2.9.* Power of chi-square difference tests to detect violations of scale-level and item-level strict invariance – meaning y_R ¸ y_F and y^*_R ¸ y^*_F in the population –are presented in *Figure 2.10.*

With the exception of when *P*=4, tests of item-level MI are generally more powerful than tests of scale-level MI. As *P* increases, power to tests of item-level MI to detect violations of MI increases, while power of tests of scale-level MI to detect violations of MI decreases. I believe that tests of scale-level MI behave in a manner more desirable to researchers interested in MI of observed composite scores. To elaborate, consider the power analysis where L_R ¸ L_F and *P*=4. In the population, $y_{comp,i,R}$ =2.8 $\chi_{i,R}$ $+\varepsilon^*_{i,R}$ for individual *i* in the reference group, and $y_{comp,j,F}$ =3.4 $\chi_{j,F}$ $+\varepsilon^*_{j,F}$ for individual *j* in the focal group. In the reference group, a 1-unit increase in $\chi_{i,R}$ predicts a 2.8 unit increase in $y_{comp,i,R}$. In the focal group, a 1-unit increase in $\chi_{j,F}$ predicts a 3.4 unit increase in $y_{comp,j,F}$. It is reasonable to say that, for a given 1-unit increase in $\chi_{j,F}$, 2.8 of the associated 3.4 unit increase in $y_{comp,j,F}$ is justified, and 0.6 of the 3.4 unit increase is due to violations of MI. Now consider the case where L_R ¸ L_F and *P*=14. In the population, $y_{comp,i,R}$ =9.8 $\chi_{i,R}$ $+\varepsilon^*_{i,R}$, and $y_{comp,j,F}$ =10.4 $\chi_{j,F}$ $+\varepsilon^*_{j,F}$. In this case, it is reasonable to say that for a given 1-unit increase in $\chi_{j,F}$, 9.8 of the associated 10.4 increase in $y_{comp,j,F}$ is fair, and 0.6 of the 10.4 unit increase is due to violations of MI. As *P* increases, the variance of $y_{comp,j,F}$ increases, while the contribution of violations of MI to $E(y_{comp,j,F} \mid \chi_{j,F})$ is constant. Thus, when *P* is greater, the proportion of variability in $y_{comp,j,F}$ due to violations of MI is smaller. Applied researchers who wish to verify MI of observed composites would certainly prefer to use methods for which power increases, rather than decreases, as a function of increasing magnitude of the impact of violations of MI.

2.4.2. Limitations

Researchers interested in applying the discussed methodology are cautioned against using item-level and scale-level MI constraints in conjunction. For example, a researcher who has used item-level MI constraints to identify a partial invariance model may be tempted to use scale-level constraints to test the invariance of the remaining items as a parcel. Post-hoc parcelling of items is known to radically reduce the power of tests of measurement invariance, particularly if researchers take the liberty of evaluating a variety of different combinations of items as parcels (Meade & Kroustalis, 2006). We encourage only placing scale-level constraints across all indicators of a single factor, with the exception of the indicator whose associated parameters are fixed in latent variable identification.

It is also worth noting that because the present manuscript uses idealized power curves based on asymptotic power calculations, results can only be assumed to generalize to multivariate normal data. Further research is necessary to evaluate the behavior of tests of scale-level MI with popular corrections for non-normality (Brace & Savalei, 2016; Nevitt & Hancock, 2000; Satorra, 2000; Satorra & Bentler 1988; Satorra & Bentler, 2001; Satorra & Bentler, 2010). Finally, all present simulations use positively loading items. We encourage researchers making use of the proposed method to use positively correlated items, or negatively keyed items, as the impact of negative loadings on model fit has not been evaluated, and may threaten the validity of the method. Other limitations to the proposed method which are also relevant to the methods proposed in Chapter 3 and Chapter 4 are addressed in Chapter 5.

2.4.3. Conclusion

It has been demonstrated that tests of scale-level MI are useful for evaluating between-group comparability of observed composite scores on psychometric instruments when measurement invariance of individual items is not present or of little interest, which is frequently the case in practice. Of primary importance, we find that tests of scale-level MI maintain expected type I error rates when item-level violations of MI that do not violate scale-level MI are present. We also find that the power of tests of scale-level MI positively correlates with the magnitude of the impact of scale-level violations of MI, while the power of tests of item-level MI is often unrelated or even negatively correlated with the magnitude of the impact of scale-level violations of MI.

Location of Violations of MI	Figure	Focal Group Parameters That Differ From Reference Group Parameters	Parameter Total Differences	M_B	M_A
L_R ¸ $_F\mathrm{L}$	1(a)	$\Lambda_F' = [.7, .6,$	$l^*_F - {}^*_R l \neq 0$	Configural	Metric
	1(b)	$\Lambda_F' = [.7, .5,$	$l^*_F - {}^*_R l \neq 0$	Configural	Metric
t_R ¸ $_F l$	1(c)	$\tau_F' = [0, -.1,$	$t^*_F - {}^*_R t \neq 0$	Metric	Scalar
	1(d)	$\tau_F' = [0, -.2,$	$t^*_F - {}^*_R t \neq 0$	Metric	Scalar
Y_R ¸ $_F\mathrm{Y}$	1(e)	diag(Ψ_F) = [.51, .41, .6]	$y^*_F - y^*_R \neq 0$	Scalar	Strict
	1(f)	diag(Ψ_F) = [.51, .31, .7]	$y^*_F - y^*_R \neq 0$	Scalar	Strict

Table 2.1. Population and Fitted Models for Power Analyses Conducted in Simulation Study 1.

In the reference group, $\mathrm{L}'_R \;ǂ.7,.7,.7,.7,.7,.7,.7,.7]$, $t'_R =[0,0,0,0,0,0,0,0]$, diag(Ψ_R) = [.51, .51, .5], $\mathrm{F}_R \;=1$, and $k_R =0$ in all power analyses. Population parameters in the focal group that differ from those in the reference group in each population are given in the second column. All other focal group parameters are invariant in the population. M_B is the baseline model and M_A the constrained model fit to population models in each power analysis. Note that the models named in the fourth and fifth columns refer to both scale-level and item-level MI models. Thus, in the first power analysis, scale-level and item-level metric invariance models are fit to a multiple-group population for which violations of metric invariance are present. The method of MacCallum, Browne, and Cai (2006) is then used to compute the power of a chi-square difference test to reject constraints imposed by M_A above and beyond those already present under M_B.

Location of Violations of MI	Figure	Focal Group Parameters That Differ From Reference Group Parameters	Parameter Total Differences	M_B	M_A
L_R ¸ $_F\mathrm{L}$	2(a)	$\Lambda_F' = [.7,.8,$	$l_F^* - {}_R^*l \,/\!\!= .7$	Configural	Metric
	2(b)	$\Lambda_F' = [.7,.7,$	$l_F^* - {}_R^*l \,/\!\!= .6$	Configural	Metric
	2(c)	$\Lambda_F' = [.7,.7,$	$l_F^* - {}_R^*l \,/\!\!= .5$	Configural	Metric
	2(d)	$\Lambda_F' = [.7,.7,$	$l_F^* - {}_R^*l \,/\!\!= .4$	Configural	Metric
	2(e)	$\Lambda_F' = [.7,.7,$	$l_F^* - {}_R^*l \,/\!\!= .3$	Configural	Metric
	2(f)	$\Lambda_F' = [.7,.7,$	$l_F^* - {}_R^*l \,/\!\!= .2$	Configural	Metric
t_R ¸ $_F t$	3(a)	$\tau_F' = [0,.1,$	$t_F^* - {}_R^*t \,/\!\!= .7$	Metric	Scalar
	3(b)	$\tau_F' = [0,0,.$	$t_F^* - {}_R^*t \,/\!\!= .6$	Metric	Scalar
	3(c)	$\tau_F' = [0,0,0$	$t_F^* - {}_R^*t \,/\!\!= .5$	Metric	Scalar
	3(d)	$\tau_F' = [0,0,0$	$t_F^* - {}_R^*t \,/\!\!= .4$	Metric	Scalar
	3(e)	$\tau_F' = [0,0,($	$t_F^* - {}_R^*t \,/\!\!= .3$	Metric	Scalar
	3(f)	$\tau_F' = [0,0,$	$t_F^* - {}_R^*t \,/\!\!= .2$	Metric	Scalar
Y_R ¸ $_F\mathrm{Y}$	4(a)	$\mathrm{diag}(\Psi_F) = [.51,.61,.6]$	$y_F^* - y_R^* \,/\!\!= .7$	Scalar	Strict
	4(b)	$\mathrm{diag}(\Psi_F) = [.51,.51,.6]$	$y_F^* - y_R^* \,/\!\!= .6$	Scalar	Strict
	4(c)	$\mathrm{diag}(\Psi_F) = [.51,.51,.5]$	$y_F^* - y_R^* \,/\!\!= .5$	Scalar	Strict
	4(d)	$\mathrm{diag}(\Psi_F) = [.51,.51,.5]$	$y_F^* - y_R^* \,/\!\!= .4$	Scalar	Strict
	4(e)	$\mathrm{diag}(\Psi_F) = [.51,.51,.5]$	$y_F^* - y_R^* \,/\!\!= .3$	Scalar	Strict
	4(f)	$\mathrm{diag}(\Psi_F) = [.51,.51,.5]$	$y_F^* - y_R^* \,/\!\!= .2$	Scalar	Strict

Table 2.2. Population and Fitted Models for Power Analyses Conducted in Simulation Study 2.

In the reference group, L'_R ‡.7,.7,.7,.7,.7,.7,.7,.7], t'_R =[0,0,0,0,0,0,0,0], $\mathrm{diag}(\Psi_R) = [.51,.51,.5]$, F_R ⫤1, and k_R =0 in all power analyses. Population parameters in the focal group that differ from those in the reference group in each population are given in the second column. All other focal group parameters are invariant in the population. M_B is the baseline model and M_A the constrained model fit to population models in each power analysis.

Location of Violations of MI	Figure	Focal Group Parameters That Differ From Reference Group Parameters	Parameter Total Differences	M_B	M_A
$\Lambda_R \neq \Lambda_F$	5(a)	$\Lambda_F' = [.7, .7,$	$l_F^* - l_R^* = .6$	Configural	Metric
	5(b)	$\Lambda_F' = [.7, .7, .7$	$l_F^* - l_R^* = .6$	Configural	Metric
	5(c)	$\Lambda_F' = [.7, .7, .7$	$l_F^* - l_R^* = .6$	Configural	Metric
	5(d)	$\Lambda_F' = [.7, .7,$	$l_F^* - l_R^* = .6$	Configural	Metric
	5(e)	$\Lambda_F' = [.7, .7$	$l_F^* - l_R^* = .6$	Configural	Metric
	5(f)	$\Lambda_F' = [.7, .7,$	$l_F^* - l_R^* = .6$	Configural	Metric
$\tau_R \neq \tau_F$	6(a)	$\tau_F' = [0, 0, .$	$t_F^* - t_R^* = .6$	Metric	Scalar
	6(b)	$\tau_F' = [0, 0, 0, .$	$t_F^* - t_R^* = .6$	Metric	Scalar
	6(c)	$\tau_F' = [0, 0, 0, 0$	$t_F^* - t_R^* = .6$	Metric	Scalar
	6(d)	$\tau_F' = [0, 0, ($	$t_F^* - t_R^* = .6$	Metric	Scalar
	6(e)	$\tau_F' = [0, 0,$	$t_F^* - t_R^* = .6$	Metric	Scalar
	6(f)	$\tau_F' = [0, 0,$	$t_F^* - t_R^* = .6$	Metric	Scalar
$\Psi_R \neq \Psi_F$	7(a)	diag(Ψ_F) = [.51, .51, .6]	$y_F^* - y_R^* = .6$	Scalar	Strict
	7(b)	diag(Ψ_F) = [.51, .51, .5]	$y_F^* - y_R^* = .6$	Scalar	Strict
	7(c)	diag(Ψ_F) = [.51, .51, .5]	$y_F^* - y_R^* = .6$	Scalar	Strict
	7(d)	diag(Ψ_F) = [.51, .51, .5]	$y_F^* - y_R^* = .6$	Scalar	Strict
	7(e)	diag(Ψ_F) = [.51, .51, .5]	$y_F^* - y_R^* = .6$	Scalar	Strict
	7(f)	diag(Ψ_F) = [.51, .51, .511]	$y_F^* - y_R^* = .6$	Scalar	Strict

Table 2.3. Population and Fitted Models for Power Analyses Conducted in Simulation Study 3.

In the reference group, $\Lambda'_R = [.7,.7,.7,.7,.7,.7,.7,.7]$, $\tau'_R = [0,0,0,0,0,0,0,0]$, diag($\Psi_R$) = [.51, .51, .5], $\Phi_R = 1$, and $\kappa_R = 0$ in all power analyses. Population parameters in the focal group that differ from those in the reference group in each population are given in the second column. All other focal group parameters are invariant in the population. M_B is the baseline model and M_A the constrained model fit to population models in each power analysis.

	1.	2.	3.	4.
1. I intend to try to convince employees of the benefits the DRG-policy	2.021(0.731), 1.828(0.732)			
2. I intend to put effort into achieving the goals of the DRG-policy	.591, .549	2.651(1.044), 2.416(1.140)		
3. I intend to reduce resistance among employees regarding the DRG-policy	.728, .737	.617, .599	2.352(0.765), 2.186(0.951)	
4. I intend to make time to implement the DRG-policy	.453, .472	.445, .490	.486, .512	2.794(0.943), 2.471(1.091)

Table 2.4. Correlations Among Items on the “Willingness to Change *for* DRG” M *Psychologists (n=570) and Psychiatrists (n=504).* Means (Variances) are presented on the diagonal

	c^2	*df*	*p*(c^2 ,*df*)	Dc^2	D*df*	*p*(Dc^2 , D*df*)
Configural (M_0)	16.727	4	.002	-	-	-
Item-Level Metric (M_1)	21.870	7	.003	5.143	3	.162
Item-Level Scalar (M_2)	36.058	10	.00008	14.335	3	.002
Scale-level Metric (M_{T1})	18.938	5	.002	2.2115	1	.137
Scale-level Scalar (M_{T2})	20.478	6	.002	1.540	1	.215

Table 2.5. Tests of Model Fit and Chi-Square Difference Tests for Item-Level and Scale-level Measurement Invariance Constraints Fit to Tum
Change ” Data.

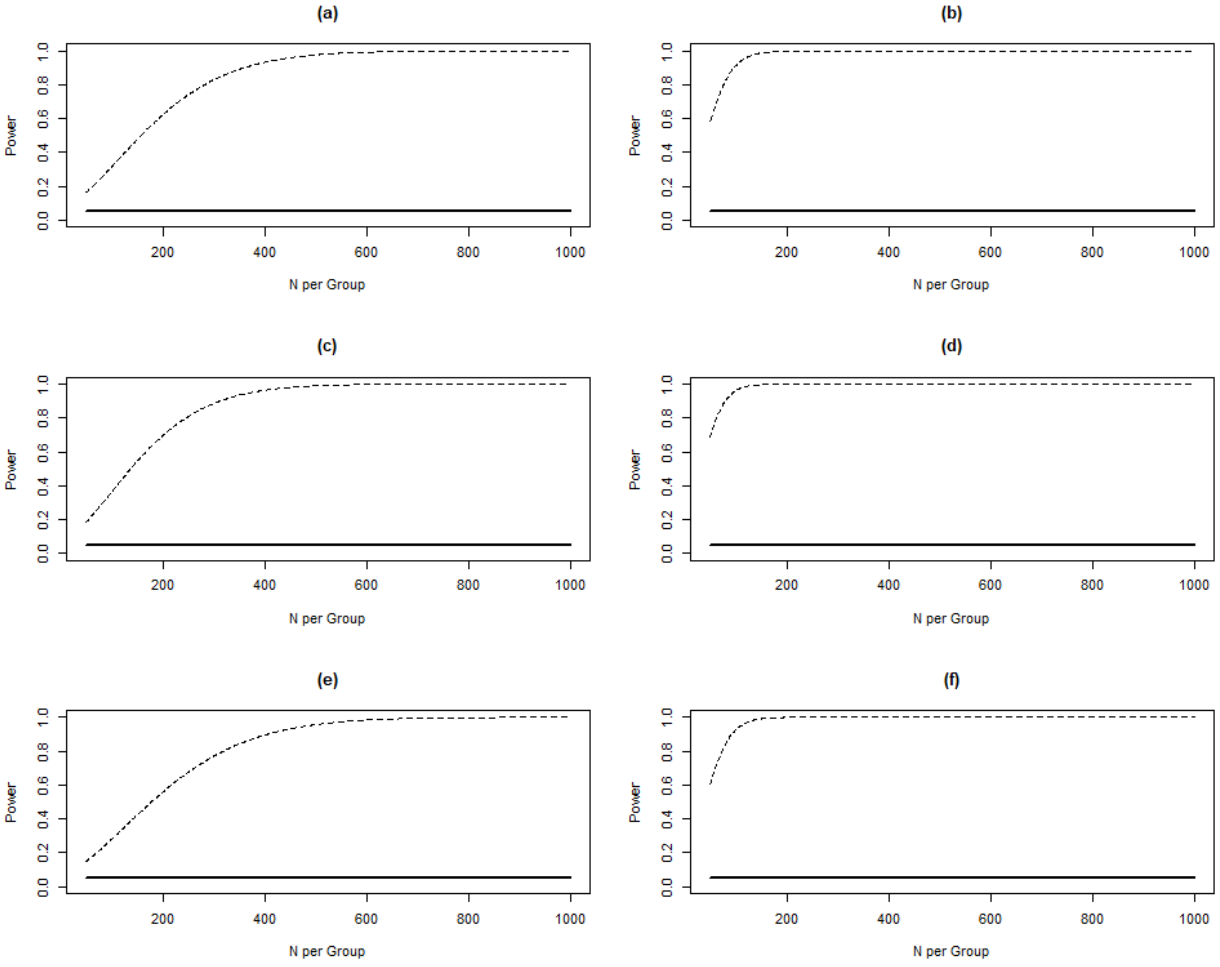

Figure 2.1. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Item-Level Violations of Measurement Invariance When Scale-Level Measurement Invariance is True in the Population.

Figures (a) and (b) present power of chi-square difference tests to detect violations of item-level metric invariance when non-invariant loadings differ by .1 and .2, respectively. Figures (c) and (d) present power of chi-square difference tests to detect violations of item-level scalar invariance when non-invariant intercepts differ by .1 and. 2, respectively. Figures (e) and (f) present power of chi-square difference tests to detect violations of item-level strict invariance when non-invariant residual variances differ by .1 and .2, respectively. The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

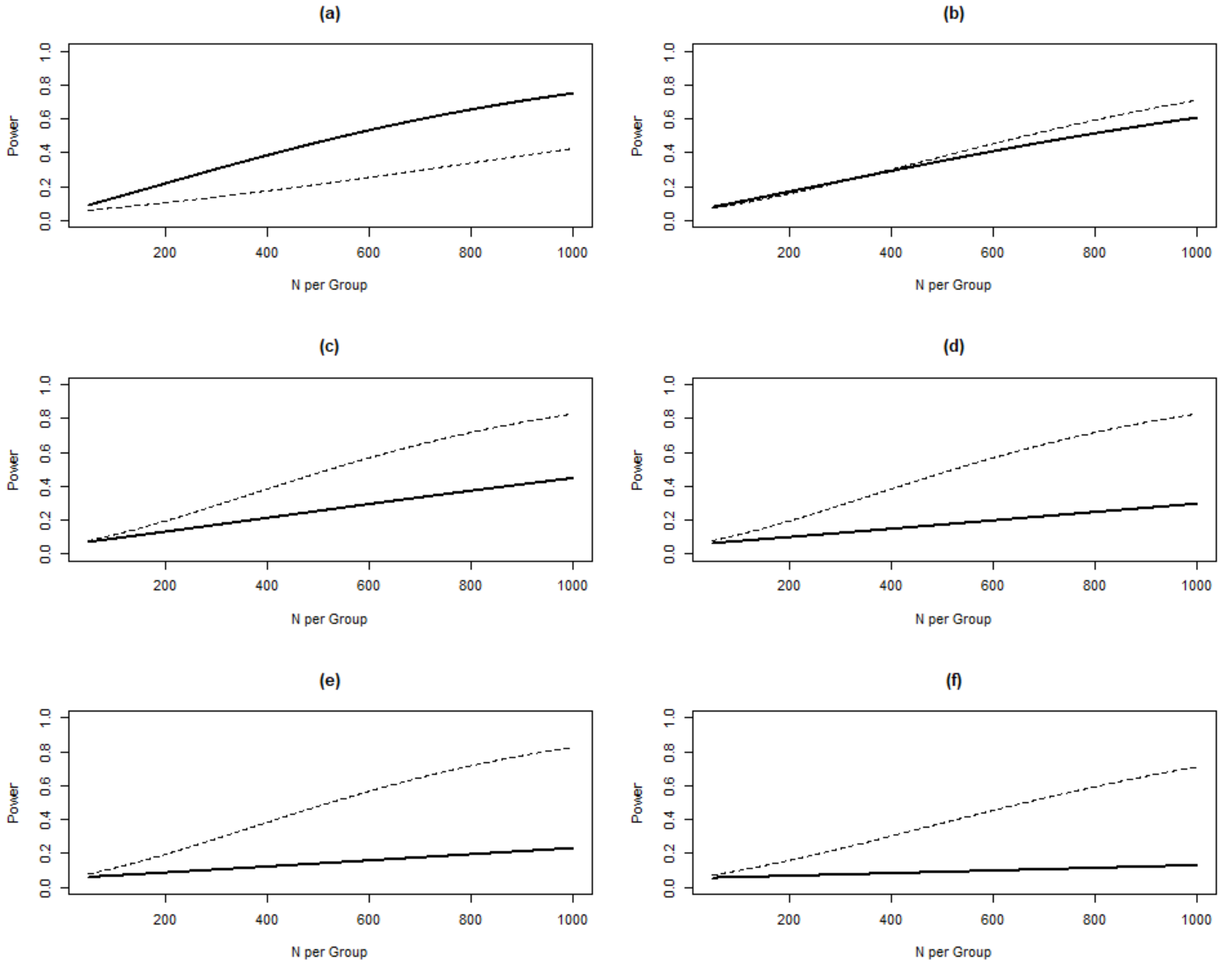

Figure 2.2. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Metric Invariance While Varying Number of Non-Invariant Loadings and Total Non-Invariance.

Figures (a) through (f) depict power to detect 7, 6, 5, 4, 3, and 2 non-invariant loadings (+.1 in group 2), respectively. The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

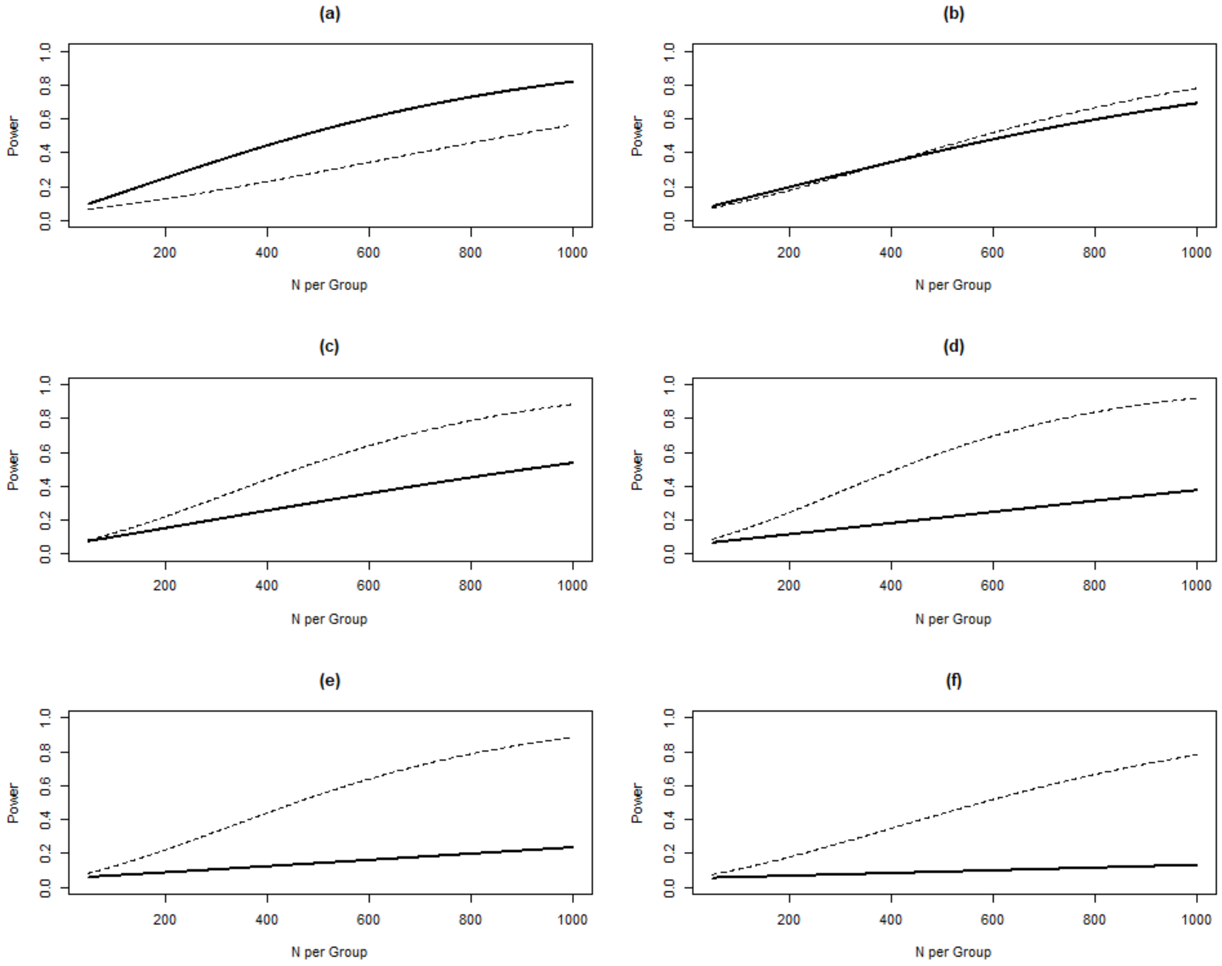

Figure 2.3. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Scalar Invariance While Varying Number of Non-Invariant Intercepts and Total Non-Invariance.

Figures (a) through (f) depict power to detect 7, 6, 5, 4, 3, and 2 non-invariant intercepts (+.1 in group 2), respectively. The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

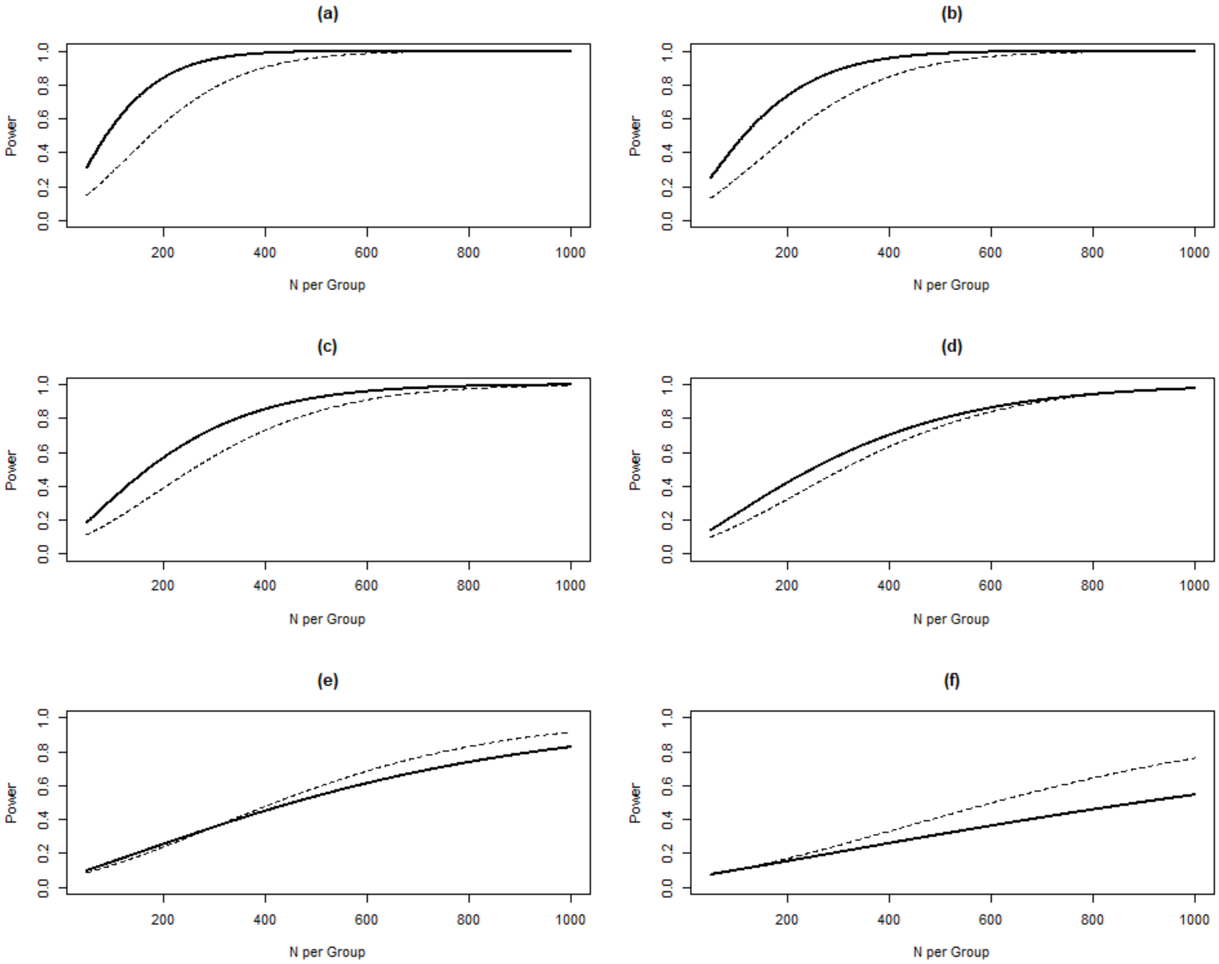

Figure 2.4. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Strict Invariance While Varying Number of Non-Invariant Residual Variances and Total Non-Invariance.

Figures (a) through (f) depict power to detect 7, 6, 5, 4, 3, and 2 non-invariant residual variance (+.1 in group 2), respectively. The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

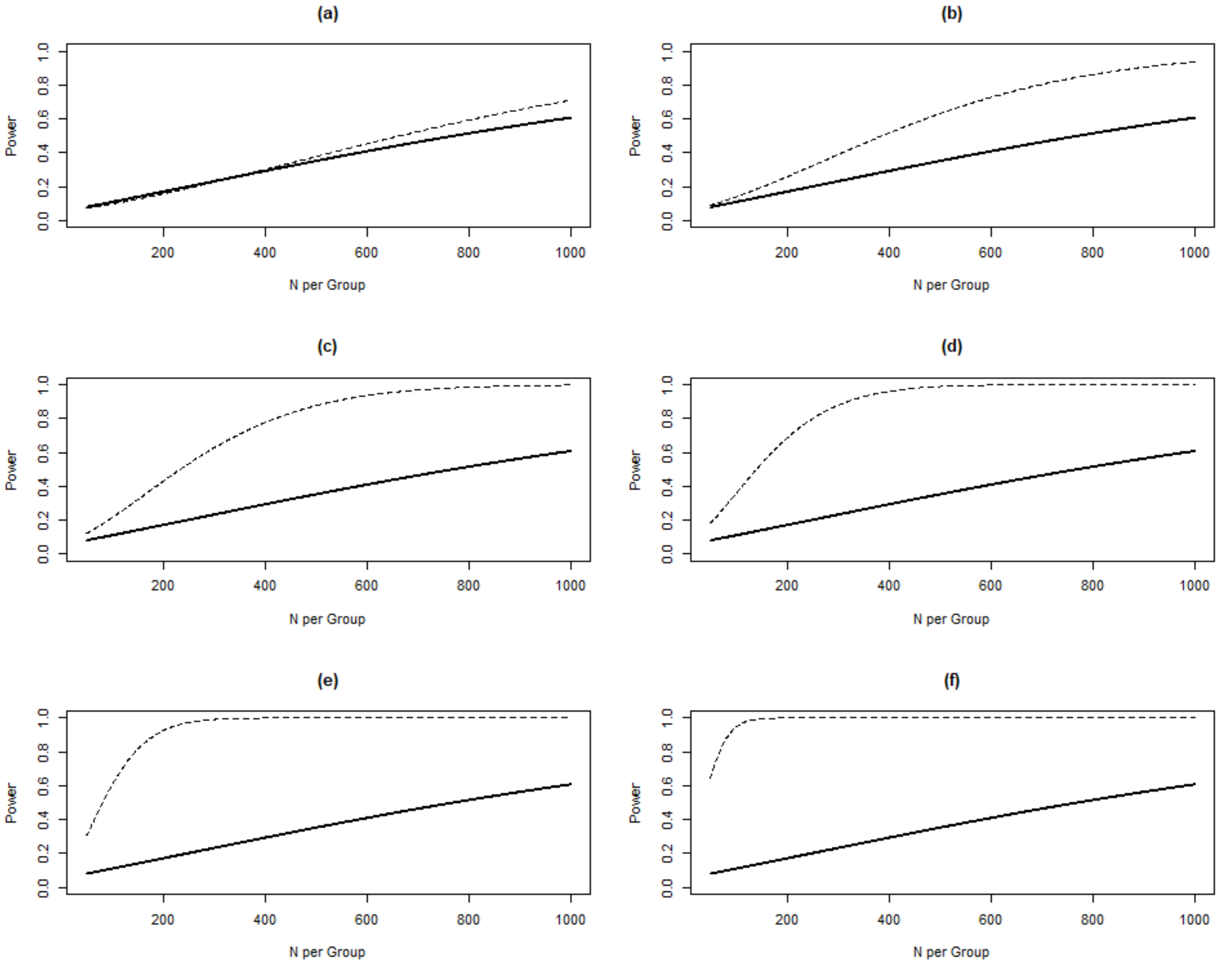

Figure 2.5. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Metric Invariance While Varying Number of Non-Invariant Loadings but Holding Total Non-Invariance Constant.

Figure (a) depicts power to detect 6 non-invariant loadings (+.1 in group 2). Figure (b) depicts power to detect 5 non-invariant loadings (+.12 in group 2). Figure (c) depicts power to detect 4 non-invariant loadings (+.15 in group 2). Figure (d) depicts power to detect 3 non-invariant loadings (+.2 in group 2). Figure (e) depicts power to detect 2 non-invariant loadings (+.3 in group 2). Figure (f) depicts power to detect 1 non-invariant loading (+.6 in group 2). The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance

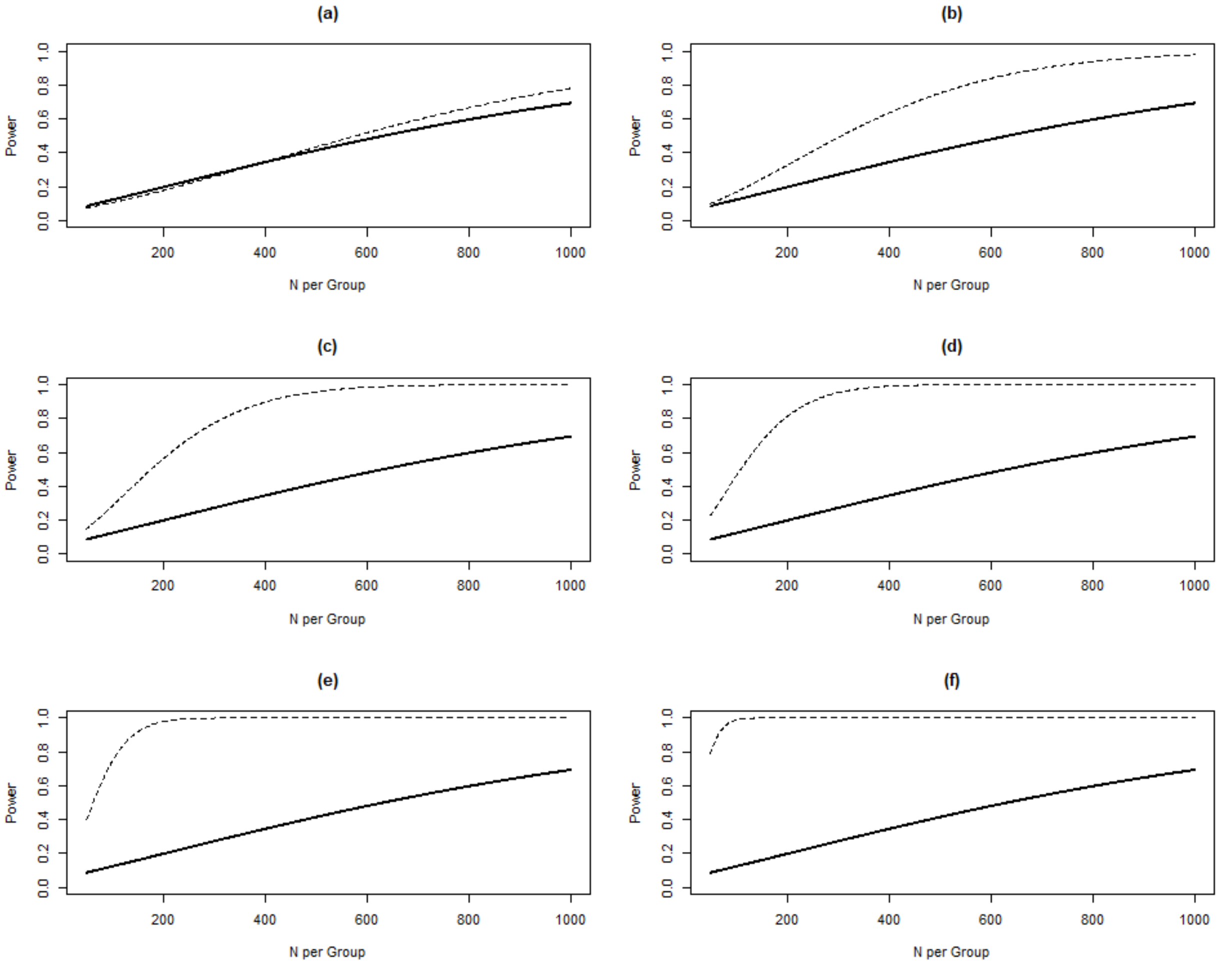

Figure 2.6. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Scalar Invariance While Varying Number of Non-Invariant Intercepts but Holding Total Non-Invariance Constant.

Figure (a) depicts power to detect 6 non-invariant intercepts (+.1 in group 2). Figure (b) depicts power to detect 5 non-invariant intercepts (+.12 in group 2). Figure (c) depicts power to detect 4 non-invariant intercepts (+.15 in group 2). Figure (d) depicts power to detect 3 non-invariant intercepts (+.2 in group 2). Figure (e) depicts power to detect 2 non-invariant intercepts (+.3 in group 2). Figure (f) depicts power to detect 1 non-invariant intercept (+.6 in group 2). The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

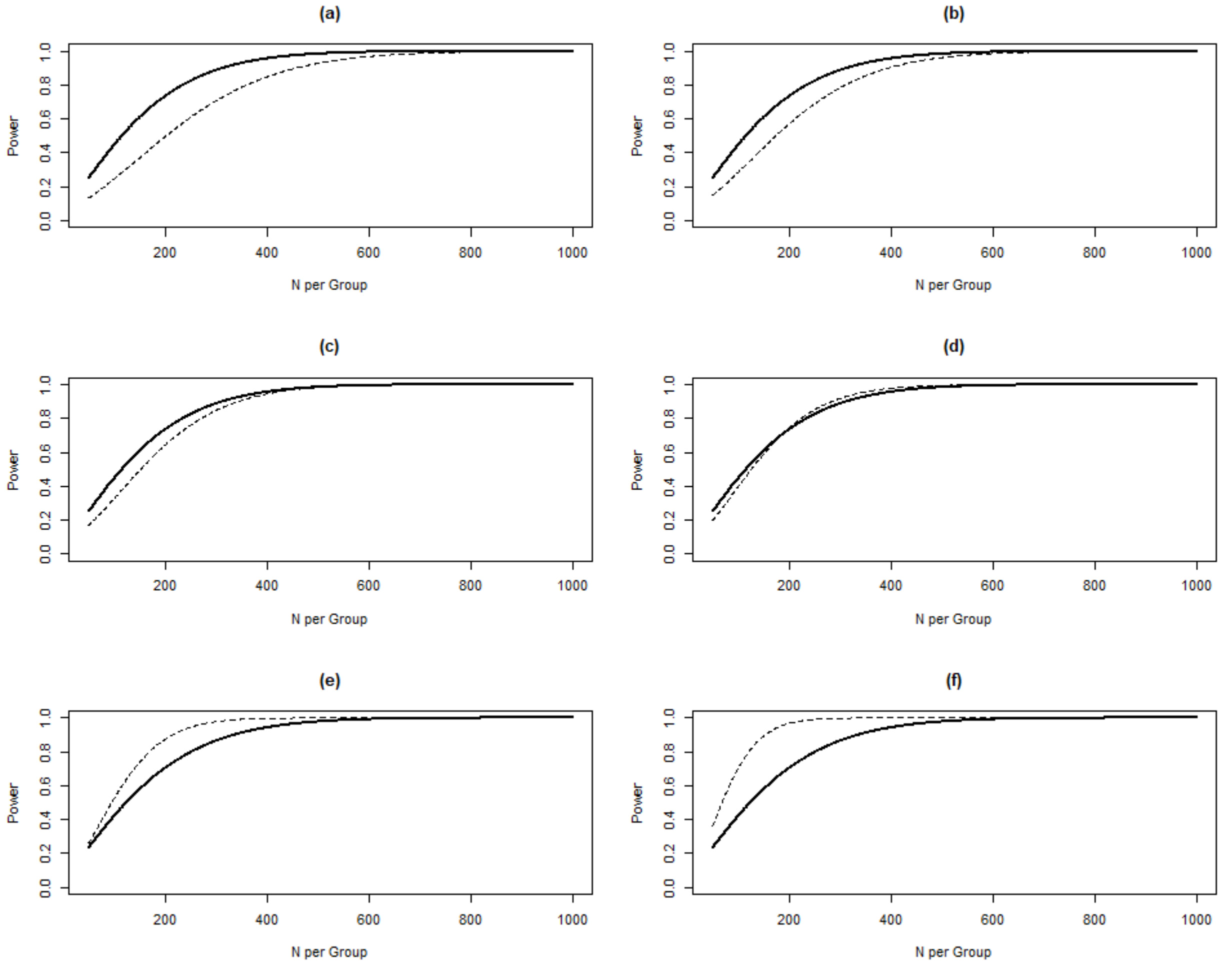

Figure 2.7. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Strict Invariance While Varying Number of Non-Invariant Residual Variances but Holding Total Non-Invariance Constant.

Figure (a) depicts power to detect 6 non-invariant residual variances (+.1 in group 2). Figure (b) depicts power to detect 5 non-invariant residual variances (+.12 in group 2). Figure (c) depicts power to detect 4 non-invariant residual variances (+.15 in group 2). Figure (d) depicts power to detect 3 non-invariant residual variances (+.2 ingroup 2). Figure (e) depicts power to detect2non-invariant residual variances (+.3 in group 2). Figure (f) depicts power to detect 1 non-invariant residual variance (+.6 in group 2). The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

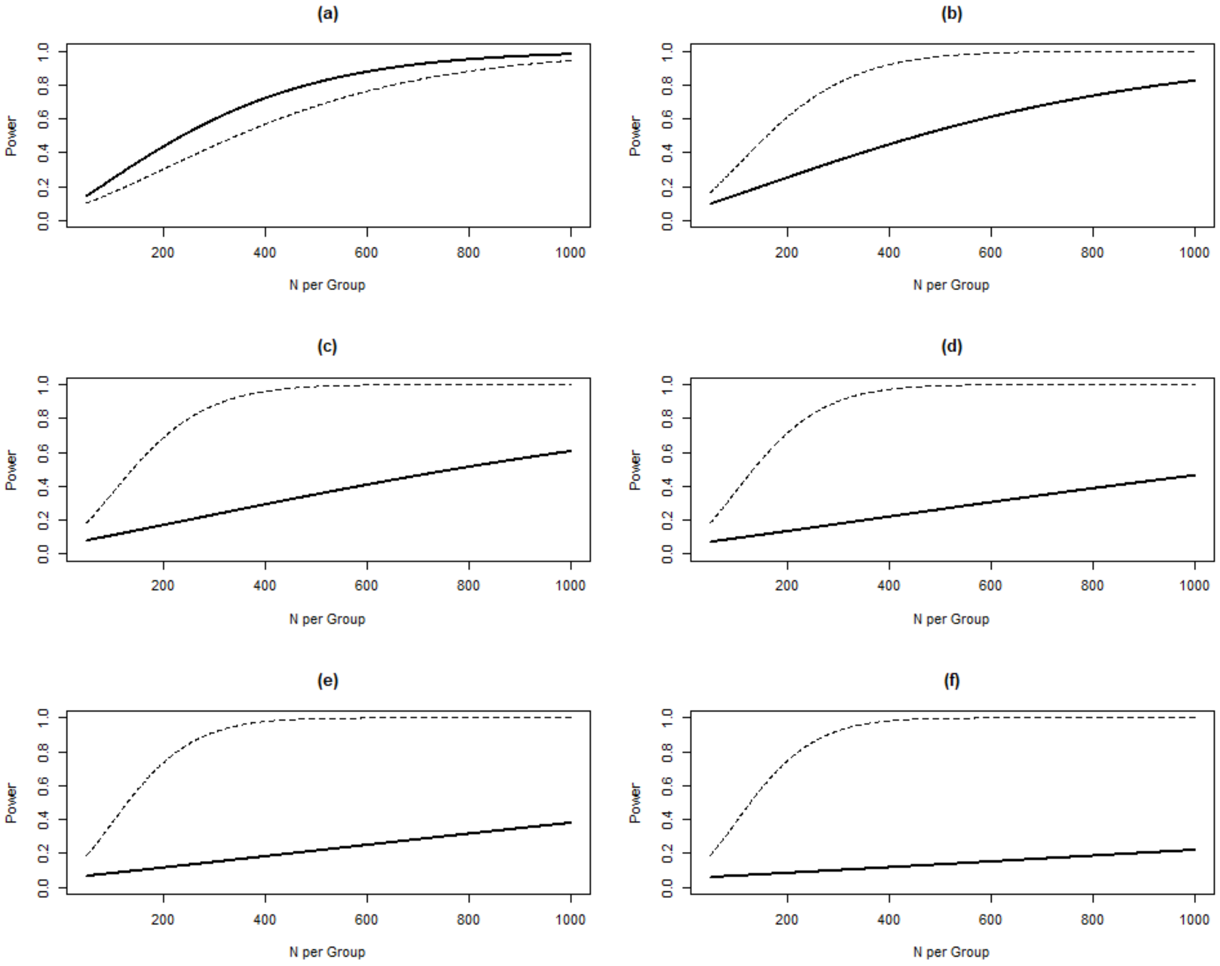

Figure 2.8. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Metric Invariance While Varying Scale Length but Holding Number of Non-Invariant Loadings and Total Non-Invariance Constant.

Figures (a) through (f) depict power of tests of scale-level and item-level metric invariance to detect 3 non-invariant loadings (+.2 in group 2) when scale lengths are p=4, p=6, p=8, p=10, p=12, and p=14, respectively. The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance

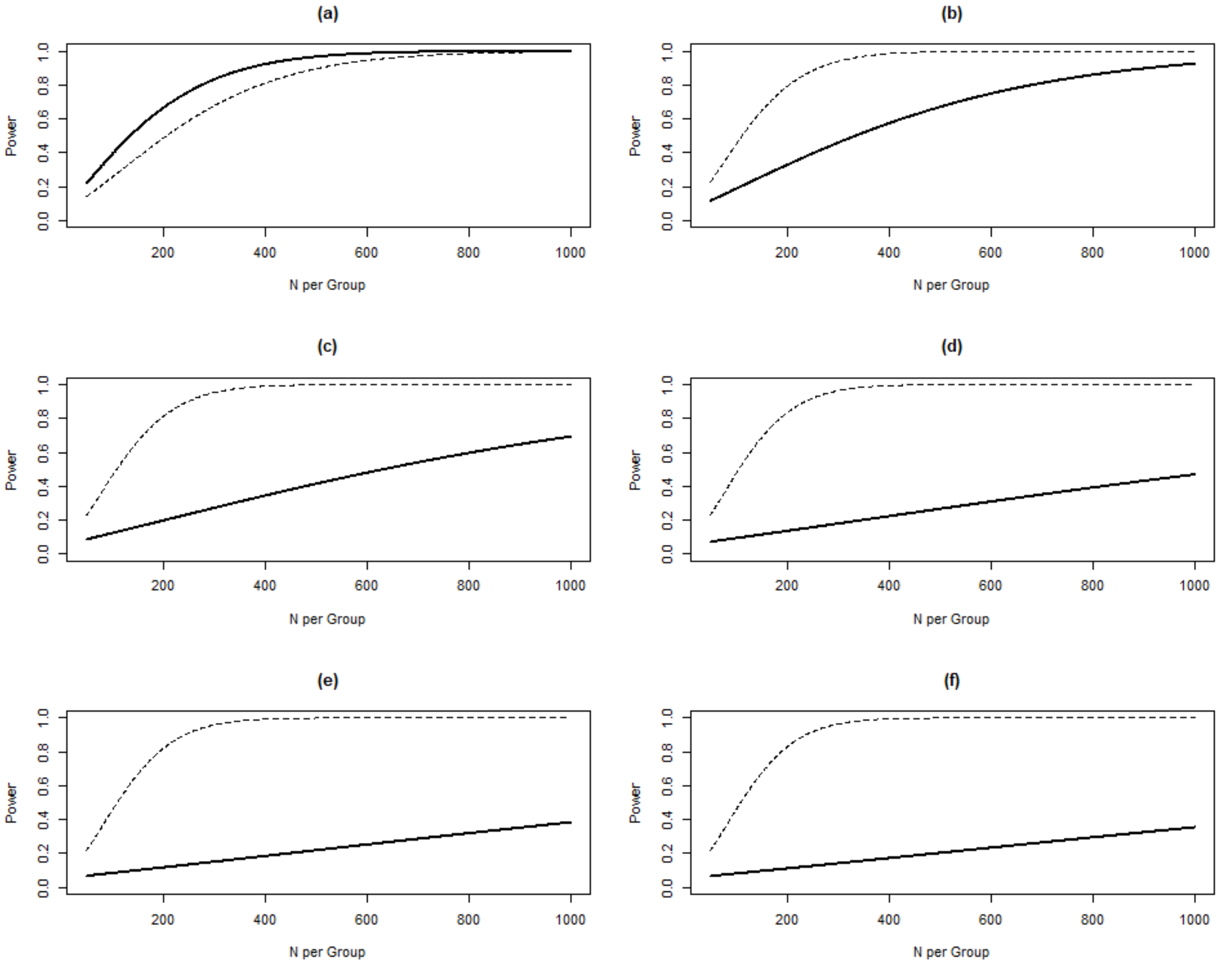

Figure 2.9. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Scalar Invariance While Varying Scale Length but Holding Number of Non-Invariant Intercepts and Total Non-Invariance Constant.

Figures (a) through (f) depict power of tests of scale-level and item-level scalar invariance to detect 3 non-invariant intercepts (+.2 in group 2) when scale lengths are p=4, p=6, p=8, p=10, p=12, and p=14, respectively. X-axis indicate po -level invariance. The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

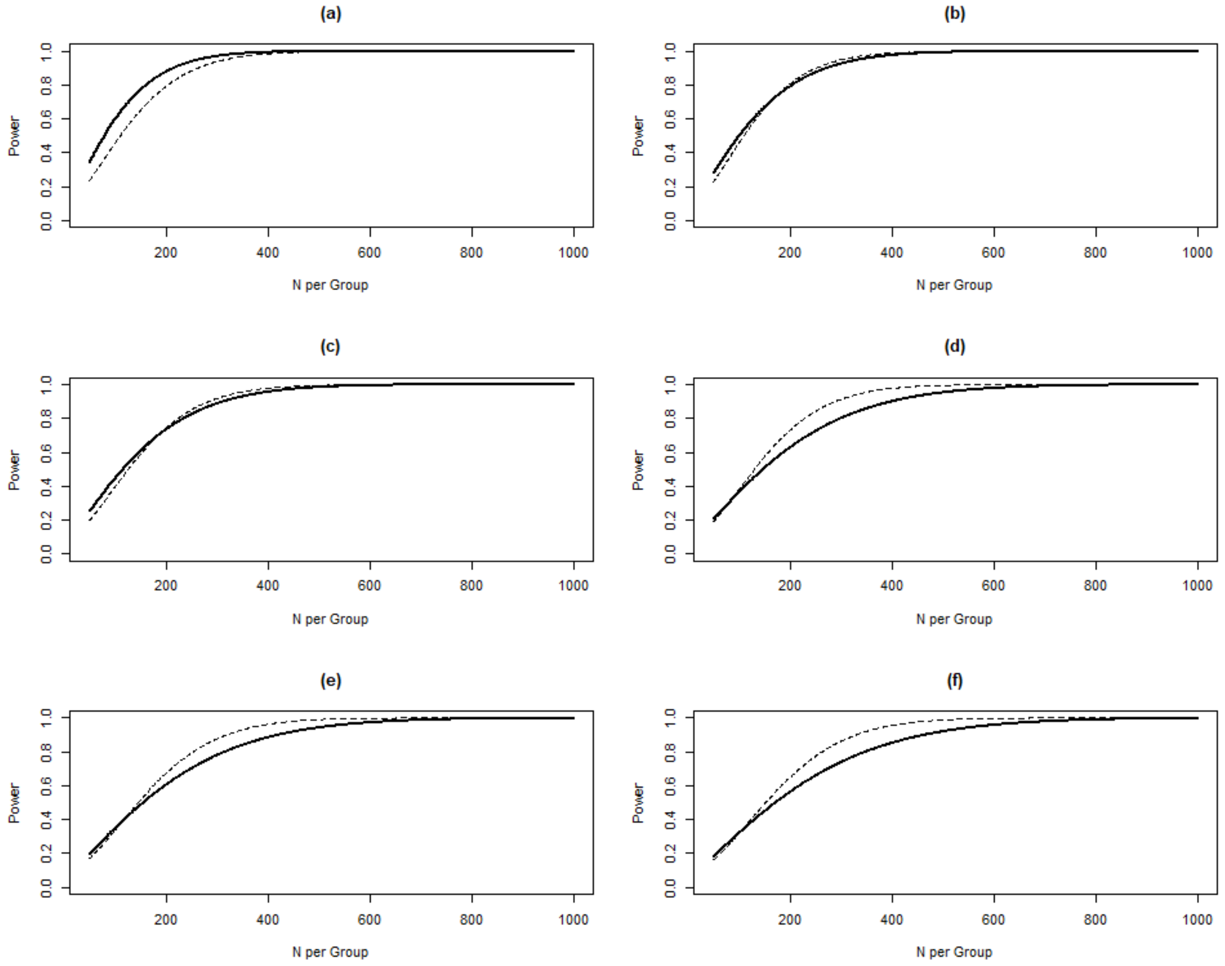

Figure 2.10. Power of Tests of Scale-Level and Item-Level Measurement Invariance to Detect Impactful Violations of Strict Invariance While Varying Scale Length but Holding Number of Non-Invariant Residual Variances and Total Non-Invariance Constant.

Figures (a) through (f) depict power of tests of scale-level and item-level strict invariance to detect 3 non-invariant residual variances (+.2 in group 2) when scale lengths are p=4, p=6, p=8, p=10, p=12, and p=14, respectively. The solid line indicates power of tests of scale-level invariance. The dashed line indicates power of tests of item-level invariance.

Chapter 3: Evaluating the Impact of Violations of Measurement Invariance on Classification Accuracy in Multiple-Group Populations

3.1. Introduction

For applied researchers and clinicians, the most common way to convert participant and client responses to psychometric instruments into observable, interpretable scores is by computing observed composite scores (e.g., Head, Allison, Lucena, Hassenstab, & Morris, 2017; Levant, Alto, McKelvey, Richmond, & McDermott, 2017; McCuish, Mathesius, Lussier, & Corrado, 2017; McDermott et al., 2017; Stevens, Blanchard, Shi, & Littlefield, 2018). Composites are generally computed as $y_{comp} = w'Y$, where w is a $p\ ^{3}1$ vector of item weights – a vector of 1's in the case of unweighted com–and Y is a $p\ ^{3}1$ vector of item scores. Retention of the null hypothesis of strong invariance across populations, as defined in Chapter 1, is generally interpreted as evidence that $E(Y \mid X, g) = E(Y \mid X)$, and retaining strict invariance is interpreted as evidence $P(Y \mid X, g) = P(Y \mid X)$. These conditional dependencies necessarily imply that $E(y_{comp} \mid X, g) = E(y_{comp} \mid X)$, and $P(y_{comp} \mid X, g) = P(y_{comp} \mid X)$, respectively, ensuring comparability of observed composite scores between populations.

When *measurement invariance* (MI) does not hold for all measurement parameters, we say that *violations of measurement invariance* are present. When violations of MI are present, but researchers are still interested in using their measures, they are often encouraged to begin a specification search for the most parsimonious partial MI model (Byrne, Shavelson & Muthen, 1989; Cheung & Rensvold, 1999) consistent with their data. Within an SEM context, partial MI is generally considered sufficient to permit comparative analyses involving latent variables. For example, partial scalar invariance is considered sufficient to permit the between-group comparison of latent means, while partial metric invariance is considered sufficient to permit between-group comparison of correlations involving latent variables (Steenkamp & Baumgartner, 1998). However, for researchers whose evaluations of MI are motivated by a desire to compute valid composite scores independent of group membership, there is little guidance for what should be done subsequent to failure to retain –at the very least –full scalar MI, as partial MI does not necessarily imply comparable observed composite scores (Steinmetz, 2013).

The purpose of the present chapter is to provide guidance to applied researchers and clinicians who wish to use observed composite scores for the purposes of selection/diagnosis, but whose instruments violate MI. When applied researchers and clinicians make use of observed composite scores, they are assuming that observed scores serve as valid proxies for unobservable true scores on the latent construct of interest. When observed composite scores, $y_{comp,i,g}$, for individual *i* in group *g*, are used for the purposes of selection/diagnosis, a participant who exceeds a particular critical threshold, $y_{comp,c}$, on observed scores –for example, the 80th percentile –is, thus, likely to exceed an analogous threshold, χ_c, on the latent construct of interest, i.e., $E(\chi_{i,g} \mid y_{comp,i,g} > y_{comp,c}) > \chi_c$ is assumed. However, when violations of MI are present such that $E(y_{comp,i} \mid \chi_i, g) \neq E(y_{comp,i} \mid \chi_i)$, $E(y_{comp,i} \mid \chi_i = \chi_c, g = R) \neq E(y_{comp,j} \mid \chi_j = \chi_c, g = F)$. To clarify, when violations of scale-level scalar invariance are present in the population, individual *i* in the reference group, *g*=*R*, and individual *j* in the focal group, *g*=*F*, having equal unobservable true scores, such that $\chi_{i,R} = \chi_{j,F}$, does not imply equal expected observed composite scores. Therefore, $P(y_{comp,i} > y_{comp,c} \mid \chi_i > \chi_c, g = R) \neq P(y_{comp,j} > y_{comp,c} \mid \chi_j > \chi_c, g = F)$: the probability of exceeding the critical threshold for selection on observed composite scores, given that one exceeds the critical threshold for selection on unobservable factor scores, is dependent on group membership when violations of MI are present. The present chapter describes a framework for identifying critical percentiles for selection from multiple groups that minimize between-group differences in measures of classification accuracy when violations of MI are present. Measures of classification accuracy of interest include *sensitivity*, $P(y_{comp,i} > y_{comp,c} \mid \chi_i > \chi_c)$, *specificity*, $P(y_{comp,i} < y_{comp,c} \mid \chi_i < \chi_c)$, *positive predictive value*, $P(\chi_i > \chi_c \mid y_{comp,i} > y_{comp,c})$, and *negative predictive value*, $P(\chi_i < \chi_c \mid y_{comp,i} < y_{comp,c})$. Measures of classification accuracy are further defined in *Table 3.1*, and will be discussed in greater detail shortly.

I begin with a summary of research by Millsap and Kwok (2004), who provided the initial framework for estimating classification accuracy in multiple populations with respect to a specified critical percentile when violations of MI are present. I then extend this methodology to a more general case by estimating measures of classification accuracy for all percentiles, and presenting results as plots. Throughout this manuscript, I make use of plots depicting classification accuracy ratios, as well as plots depicting separate estimates of classification accuracy for each group. I demonstrate how population classification accuracy ratio plots vary depending on which measures of classification accuracy are of interest, and which parameters violate MI, illustrating that violations of measurement invariance across multiple parameters can have additive or compensatory effects on between-group differences in classification accuracy. To my knowledge, no other approaches to evaluating MI within a factor analytic framework allow, for example, violations of scalar invariance to compensate for violations of metric invariance. Next, I conduct a series of simulations illustrating the variability of estimated classification accuracy ratio plots as a function of sample size and reliability. I also illustrate how researchers can estimate the finite-sample variability of classification accuracy ratio plots for their own data using a Bollen-Stine (1992) bootstrapping procedure. Finally, I use the proposed method to determine critical percentiles for selection that mitigate discrepancies in classification accuracy across groups in a real data example.

3.1.1. Millsap and Kwok (2004)

Millsap and Kwok (2004) provide a method for estimating classification accuracy about a specified percentile in multiple groups when violations of MI are present. Their three step procedure involves 1) estimating the critical value for selection, ξ_c, in the pooled distribution of unobservable factor scores, $\xi_{i,g}$, 2) estimating the critical value for selection, $y_{comp,c}$, in the pooled distribution of observed composite scores, $y_{comp,i,g}$, and 3) estimating the proportion of the bivariate distribution of $\xi_{i,g}$ and $y_{comp,i,g}$ in each group that falls in each quadrant defined by ξ_c and $y_{comp,c}$ in the population. Note that this methodology assumes that individual item scores and factor scores are normally distributed within each group in the population, but makes no such assumptions about the pooled distributions of observed composite scores and unobservable factor scores.

In step 1, latent mean estimates, κ_g, latent variance estimates, Φ_g, and a population mixing proportion –the proportion of the pooled population accounted for by the reference group –are used to estimate ξ_c via an iterative search procedure. This procedure iteratively computes the percentile associated with a particular value of x separately for each group, and aggregates that information to estim in the pooled population.

This value is then updated until the value associated with the desired percentile is obtained. Note that at least partial scalar invariance must be retained before latent means become estimable (Vandenberg & Lance, 2000). In step 2, the same iterative search procedure is used to estimate $y_{comp,c}$. In this step, model-implied observed composite means and variances are used instead of estimated latent means and variances. Model-implied observed composite means are estimated as $\hat{\mu}_{comp,g} = \hat{\lambda}^{*}_{g}\hat{\kappa}_{g} + \hat{\tau}^{*}_{g}$, where $\hat{\lambda}^{*}_{g} = w'\hat{\Lambda}_{g}$, and $\hat{\tau}^{*}_{g} = w'\hat{\tau}_{g}$ in group *g*. Observed composite variances are estimated as $\hat{\sigma}^{2}_{comp,g} = (\hat{\lambda}^{*}_{g})^{2}\hat{\Phi}_{g} + \hat{\psi}^{*}_{g}$, where $\hat{\psi}^{*}_{g} = w'\hat{\Psi}_{g}w$.

In step 3, $\hat{\kappa}_{g}$, $\hat{\Phi}_{g}$, $\hat{\mu}_{comp,g}$, and $\hat{\sigma}^{2}_{comp,g}$ as well as the estimated correlation between $\xi_{i,g}$ and $y_{comp,i,g}$ in each group, $\hat{\rho}_{y_{comp,i,g}\xi_{i,g}} = \frac{\hat{\lambda}^{*}_{g}\sqrt{\hat{\Phi}_{g}}}{\sqrt{(\hat{\lambda}^{*}_{g})^{2}\hat{\Phi}_{g} + \hat{\psi}^{*}_{g}}}$, are used to estimate the bivariate distribution of $\xi_{i,g}$ and $y_{comp,i,g}$ in each group in the multiple-group population. Subsequently, ξ_{c} and $y_{comp,c}$ are used to define 4 quadrants on the bivariate distributions of $\xi_{i,g}$ and $y_{comp,i,g}$ in each group. *Figure 3.1*, generated using software provided by Lai, Kwok, Yoon and Hsaio (2017), gives a visual representation of the bivariate distributions of $\xi_{i,g}$ and $y_{comp,i,g}$ in each group, as well as the quadrants defined by ξ_{c} and $y_{comp,c}$ over these distributions. The density of the quadrant for which $\xi_{i,g} > \xi_{c}$ and $y_{comp,i,g} > y_{comp,c}$ –Quadrant A in *Figure 3.1* –gives the estimated proportion of *true positives* (*TP*) in a given group when the specified percentile is used for selection. The density of the quadrant for which $\xi_{i,g} < \xi_{c}$ and $y_{comp,i,g} < y_{comp,c}$ –Quadrant C in *Figure 3.1* –gives the estimated proportion of *true negatives* (*TN*) in a given group. The density of the quadrant for which $\xi_{i,g} < \xi_{c}$ and $y_{comp,i,g} > y_{comp,c}$ –Quadrant B in *Figure 3.1* –gives the estimated proportion of *false positives* (*FP*) in a given group. The density of the quadrant for which $\chi_{i,g} > \chi_c$ and $y_{comp,i,g} < y_{comp,c}$ –Quadrant D in *Figure 3.1* –gives the estimated proportion of *false negatives* (*FN*) in a given group. The densities associated with each quadrant of *Figure 3.1* in each group, as well as estimated measures of classification accuracy in each group, are given in *Table 3.2*.

The 2×2 matrix of true positives, true negatives, false positives and false negatives is often referred to as a *confusion matrix*, which serves as input for the computation of measures of classification accuracy. For example, *Sensitivity*, $P(y_{comp,i,g} > y_{comp.c} \mid \chi_{i,g} > \chi_c)$, can be estimated as $\eta_{Sens,g} = \frac{TP_g}{TP_g + FN_g}$, the ratio of true positives to true positives and false negatives. *Specificity*, $P(y_{comp,i,g} < y_{comp.c} \mid \chi_{i,g} < \chi_c)$, can be estimated as $\eta_{Spec,g} = \frac{TN_g}{TN_g + FP_g}$, the ratio of true negatives to true negatives and false positives. While Millsap and Kwok (2004) were primarily interested in examining between-group differences in sensitivity and specificity, there exists a series of 8 simple measures of classification accuracy that condition on each row and column of the confusion matrix, which is presented in *Table 3.1*.

Real Data Illustration. Everson, Millsap and Rodriguez (1991) conducted a multiple-group analysis comparing men (N=219) and women (N=282) on the Test Anxiety Inventory (TAI; Spielberger et al., 1980). Millsap and Kwok (2004) evaluated MI of the 8-item "worry" subscale of this test using Everson et al.'s (1991) data Their most parsimonious partial scalar invariance model –obtained using the method of Byrne, Shavelson, and Muthen (1989) –produced the following total parameter estimates, with female being the reference group: $\hat{\lambda}^*_R = 7.279$, $\hat{\lambda}^*_F = 7.679$, $\hat{\tau}^*_R = 16.621$, $\hat{\tau}^*_F = 16.581$, $\hat{\psi}^*_R = 4.329$,

$\hat{y}^*_F = 3.405$, $\hat{F}^2_R = .544$, $\hat{F}^2_F = .477$, $k_R = 0$, and $\hat{k}_F = -0.126$. Millsap and Kwok then used their method to evaluate sensitivity at the 90th percentile, assuming an equal mixture of men and women in the population. A replication of this analysis is presented in the left panels of *Figure 3.1* and *Table 3.2*. Millsap and Kwok (2004) did not explicitly report their final x_c and $y_{comp,c}$, but estimated $\hat{n}_{Sens,R} = .743$ and $\hat{n}_{Sens,F} = .775$. My replication estimates $x_c = .857$, $y_{comp,c} = 23.457$, $\hat{n}_{Sens,R} = .742$, and $\hat{n}_{Sens,F} = .774$.

After estimating sensitivity in both popul

whether observed differences are problematic, and whether they are a function of violations of MI. Millsap and Kwok (2004) conducted a follow-up analysis using the same procedure, but with the added assumption that all measurement parameters –loadings, intercepts, and residual variances –are MI, while still allowing latent means and latent variances to differ. This assumption is made by selecting a single value for each measurement parameter across groups, and by extension, a single value for l^*, t^* and y^* across groups. For measurement parameters constrained to equality in the retained partial invariance model, the constrained estimate is treated as the parameter value in both populations. For measurement parameters where the null hypothesis of invariance is not tenable, the sample-weighted mean of parameter estimates is treated as the parameter estimate in both groups, such that $\hat{q} = \frac{\hat{q}_R N_R + \hat{q}_F N_F}{N_R + N_F}$, where $\hat{q}_R$ and $\hat{q}_F$ are estimates for any non-invariant parameter taken from the partial measurement invariance model output in the reference group and focal group, respectively. Millsap and Kwok (2004) argue that if estimated sensitivities observed under violations of MI are largely unchanged under assumed MI, between-group differences in estimated measures of classification accuracy cannot be attributed to violations of MI.

Under the assumption of measurement invariance, estimated measurement parameter totals are $\hat{l}^* = 7.454$, $\hat{t}^* = 16.604$, and $\hat{y}^* = 3.925$ in both groups, while latent mean and latent variance estimates are unchanged, $k_R = 0$, $\hat{k}_F = -.126$, $\hat{F}_R = .544$, and $\hat{F}_F = .477$. A replication of Millsap and Kwok's analysis based on these parameter estimates is presented in the right panels of *Figure 3.1* and *Table 3.2*. Again, Millsap and Kwok (2004) did not explicitly report x_c and $y_{comp,c}$ in their manuscript, however my replication estimates $x_c = .857$ and $y_{comp,c} = 23.447$. When MI is assumed, Millsap and Kwok (2004) estimated $\hat{n}_{Sens,R} = .767$ and $\hat{n}_{Sens,F} = .739$ at the 90th percentile. My replication produces similar results, with $\hat{n}_{Sens,R} = .767$ and $\hat{n}_{Sens,F} = .738$. The researchers argue that these observed changes – a .024 increase in $\hat{v}_{Sens,R}$, and a .036 decrease in $\hat{v}_{Sens,F}$ – are reasonably small, and by extension, conclude that group differences in factor structure do not meaningfully reduce classification accuracy. It is worth noting that between-group differences on latent means and latent variances can lead to unequal measures of classification accuracy across populations, even when no violations of MI are present. Bivariate density plots when MI is assumed are presented in the right panel of *Figure 3.1*. *Table 3.2* gives estimated quadrant densities and estimated measures of classification accuracy in each group with and without the assumption of MI.

Despite its merits, the method of Millsap and Kwok (2004) has seen little use in the behavioral sciences. While similar methods exist in the regression and item response theory literature, such as analyses of differential validity and differential prediction (Drasgow & Kang, 1984; Gresham, MacMillan & Bocian, 1987), and analyses of differential functioning of items and tests (Raju, van der Linden & Fleer, 1995 tie differential classification accuracy to latent variable models. While frequently cited, Millsap and Kwok's (2004) paper is generally invoked group differences in psychometric instrument functioning in manuscripts that then go on to assess invariance in a more conventional manner. Lai, Kwok, Yoon and Hsaio (2017) argue that one reason the method has seen little use among applied researchers is because software that made it straightforward to implement was never developed. While Millsap and Kwok (2004) did include S-PLUS syntax –a precursor to R –in an appendix, this syntax was limited in that it did not illustrate their complete routine, as well as the fact that it made use of Danish-language software which is no longer hosted online (Holst, Jorgensen, & Natolski, 2001). In response to this issue, Lai et al. (2017) developed software in R which implements the methodology of Millsap and Kwok (2004), taking parameter estimates, a critical percentile, and an estimated mixing proportion in the pooled population as inputs, and giving confusion matrices and measures of classification accuracy in each population as output, as well as a graphical representation of the bivariate sampling distributions of $\chi_{i,g}$ and $y_{comp,i,g}$ in each group. *Figure 3.1* and *Table 3.2* were generated using this software. See *Appendix C* for Lai et al.'s software, and *Appendix D* for sample code illustrating its use to generate *Figure 3.1* and *Table 3.2*.

3.1.2. Extending Millsap and Kwok (2004) to a More General Case

The purpose of the present research is to extend the method of Millsap and Kwok (2004) to a more general case. Rather than require researchers to specify a single candidate critical value about which to estimate classification accuracy, measures of classification accuracy are estimated for every percentile from 1 to 99, and results are presented as plots. The measures of classification accuracy of greatest interest –taken from *Table 3.1* –of are sensitivity, specificity, positive predictive value, and negative predictive value. Sensitivity is computed as

$$\eta_{Sens} = \frac{TP}{TP + FN}, \tag{3.1}$$

which gives $P(y_{comp,i} > y_{comp,c} \mid \chi_i > \chi_c)$, the probability of exceeding the threshold for selection on observed composite scores given that one exceeds the threshold for selection on unobservable factor scores. Specificity is computed as

$$\eta_{Spec} = \frac{TN}{TN + FP}, \tag{3.2}$$

which gives $P(y_{comp,i} < y_{comp,c} \mid \chi_i < \chi_c)$, the probability of falling below the threshold for selection on observed composite scores given that one falls below the threshold for selection on unobservable factor scores. Positive predictive value (PPV) is computed as

$$\eta_{PPV} = \frac{TP}{TP + FP}, \tag{3.3}$$

Which gives $P(\chi_i > \chi_c \mid y_{comp,i} > y_{comp,c})$, the probability of exceeding the threshold for selection on unobservable factor scores given that one exceeds the threshold for selection on observed composite scores. Negative predictive value is computed as

$$\eta_{NPV} = \frac{TN}{TN + FN}, \tag{3.4}$$

which gives $P(\chi_i < \chi_c \mid y_{comp,i} < y_{comp,c})$, the probability of falling below the threshold for selection on unobservable factor scores given that one falls below the threshold for selection on observed composite scores.

Sensitivity and positive predictive value are both measures of classification accuracy concerned with identifying participants *above* a given threshold. If classification is perfect, $\eta_{Spec} = \eta_{PPV} = 1$. If classification is not perfect, η_{Sens} decreases with increasing number of false negatives, while η_{PPV} decreases with increasing number of false positives. Similarly, specificity and negative predictive value are concerned with identifying participants *below* a given threshold, with η_{Spec} decreasing with increasing false positives, and η_{NPV} decreasing with increasing false negatives. Thus, researchers can determine which measures of classification accuracy are of greatest interest to them based on 1) whether they are interested in identifying individuals who are particularly high or particularly low on their construct of interest, and 2) whether they are more interested controlling for false positives or false negatives.

When presenting measures of classification accuracy as plots, one can examine measures of classification accuracy separately for each group, or examine *classification accuracy ratios*. Ratios of estimated measures of classification accuracy, $\iota = \frac{\eta_R}{\eta_F}$ –where η_R and η_F are a given measure of classification accuracy in the reference and focal group, respectively –are useful because they give the multiplicative increase in the probability of a particular decision for someone who in in the reference group rather than the focal group, all else being equal. For example, if $\eta_{Sens,R} = .9$ and $\eta_{Sens,F} = .75$, $\iota_{Sens} = 1.2$, meaning $P(y_{comp,i} > y_{comp,c} \mid x_i > x_c, g = R) = 1.2 P(y_{comp,j} > y_{comp,c} \mid x_j > x_c, g = F)$: the probability of exceeding $y_{comp,c}$ for someone who exceeds x_c is 1.2 times greater in the reference group than in the focal group. If $\eta_{PPV,R} = .9$ and $\eta_{PPV,F} = .75$, $\iota_{PPV} = 1.2$, meaning $P(x_i > x_c \mid y_{comp,i} > y_{comp,c}, g = R) = 1.2 * P(x_j > x_c \mid y_{comp,j} > y_{comp,c}, g = F)$: the probability of exceeding χ_c for someone who exceeds $y_{comp,c}$ is 1.2 times greater in the reference group than in the focal group.

When presenting classification accuracy ratio plots, I recommend plotting $\log_{10}(i)$, rather than untransformed i, against critical percentiles. This can be justified by examining *Figure 3.2*. When i is plotted against critical percentiles, the magnitude of the departure from a flat line at $i=1.0$ is greater when the favored group –the group with the larger estimated η –is the reference group rather than the focal group. Panels a) and b) of *Figure 3.2* give the untransformed i_{Sens} plots when a group with higher loadings is the focal group and the reference group, respectively. Panels c) and d) of *Figure 3.2* give the $\log_{10}(i_{Sens})$ plots when a group with higher loadings is the focal group and the reference group, respectively. We see that the absolute magnitude of deviations from $\log_{10}(i)=0$ shown in *Figure 3.2(c)* are identical to those shown in *Figure 3.2(d)*, as $|\log_{10}(\frac{\eta_R}{\eta_F})-0|=|\log_{10}(\frac{\eta_F}{\eta_R})-0|$ for all possible values of η_R and η_F. Alternatively, the absolute magnitude of deviations from $i=1$ shown in *Figure 3.2(a)* are consistently greater than those shown in *Figure 3.2(b)*, as $|\frac{\eta_R}{\eta_F}-1|>|\frac{\eta_F}{\eta_R}-1|$ when $\eta_R > \eta_F$. Thus, it is plausible that applied researchers may be more likely to interpret a given i as sufficiently close to 1.0 when the favored group is the focal group rather than the reference group. No such issue is present when $\log_{10}(i)$ is presented on the *y*-axis. A reference to facilitate conversions between major $\log_{10}(i)$ values and i values, and vice-versa, is given in *Table 3.3.*

For the sake of proposing benchmarks, let us define violations of measurement invariance for which $-.05 \le \log_{10}(i) \le .05$ at a given percentile as having a *minimal impact* on classification accuracy ratios. This range implies $.891 < i \leq 1.122$. Thus, when $-.05 \leq \log_{10}(i) \leq .05$, the increased probability of selection in the favored group relative to the other is, at most, 12%. Let us define violations of MI for which $-.1 \leq \log_{10}(i) \leq .1$ but not $-.05 \leq \log_{10}(i) \leq .05$ as having a *modest impact* on classification accuracy ratios. This range implies $.794 < i \leq 1.259$. Thus, when $-.1 \leq \log_{10}(i) \leq .1$, the increased probability of selection in the favored group relative to the other is, at most, 25%. Finally, let us define violations of measurement invariance for which $\log_{10}(i)$ falls outside of both discussed ranges as having a *large impact* on classification accuracy ratios. See *Appendix E* for software which generates $\log_{10}(i)$ plots from partial measurement invariance model parameter estimates.

3.2. Illustrating the Behavior of Classification Accuracy Ratio Plots When Violations of Measurement Invariance are Present and Population Parameters are Known

The purpose of the following section is to illustrate how $\log_{10}(i)$ plots vary as a function of the location of violations of measurement invariance, and how these patterns vary as a function of which measures of classification accuracy are of interest. For each of four measures of classification accuracy –sensitivity, specificity, PPV, and NPV –a single figure is generated which includes separate population $\log_{10}(i)$ plots for each of 5 multiple-group population conditions for which violations of MI are present. Population conditions include one where only violations of metric invariance are present, one where only violations of scalar invariance are present, one where only violations of strict invariance are present, and two where violations of both metric and scalar invariance are present. In all population conditions, $l^*_R = 4.2$, $t^*_R = 0$,

$y_R^* = 3.06$, $\Phi_R = 1$, and $\kappa_R = 0$, with a population mixing proportion of .5. Population parameters for the focal group in each population condition are given in *Table 3.4*.

Sensitivity. Figure 3.3 presents $\log_{10}(i_{Sens})$ plots for each of the five measurement non-invariant multiple-group population conditions given in *Table 3.4*. In condition 1, where factor loadings are greater in the focal group, sensitivity ratios decrease with increasing critical percentile, meaning $P(y_{comp,j} > y_{comp,c} \mid \xi_j > \xi_c, g = F) > P(y_{comp,i} > y_{comp,c} \mid \xi_i > \xi_c, g = R)$. Conversely, greater factor loadings in the reference group implies $P(y_{comp,i} > y_{comp,c} \mid \xi_i > \xi_c, g = R) > P(y_{comp,j} > y_{comp,c} \mid \xi_j > \xi_c, g = F)$. In condition 2, where intercepts are greater in the focal group, sensitivity ratios decrease with increasing critical percentile, meaning $P(y_{comp,j} > y_{comp,c} \mid \xi_j > \xi_c, g = F) > P(y_{comp,i} > y_{comp,c} \mid \xi_i > \xi_c, g = R)$. Conversely, greater intercepts in the reference group implies $P(y_{comp,i} > y_{comp,c} \mid \xi_i > \xi_c, g = R) > P(y_{comp,j} > y_{comp,c} \mid \xi_j > \xi_c, g = F)$. In condition 3, where residual variances are greater in the focal group, sensitivity is largely equal across groups at all percentiles.

In condition 4, where both loadings and intercepts are greater in the focal group, sensitivity ratios decrease with increasing critical percentile. Notably, $\log_{10}(i_{Sens})$ in this condition is more extreme than in conditions 1 or 2, where only loadings or only intercepts are greater in the focal group, respectively. This suggests that the present violations of metric and scalar invariance have an additive effect on sensitivity ratios, resulting in sensitivity ratios that strongly favor the focal group. This is consistent with the fact that these violations of metric and scalar invariance, separately, both result in greater sensitivity in the focal group. In condition 5, where loadings are greater in the focal group and intercepts are greater in the reference group, sensitivity ratios, again, decrease with increasing critical percentile. In this condition, $\log_{10}(j_{Sens})$ generally falls closer to zero than in conditions 1 or 2, where only loadings and only intercepts violate MI, respectively. This suggests that the present violations of metric and scalar invariance have a compensatory effect on sensitivity ratios, resulting in sensitivity ratios that favor the focal group to a lesser extent. This is consistent with the fact that these violations of metric and scalar invariance, separately, result in greater sensitivity in the focal group and the reference group, respectively. Examining the 90th percentile of *Figure 3.3*, if one recalls the previously defined benchmarks, it can be seen that conditions 1 and conditions 2 both have a modest impact on $\log_{10}(j_{Sens})$, while condition 4 has a large impact and condition 5 has a minimal impact on $\log_{10}(j_{Sens})$. The impact of different locations and directions of violations of measurement invariance on sensitivity ratios is summarized in *Table 3.5*. Interestingly, varying the population mixing proportion such that the reference group is four times larger than the focal group negligibly affects sensitivity ratios.

Specificity. *Figure 3.4* gives population specificity ratio plots for each of the five measurement non-invariant multiple-group population conditions given in *Table 3.4*. In condition 1, where factor loadings are greater in the focal group, specificity ratios decrease with decreasing critical percentile, meaning $P(y_{comp,j} < y_{comp,c} \mid x_j < x_c, g = F) > P(y_{comp,i} < y_{comp,c} \mid x_i < x_c, g = R)$. By extension, greater factor loadings in the reference group implies $P(y_{comp,i} < y_{comp,c} \mid x_i < x_c, g = R) > P(y_{comp,j} < y_{comp,c} \mid x_j < x_c, g = F)$. In condition 2, where intercepts are greater in the focal group, specificity ratios increase with decreasing critical percentile, meaning $P(y_{comp,i} < y_{comp,c} \mid x_i < x_c, g = R) > P(y_{comp,j} < y_{comp,c} \mid x_j < x_c, g = F)$.

Conversely, greater intercepts in the reference group implies $P(y_{comp,j} < y_{comp,c} \mid x_j < x_c, g = F) > P(y_{comp,i} < y_{comp,c} \mid x_i < x_c, g = R)$. In condition 3, where residual variances are greater in the focal group, specificity is largely equal across groups at all percentiles.

In condition 4, where both loadings and intercepts are greater in the focal group, specificity ratios decrease with decreasing critical percentile. In this condition, unlike with sensitivity ratios, $\log_{10}(i_{Spec})$ generally falls closer to 0 than it did in conditions 1 or 2. This suggests that the present violations of metric and scalar invariance have a compensatory effect on specificity ratios, resulting in specificity ratios that favor the focal group to a lesser extent. This is consistent with the fact that the present violations of metric and scalar invariance, separately, result in greater specificity in the focal and reference group, respectively. In condition 5, where loadings are greater in the focal group and intercepts are greater in the reference group, sensitivity ratios decrease with increasing critical percentile. Unlike in condition 4, $\log_{10}(i_{Spec})$ is generally more extreme in condition 5 than in conditions 1 or 2, where only violations of metric invariance and only violations of scalar invariance are present, respectively. This suggests that the present violations of metric and scalar invariance have an additive effect on specificity ratios, resulting in specificity ratios that strongly favor the focal group. This is consistent with the fact that these violations of metric and scalar invariance, separately, both result in greater specificity in the focal group. Examining the 90th percentile of *Figure 3.4*, one can see that conditions 1 and conditions 2 both have a modest impact on $\log_{10}(i_{Spec})$, while condition 4 has a minimal impact and condition 5 has a large impact on $\log_{10}(i_{Spec})$. The impact of different locations and directions of violations of MI on specificity ratios is summarized in *Table 3.5*. As was the case with sensitivity ratios, varying the population mixing proportion such that the reference group is four times larger than the focal group minimally affects specificity ratios.

Positive Predictive Value (PPV). Figure 3.5 gives population PPV ratio plots for each of the five measurement non-invariant multiple-group population conditions given in *Table 3.4*. In condition 1, where factor loadings are greater in the focal group, PPV ratios increase with increasing critical percentile, meaning $P(x_i > x_c | y_{comp,i} > y_{comp,c}, g = R) > P(x_j > x_c | y_{comp,j} > y_{comp,c}, g = F)$. Conversely, greater factor loadings in the reference group implies $P(x_j > x_c | y_{comp,j} > y_{comp,c}, g = F) > P(x_i > x_c | y_{comp,i} > y_{comp,c}, g = R)$. In condition 2, where intercepts are greater in the focal group, PPV ratios increase with increasing critical percentile, meaning $P(x_i > x_c | y_{comp,i} > y_{comp,c}, g = R) > P(x_j > x_c | y_{comp,j} > y_{comp,c}, g = F)$. Conversely, greater intercepts in the reference group implies $P(x_j > x_c | y_{comp,j} > y_{comp,c}, g = F) > P(x_i > x_c | y_{comp,i} > y_{comp,c}, g = R)$. In condition 3, where residual variances are greater in the focal group, PPV ratios increase with increasing critical percentile, meaning $P(x_i > x_c | y_{comp,i} > y_{comp,c}, g = R) > P(x_j > x_c | y_{comp,j} > y_{comp,c}, g = F)$. Conversely, greater residual variances in the reference group implies $P(x_j > x_c | y_{comp,j} > y_{comp,c}, g = F) > P(x_i > x_c | y_{comp,i} > y_{comp,c}, g = R)$.

In condition 4, where both loadings and intercepts are greater in the focal group, PPV ratios increase with increasing critical percentile. Notably, $\log_{10}(\hat{\phi}_{PPV})$ in this condition is generally more extreme than in conditions 1 or 2, where only loadings and only intercepts favor the reference group, respectively. This suggests that the present violations of metric and scalar invariance have an additive effect on PPV ratios, resulting in PPV ratios that strongly favor the reference group. This is consistent with the fact that these violations of metric and scalar invariance, separately, both result in greater PPV in the reference group. In condition 5, where loadings are greater in the focal group and intercepts are greater in the reference group, PPV ratios, again, increase with increasing critical percentile. In this condition, $\log_{10}(\phi_{PPV})$ generally falls closer to zero than in conditions 1 or 2, where only loadings and only intercepts violate MI, respectively. This suggests that the present violations of metric and scalar invariance have a compensatory effect on PPV ratios, resulting in PPV ratios that favor the reference group to a lesser extent. This is consistent with the fact that these violations of metric and scalar invariance, separately, both result in greater PPV in the reference group. Examining the 90th percentile of *Figure 3.5*, we see that conditions 1 and conditions 2 both have a minimal impact on $\log_{10}(\phi_{PPV})$, while condition 4 has a modest impact and condition 5 has a minimal impact on $\log_{10}(\phi_{PPV})$. The impact of different locations and directions of violations of MI on PPV ratios is summarized in *Table 3.5*. As is the case with previously discussed measures of classification accuracy, varying the population mixing proportion such that the reference group is four times larger than the focal group minimally affects PPV ratios.

Negative Predictive Value (NPV). *Figure 3.6* gives population NPV ratio plots for each of the five measurement non-invariant multiple-group populations conditions given in *Table 3.4*. In condition 1, where factor loadings are greater in the focal group, NPV ratios increase with decreasing critical percentile, meaning

$P(X_i < X_c \mid y_{comp,i} \le y_{comp,c}, g = R) > P(X_j < X_c \mid y_{comp,j} \le y_{comp,c}, g = F)$. Conversely, greater factor loadings in the reference group implies

$P(X_j < X_c \mid y_{comp,j} \le y_{comp,c}, g = F) > P(X_i < X_c \mid y_{comp,i} \le y_{comp,c}, g = R)$. In condition 2, where intercepts are greater in the focal group, NPV ratios decrease with decreasing critical percentile, meaning $P(X_j < x_c \mid y_{comp,j} \le y_{comp,c}, g = F) > P(X_i < x_c \mid y_{comp,i} \le y_{comp,c}, g = R)$. Conversely, greater intercepts in the reference group implies $P(X_i < x_c \mid y_{comp,i} \le y_{comp,c}, g = R) > P(X_j < x_c \mid y_{comp,j} \le y_{comp,c}, g = F)$. In condition 3, where residual variances are greater in the focal group, NPV ratios increase with decreasing critical percentile, meaning $P(X_i < x_c \mid y_{comp,i} \le y_{comp,c}, g = R) > P(X_j < x_c \mid y_{comp,j} \le y_{comp,c}, g = F)$. Conversely, greater residual variances in the reference group implies $P(X_j < x_c \mid y_{comp,j} \le y_{comp,c}, g = F) > P(X_i < x_c \mid y_{comp,i} \le y_{comp,c}, g = R)$.

In condition 4, where both loadings and intercepts are greater in the focal group, NPV ratios increase with decreasing critical percentile. In this condition, $\log_{10}(\lambda_{NPV})$ generally falls closer to zero than in conditions 1 or 2, where only loadings and only intercepts violate MI, respectively. This suggests that the present violations of metric and scalar invariance have a compensatory effect on NPV ratios, resulting in NPV ratios that favor the reference group to a lesser extent. This is consistent with the fact that the present violations of metric and scalar invariance, separately, result in greater NPV in the reference and focal group, respectively. In condition 5, where loadings are greater in the focal group and intercepts are greater in the reference group, NPV ratios decrease with decreasing critical percentile. Unlike in condition 4, $\log_{10}(\lambda_{NPV})$ is generally more extreme than in conditions 1 or 2, where only violations of metric invariance and only violations of scalar invariance are present, respectively. This suggests that the present violations of metric and scalar invariance have an additive effect on NPV ratios, resulting in specificity ratios that strongly favor the focal group. This is consistent with the fact that these violations of metric and scalar invariance, separately, both result in greater NPV in the reference group. Examining the 90th percentile of *Figure 3.6*, we see that conditions 1 and 2 both have a minimal impact on $\log_{10}(j_{NPV})$, while condition 4 has a minimal impact and condition 5 has a modest impact on $\log_{10}(j_{NPV})$. The impact of different locations and directions of violations of MI on NPV ratios is summarized in *Table 3.5*. As is the case with previously discussed measures of classification accuracy, varying the population mixing proportion such that the reference group is four times larger than the focal group minimally affects NPV ratios.

Summary. When violations of metric invariance are present, sensitivity and specificity ratios favor the group with greater loadings, while PPV and NPV ratios favor the group with lower loadings. When violations of scalar invariance are present, sensitivity and NPV ratios favor the group with greater intercepts, while specificity and PPV ratios favor the group with lower intercepts. When violations of strict invariance are present, PPV and NPV ratios favor the group with lower residual variances, while sensitivity and specificity ratios are largely unaffected. See *Table 3.5* for a summary of the impacts different violations of measurement invariance have on different classification accuracy ratios. Overall, we have illustrated how the locations and directions of violations of MI differentially impact different measures of classification accuracy, allowing violations of MI at multiple locations to have additive or compensatory effects on classification accuracy ratios depending on the nature of those violations. See *Appendix F* for sample code which generates *Figure 3.3* using the software provided in *Appendix E*.

3.3. Simulation: Illustrating the Variability of Classification Accuracy Ratio Plots in Finite Samples When Violations of Measurement Invariance are Present.

Section 3.2 of this chapter illustrates how $\log_{10}(i)$ plots vary across multiple-group populations featuring different violations of MI. In practice, however, applied researchers never have access to exact population values for measurement parameters. Rather, they have parameter estimates based on fitting partial MI models to finite samples, which necessarily vary from sample to sample. The purpose of the present section is to illustrate the degree to which estimated classification accuracy ratio plots vary as a function of sample size and reliability. This is done by generating a large number of finite samples from the multiple-group population models given in *Table 3.6*, fitting the correct partial invariance model to each dataset, using parameter estimates from each replication to estimate classification accuracy ratios at each percentile for each replication, and finally presenting the empirically estimated confidence interval around the population $\log_{10}(i)$ plot. For simplicity, only the variability of $\log_{10}(i_{Sens})$ plots are examined in this simulation. Population models used in data generation are now discussed.

3.3.1. Population Models and Data Generation

Multivariate normal data are generated in R from each of four multiple-group population models: two based on condition 4, and two based on condition 5 in *Table 3.4*. Recall that in condition 4, violations of both metric and scalar invariance are present, with both loadings and intercepts greater in the focal group. In condition 5, again, violations of both metric and scalar invariance are present, but with loadings greater in the focal group and intercepts greater in the reference group. The two population models associated with each condition have moderate and high reliability. Within each population model, four different sample sizes are used in data generation, for a total of 16 plots, which are presented in *Figure 3.7* through *Figure 3.10*.

Multiple-group population models based on condition 4 from *Table 3.4* have 1 factor with 6 indicator variables. In the reference group, $\Lambda_R = [.7,.7,.7,.7,.7,.7]$, $\tau_R = [0,0,0,0,0,0]$, $\Phi = I_R$, and $\kappa_R = 0$. For the population with moderate reliability, $\Psi_R = diag([.51,.51,.51,.51,.51,.51])$, resulting in a reliability of $\alpha_R = .85$. For the population with high reliability, $\Psi_R = diag([.21,.21,.21,.21,.21,.21])$, resulting in a reliability of $\alpha_R = .93$. In the focal group, $\Lambda_F = [.7,.9,.7,.9,.7,.9]$, $\tau_F = [0,.2,0,.2,0,.2]$, with all other focal group parameters being equal to those in the reference group. Note that because factor loadings differ across groups, reliability will be higher in the focal group, with $\alpha_F = .88$ and $\alpha_F = .95$ in the moderate and high reliability groups, respectively.

Multiple-group population models based on condition 5 from *Table 3.4* also have 1 factor with 6 indicator variables. In the reference group, again, $\Lambda_R = [.7,.7,.7,.7,.7,.7]$, $\tau_R = [0,0,0,0,0,0]$, $\Phi = I_R$, and $\kappa_R = 0$. For the population with moderate reliability, $\Psi_R = diag([.51,.51,.51,.51,.51,.51])$, resulting in a reliability of $\alpha_R = .85$. For the population with high reliability, $\Psi_R = diag([.21,.21,.21,.21,.21,.21])$, resulting in a reliability of $\alpha_R = .93$. In the focal group, $\Lambda_F = [.7,.9,.7,.9,.7,.9]$, $\tau_F = [0,-.2,0,-.2,0,-.2]$, with all other focal group parameters being equal to those in the reference group. Note that because factor loadings differ across groups, reliability will be higher in the focal group, with $\alpha_F = .88$ and $\alpha_F = .95$ in the moderate and high reliability populations, respectively. For each of our 4 multiple-group population models, 4 sample sizes will be used for data generation: $N_1 = N_2 = 200$,

$N_1 = N_2 = 300$, $N_1 = N_2 = 500$, and $N_1 = N_2 = 1000$. For each multiple-group population model, a total of 100 samples are generated for each sample size. Population parameters are summarized in *Table 3.6*.

3.3.2. Analysis

For each multiple-group population model, the most parsimonious partial measurement invariance model that is consistent with the data-generation model is fit to data using the SEM software package *lavaan* (Rosseel, 2012). In all conditions, the most parsimonious partial MI model is a 2-group, 1-factor model with 6 indicators, and between-group equality constraints on the first, third, and fifth loading, intercept, and residual, with all measurement parameters on the second, fourth, and sixth item freely estimated in both groups. Note that between-group equality constraints are not placed on the second, fourth, and sixth residual variances, despite them being equal in the population, because when specifying partial MI models, invarian loading is considered pre-requisite to testing invariance of its intercept, and invariance of both the intercept and loading is pre-requisite to testing invariance of its residual variance (Byrne, Shavelson, and Muthen, 1989).

Subsequent to model fitting, total loading ($\hat{\lambda}^*_g$), total intercept ($\hat{\tau}^*_g$), total residual variance ($\hat{\psi}^*_g$), latent variance, and latent mean estimates in each group are used to estimate $\log_{10}(\hat{\varphi}_{sens})$ at each percentile in each replication. The 95% confidence interval around the population $\log_{10}(\varphi_{sens})$ can then be empirically estimated at each percentile. Results are presented as a series of four 2×2 panel plots, presented in *Figure 3.7* through *Figure 3.10*, with each figure representing sensitivity ratios associated with a particular population model, and each panel giving the confidence interval around $\log_{10}(i_{Sens})$ at each percentile for a particular sample size.

3.3.3. Results

Figure 3.7 and *Figure 3.8* give the range of finite-sample estimates of $\log_{10}(i_{Sens})$ plots in condition 4 when reliability is moderate, and high, respectively. Examining *Figure 3.7* - where $l^*_R = 4.2$, $l^*_F = 4.8$, $t^*_R = 0$, $t^*_F = .6$, $y^*_R = y^*_F = 3.06$, $k_R = k_F = 0$, $F_R = F_F = 1$ and the population mixing proportion is .5 –the variability of sensitivity ratio estimates is very small near the lowest percentiles, and increases towards higher percentiles. As expected, variability of sensitivity ratio estimates decreases with increasing sample size, with the greatest variability still seen at the highest percentiles. At all studied sample sizes, the impact of violations of MI on i_{Sens} is minimal up to the 50th percentile, meaning one could reasonably conduct a median split at any studied sample size and be confident that $.891 < i_{Sens} < 1.122$ in the population, as the confidence interval around $\log_{10}(i_{Sens})$ at this percentile is fully contained within the range $-.05 < \log_{10}(i_{Sens}) < .05$. Further, at all studied sample sizes, the impact of violations of MI on i_{Sens} is not large up to the 70th percentile, meaning one could use the 70th percentile for selection and be reasonably confident that $.794 < i_{Sens} < 1.259$ in the population, as the confidence interval around $\log_{10}(i_{Sens})$ at this percentile is fully contained within the range $-.1 < \log_{10}(i_{Sens}) < .1$. Beyond the 70th percentile, it is plausible that the impact of violations of MI on i_{Sens} is large, as the confidence interval around $\log_{10}(i_{Sens})$ falls outside of the range $-.1 < \log_{10}(i_{Sens}) < .1$.

Examining *Figure 3.8*, which represents the same population as *Figure 3.7* but with greater reliability in each group, much less variability in estimated sensitivity ratios is observed. That said, the largest critical percentiles that ensure a minimal or modest impact on i_{Sens} are largely unchanged between the population represented in *Figure 3.7* and the population represented in *Figure 3.8*, as the present violations of metric and scalar invariance have an additive effect on population sensitivity ratios, as illustrated in *Table 3.5*, resulting in sensitivity ratios that strongly favor the focal group in the population.

Figure 3.9 and *Figure 3.10* give the range of finite sample estimates of $\log_{10}(i_{Sens})$ plots in condition 5 when reliability is moderate, or high, respectively. Examining *Figure 3.9* - where $l_R^* = 4.2$, $l_F^* = 4.8$, $t_R^* = 0$, $t_F^* = .6$, $y_R^* = y_F^* = 3.06$, $k_R = k_F = 0$, $F_R = F_F = 1$ and the population mixing proportion is .5 – the variability of sensitivity ratio estimates about population values are similar to those observed in *Figure 3.7*. Recalling that the violations of metric and scalar invariance present in condition 5 have a compensatory effect on population sensitivity ratios, this suggests that the magnitude of population sensitivity ratios, at a given percentile, is unrelated to the variability of sensitivity ratio estimates. When $N_R = N_F = 200$, 300, or 500, the impact of violations of MI on i_{Sens} is minimal up to the 80th percentile, meaning one could use the 80th percentile for selection at these sample sizes and be reasonably confident that $.891 < i_{Sens} \leq 1.122$ in the population, as the confidence interval around $\log_{10}(i_{Sens})$ at this percentile is fully contained within the range $-.05 \leq \log_{10}(i_{Sens}) \leq .05$. When $N_R = N_F = 1000$, the impact of violations of MI on i_{Sens} is minimal up to the 90th percentile. Further, at all studied sample sizes, the impact of violations of MI on i_{Sens} is not large up to the 90th percentile, meaning one could use the 90th percentile for selection at all sample sizes, and be reasonably confident that $.794 < \lambda_{Sens} \leq 1.259$ in the population, as the confidence interval around $\log_{10}(\lambda_{Sens})$ at this percentile is fully contained within the range $-.1 \leq \log_{10}(\lambda_{Sens}) \leq .1$. Similar results are observed in *Figure 3.10*, which represents the same population as *Figure 3.9*, but with greater reliability in each group.

In practice, researchers can use Bollen-Stine (1993) bootstrapping to empirically estimate the 95% confidence interval around their estimated $\log_{10}(\hat{\lambda})$ plots. This technique involves identifying the most parsimonious partial MI simulating a large number of datasets from the model-implied covariance matrix and model-implied means. Next, the retained partial MI model is fit to each replicated dataset. Parameter estimates can then be used to estimate $\log_{10}(\hat{\lambda})$ at each percentile for each replication, allowing the computation of an empirically estimated confidence interval around $\log_{10}(\hat{\lambda})$ at each percentile. If the estimated confidence interval around $\log_{10}(\hat{\lambda})$ falls within a specified range at a given percentile —such as $-.05 \leq \log_{10}(\hat{\lambda}) \leq .05$, reflecting a minimal impact of violations of measurement invariance on λ —one has much stronger evidence that the true population λ falls within that range than if one simply interprets the classification accuracy ratio plot associated with initial parameter estimates alone. Researchers may then safely infer that using the specified percentile as a critical value for selection minimizes between-group differences in classification accuracy. This technique will now be illustrated by applying the proposed method to Everson et al.'s (1991) data. Sample code for generating *Figure 3.7* through *Figure 3.10* is included in *Appendix G*.

3.4. Real Data Example Revisited

Classification accuracy ratio plots can be used to take a more detailed look at Millsap and Kwok's (2004) analysis of the worry subscale (TAI) data. Recall that their estimated parameter totals are : $\hat{\lambda}^*_R = 7.279$, $\hat{\lambda}^*_F = 7.679$, $\hat{\tau}^*_R = 16.621$, $\hat{\tau}^*_F = 16.581$, $\hat{\psi}^*_R = 4.329$, $\hat{\psi}^*_F = 3.405$, $\hat{\Phi}_R = .544$, $\hat{\Phi}_F = .477$, $\kappa_R = 0$, and $\hat{\kappa}_F = -0.126$, with female respondents being the reference group. *Figure 3.11* gives $\log_{10}(\hat{\rho})$ plots associated with these parameter estimates for four measures of classification accuracy. *Figure 3.12* supplements with separate $\hat{\eta}$ plots for each group for each measure of classification accuracy.

Examining *Figure 3.11*, one can see that sensitivity and specificity are slightly greater in the focal group, while PPV and NPV are slightly greater in the reference group. For all measures of classification accuracy, deviations from $\log_{10}(\rho) = 0$ are generally very slight. Sensitivity ratios reflect minimal impact of violations of MI on classification accuracy at all percentiles, meaning $.891 < \hat{\rho}_{Sens} < 1.122$, suggesting that sensitivity is, at most, 12% higher in the focal group than reference group at all percentiles. NPV ratios also reflect minimal impact of violations of MI on classification accuracy at all percentiles. Specificity and PPV ratios reflect minimal impact of violations of MI on selection accuracy at most percentiles, while showing some modest impact at extreme percentiles. In general, classification accuracy ratio plots indicate that violations of MI have a minimal impact on all measures of classification accuracy at most percentiles.

Figure 3.12 gives observed measures of classification accuracy for each group at each percentile, with the solid line indicating the reference group, women, and the dashed line indicating the focal group, men. Consistent with classification accuracy ratio plots, estimates of classification accuracy are generally very close between populations at all percentiles. Specificity appears to be the measure of classification accuracy which is the most discrepant across groups, with $\eta_{Spec,R} = .6$ and $\eta_{Spec,F} = .7$ at the 1st percentile. Because between-group differences in η_{Sens} and η_{PPV} appear very small at all percentiles, it seems slightly more appropriate to use the worry subscale of the test anxiety inventory to identify participants who are high in worry rather than low in worry, although the impact of violations of measurement invariance on λ_{Spec} and λ_{NPV} are still generally minimal at most percentiles.

Assuming Measurement Invariance. Recall that Millsap and Kwok (2004) also supplement their analysis of Everson et al.'s

under the assumption of MI. This allows them to compare the between-group differences in classification accuracy present under violations of MI to the amount introduced by between-group differences in latent means and latent variances alone. A similar analysis can be conducted using the proposed method by generating $\log_{10}(\hat{\lambda})$ plots and $\hat{\eta}$ plots for Millsap a

(2004) estimated parameter totals under the assumption of MI: $\lambda^* = 7.454$, $\tau^* = 16.604$, and $\psi^* = 3.925$ in each group, while $\hat{\Phi}_R = .544$, $\hat{\Phi}_F = .477$, $\hat{\kappa}_R = 0$, and $\hat{\kappa}_F = -0.126$. Examining *Figure 3.13*, which gives $\log_{10}(\hat{\lambda})$ plots under the assumption of invariance, the slight favoring of the focal group by the $\log_{10}(\hat{\lambda}_{Sens})$ plot seen in *Figure 3.11* becomes a slight favoring of the reference group, although the impact remains minimal. Note also that the modest favoring of the reference group by the $\log_{10}(\hat{\lambda}_{PPV})$ plot near the 99th percentile observed in *Figure 3.11* remains under the assumption of MI, suggesting that violations of MI do not influence between-group differences in PPV in this context. The $\log_{10}(\hat{\lambda}_{NPV})$ plot also shows negligible change as a function of the assumption of MI. The $\log_{10}(\hat{\lambda}_{Spec})$ plot in *Figure 3.13* shows that, under the assumption of MI, no between-group differences in specificity are expected, while *Figure 3.11* shows that when MI is not assumed, violations of measurement invariance appear to have a modest impact on $\hat{\lambda}_{Spec}$ in favor of the focal group at the 10th percentile and below. That said, the estimated impact of violations of MI on λ_{Spec} is still minimal at most percentiles, and is unlikely to be problematic unless extreme lower percentiles are used as critical values for selection.

Comparing *Figure 3.12* and *Figure 3.14*, which give classification accuracy plots separately for each group without and with the assumption of MI, respectively, the same patterns emerge. Almost no between-group differences in $\hat{\eta}_{Sens}$ and $\hat{\eta}_{NPV}$ are seen at all percentiles, with and without the assumption of MI. Slight between-group differences in $\hat{\eta}_{PPV}$ are observed at high percentiles with and without the assumption of MI. Slight between-group differences in $\hat{\eta}_{Spec}$ appear to be introduced by violations of MI, as $\hat{\eta}_{Spec}$ is greater in the focal group when violations of MI are present, while no between-group differences in $\hat{\eta}_{Spec}$ are expected under the assumption of MI.

Confidence Intervals. One final consideration when evaluating the impact of violations of MI on classification accuracy is that parameter totals used to compute $\log_{10}(\hat{\lambda})$ plots are estimates based on finite samples, and thus, $\log_{10}(\hat{\lambda})$ plots serve only as estimates of the true population-level $\log_{10}(\lambda)$ plot. Simulations conducted in section 3.3 illustrate that, when sample sizes are small, and reliability is poor, there can be tremendous variability in estimated $\log_{10}(\hat{\lambda})$ plots, particularly near extreme critical percentiles. In the TAI example, $N_R = 282$, $N_F = 219$, $\alpha_R = .869$, and $\alpha_R = .892$.

Figure 3.15 gives Bollen-Stine bootstrapped $\log_{10}(\hat{\lambda})$ plots for each measure of classification accu r2a0c0y4)f oarn aMI iylslissa po fa nEdv eKrwsookn' data. The confidence interval around $\log_{10}(\hat{\lambda}_{Sens})$ indicates a minimal impact of violations of MI for all percentiles up to and including the 80th, and no greater than modest impact for all percentiles up to and including the 90th. Researchers can safely select participants above the 80th or 90th percentile on the worry subscale of the TAI and be reasonably confident that group membership has a minimal or modest impact, respectively on $P(y_{comp,i,g} > y_{comp,c} \mid X_{i,g} > X_c)$. The confidence interval around $\log_{10}(\hat{\lambda}_{Spec})$ indicates a minimal impact of violations of MI for all percentiles above and including the 30th percentile, and no greater than modest impact for all percentiles above and including the 10th. Researchers can safely select participants below the 30th or 10th percentile on the worry subscale of the TAI and be reasonably confident that group membership has a minimal or modest impact, respectively, on $P(y_{comp,i,g} < y_{comp,c} \mid X_{i,g} < X_c)$. The confidence interval around $\log_{10}(\hat{\lambda}_{PPV})$ indicates a minimal impact of violations of MI for all percentiles up to and including the 90th percentile, and no greater than modest impact for all percentiles. Researchers can safely select participants above the 90th percentile on this psychometric instrument and be reasonably confident that group membership has a minimal impact on $P(X_{i,g} > X_c \mid y_{comp,i,g} > y_{comp,c})$. The confidence interval around $\log_{10}(\hat{\lambda}_{NPV})$ indicates a minimal impact of violations of MI for all percentiles above and including the 20th percentile, and no greater than modest impact for all percentiles above and including the 5th percentile. Researchers can safely select participants below the 20th or 5th percentile on this psychometric instrument and be reasonably confident that group membership has a minimal or modest impact, respectively, on $P(X_{i,g} < X_c \mid y_{comp,i,g} < y_{comp,c})$. Sample code for generating *Figure 3.11* and *Figure*

3.12 is included in *Appendix H.* Sample code using Bollen-Stine bootstrapping to estimate the confidence intervals around classification accuracy ratio plots presented in *Figure 3.15* is included in *Appendix I.*

3.5. Discussion

3.5.1. Summary of Research

I believe my exploration of the behavior of classification accuracy ratio plots has effectively illustrated their utility for the purposes of identifying critical percentiles for selection that minimize between-group differences in classification accuracy due to violations of MI. I began by introducing Millsap and Kwok's (2004) method, which is u

classification accuracy about a specified percentile for a multiple-group population for which MI does not necessarily hold. This involves estimating x_c and $y_{comp,c}$, the critical values associated with the specified percentile in the pooled population of factor scores and observed composite scores, respectively. These values are then used to divide the bivariate distribution of $x_{i,g}$ and $y_{comp,i,g}$ in each group into four quadrants. The densities of each group in each quadrant are used to estimate measures of classification accuracy for that group. This method can also be used to illustrate expected measures of classification accuracy under the assumption of MI for the purposes of addressing whether between-group differences on measures of classification are due to violations MI. Unfortunately Millsap and K

the behavioral sciences.

I propose a modification to Millsap and Kwok's (2004)

requiring a specified critical percentile of interest, presents estimated classification accuracy ratios, $\acute{l}$, for all percentiles, allowing researchers to choose critical values for selection based on which percentiles show acceptably low between-group differences in measures of classification accuracy as a function of violations of measurement invariance. Our preferred presentation of classification accuracy ratios is a 2 ³2 panel plot displaying classification accuracy ratios for all percentiles for each of 4 measures of classification accuracy: sensitivity, specificity, positive predictive value, and negative predictive value. Our preferred presentation of classification accuracy ratio plots also places ratios on a $\log_{10}$ scale for the reasons illustrated in *Figure 3.2*. *Figure 3.11* and *Figure 3.13* are examples of this presentation.

To illustrate the impact of population-level violations of MI on classification accuracy – and how classification accuracy ratio plots allow us to readily decode that impact –a series of classification accuracy ratio plots based on multiple-group populations featuring different patterns of violations of MI was generated. Population parameters used in generating classification accuracy ratio plots are presented in *Table 3.4*. When violations of metric invariance are present, and favor the focal group, meaning factor loadings are higher in the focal group, i_{Sens} and i_{Spec} favor the focal group, meaning sensitivity and specificity are higher in the focal group than reference group. When violations of metric invariance favor the focal group, i_{PPV} and i_{NPV}, however, favor the reference group. More generally, when violations of metric invariance are present, with all else being equal, the group with higher loadings will have higher sensitivity and specificity, and the group with lower loadings will have higher PPV and NPV. When violations of scalar invariance are present and favor the focal group, meaning intercepts are higher in the focal group, i_{Sens} and i_{NPV} favor the focal group, while i_{Spec} and i_{PPV} favor the reference group. More generally, when violations of scalar invariance are present, with all else being equal, the group with higher intercepts will have higher sensitivity and NPV, and the group with lower intercepts will have higher specificity and PPV. When violations of strict invariance are present, and residual variances are higher in the focal group, all else being equal, i_{Sens} and i_{Spec} reflect minimal bias at all percentiles, while i_{PPV} and i_{NPV} favor the reference group. More specifically, when violations of strict invariance are present, the group with lower residual variances will have higher PPV and NPV. The direction of the impact of violations of MI on each measure of classification accuracy in each of these three conditions is summarized in *Table 3.5*.

The primary takeaway from these first three illustrations is that different patterns of violations of MI differentially impact different measures of classification accuracy. Thus, when violations of measurement invariance are present for more than one group of measurement parameters —loadings, intercepts, or residuals —one might observe additive or compensatory effects on classification accuracy. For example, in condition 4, where both factor loadings and intercepts are higher in the focal group, sensitivity strongly favors the focal group, while specificity minimally favors the focal group, as shown in *Figure 3.3* and *Figure 3.4*. If we refer to *Table 3.5*, this pattern of results is consistent with the fact that the group with higher loadings will have higher sensitivity and specificity, while the group with higher intercepts will have higher sensitivity and lower specificity, resulting in an additive effect on sensitivity ratios and a compensatory effect on specificity ratios. Similarly, in condition 5, where factor loadings are higher in the focal group and intercepts are higher in the reference group, specificity strongly favors the focal group, while sensitivity minimally favors the focal group, as shown in *Figure 3.3* and *Figure 3.4*. These results are also consistent with the patterns presented in *Table 3.5*.

3.5.2. Between-Group Differences in Latent Means and Latent Variances

Beyond measurement parameters, it is also worth addressing the impact that other parameter inequalities have on classification accuracy ratio plots. As previously highlighted in our real data example, between-group inequalities on latent variances and latent means can lead to between-group inequalities on measures of classification accuracy, even when MI holds for all parameters. While the effect on classification accuracy of interactions between latent variance inequalities, latent mean inequalities, and violations of MI has not been the focus of this research, it is worth addressing the impact of unequal latent means and unequal latent variances on classification accuracy ratio plots when MI holds. *Figure 3.16* gives classification accuracy ratio plots for a multiple-group population where $\lambda_R^* = \lambda_F^* = 4.2$, $\tau_R^* = \tau_F^* = 0$, $\psi_R^* = \psi_F^* = 3.06$, $\kappa_R = 0$, $\kappa_F = .5$, $\Phi_R = \Phi_F = 1$, with a population mixing proportion of .5. Examining *Figure 3.16*, we can see that the group with the higher latent mean has slightly higher sensitivity and PPV, while the group with the lower latent mean has slightly higher specificity and NPV. *Figure 3.17* gives classification accuracy ratio plots for a multiple-group population where $\lambda_R^* = \lambda_F^* = 4.2$, $\tau_R^* = \tau_F^* = 0$, $\psi_R^* = \psi_F^* = 3.06$, $\kappa_R = \kappa_F = 0$, $\Phi_R = 1$, and $\Phi_F = 1.5$, with a population mixing proportion of .5. Examining *Figure 3.17*, it can be seen that the group with the higher latent variance has slightly higher sensitivity, specificity, PPV, and NPV. Between-group inequalities in classification accuracy introduced by between-group inequalities in latent means and latent variances are generally quite small compared to those introduced by violations of metric or scalar invariance.

Given that measurement invariance is the equivalent functioning of psychometric instruments when applied across multiple groups, it is not unreasonable that some applied researchers may intuit that full strict invariance of measures should necessarily imply invariance of measures of classification accuracy. The purpose of section 3.5.2. is to illustrate that this is not the case when between-group differences on latent means or latent variances are present in the population. This leads to an interesting question: should we aspire to create measures that have equal measures of classification accuracy at all percentiles? Or should we aspire to create measures that approximate the classification accuracy ratio plots that are expected under MI? Ideally we would have both, but which is preferable if we only have one? While the latter makes more intuitive sense, it is worth considering that often the motivation for evaluating MI in the first place is ensuring equivalent instrument behavior across groups. It is reasonable to consider that a “technically unbiased” instrument that still l ication accuracy might be a difficult sell in many scenarios, such as when psychometric instruments are used in employee recruitment. While an extreme example unlikely to occur in practice, *Figure 3.18* illustrates how between-group bias in measures of classification accuracy can theoretically be severe, even when MI holds in the population.

3.5.3. Finite Sample Considerations

I have discussed how in practice, researchers will never know exact population values for model parameters, and thus, all classification accuracy ratio plots will be based on finite sample estimates. Since finite sample parameter estimates will necessarily vary from sample to sample, estimated classification accuracy ratio plots will also vary from sample to sample. *Figure 3.7* through *Figure 3.10* illustrate how classification accuracy ratio plots for multiple-group populations vary under different violations of MI, different sample sizes, and different reliabilities of measures. Examining such plots may be useful for sample size planning, particularly if estimates of reliability are available. Applied researchers might plan for a sample size based on the width of empirically estimated confidence intervals at a particular critical percentile at a particular sample size. This approach is, of course, limited in that it assumes that variability of classification accuracy ratios at a given critical percentile is only dependent on reliability and sample size, and is otherwise largely model independent. Further research is necessary to determine when/if this is necessarily true.

One potential avenue for researchers interested in examining classification accuracy ratio plots, but hesitant to do so based on uncertainty pertaining to finite sample variability, would be to use Bollen-Stine bootstrapping (Bollen & Stine, 1992). Bollen-Stine bootstrapping is a technique which involves simulating data from observed parameter estimates, and subsequently fitting one's model to each replication to ev

Figure 3.15 illustrates the results of applying Bollen-Stine bootstrapping to our real data example making use of [illegible] (1991) TAI data. To produce this plot, 100 datasets are generated from the parameter estimates obtained by Mill [illegible] most parsimonious partial MI model. Subsequently, that partial MI model is fit to each simulated dataset. Finally, classification accuracy ratios are estimated at each percentile for each replication, allowing the computation of an empirical confidence interval around $\log_{10}(\hat{\lambda})$ at each percentile, which are plotted alongside the original $\log_{10}(\hat{\lambda})$ estimates. In this example, the confidence interval appears much narrower than is observed for similar sample sizes in the simulations presented in *Figure 3.7* through *Figure 3.10*, possibly because the impact of present violations of MI on observed for this data are very small. If one examines, for example, the 80th percentile of *Figure 3.15*, we see that the impact of violations of MI on $\log_{10}(\hat{\lambda}_{Sens})$ appears minimal. A researcher who considers all values in this particular range sufficiently unbiased can safely use the 80th percentile of the worry subscale of the test anxiety inventory as a critical value for selection and be confident that group membership has a minimal impact on $P(y_{comp,i,g} > y_{comp,c} \mid \chi_{i,g} > \chi_c)$. See *Appendix I* for sample code illustrating this procedure.

3.5.4. The Issue of Reliability

One final, but important consideration, is the impact of reliability of measures on classification accuracy ratio plots. Reliability is the proportion of variance in $y_{comp,i,g}$ accounted for by one's model, and thus, reliability is en residual variances are higher. It is reasonable to assume that when residual variances are higher, the number of false positives and false negatives in each population at each critical percentile will increase. For example, a participant with a factor score just below χ_c is more likely to exceed $y_{comp,c}$ on their observed composite score if the sampling distribution of plausible error scores includes a greater density of values that would push that participant past the threshold, producing a false positive. If lower reliability leads to greater proportions of false positives and false negatives, we would also expect it to lead to lower sensitivity, specificity, PPV, and NPV. If measures of classification accuracy decrease as a function of decreasing reliability in both groups, classification accuracy ratios could theoretically be unaffected by variations in reliability.

Figure 3.19 illustrates how sensitivity ratio plots vary as a function of increasing residual variances in a multiple-group population that is otherwise identical to condition 1 as described in *Table 3.4*. Panels (a) through (d) of *Figure 3.19* give classification accuracy ratios when $y^* = 1$, 3, 5, and 7 in both groups, respectively. In these populations, reliabilities in the reference group are .94, .85, .78, and .72, respectively, and reliabilities in the focal group are .96, .88, .82, and .77, respectively. Examining *Figure 3.19*, we see very little variation in $\log_{10}(i_{Sens})$ plots as a function of reliability: the largest percentile indicating a minimal impact of violations of MI on

i_{Sens} is the 80th for all reliabilities, and the largest percentile indicating a modest impact of violations of MI on i_{Sens} is the 90th for all reliabilities. If we turn to *Figure 3.20*, which gives individual group η_{Sens} at each percentile, we see that sensitivity steadily decreases in both groups as residual variance increases. This illustrates an important limitation of classification accuracy ratio plots: classification accuracy ratio plots only reflect between-group differences in classification accuracy, and are in no way indicative of whether classification accuracy in any particular group is adequate. Supplementing analyses of measurement invariance via classification accuracy ratio plots with individual group estimates of classification accuracy is helpful if one wishes to avoid conflating MI with overall adequacy of a measure.

3.5.5. Additional Limitations

Further limitations to the proposed research include the fact that the method is only evaluated for normally distributed data, and for equal sample sizes across groups. Theoretically, the proposed methodology should function properly for non-normally distributed data as well, so long as the probability density function of y_{comp} is estimable. If the density of the sampling distribution of y_{comp} that exceeds a particular value can be accurately estimated, $y_{comp,c}$ should still be retrievable by Millsap and Kwok's (2

computation of quadrant densities defined by $y_{comp,c}$ and χ_c over the bivariate distributions of χ and y_{comp} in each group. It should be noted that while extensions of the proposed method to non-normal data are theoretically possible, all presently existing software only offers support for normally distributed data. Further research should be conducted to evaluate the behavior of the proposed method when data are not normally distributed, particularly given that violations of the assumption of normality will also impact model fit, and by extension, the final retained partial invariance model.

While it was addressed that unbalanced population mixing proportions have a negligible impact on classification accuracy ratio plots, it should be noted that this is not equivalent to saying sample size evenness has no impact on classification accuracy ratio plots. When sample sizes are unequal in MGCFA, final parameter estimates for parameters constrained to equality will generally be weighted towards the unconstrained estimates in the larger of the two groups, with this weighting generally increasing as samples become more uneven. Thus, uneven sample sizes may result in inaccurate l^*, t^*, and y^* estimates, particularly in the smaller of the two groups. Further research is necessary to evaluate the behavior of estimated classification accuracy ratio plots when sample sizes are uneven across groups.

	$x_i < x_c$ Condition Negative	$x_i > x_c$ Condition Positive	
$y_{comp,i} < y_{comp,c}$ Selection Negative	True Negative (TN)	False Negative (FN)	$NPV = \frac{TN}{TN + FN}$, $FOR = \frac{FN}{FN + TN}$
$y_{comp,i} > y_{comp,c}$ Selection Positive	False Positive (FP)	True Positive (TP)	$PPV = \frac{TP}{TP + FP}$, $FDR = \frac{FP}{TP + FP}$
	$Specificity = \frac{TN}{TN + FP}$ $FPR = \frac{FP}{TN + FP}$	$Sensitivity = \frac{TP}{TP + FN}$ $FNR = \frac{FN}{TP + FN}$	

Table 3.1. Measures of Classification Accuracy and How They Relate to the Confusion Matrix.

NPV=Negative Predictive Value, FOR=False Omission Rate, PPV=Positive Predictive Value, FDR=False Discovery Rate, FNR=False Negative Rate, FPR=False Positive Rate. Note that each row and column of the confusion matrix has two associated measures of classification accuracy which sum to 1.

	Original Model		*Invariance Assumed Model*	
	Reference Group	*Focal Group*	*Reference Group*	*Focal Group*
A (True Positives)	.091	.060	.094	.057
B (False Positives)	.027	.022	.027	.022
C (True Negatives)	.850	.901	.851	.901
D (False Negatives)	.032	.017	.029	.020
Sensitivity	.742	.774	.767	.738
Specificity	.969	.976	.969	.976

Table 3.2. Sample Quadrant Densities and Classification Accuracies About the 90th Percentile for Women and Men's "Worry" Subscale of *the Test Anxiety Inventory, Generated Using Lai* et al.'s (2017) Software.

Ratio	Log10 Ratio
0.501	-.30
0.562	-.25
0.631	-.20
0.798	-.15
0.794	-.10
0.891	-.05
1.000	0
1.122	.05
1.259	.10
1.413	.15
1.585	.20
1.778	.25
1.995	.30

Table 3.3. Reference Table for Converting Between Classification Accuracy Ratios and Log10 Classification Accuracy Ratios.

We can interpret $\log_{10}( ) = .05$ as indicating that a measure of classification accuracy is roughly 10% higher in the reference group, while $\log_{10}( ) = -.05$ indicates that classification accuracy is roughly 10% higher in the focal group. We can interpret $\log_{10}( ) = .1$ as indicating that a measure of classification accuracy is roughly 25% higher in the reference group, while $\log_{10}( ) = -.1$ indicates that classification accuracy is roughly 25% higher in the focal group. We can interpret $\log_{10}( ) = .2$ as indicating that a measure of classification accuracy is roughly 50% higher in the reference group, while $\log_{10}( ) = -.2$ indicates that classification accuracy is roughly 50% higher in the focal group. We can interpret $\log_{10}( ) = .3$ as indicating that a measure of classification accuracy is roughly 100% higher in the reference group, while $\log_{10}( ) = -.3$ indicates that classification accuracy is roughly 100% higher in the focal group.

Illustration Condition	Population Parameters in Focal Group
Condition 1: Higher Loadings in Focal Group	$\lambda_F^* = 4.8$, $\tau_F^* = 0$, $\psi_F^* = 3.06$, $\Phi_F = 1$, $\kappa_F = 0$
Condition 2: Higher Intercepts in Focal Group	$\lambda_F^* = 4.2$, $\tau_F^* = .6$, $\psi_F^* = 3.06$, $\Phi_F = 1$, $\kappa_F = 0$
Condition 3: Higher Residual Variances in Focal Group	$\lambda_F^* = 4.2$, $\tau_F^* = 0$, $\psi_F^* = 4.86$, $\Phi_F = 1$, $\kappa_F = 0$
Condition 4: Higher Loadings and Intercepts in Focal Group	$\lambda_F^* = 4.8$, $\tau_F^* = .6$, $\psi_F^* = 3.06$, $\Phi_F = 1$, $\kappa_F = 0$
Condition 5: Higher Loadings and Lower Intercepts in Focal Group	$\lambda_F^* = 4.8$, $\tau_F^* = -.6$, $\psi_F^* = 3.06$, $\Phi_F = 1$, $\kappa_F = 0$

Table 3.4. Focal Group Population Parameters for Population Classification Accuracy Ratio Plot Illustrations.

In all conditions, $\lambda_R^* = 4.2$, $\tau_R^* = 0$, $\psi_R^* = 3.06$, $\Phi_R = 1$, and $\kappa_R = 0$, with a population mixing proportion of .5.

	Sensitivity Ratio	Specificity Ratio	PPV Ratio	NPV Ratio
$l_F^* > {}_R^*$	<1	<1	>1	>1
$t_F^* > {}_R^*$	<1	>1	>1	<1
$y_F^* > y_R^*$	1	1	>1	>1
$l_F^* < {}_R^*$	>1	>1	<1	<1
$t_F^* < {}_R^*$	>1	<1	<1	>1
$y_F^* < y_R^*$	1	1	<1	<1

Table 3.5. Expected Impact of Violations of Measurement Invariance on Classification Accuracy Ratios.

Cells colored blue indicate that the violation of measurement invariance specified in the far left column results in classification accuracy ratios favoring the focal group, while red cells indicate that the reference group is favored. We can use this table to estimate the aggregate impact of violations of measurement invariance on classification accuracy ratios. For example, if loadings, intercepts, and residual variances are higher in the focal group, we would expect PPV ratios to severely favor the reference group, as all 3 violations of measurement invariance bias PPV ratio in favor of the reference group, while NPV ratios would be expected to less dramatically favor the reference group, as 2 violations of measurement bias NPV ratio in favor of the reference group, while 1 violation biases NPV in favor of the focal group.

Condition	Reference Group	Focal Group
Non-Invariant Loadings, Moderate Reliability (*Figure 3.7*)	$\Lambda = [.7,.7,.7,.7,.7,.7]$	$\Lambda = [.7,.9,.7,.9,.7,.9]$
	$\tau = [0,0,0,0,0,0]$	$\tau = [0,.2,0,.2,0,.2]$
	$\Psi = diag([.51,.51,.51,.51,.51,.51])$	$\Psi = diag([.51,.51,.51,.51,.51,.51])$
	$\kappa = 0$	$\kappa = 0$
	$\Phi = 1$	$\Phi = 1$
Non-Invariant Loadings, High Reliability (*Figure 3.8*)	$\Lambda = [.7,.7,.7,.7,.7,.7]$	$\Lambda = [.7,.9,.7,.9,.7,.9]$
	$\tau = [0,0,0,0,0,0]$	$\tau = [0,.2,0,.2,0,.2]$
	$\Psi = diag([.21,.21,.21,.21,.21,.21])$	$\Psi = diag([.21,.21,.21,.21,.21,.21])$
	$\kappa = 0$	$\kappa = 0$
	$\Phi = 1$	$\Phi = 1$
Non-Invariant Intercepts, Moderate Reliability (*Figure 3.9*)	$\Lambda = [.7,.7,.7,.7,.7,.7]$	$\Lambda = [.7,.9,.7,.9,.7,.9]$
	$\tau = [0,0,0,0,0,0]$	$\tau = [0,-.2,0,-.2,0,-.2]'$
	$\Psi = diag([.51,.51,.51,.51,.51,.51])$	$\Psi = diag([.51,.51,.51,.51,.51,.51])$
	$\kappa = 0$	$\kappa = 0$
	$\Phi = 1$	$\Phi = 1$
Non-Invariant Intercepts, High Reliability (*Figure 3.10*)	$\Lambda = [.7,.7,.7,.7,.7,.7]$	$\Lambda = [.7,.9,.7,.9,.7,.9]$
	$\tau = [0,0,0,0,0,0]$	$\tau = [0,-.2,0,-.2,0,-.2]'$
	$\Psi = diag([.21,.21,.21,.21,.21,.21])$	$\Psi = diag([.21,.21,.21,.21,.21,.21])$
	$\kappa = 0$	$\kappa = 0$
	$\Phi = 1$	$\Phi = 1$

Table 3.6. Population Parameters Used to Illustrate Finite Sample Variability of Classification Accuracy Ratio Plots.

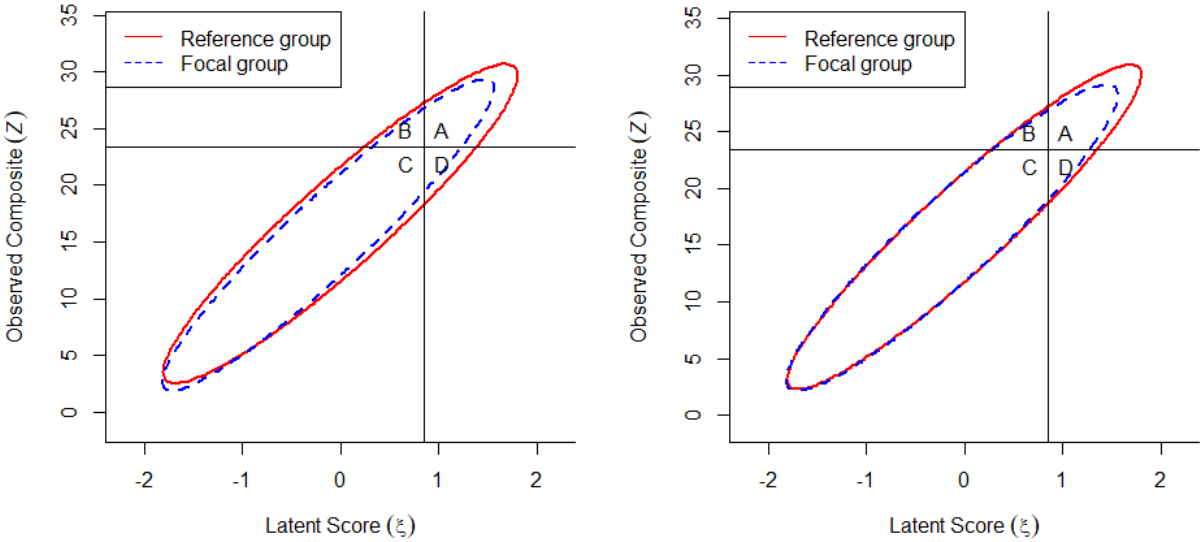

Figure 3.1. Replication of Millsap and Kwok (2004) Bivariate Density Plot Evaluating Measurement Invariance Between Men and Women on the Test Anxiety Inventory.

Population parameters in the left panel are : $\hat{l}_R$ =7.279 , $\hat{l}_F$ =7.679 , $\hat{t}^*_R$ =16.621 , $\hat{t}^*_F$ =16.581 , $\hat{y}^*_R$ =4.329 , $\hat{y}^*_F$ =3.405 , $\hat{F}_R$ =.544 , $\hat{F}_F$ =.477 , k_R =0 , and $\hat{k}_F$ = −0.126 , with women being the reference group. The right panel illustrates the bivariate density in each population when measurement invariance is assumed, in which case, population parameters are treated as $\hat{l}$ =7.454 , $\hat{t}^*$ =16.604 , $\hat{y}^*$ =3.925 , $\hat{F}_R$ =.544 , $\hat{F}_F$ =.477 , k_R =0 , and $\hat{k}_F$ = −0.126 . These plots were generated using software provided by Lai et al. (2017). Note that Millsap and Kwok (2004) use *Z* as the symbol for observed composite scores, while I prefer to use y_{comp} .

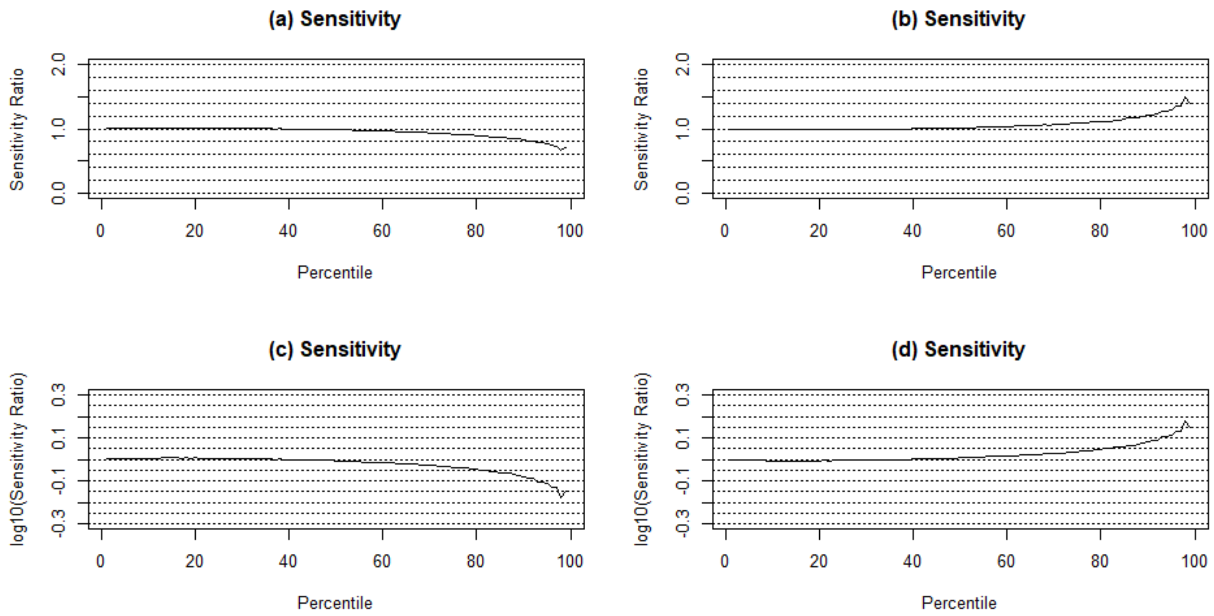

Figure 3.2. Comparison of Sensitivity Ratio Plots With and Without a Log10 Scale.

Plots (a) and (b) give untransformed sensitivity ratios, and plots (c) and (d) give log10 sensitivity ratios. Plots (a) and (c) are generated from a multiple-group population where $l^*_R = 4.2$, $l^*_F = 4.8$, $t^*_R = t^*_F = 0$, $y^*_R = y^*_F = .06$, $k_R = k_F = 0$, and $F_R = F_F = 1$. Plots (b) and (d) are generated from a multiple-group population where $l^*_R = 4.8$, $l^*_F = 4.2$, $t^*_R = t^*_F = 0$, $y^*_R = y^*_F = .06$, $k_R = k_F = 0$, and $F_R = F_F = 1$. Given that the only difference between these two populations is which group has $l^* = 4.2$ and which has $l^* = 4.8$, we know that the absolute value of the between-group differences in sensitivity at any given percentile is the same in both populations, with the only difference being which group is favored. If we examine the plots with untransformed sensitivity ratios, (a) and (b), it appears as though the reference group is more strongly favored in panel (b) than the focal group is in panel (a), as the departure from a flat line at $i_{Sens} = 1$ is greater in panel (b) than in panel (a). This is, of course, an artifact of the fact that $|\frac{x_1}{x_2} - 1| > |\frac{x_2}{x_1} - 1|$ when $x_1 > x_2$. A ratio of 1.5 indicates that the numerator is 50% greater than the denominator, but a ratio of .5 indicates that the denominator is 100% greater than the numerator. When ratios are placed on a log10 scale, however, the magnitude of departure from a flat line at $\log_{10}(i_{Sens}) = 0$ is the same in both panels (c) and (d), differing only in terms of which population is favored. See *Table 3.3* for conversions between $\imath$ and $\log_{10}(\jmath)$.

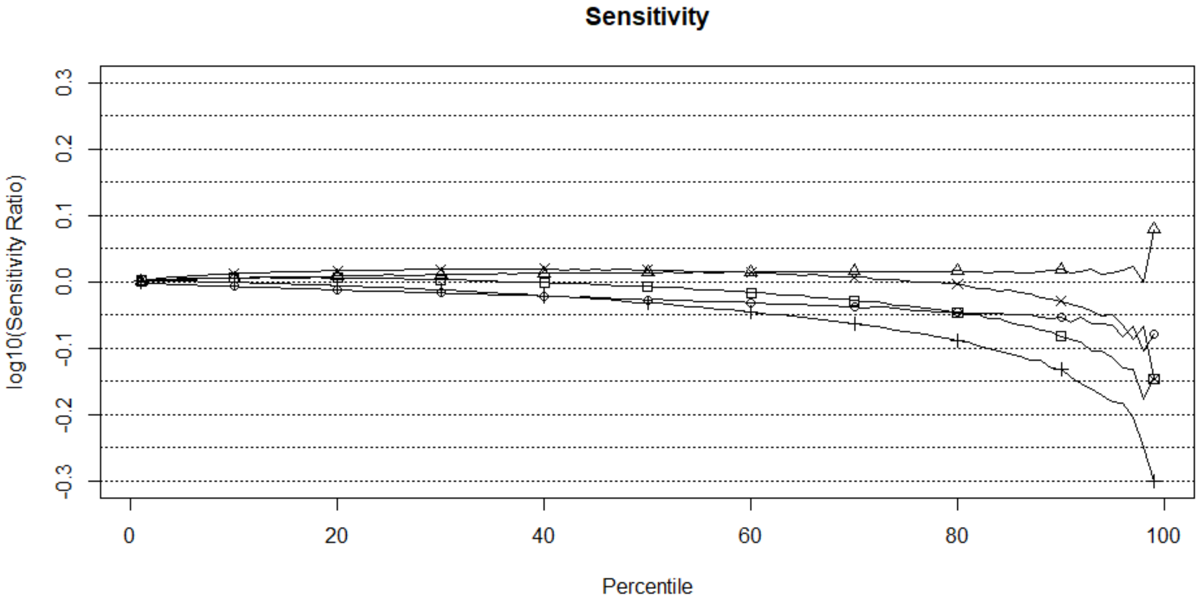

Figure 3.3. Population Log$_{10}$ Sensitivity Ratio Plots Under Different Violations of Measurement Invariance.

Note that $\lambda_{Sens} = \frac{\eta_{Sens,R}}{\eta_{Sens,F}}$, meaning positive values of $\log_{10}(\lambda_{Sens})$ reflect greater sensitivity in the reference group, and negative values of $\log_{10}(\lambda_{Sens})$ reflect greater sensitivity in the focal group. Lines are supplemented with an identifying glyph at the 1st, 99th, and every tenth percentile. The line with squares (□) gives $\log_{10}(\lambda_{Sens})$ in *condition 1*. The line with circles ( ) gives $\log_{10}(\lambda_{Sens})$ in *condition 2*. The line with triangles ( ) gives $\log_{10}(\lambda_{Sens})$ in *condition 3*. The line with plus signs (+) gives $\log_{10}(\lambda_{Sens})$ in *condition 4*. The line with x's (x) gives $\log_{10}(\lambda_{Sens})$ in *condition 5*. Population parameters for all conditions are given in *Table 3.4*.

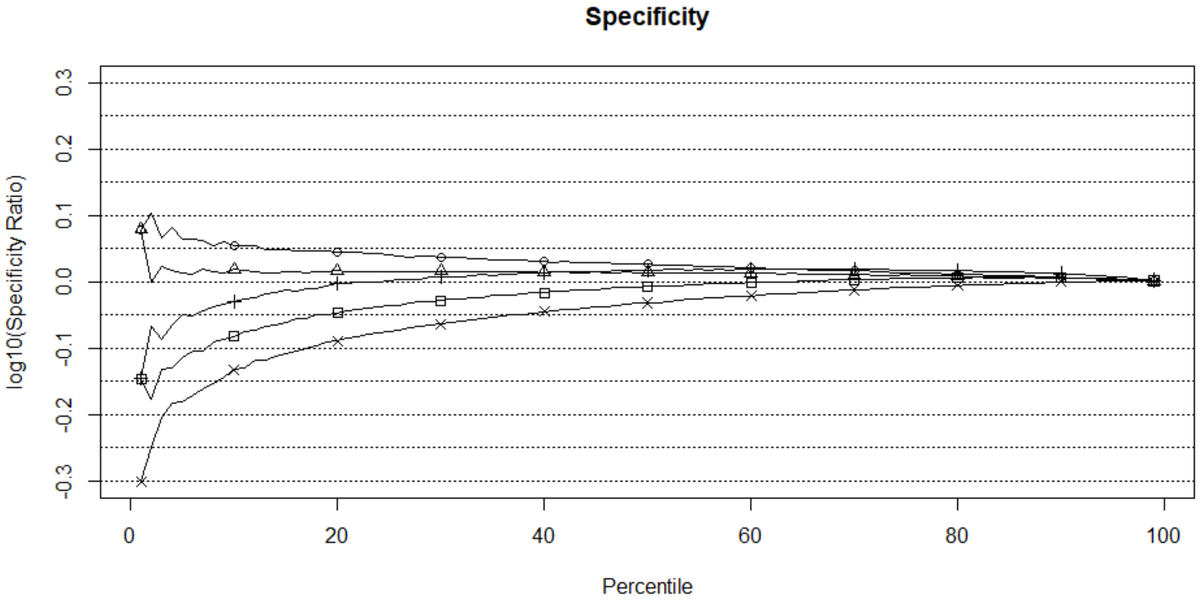

Figure 3.4. Population Log$_{10}$ Specificity Ratio Plots Under Different Violations of Measurement Invariance.

Note that $i_{Spec} = \frac{\eta_{Spec,R}}{\eta_{Spec,F}}$, meaning positive values of $\log_{10}(i_{Spec})$ reflect greater specificity in the reference group, and negative values of $\log_{10}(i_{Spec})$ reflect greater specificity in the focal group. Lines are supplemented with an identifying glyph at the 1st, 99th, and every tenth percentile. The line with squares (□) gives $\log_{10}(i_{Spec})$ in *condition 1*. The line with circles $\log_{10}(i_{Spec})$ in *condition 2*. The line with triangles ( ) gives $\log_{10}(i_{Spec})$ in *condition 3*. The line with plus signs (+) gives $\log_{10}(i_{Spec})$ in *condition 4*. The line with x's (x) gives $\log_{10}(i_{Spec})$ in *condition 5*. Population parameters for all conditions are given in *Table 3.4*.

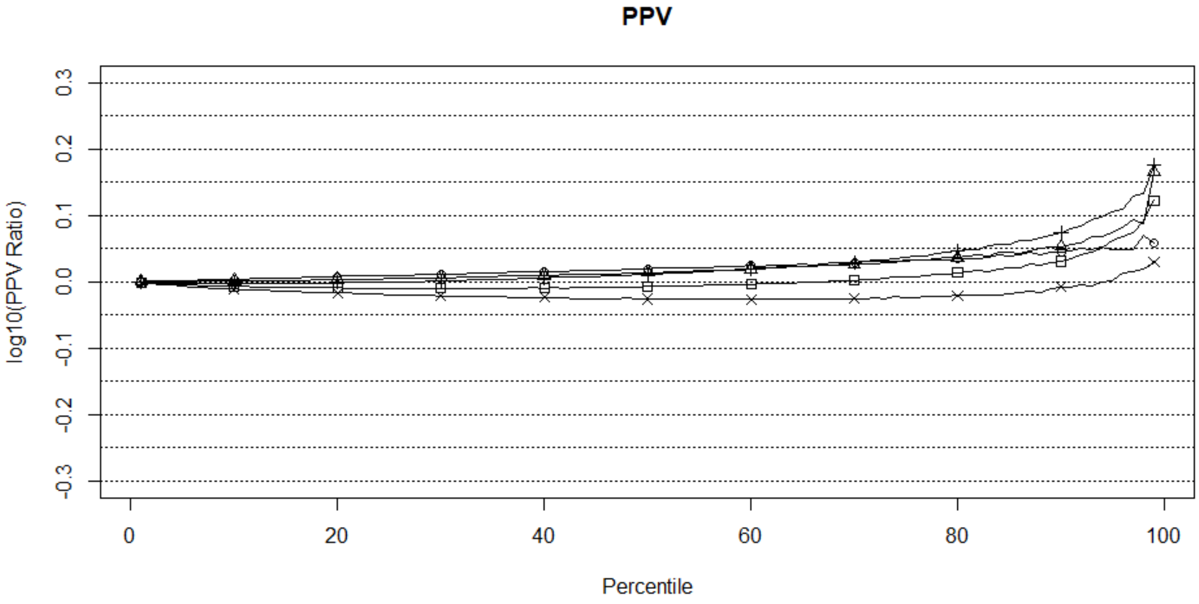

Figure 3.5. Population Log₁₀ Positive Predictive Value Ratio Plots Under Different Violations of Measurement Invariance.

Note that $j_{PPV} = \frac{\eta_{PPV,R}}{\eta_{PPV,F}}$, meaning positive values of $\log_{10}(j_{PPV})$ reflect greater PPV in the reference group, and negative values of $\log_{10}(j_{PPV})$ reflect greater PPV in the focal group. Lines are supplemented with an identifying glyph at the 1st, 99th, and every tenth percentile. The line with squares (□) gives $\log_{10}(j_{PPV})$ in *condition 1*. The line with circles (○ gives $\log_{10}(j_{PPV})$ in *condition 2*. The line with triangles ( ) gives $\log_{10}(j_{PPV})$ in *condition 3*. The line with plus signs (+) gives $\log_{10}(j_{PPV})$ in *condition 4*. The line with x's (×) gives $\log_{10}(j_{PPV})$ in *condition 5*. Population parameters for all conditions are given in *Table 3.4*.

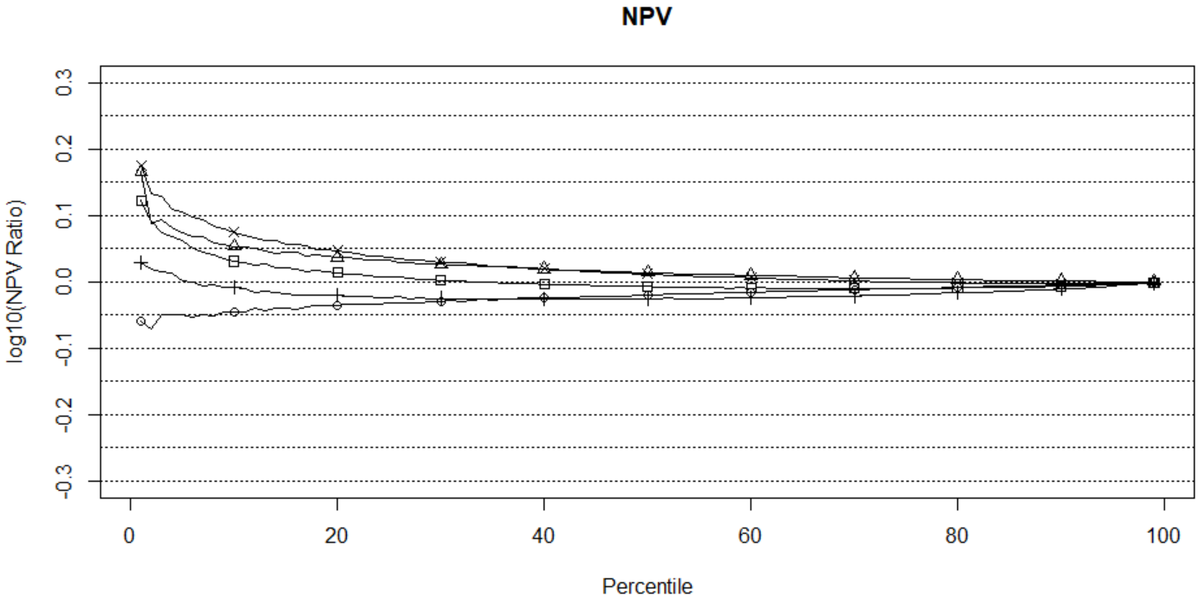

Figure 3.6. Population Log$_{10}$ Negative Predictive Value Ratio Plots Under Different Violations of Measurement Invariance.

Note that $i_{NPV} = \frac{\eta_{NPV,R}}{\eta_{NPV,F}}$, meaning positive values of $\log_{10}(i_{Sens})$ reflect greater sensitivity in the reference group, and negative values of $\log_{10}(i_{Sens})$ reflect greater sensitivity in the focal group. Lines are supplemented with an identifying glyph at the 1st, 99th, and every tenth percentile. The line with squares (□) gives $\log_{10}(i_{NPV})$ in *condition 1*. The line with circles $\log_{10}(i_{NPV})$ in *condition 2*. The line with triangles ( ) gives $\log_{10}(i_{NPV})$ in *condition 3*. The line with plus signs (+) gives $\log_{10}(i_{NPV})$ in *condition 4.* The line with x's (x) gives $\log_{10}(i_{NPV})$ in *condition 5.* Population parameters for all conditions are given in *Table 3.4*.

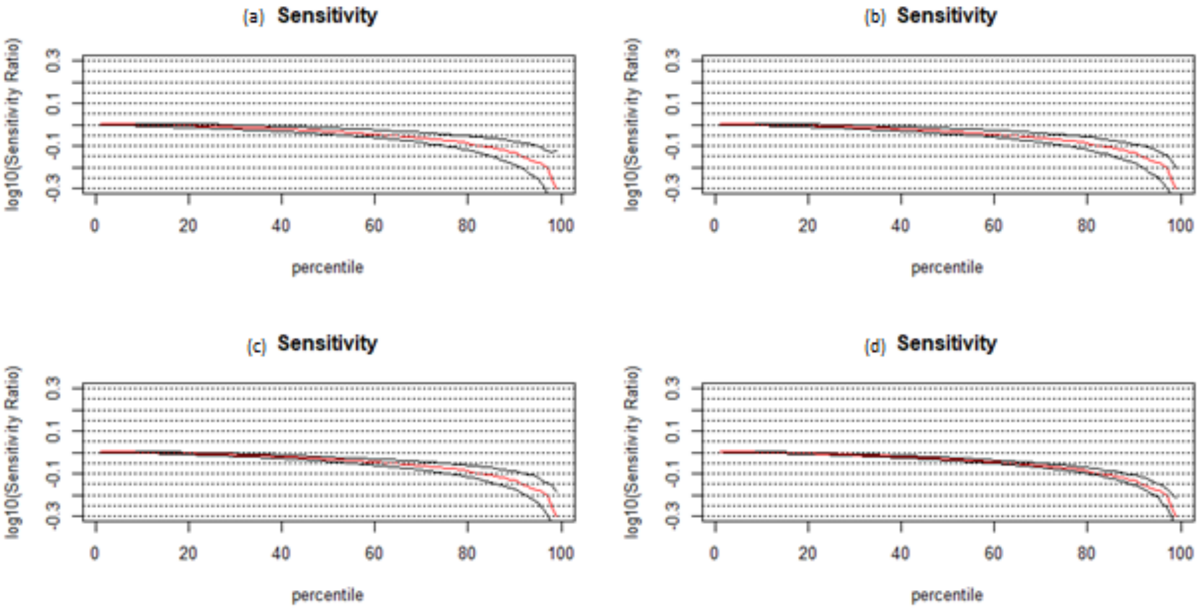

Figure 3.7. Variability of Finite Sample Estimates of Classification Accuracy Ratio Plots When Loadings and Intercepts are Greater in the Focal Group and Reliability is Moderate.

In each panel, the top and bottom curves give the estimated 95% confidence intervals around $\log_{10}(j_{Sens})$ at each percentile, while the red curve between them gives the population $\log_{10}(j_{Sens})$ plot. Note that $j_{Sens} = \frac{\eta_{Sens,R}}{\eta_{Sens,F}}$, meaning positive values of $\log_{10}(j_{Sens})$ reflect greater sensitivity in the reference group, and negative values of $\log_{10}(j_{Sens})$ reflect greater sensitivity in the focal group. Population parameter totals used in data generation are $l_R^* = 4.2$, $l_F^* = 4.8$, $t_R^* = 0$, $t_F^* = .6$, $y_R^* = y_F^* = 3.06$, $k_R = k_F = 0$, and $F_R = F_F = 1$, with a population mixing proportion of .5. Sample sizes represented in panels (a) through (d) are $N_1 = N_2 = 200$, $N_1 = N_2 = 300$, $N_1 = N_2 = 500$, and $N_1 = N_2 = 1000$, respectively. Population reliabilities are $a_R = .85$ and $a_F = .88$.

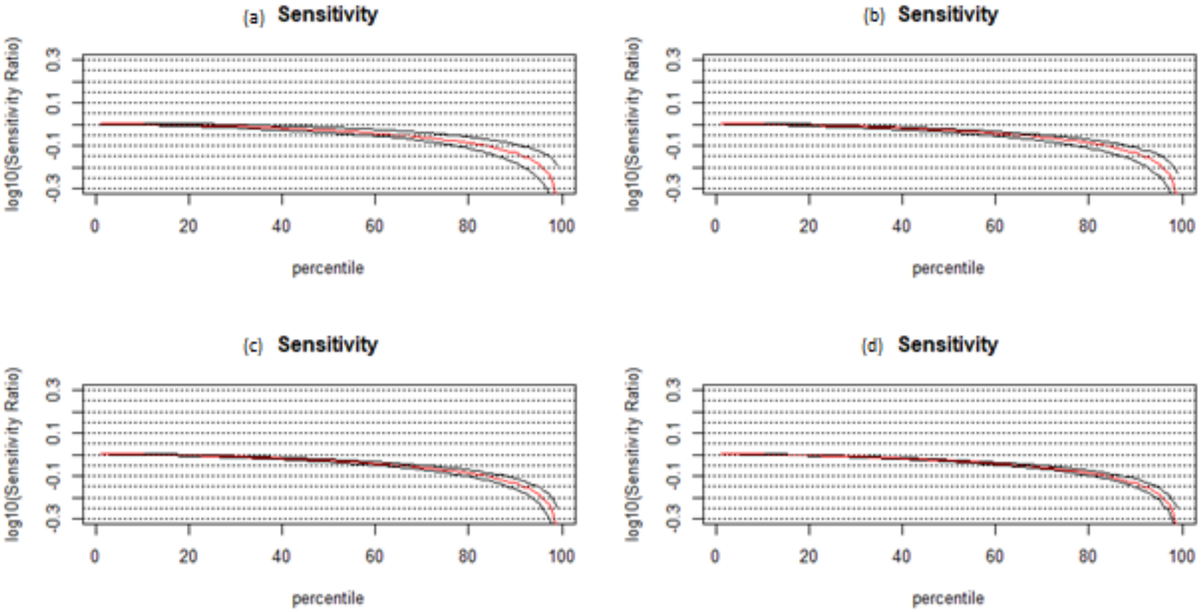

Figure 3.8. Variability of Finite Sample Estimates of Classification Accuracy Ratio Plots When Loadings and Intercepts are Greater in the Focal Group and Reliability is High.

In each panel, the top and bottom curves give the estimated 95% confidence intervals around $\log_{10}(\varphi_{Sens})$ at each percentile, while the red curve between them gives the population $\log_{10}(\varphi_{Sens})$ plot. Note that $\varphi_{Sens} = \frac{\eta_{Sens,R}}{\eta_{Sens,F}}$, meaning positive values of $\log_{10}(\varphi_{Sens})$ reflect greater sensitivity in the reference group, and negative values of $\log_{10}(\varphi_{Sens})$ reflect greater sensitivity in the focal group. Population parameter totals used in data generation are $\lambda_R^* = 4.2$, $\lambda_F^* = 4.8$, $\tau_R^* = 0$, $\tau_F^* = .6$, $\psi_R^* = \psi_F^* = 1.26$, $\kappa_R = \kappa_F = 0$, and $\Phi_R = \Phi_F = 1$, with a population mixing proportion of .5. Sample sizes represented in panels (a) through (d) are $N_1 = N_2 = 200$, $N_1 = N_2 = 300$, $N_1 = N_2 = 500$, and $N_1 = N_2 = 1000$, respectively. Population reliabilities are $\alpha_R = .93$ and $\alpha_F = .95$.

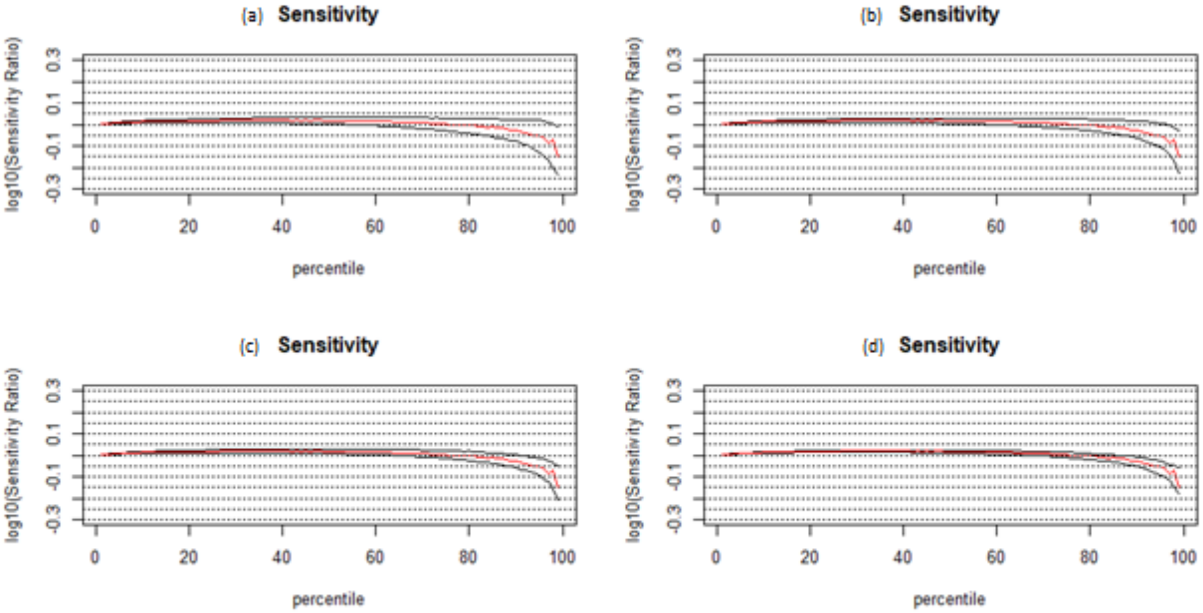

Figure 3.9. Variability of Finite Sample Estimates of Classification Accuracy Ratio Plots When Loadings are Greater in the Focal Group, Intercepts are Greater in the Reference Group, and Reliability is Moderate.

In each panel, the top and bottom curves give the estimated 95% confidence intervals around $\log_{10}(\varphi_{Sens})$ at each percentile, while the red curve between them gives the population $\log_{10}(\varphi_{Sens})$ plot. Note that $\varphi_{Sens} = \frac{\eta_{Sens,R}}{\eta_{Sens,F}}$, meaning positive values of $\log_{10}(\varphi_{Sens})$ reflect greater sensitivity in the reference group, and negative values of $\log_{10}(\varphi_{Sens})$ reflect greater sensitivity in the focal group. Population parameter totals used in data generation values are $\lambda^*_R = 4.2$, $\lambda^*_F = 4.8$, $\tau^*_R = 0$, $\tau^*_F = .6$, $\psi^*_R = \psi^*_F = 3.06$, $\kappa_R = \kappa_F = 0$, and $\Phi_R = \Phi_F = 1$, with a population mixing proportion of .5. Sample sizes represented in panels (a) through (d) are $N_1 = N_2 = 200$, $N_1 = N_2 = 300$, $N_1 = N_2 = 500$, and $N_1 = N_2 = 1000$, respectively. Population reliabilities are $\alpha_R = .85$ and $\alpha_F = .88$.

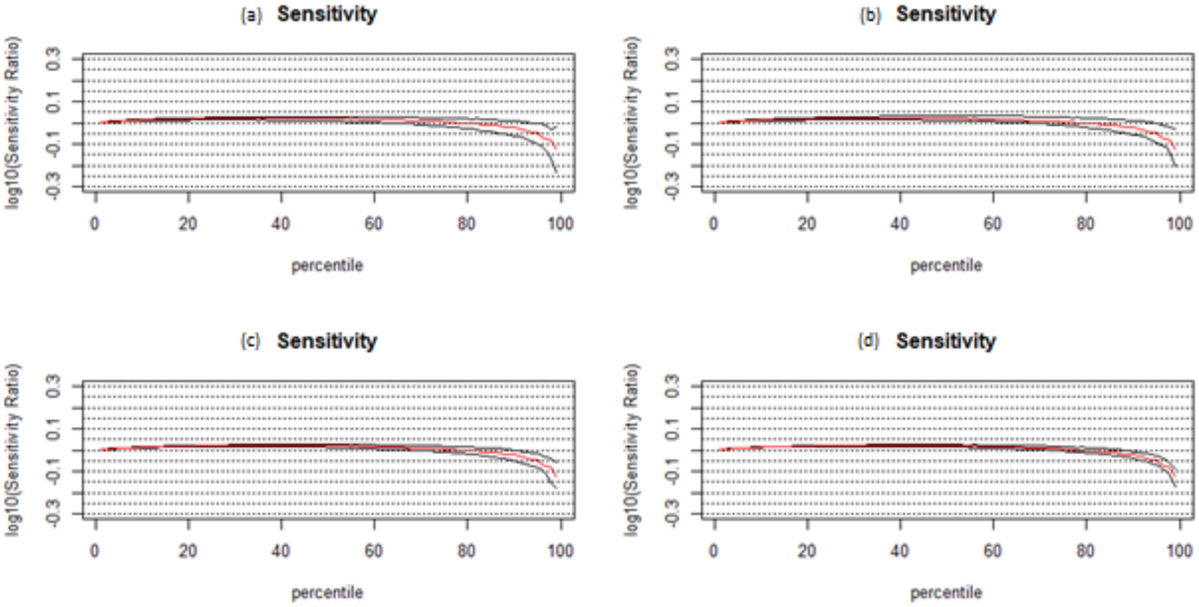

Figure 3.10. Variability of Finite Sample Estimates of Classification Accuracy Ratio Plots When Loadings are Greater in the Focal Group, Intercepts are Greater in the Reference Group, and Reliability is High.

In each panel, the top and bottom curves give the estimated 95% confidence intervals around $\log_{10}(\varphi_{Sens})$ at each percentile, while the red curve between them gives the population $\log_{10}(\varphi_{Sens})$ plot. Note that $\varphi_{Sens} = \frac{\eta_{Sens,R}}{\eta_{Sens,F}}$, meaning positive values of $\log_{10}(\varphi_{Sens})$ reflect greater sensitivity in the reference group, and negative values of $\log_{10}(\varphi_{Sens})$ reflect greater sensitivity in the focal group. Population parameter totals used in data generation values are $\lambda^*_R = 4.2$, $\lambda^*_F = 4.8$, $\tau^*_R = 0$, $\tau^*_F = .6$, $\psi^*_R = \psi^*_F = 3.06$, $\kappa_R = \kappa_F = 0$, and $\Phi_R = \Phi_F = 1$, with a population mixing proportion of .5. Sample sizes represented in panels (a) through (d) are $N_1 = N_2 = 200$, $N_1 = N_2 = 300$, $N_1 = N_2 = 500$, and $N_1 = N_2 = 1000$, respectively. Population reliabilities are $\alpha_R = .93$ and $\alpha_F = .95$.

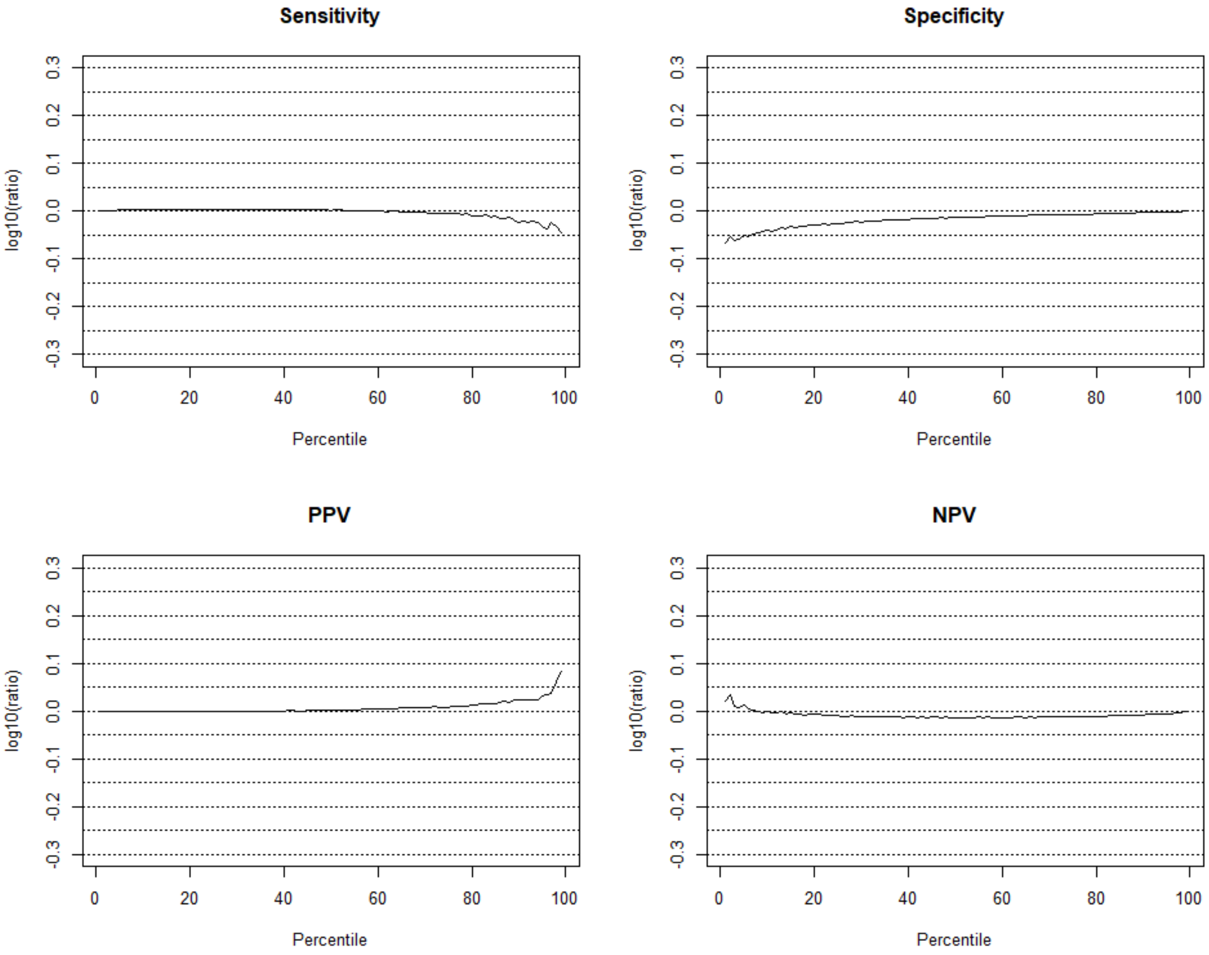

*Figure 3.11. Estimated Classification Accuracy Ratio Plots f*or Millsap and Kwok's Analysis of Everson et al.'s (1991) Test Anxi

Note that $\hat{\imath} = \frac{\hat{\eta}_R}{\hat{\eta}_F}$, meaning positive values of $\log_{10}(\hat{\jmath})$ reflect greater measures of classification accuracy in the reference group, women, and negative values of $\log_{10}(\,)\!\jmath$ reflect greater measures of classification accuracy in the focal group, men. Estimated population parameter totals are $\hat{l}^*_R = 7.279$, $\hat{l}^*_F = 7.679$, $\hat{t}^*_R = 16.621$, $\hat{t}^*_F = 16.581$, $\hat{y}^*_R = 4.329$, $\hat{y}^*_F = 3.405$, $\hat{\mathsf{F}}_R = .544$, $\hat{\mathsf{F}}_F = .477$, $k_R = 0$, and $\hat{k}_F = \theta.126$, with a population mixing proportion of .5. Estimated reliabilities in each group are $\hat{a}_R = .869$ and $\hat{a}_F = .892$.

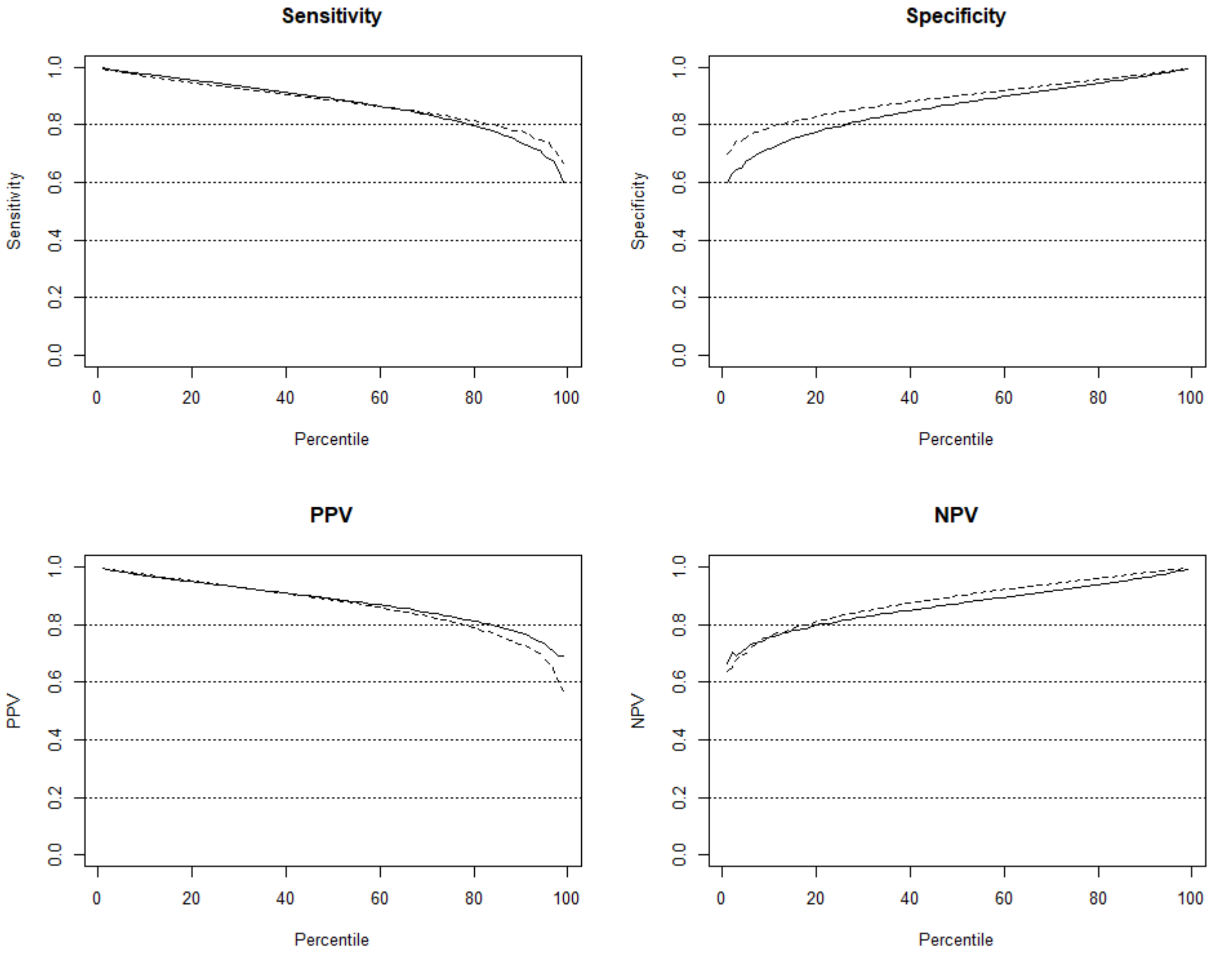

Figure 3.12. Estimated Classification Accuracy Plots in Each Group for Millsap and Kwok's (2004) Analysis of Everson et al.'s (1991) Te

Solid lines give estimated measures of classification accuracy in the reference group, women, and dashed lines give estimated measures of classification accuracy in the focal group, men. Estimated population parameter totals are $\hat{l}^*_R$ =7.279 , $\hat{l}^*_F$ =7.679 , $\hat{t}^*_R$ =16.621 , $\hat{t}^*_F$ =16.581 , $\hat{y}^*_R$ =4.329 , $\hat{y}^*_F$ =3.405 , $\hat{F}_R$ =.544 , $\hat{F}_F$ =.477 , $\hat{k}_R$ =0 , and $\hat{k}_F$ = θ.126 , with a population mixing proportion of .5. Estimated reliabilities in each group are $\hat{a}_R$ =.869 and $\hat{a}_F$ =.892 .

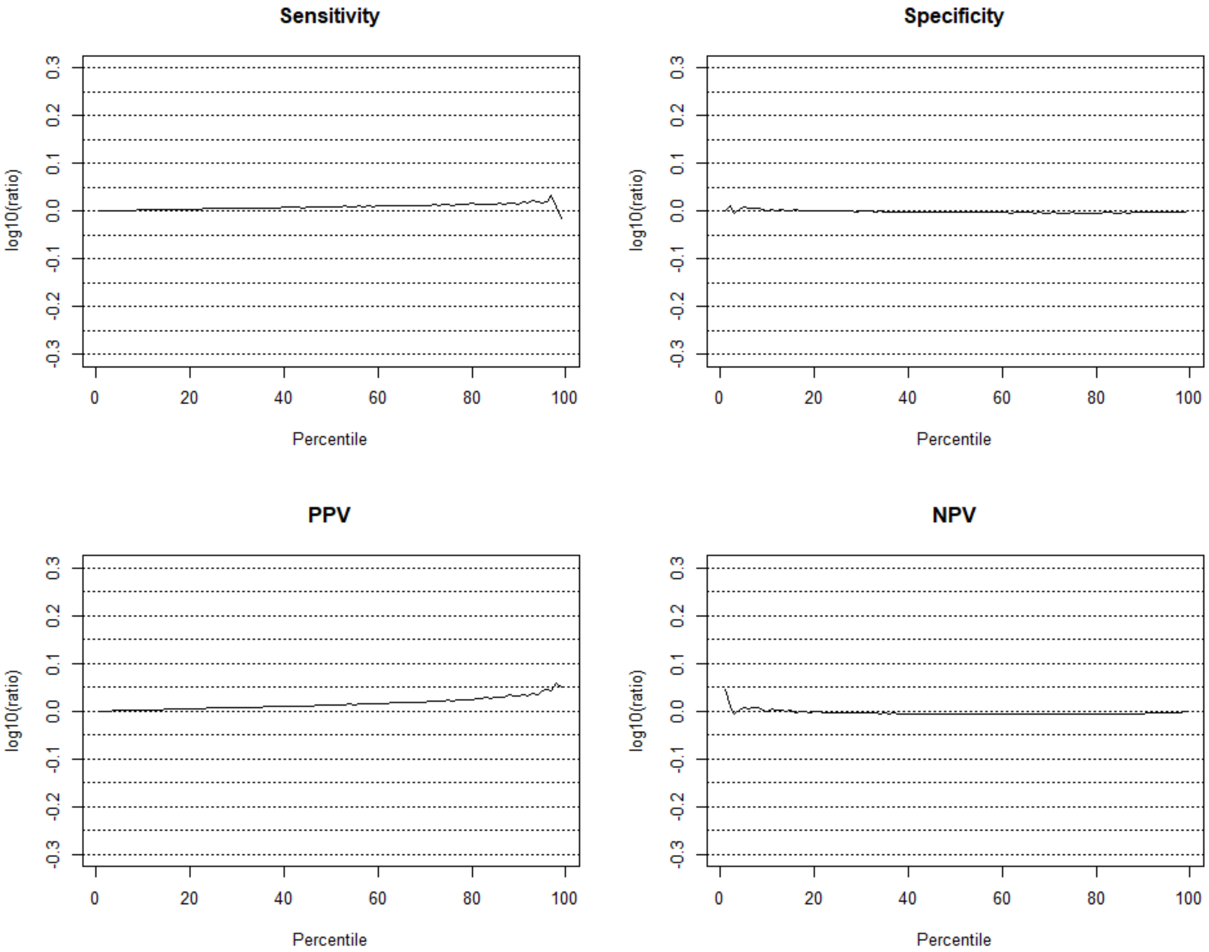

*Figure 3.13. Estimated Classification Accuracy Ratio Plots f*or Millsap and Kwok's Analysis of Everson et al.'s (1991) Test Anxi *When Measurement Invariance is Assumed.*

Note that $\hat{\imath} = \frac{\hat{\eta}_R}{\hat{\eta}_F}$, meaning positive values of $\log_{10}(\hat{\imath})$ reflect greater measures of classification accuracy in the reference group, women, and negative values of $\log_{10}(\hat{\imath})$ reflect greater measures of classification accuracy in the focal group, men. Estimated population parameter totals are $\hat{l}^* = 7.454$, $\hat{t}^* = 16.604$, and $\hat{y}^* = 3.925$, $\hat{F}_R = .544$, $\hat{F}_F = .477$, $\hat{k}_R = 0$, and $\hat{k}_F = -0.126$, with a population mixing proportion of .5. Estimated reliabilities in each group under the assumption of measurement invariance are $\hat{a}_R = .885$ and $\hat{a}_F = .871$.

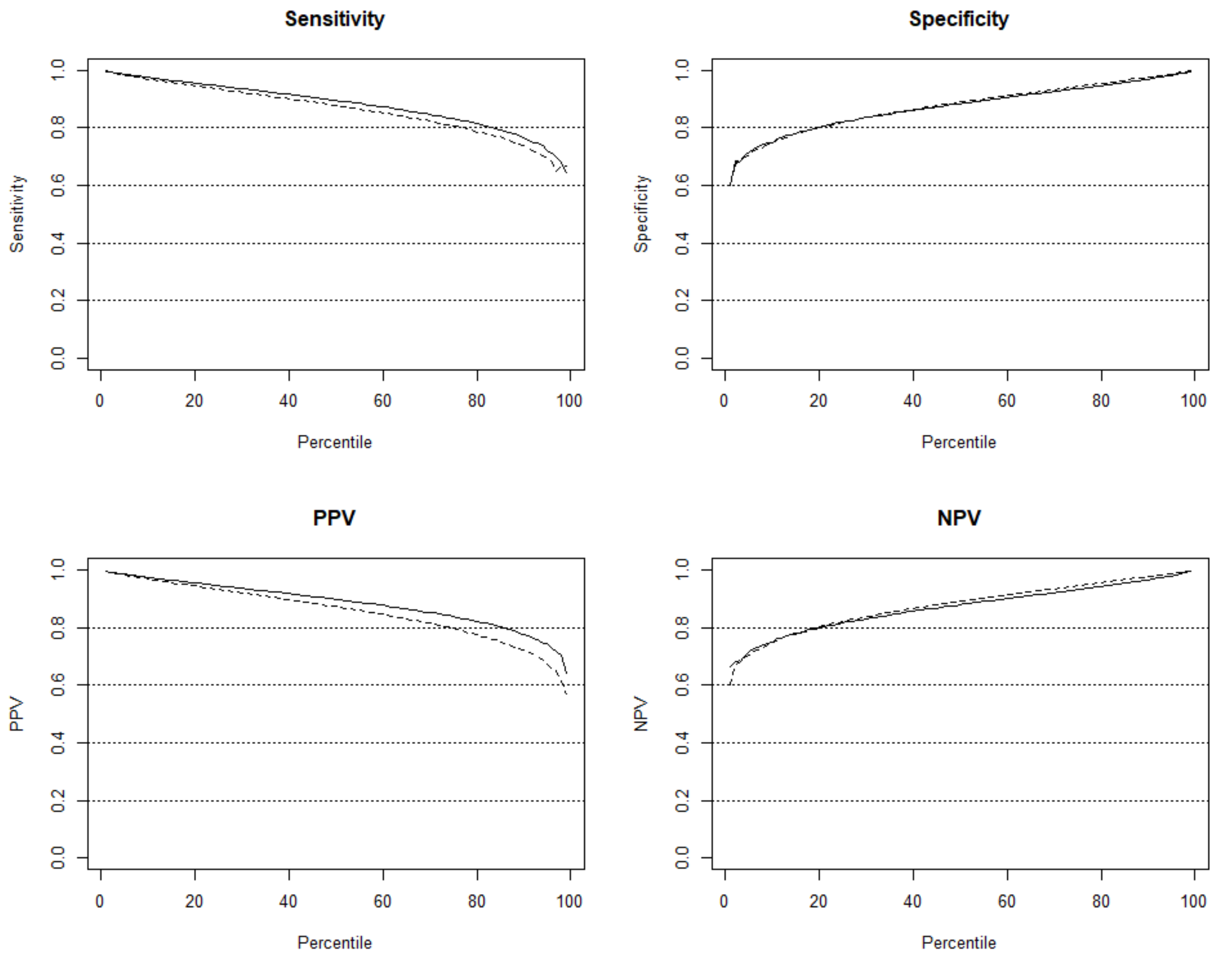

Figure 3.14. Estimated Classification Accuracy Plots in Each Group for Millsap and Kwok *(2004) Analysis of Evers*on et al.'s (1991) Test Anxiety Inv *When Measurement Invariance is Assumed.*

Solid lines give estimated measures of classification accuracy in the reference group, women, and dashed lines give estimated measures of classification accuracy in the focal group, men. Estimated population parameter totals are $\hat{l} = 7.454$, $\hat{t}^{*} = 16.604$, and $\hat{y}^{*} = 3.925$, $\hat{F}_R = .544$, $\hat{F}_F = .477$, $k_R = 0$, and $\hat{k}_F = -0.126$, with a population mixing proportion of .5. Estimated reliabilities in each group under the assumption of measurement invariance are $\hat{a}_R = .885$ and $\hat{a}_F = .871$.

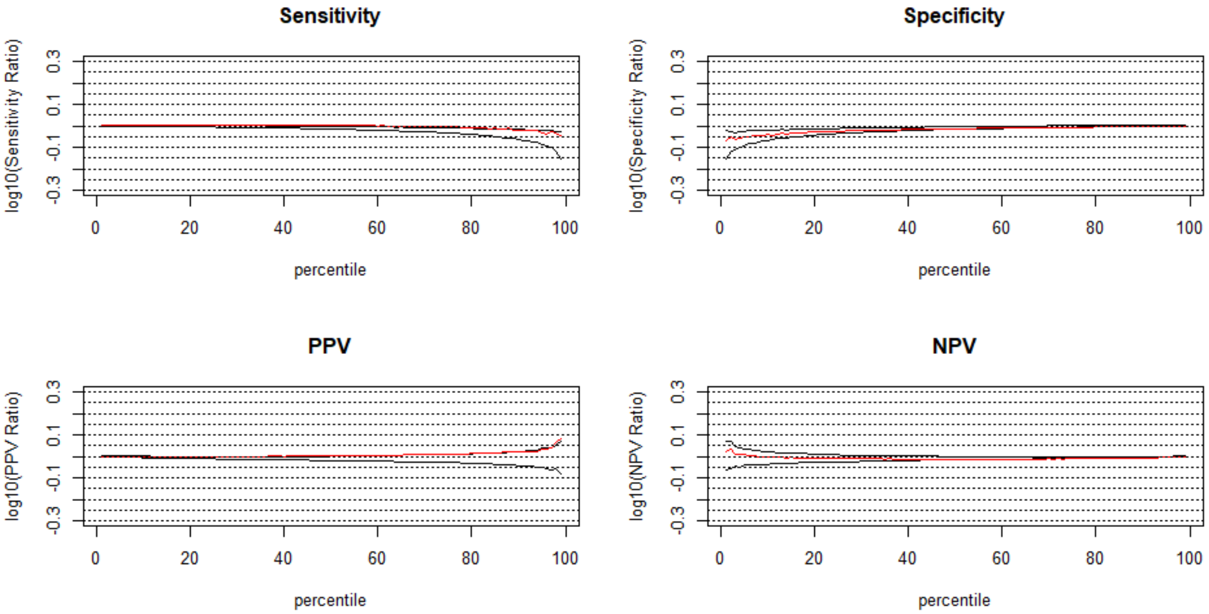

Figure 3.15. Finite Sample Classification Accuracy Ratio Plots Generated by Using Bollen-Stine Bootstrapping on Everson et al.'s (1991) Data.

In each panel, the top and bottom curves give the estimated 95% confidence intervals around $\log_{10}(\hat{l})$ at each percentile, while the red curve between them gives the population $\log_{10}(\hat{l})$ plot. Note that $\hat{l} = \frac{\hat{\eta}_R}{\hat{\eta}_F}$, meaning positive values of $\log_{10}(\hat{l})$ reflect greater measures of classification accuracy in the reference group, women, and negative values of $\log_{10}(\hat{l})$ reflect greater measures of classification accuracy in the focal group, men. Estimated population parameter totals are $\hat{l}^* = 7.454$, $\hat{t}^* = 16.604$, and $\hat{y}^* = 3.925$, $\hat{F}_R = .544$, $\hat{F}_F = .477$, $k_R = 0$, and $\hat{k}_F = -0.126$, with a population mixing proportion of .5. Estimated reliabilities in each group are $\hat{a}_R = .869$ and $\hat{a}_F = .892$. R code for generating this plot is included in *Appendix I.*

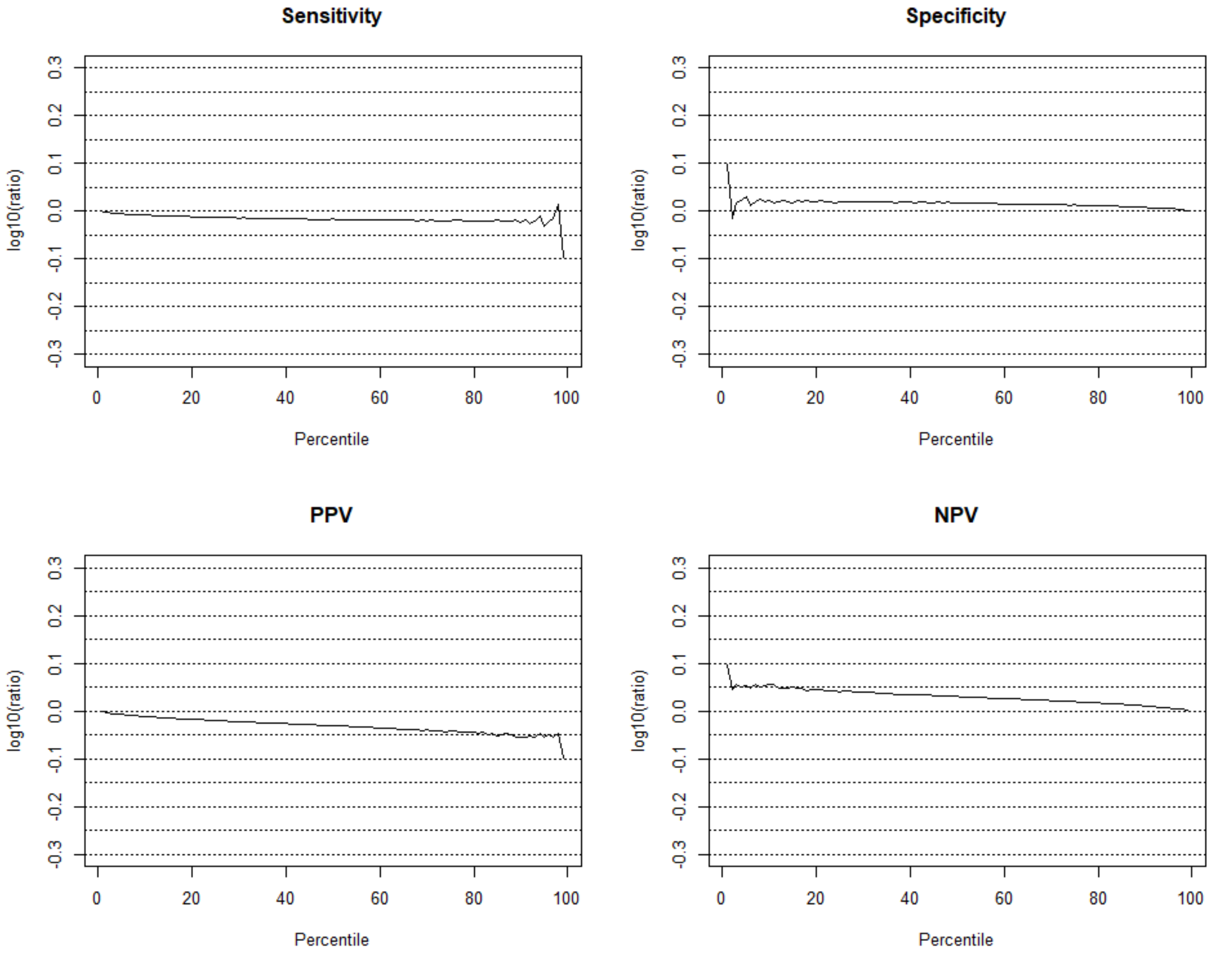

Figure 3.16. Population Classification Accuracy Ratio Plots for Measurement Invariant Populations With Unequal Latent Means.

Note that $\hat{\iota} = \frac{\hat{\eta}_R}{\hat{\eta}_F}$, meaning positive values of $\log_{10}(\hat{\iota})$ reflect greater measures of classification accuracy in the reference group, and negative values of $\log_{10}(\hat{\iota})$ reflect greater measures of classification accuracy in the focal group. Population parameter totals are $\lambda_R^* = \lambda_F^* = 4.2$, $\tau_R^* = \tau_F^* = 0$, $\psi_R^* = \psi_F^* = 3.06$, $\kappa_R = 0$, $\kappa_F = .5$, and $\Phi_R = \Phi_F = 1$;, with a population mixing proportion of .5.

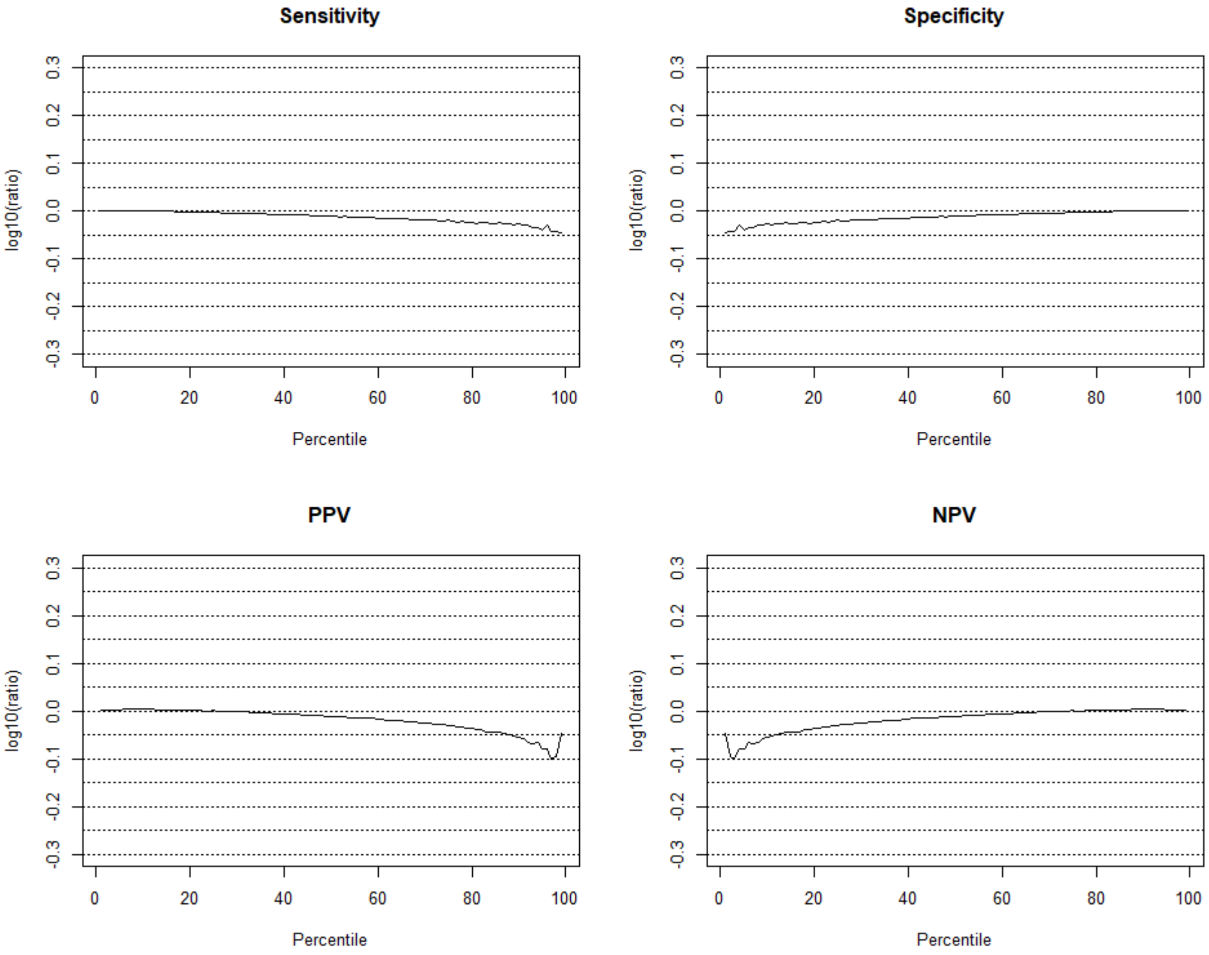

Figure 3.17. Population Classification Accuracy Ratio Plots for Measurement Invariant Populations With Unequal Latent Variances.

Note that $\hat{\imath} = \frac{\hat{\eta}_R}{\hat{\eta}_F}$, meaning positive values of $\log_{10}(\hat{\jmath})$ reflect greater measures of classification accuracy in the reference group, and negative values of $\log_{10}(\hat{\jmath})$ reflect greater measures of classification accuracy in the focal group. Population parameter totals are $l_R^* = {}_F^* l = 4.2$, $t_R^* = {}_F^* t = 0$, $y_R^* = y_F^* = 3.06$, $k_R = {}_F k = 0$, $F_R = 1$, and $F_F = 1.5$, with a population mixing proportion of .5.

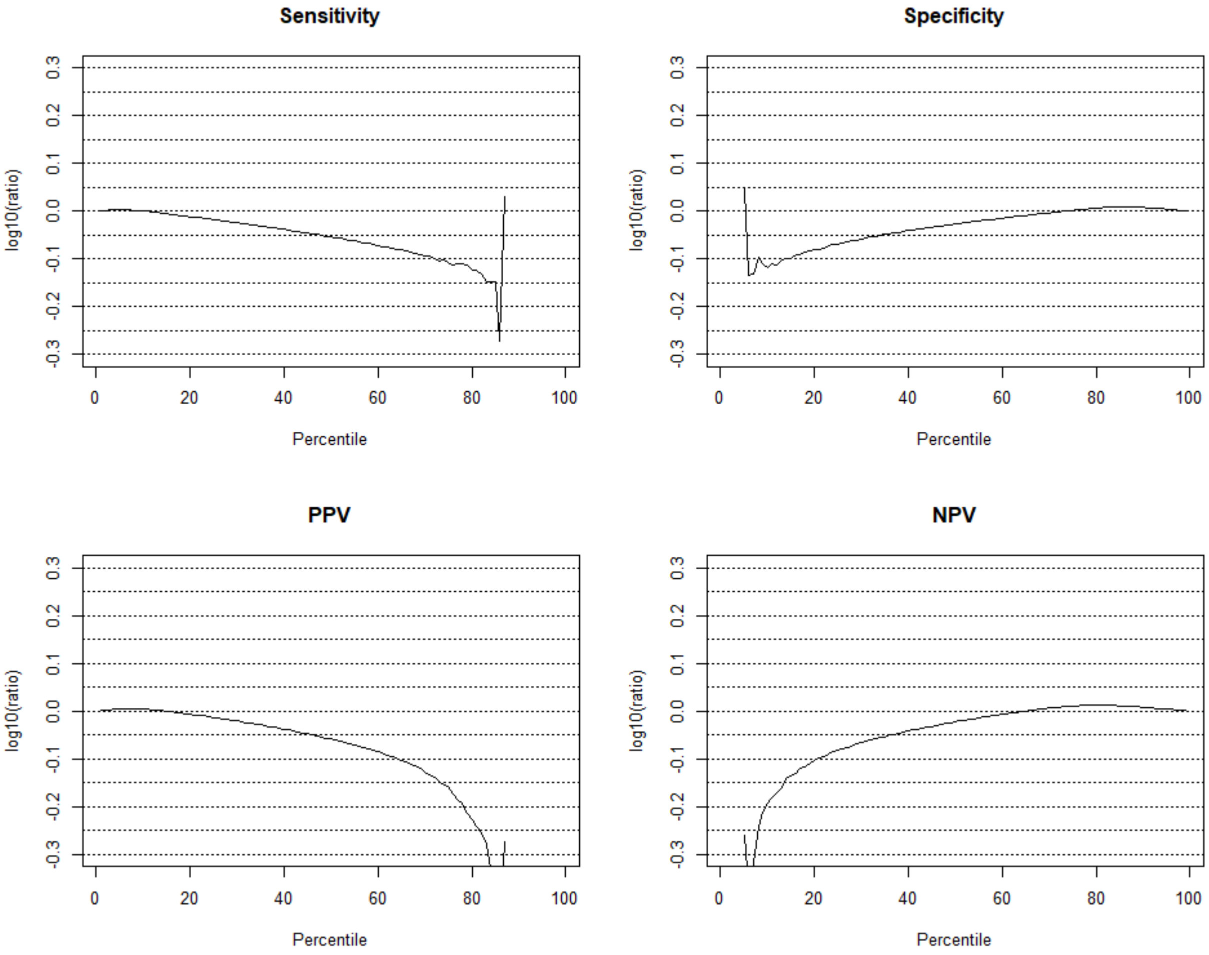

Figure 3.18. Population Classification Accuracy Ratio Plots for Measurement Invariant Populations With Extremely Unequal Latent Means and Latent Variances.

Note that $\hat{\imath} = \frac{\hat{\eta}_R}{\hat{\eta}_F}$, meaning positive values of $\log_{10}(\hat{\imath})$ reflect greater measures of classification accuracy in the reference group, and negative values of $\log_{10}(\hat{\imath})$ reflect greater measures of classification accuracy in the focal group. Population parameter totals are $l_R^* = l_F^* = 4.2$, $t_R^* = t_F^* = 0$, $y_R^* = y_F^* = 3.06$, $k_R = 0$, $k_F = 1$, $F_R = 1$, and $F_F = 10$, with a population mixing proportion of .5.

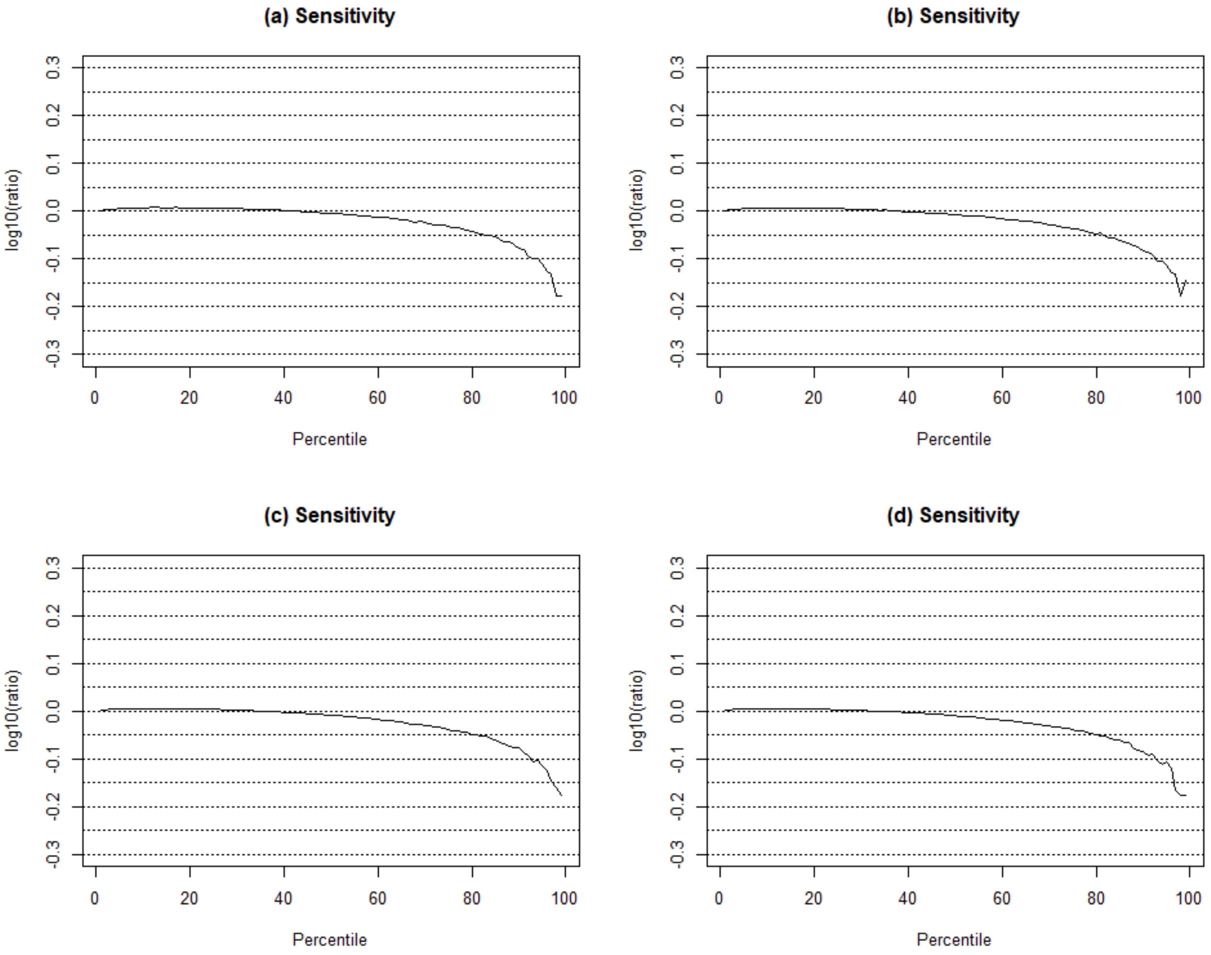

Figure 3.19. Population Sensitivity Ratio Plots as A Function of Reliability.

Note that $\hat{\imath} = \frac{\hat{\eta}_R}{\hat{\eta}_F}$, meaning positive values of $\log_{10}(\hat{\imath})$ reflect greater measures of classification accuracy in the reference group, and negative values of $\log_{10}(\hat{\imath})$ reflect greater measures of classification accuracy in the focal group. Population parameter totals are $l_R^* = 4.2$, $l_F^* = 4.8$, $t_R^* = t_F^* = 0$, $k_R = k_F = 0$, and $F_R = F_F = 1$, with residual variances varying across panels. In panels (a) through (d), residual variances are $y_R^* = y_F^* = 1$, $y_R^* = y_F^* = 3$, $y_R^* = y_F^* = 5$, and $y_R^* = y_F^* = 7$, respectively. Reliabilities in the reference group are .94, .85, .78, and .72, respectively, and reliabilities in the focal group are .96, .88, .82, and .77, respectively.

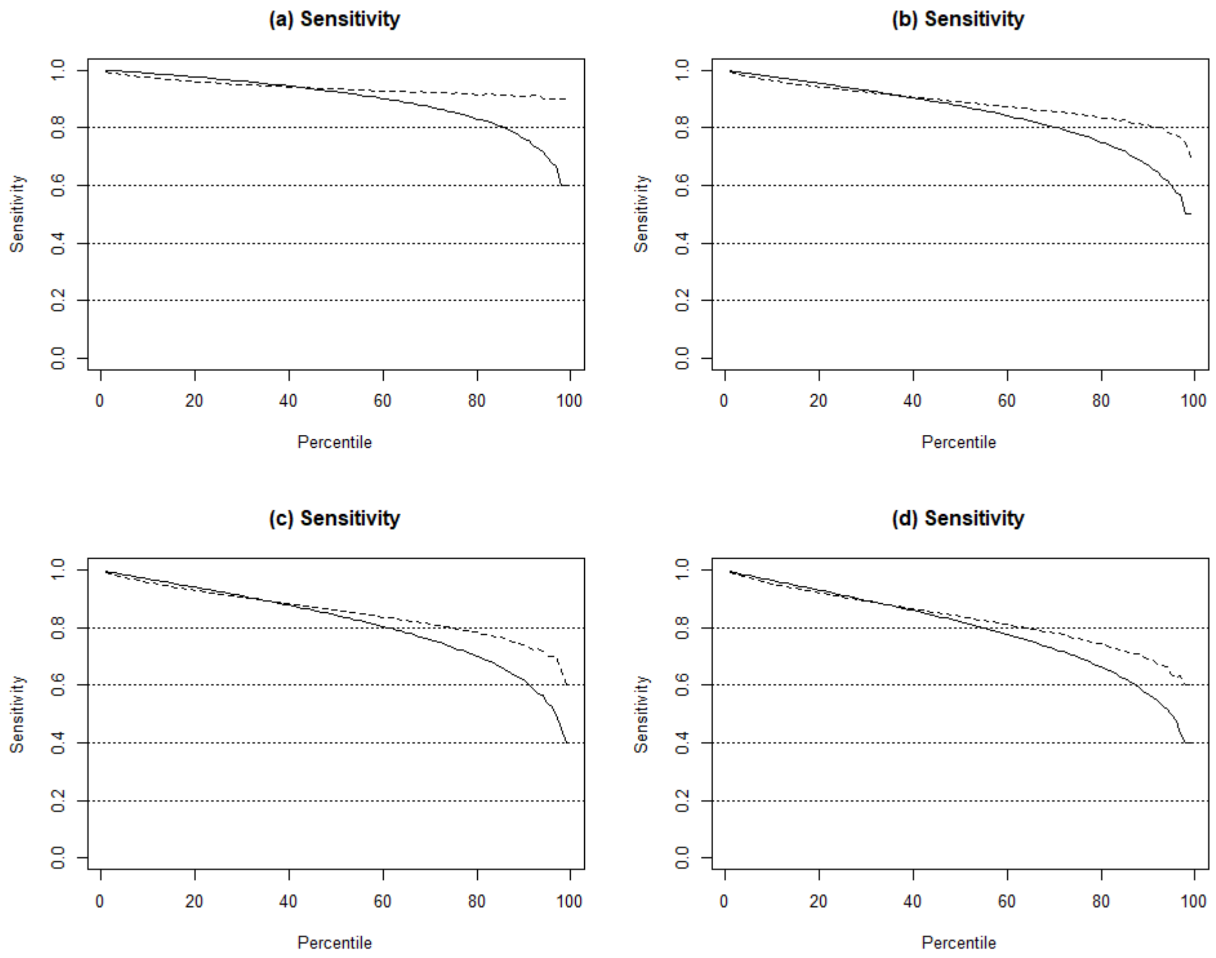

Figure 3.20. Population Sensitivity Plots for Each Group as a Function of Reliability.

Solid lines give population measures of classification accuracy in the reference group, women, and dashed lines give population measures of classification accuracy in the focal group, men. Population parameter totals are $\lambda_R^* = 4.2$, $\lambda_F^* = 4.8$, $\tau_R^* = \tau_F^* = 0$, $\kappa_R = \kappa_F = 0$, and $\Phi_R = \Phi_F = 1$, with residual variances varying across panels. In panels (a) through (d), residual variances are $\psi_R^* = \psi_F^* = 1$, $\psi_R^* = \psi_F^* = 3$, $\psi_R^* = \psi_F^* = 5$, and $\psi_R^* = \psi_F^* = 7$, respectively. Reliabilities in the reference group are .94, .85, .78, and .72, respectively, and reliabilities in the focal group are .96, .88, .82, and .77, respectively.

Chapter 4: On the Use of RMSEA in the Comparison of Nested Measurement Invariance Models

4.1. Introduction

Measurement invariance (MI) is the property of psychometric instruments indicating equivalent functioning when applied across different groups (Millsap, 2011). When psychometric instruments that violate MI are used to make comparisons across groups, observed differences in scores may be spurious, and true differences may be attenuated (Millsap & Kwok, 2004; Steenkamp & Baumgartner, 1998; Steinmetz, 2013). This is because when MI does not hold, observed scores are a function of both unobservable true scores on the construct of interest, and group-specific variations in measurement properties. Thus, in order to use a psychometric instrument to make comparisons across groups, verifying MI is essential.

In the psychological sciences, MI is most frequently statistically evaluated within a structural equation modelling (SEM) framework called multiple group confirmatory factor analysis (MGCFA). Evaluating MI with MGCFA involves fitting the same confirmatory factor analytic (CFA) model to data across multiple groups, and subsequently testing between-group equality constraints on measurement parameter estimates (e.g., factor loadings, indicator intercepts, residual variances). If equality constraints lead to a significant loss of model fit, the null hypothesis of MI is rejected.

Fit of MI models –and SEMs in general –is typically evaluated by chi-square tests of model fit, which are limited in that they evaluate a null hypothesis of exact fit. This is a limitation because any negligibly small misspecification in an otherwise well specified model will lead to model rejection with a sufficiently large sample size. Given this issue, many researchers prefer to evaluate model fit by goodness-of-fit indices (GFI's)–alternative measures of model fit which quantify the degree to which a given model fits the data. Most GFIs are designed to be independent of sample size and model complexity. Unfortunately for researchers interested in MI, most GFIs are developed with the intention of evaluating fit of a single model, and not comparing nested models, which is the framework in which measurement invariance is typically evaluated (Vandenberg & Lance, 2000).

There exists a small literature on adapting GFIs to nested model comparison (Chen, 2007; Cheung & Rensvold, 2002), advocating for the use of $\Delta GFIs$, computed as $\Delta GFI = GFI_A - GFI_B$. For a pair of nested models, M_A and M_B, where M_A is nested within M_B, GFI_A is the measured fit on a particular GFI when M_A is fit to data, and GFI_B is the measured fit on a particular GFI when M_B is fit to data. Thus, $\Delta GFIs$ are simply the arithmetic difference in model fit when a pair of nested models are fit to the same data. As will be discussed, $\Delta GFIs$ are limited by the fact that they do not necessarily retain the same metric as the GFIs they are extensions of. As an alternative to $\Delta GFIs$, I propose $RMSEA_D$, an adjustment to the root-mean-square error of approximation (RMSEA) which quantifies change in model fit in terms of change in the minimized fit function value per added degree of freedom. $RMSEA_D$ retains the original RMSEA metric, allowing us to use familiar critical values (Browne & Cudeck, 1993; Hu & Bentler, 1999; MacCallum, Browne & Sugawara, 1996) to evaluate fit of constraints, produce a confidence interval around the sample estimate, and conduct tests of close or not-close fit (Browne & Cudeck, 1993). While this discussion primarily pertains to the RMSEA, the logic should extend to other GFIs based on the minimized fit function value. Further, while this research is motivated by facilitating MI analysis, its logic should apply to any comparison of nested SEMs.

4.1.1. The Issue of Model Fit

Fit of MGCFA models, as discussed in Chapter 1, is typically evaluated by chi-square tests. Fit of the baseline configural invariance model is assessed by the overall chi-square test of model fit, with the test statistic computed as $T_{ML} = \hat{F}_{ML}(\frac{N - G}{G})$, where $\hat{F}_{ML}$ is the minimized value of the maximum likelihood (ML) fit function, *N* is the total sample size, and *G* is the number of groups. T_{ML} is tested against a central chi-square distribution with *df* degrees of freedom. A non-significant *p*-value is interpreted as an absence of evidence against invariance, and thus, the null hypothesis of MI is retained. Fit of more restrictive MI models is then assessed via chi-square difference tests, where the test statistic is computed as $D_{ML} = T_{ML,A} - T_{ML,B}$, with $T_{ML,A}$ and $T_{ML,B}$ being the chi-square test statistics associated with the more restrictive model, M_A , and the less restrictive model, M_B, respectively. D_{ML} is then tested against a central chi-square distribution with $df_D = df_A - df_B$ degrees of freedom, where df_A and df_B are the degrees of freedom associated with M_A and M_B, respectively. If D_{ML} is not statistically significant, the null hypothesis that constrained parameters are invariant is retained. Chi-square difference tests are typically performed sequentially, such that the test of configural invariance is followed by a chi-square difference test evaluating the misfit introduced by the metric invariance model, which is subsequently followed with a difference test evaluating misfit introduced by the scalar invariance model, and so on.

A major limitation to evaluating model fit by the chi-square is that it tests a null hypothesis of perfect fit, meaning even trivial misspecifications can lead to the rejection of an otherwise adequate model, particularly when sample size and model degrees of freedom are large (Browne & Cudeck, 1993; Chen, Curran, Bollen, Kirby & Paxton, 2008). Large samples are frequently necessary for evaluations of MI, as minimum sample size requirements —which increase with p —must be met for all studied populations. Further, evaluations of MI frequently involve models with large degrees of freedom, as *df* increases multiplicatively as a function of *G*. Despite its sensitivity to trivial misspecifications, the chi-square remains the default index of model fit in SEM literature and SEM software (e.g., *lavaan* (Rosseel, 2012), *Mplus* (Muthen & Muthen, 1998-2017), *EQS* (Bentler, 2006)).

4.1.2. Goodness-of-Fit Indices

In response to the issue of negligibly small misspecifications leading to model rejection in large samples, psychometric researchers have developed goodness-of-fit indices (GFIs) which are meant to quantify the degree to which a given model does or does not fit the data. In SEM, many GFIs are transformations of the ML fit function minimum, $\hat{F}_{ML}$, onto metrics meant to be directly interpretable and independent of sample size and model degrees of freedom. Such GFIs include the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) the Comparative Fit Index (CFI; Bentler, 1990), and the Root-Mean-Square Error of Approximation (RMSEA; Steiger & Lind, 1980). Unfortunately, little guidance exists in terms of how to use and interpret GFIs for the purposes of nested model comparison, which is critical in the evaluation of MI.

Some researchers have advocated the use of differences of GFIs, or $\Delta GFIs$, for the purposes of comparing fit of nested MI models (e.g., Chen, 2007; Cheung & Rensvold, 2002).

$\Delta GFIs$ are computed as the unweighted difference in fit indices associated with a pair of nested models, M_A and M_B where M_A is nested within M_B, such that $\Delta GFI = GFI_A - GFI_B$, where GFI_A and GFI_B are the measured fit of M_A and M_B, respectively, on a particular GFI. While use of $\Delta GFIs$ to compare fit of MI models has become popular (e.g., Carter & Perez, 2016; Hukkelberg, 2016; Innstrand, 2016), there are associated issues researchers should be made aware of before using them.

The primary limitation of $\Delta GFIs$ is that they do not necessarily preserve their original metrics as transformations of $\hat{F}_{ML}$ onto a particular GFI scale. To illustrate: there are infinitely many pairs of fit function minima, $\hat{F}_{ML,A}$ and $\hat{F}_{ML,B}$, for a given pair of nested models, that can produce a particular value of $\Delta RMSEA_{ML} = RMSEA_{A,ML} - RMSEA_{B,ML}$. Consider a pair of nested models, M_A and M_B, where M_A is nested within M_B, $df_A = 15$, and $df_B = 10$. If $RMSEA_{B,ML} = .01$ and $RMSEA_{A,ML} = .04$ in the population, then $\hat{F}_{ML,B} = 10 * .01^2 = .001$ and $\hat{F}_{ML,A} = 15 * .04^2 = .024$, in which case $\Delta RMSEA_{ML} = .03$ corresponds to a .023 increase in $\hat{F}_{ML}$. Now, compare this to a pair of nested models with the same degrees of freedom, but where $RMSEA_{B,ML} = .04$ and $RMSEA_{A,ML} = .07$ in the population. In this case, $\hat{F}_{ML,B} = 10 * .04^2 = .016$, and $\hat{F}_{ML,A} = 15 * .07^2 = .0735$, meaning $\Delta RMSEA_{ML} = .03$ corresponds to an .0575 increase in $\hat{F}_{ML}$. Similar variation in $\hat{F}_{ML}$ occurs when $RMSEA_{A,ML}$ and $RMSEA_{B,ML}$ are held constant, but df_A and df_B are varied. For example, if $RMSEA_{B,ML} = .01$ and $RMSEA_{A,ML} = .04$, but $df_B = 20$ and $df_A = 25$, then $\hat{F}_{ML,B} = 20 * .01^2 = .002$ and $\hat{F}_{ML,A} = 25 * .04^2 = .040$, in which case $\Delta RMSEA_{ML} = .03$ corresponds to a .038 increase in $\hat{F}_{ML}$. In each of these examples,

$\Delta RMSEA_{ML} = .03$ and $df_D = df_B - df_A = 5$, but the amount of misfit introduced by M_A above and beyond that already present under M_B varies from example to example. Thus, whether a given magnitude of $\Delta RMSEA_{ML}$ could be considered large is necessarily contingent on df_A, df_D, and $RMSEA_{ML,A}$. This contrasts with single model RMSEA computations, where a given RMSEA value and df are associated with a single value of $\hat{F}_{ML}$. This property is what allows RMSEA to have its conventional interpretation as the amount of misfit present per model degree of freedom. Conversely, given its many dependencies, $\Delta RMSEA_{ML}$ can only be interpreted as an unweighted mean difference in misfit for a pair of nested models.

There are several negative consequences to $\Delta GFIs$ failing to preserve the $\hat{F}_{ML}$ metric beyond their inconsistent and ambiguous interpretation. One such issue is that conventional criteria for model rejection or retention (e.g., Browne & Cudeck, 1993; Browne & Mels ,1990; Hu & Bentler, 1999; MacCallum, Browne & Sugawara, 1996; Steiger, 1989) can no longer be used, and new critical values should be developed. Chen (2007) proposed a simultaneous $\Delta RMSEA_{ML} > .01$ and $\Delta CFI_{ML} < -.005$ for model rejection when $N \leq 300$, and a simultaneous $\Delta RMSEA_{ML} > .015$ and $\Delta CFI_{ML} < -.01$ for model rejection when $N > 300$. The author, however, notes that these criteria should be used with caution, as a considerable number of factors affect the magnitude of $\Delta GFIs$, including number of non-invariant parameters, whether or not non-invariant parameters are consistently greater in one group, sample size, sample size ratio between groups, and scale length (Chen, 2007). The author notes that their specific recommendations do not necessarily generalize to model conditions outside of the scope of their simulation, suggesting that simulations varying other factors such as latent variances or the number of latent variables are necessary if researchers wish to apply $\Delta GFIs$ appropriately. This variation in the recommended critical values of $\Delta GFIs$ is not surprising when we recall that whether a given value of $\Delta RMSEA$ could be interpreted as large is necessarily contingent on df_A, df_D, and $RMSEA_{A,ML}$.

Further limitations to $\Delta GFIs$ pertain exclusively to the RMSEA, the most popular GFI in the behavioral sciences. Specifically, it is not clear how to compute a confidence interval around $\Delta RMSEA_{ML}$, or how to use it to compute a test of close or not-close fit (Browne & Cudeck, 1993). The ability to easily compute a confidence interval or test of close/not-close fit for a single $RMSEA_{ML}$ is related to the fact that $RMSEA_{ML}$ values –given *df, N,* and *G* –are associated with a single non-central chi-square distribution (MacCallum, Browne, & Sugawara, 1996). Thus, it is straightforward to convert computed $RMSEA_{ML}$ values to non-centrality parameters for non-central chi-square distributions, and critical percentiles on those distributions can be converted back to the $RMSEA_{ML}$ metric (See *Appendix J*). In the case of $\Delta RMSEA_{ML}$, because there are two non-central, non-independent chi-square distributions involved in its computation, it is not immediately intuitive, if possible, how one would convert $\Delta RMSEA_{ML}$ back to the chi-square metric in order to compute confidence interval or test of close fit.

4.2 Recommended Alternative to $\Delta GFIs$: $RMSEA_D$

In the population, the ML root-mean-square error of approximation (RMSEA) is given by:

$$RMSEA_{ML} = \sqrt{\frac{G * \hat{F}_{ML}}{df}} \text{, (Steiger \& Lind, 1980; Steiger, 1998)} \qquad (4.1)$$

The sample estimate of the RMSEA is given by:

$$RMSEA_{ML,N} = \sqrt{\frac{\hat{F}_{ML}}{df} - \frac{1}{(\frac{N-G}{G})}} = \sqrt{\frac{T_{ML}-df}{df(\frac{N-G}{G})}} = \sqrt{\frac{\hat{\lambda}_{ML}}{df(\frac{N-G}{G})}}, \qquad (4.2)$$

where N is the total sample size, and G is the number of groups being studied. Note that $\frac{N-G}{G}$ simplifies to $N-1$ when fitting single-group models. The RMSEA quantifies the average contribution to $\hat{F}_{ML}$ per degree of freedom when the specified model is not correct (Browne & Cudeck, 1993; Steiger, 1998; Steiger & Lind,

by the fact that a confidence interval around its finite sample estimate can easily be obtained. This property also permits the computation of tests of close and not-close fit (See *Appendix J*).

Steiger, Shapiro and Browne (1985) demonstrated that when nested SEMs are properly specified, $D_{ML} = T_{ML,A} - T_{ML,B}$ follows a central chi-square distribution with degrees of freedom $df_D = df_A - df_B$. Further, when M_B is correctly specified, but M_A is not, D_{ML} follows a non-central chi-square distribution whose non-centrality parameter is estimated as $\hat{\lambda}_{D,ML} = D_{ML} - df_D$ (Steiger, Shapiro & Browne, 1985). Given that the sample RMSEA computation as defined in Equation (4.2) is a transformation of the estimated non-centrality parameter for a non-central chi-square distribution, and D_{ML} follows a non-central chi-square distribution whose non-centrality parameter is estimated as the difference between the observed test statistic and degrees of freedom, Equation (4.2) can also be used to quantify the amount of non-centrality introduced by M_A above and beyond that already present in M_B, such that

$$RMSEA_{D,ML,N} = \sqrt{\frac{D_{ML}-df_D}{(\frac{N-G}{G})df_D}} = \sqrt{\frac{\hat{\lambda}_{D,ML}}{(\frac{N-G}{G})df_D}}. \qquad (4.3)$$

To express $RMSEA_{D,ML,N}$ in terms of fit function minimums, if $T_{ML,A} = \hat{F}_{ML,A}(N - G)$ and $T_{ML,B} = \hat{F}_{ML,B}(N - G)$ in multiple groups,

$D_{ML} = T_{ML,A} - T_{ML,B} = \hat{F}_{ML,A}(N - G) - \hat{F}_{ML,B}(N - G) = (N - G)(\hat{F}_{ML,A} - \hat{F}_{ML,B})$, and

$$RMSEA_{D,ML,N} = \sqrt{\frac{(N-G)(\hat{F}_{ML,A} - \hat{F}_{ML,B}) - df_D}{df_D(\frac{N-G}{G})}} = \sqrt{\frac{G(\hat{F}_{ML,A} - \hat{F}_{ML,B})}{df_D} - \frac{1}{N-G}} \tag{4.4}$$

in finite samples. As N increases, $RMSEA_{D,ML,N}$ approaches the population value:

$$RMSEA_{D,ML} = \sqrt{\frac{G*(\hat{F}_{ML,A} - \hat{F}_{ML,B})}{df_D}} = \sqrt{\frac{G*(\hat{F}_{ML,D})}{df_D}}. \tag{4.5}$$

$RMSEA_{D,ML}$, as given in Equation (4.5), has a number of useful properties that are absent from $\Delta RMSEA$. Of primary importance $RMSEA_{D,ML}$ retains the original RMSEA metric and its original interpretation: $RMSEA_{D,ML}$ is the amount of misfit introduced by additional constraints per introduced degree of freedom. This allows $RMSEA_{A,ML}$ –the $RMSEA$ associated with M_A, the more restrictive of a pair of nested models –to be broken down into two components: the amount of misfit uniquely introduced by the df_B degrees of freedom associated with M_B, and the amount of misfit uniquely introduced by the df_D degrees of freedom introduced by M_A above and beyond that already associated with M_B. Consider Equation (4.5):

$$RMSEA_{D,ML} = \sqrt{\frac{G*(\hat{F}_{ML,A} - \hat{F}_{ML,B})}{df_D}} = \sqrt{\frac{G*(\frac{df_A * RMSEA^2_{A,ML}}{G} - \frac{df_B * RMSEA^2_{B,ML}}{G})}{df_D}} = \sqrt{\frac{df_A * RMSEA^2_{A,ML} - df_B * RMSEA^2_{B,ML}}{df_D}},$$

which can be rearranged as

$df_B * RMSEA_B^2 + df_D * RMSEA_D^2 = df_A * RMSEA_A^2 = G*(\hat{F}_{ML,B} + \hat{F}_{ML,D}) = G*\hat{F}_{ML,A}$. Thus, we can think of $\frac{df_B * RMSEA_B^2}{G} = \hat{F}_{ML,B}$ as the contribution to $\hat{F}_{ML,A}$ by M_B, and $\frac{df_D * RMSEA_D^2}{G} = \hat{F}_{ML,D}$ as the contribution to $\hat{F}_{ML,A}$ by M_A above and beyond that already introduced by M_B, with $\frac{df_A * RMSEA_{ML,A}^2}{G} = \hat{F}_{ML,A}$ being the total misfit associated with M_A.

Given that $RMSEA_{D,ML}$, like $RMSEA_{ML}$, is a conversion of a non-central chi-square distribution to the *RMSEA* metric, it is identically distributed to the single model *RMSEA* for a given df, *G* and $\hat{F}_{ML}$. Therefore, new decision rules for model rejection are not necessarily needed. Researchers can retain or reject their constraints using the same criteria as when assessing fit of a single model by the RMSEA (e.g., See Browne and Cudeck (1993), Chen, Curran, Bollen, Kirby and Paxton (2008), Hu and Bentler (1999), and Yuan and Chan (2016) for differing perspectives on RMSEA decision rules). Further, because $RMSEA_{D,ML}$ is on the RMSEA metric, one can compute a confidence interval about $RMSEA_{D,ML}$, as well as tests of close and not-close fit using the logic of Browne and Cudeck (1993) (See *Appendix K*).

It should be noted that similar finite sample estimations of the RMSEA for the purposes of comparing nested models, as given in Equation (4.3), have previously appeared in the psychometric literature. Browne and Du Toit (1992) first defined Equation (4.3) as the Root Deterioration per Restriction (RDR) for the purposes of assessing change in fit for nested single-group models. McDonald and Ho (2002) also used the computation given in Equation (4.3) to partition fit of single-group structural equation models by the RMSEA into a component due to misfit in the measurement model, and a component due to misfit in the structural model. While this computation is theoretically useful for the purposes of assessing fit of any set of nested models, it has primarily only been cited in the context of decomposing model fit into measurement model and structural model components (e.g., O'Boyle and Wi

and Bentler (2011) proposed a similar computation to Equation (4.3), but recommended that $\hat{F}_{ML,D}$ be averaged across all df_A degrees of freedom associated with M_A, rather than the df_D degrees of freedom introduced by M_A above and beyond those already present under M_B. The present research is the first to extend this methodology to analyses of MI, as well as the first to extend it to multiple-group models in general.

4.2.1. Comparison of Population Values of $RMSEA_{D,ML}$ and $\Delta RMSEA_{ML}$

To further illustrate the utility of $RMSEA_{D,ML}$ relative to $\Delta RMSEA_{ML}$, population values are examined under three circumstances: 1) when M_A and M_B are both correctly specified, 2) when M_B is correctly specified and M_A is misspecified, and 3) when both M_A and M_B are misspecified.

4.2.1.1. When M_A and M_B are Both Correctly Specified

Consider two nested SEMs, M_A and M_B, where M_A is nested within M_B. In the case where both models are correctly specified, $E(RMSEA_{A,ML})=0$ and $E(RMSEA_{B,ML})=0$. If we assume $E(RMSEA_{D,ML})=\sqrt{\dfrac{df_A * E(RMSEA_{A,ML})^2 - df_B * E(RMSEA_{B,ML})^2}{df_D}}$, then

$E(RMSEA_{D,ML} \mid E(RMSEA_{A,ML})=0, E(RMSEA_{B,ML}=0), df_D > 0) = 0$. Similarly, if

$E(\Delta RMSEA_{ML}) = E(RMSEA_{A,ML}) - E(RMSEA_{B,ML})$, then

$E(\Delta RMSEA_{ML} \mid E(RMSEA_{A,ML}) = 0, E(RMSEA_{B,ML}) = 0) = 0$. As one might intuit, an RMSEA difference of 0 is expected for both methods when both M_A and M_B are correctly specified.

4.2.1.2. When M_B is Correctly Specified and M_A is Incorrectly Specified

The logic of difference testing is typically that once one has verified adequate fit of the baseline model, M_B, the next step is to evaluate fit of constraints introduced by M_A above and beyond those already present under M_B. In ideal cases where M_B is exactly true $E(RMSEA_{D,ML} \mid E(RMSEA_{B,ML}) = 0) = \sqrt{\frac{df_A * E(RMSEA_{A,ML})^2}{df_D}}$. Consider the case where $RMSEA_{B,ML} = 0$ and $RMSEA_{A,ML} = .05$ in the population. When $RMSEA_{A,ML} = .05$, the average contribution to $\hat{F}_{ML,A}$ by each of df_A degrees of freedom is $\frac{RMSEA_A{}^2}{G} = \frac{\hat{F}_{ML,A}}{df_A} = .00125$. When $RMSEA_B = 0$, the average contribution to $\hat{F}_{ML,B}$ by each of the first df_B degrees of freedom is $\frac{RMSEA_B{}^2}{G} = \frac{\hat{F}_{ML,B}}{df_B} = 0$. Thus, while the total misfit associated with M_A is $\hat{F}_{ML,A} = \frac{df_A * .0025}{G}$, we know that the first df_B degrees of freedom do not contribute to $\hat{F}_{ML,A}$. Therefore, 100% of $\hat{F}_{ML,A}$ is associated with the df_D constraints introduced by M_A above and beyond those already present under M_B. The average contribution to misfit by these df_D constraints is $\frac{G * \hat{F}_{ML,A}}{df_D} = RMSEA_D{}^2$.

If $\Delta RMSEA$ is considered under the same circumstances: $E(\Delta RMSEA_{ML} \mid E(RMSEA_{A,ML}) = 0) = E(RMSEA_{B,ML})$. Thus, when M_B is correctly specified, there is no logical difference in interpretation between $\Delta RMSEA_{ML}$ and $RMSEA_{A,ML}$. Therefore, $\Delta RMSEA_{ML}$ tells the researcher nothing specific about the misfit introduced by the df_D constraints imposed by M_A above and beyond those already associated with M_B, despite the fact that we know the amount of misfit associated with M_B and the first df_B degrees of freedom.

4.2.1.3. When M_A and M_B are Both Misspecified

In practice, when evaluating fit of nested models using *GFIs*, one is not looking for evidence that M_A fits perfectly, but for evidence that M_A fits adequately, as evidenced by the estimated value of the fit index falling above or below a certain threshold. When one retains M_A on the grounds that fit is adequate, but not perfect, higher order difference tests for which M_A is the new baseline will, thus, have a slightly misspecified baseline. Given that this scenario is more likely to be the norm than the exception (Edwards, 2013), it is important to explore the behavior of $\Delta RMSEA_{ML}$ and $RMSEA_{D,ML}$ under these sub-optimal circumstances.

Consider a pair of nested models, M_A and M_B, where M_A is nested within M_B, $RMSEA_{B,ML} = .01$, and $RMSEA_{A,ML} = .04$. The expected value of $\Delta RMSEA_{ML}$ under these circumstances is $E(\Delta RMSEA_{ML}) = .03$. But what does this number actually quantify? We know that the average root contribution of the first df_B degrees of freedom to $G * \hat{F}_{ML,B}$ is .01, and that the average root contribution of the first df_A degrees of freedom to $G * \hat{F}_{ML,A}$ is .04. $\Delta RMSEA_{ML} = .03$ can be interpreted as the change in the average contribution to $G * \hat{F}_{ML}$ by the first df_B degrees of freedom if we instead average across all df_A degrees of freedom. Despite the fact that we know all of this new misfit is associated with the second df_D degrees of freedom,

$\mathrm{D}RMSEA$ implicitly attributes it to all df_A degrees of freedom, and tells us nothing specific about the misfit introduced by the second df_D degrees of freedom.

Now consider comparing the same pair of models by $RMSEA_D$. Given a non-zero $RMSEA$ associated with both M_A and M_B, the expected value of $RMSEA_{D,ML}$ is $E(RMSEA_{D,ML}) = \sqrt{\dfrac{df_A * E(RMSEA_{A,ML})^2 - df_B * E(RMSEA_{B,ML})^2}{df_D}}$. The expected value of $RMSEA_{D,ML}$ given that $RMSEA_{B,ML} = .01$ and $RMSEA_{A,ML} = .04$, is dependent on df_A and df_B. If we know that the average root contribution to misfit by the df_A degrees of freedom associated with M_A is .04, and that the average root contribution to misfit by the first df_B degrees of freedom is .01, then the average root contribution to misfit by the second df_D degrees of freedom is necessarily greater than .04, with magnitude generally increasing as a function of increasing $\dfrac{df_B}{df_A}$. To illustrate, if $df_A = 100$ and $df_B = 10$, $RMSEA_{D,ML} = .042$. Alternatively, if $df_A = 100$ and $df_B = 90$, $RMSEA_{D,ML} = .123$. This makes sense if we consider that in the latter case, a .03 increase in the average contribution to misfit across a large df_A is accounted for by a small number of constraints, and therefore, these constraints must be severely misspecified. A useful index of change in model fit should be able to differentiate between these two cases, otherwise, it will fail to detect even large misspecifications with a sufficiently large df_B and sufficiently small df_D. This threat to power is particularly great in evaluations of MI, as MI models frequently have large degrees of freedom $-$increasing as a function of G $-$and investigations into MI frequently involve a small number of constraints when specifying nested models, such as during specification searches for the most parsimonious partial measurement invariance model (Byrne, Shavelson & Muthen, 1989), which often involve $df_D = 1$ comparisons.

4.3. Discussion

4.3.1. Summary of Advantages of RMSEA$_D$

Comparing fit of nested MGCFA models by $RMSEA_D$ has several advantages over using chi-square difference tests or $\Delta GFIs$:

1) The sensitivity of the chi-square difference test to a misspecified M_A increases with sample size and degrees of freedom, even if the magnitude of the misspecification is negligibly small. Conversely, the sensitivity of $RMSEA_D$, as a modification of the single model *RMSEA*, is independent of sample size and degrees of freedom.
2) $RMSEA_D$ preserves the RMSEA metric, allowing conventional criteria for model rejection (e.g., Browne and Cudeck (1993), Chen, Curran, Bollen, Kirby and Paxton (2008), Hu and Bentler (1999), and Yuan and Chan (2016)) to be used. Conversely, model rejection by $\Delta GFIs$ requires new critical values. Chen (2007) provides some recommendations for how to evaluate model fit by $\Delta GFIs$, however, these recommendations are difficult to generalize beyond the author's simu conditions.
3) Retaining the RMSEA metric permits the analytic estimation of a confidence interval about $RMSEA_D$, as well as the specification of tests of close and not-close fit. To our knowledge, neither is possible with $\Delta GFIs$.

4) Retaining the RMSEA metric gives $RMSEA_D$ the same direct interpretation as RMSEA: $RMSEA_D$ is the amount of misfit introduced by M_B above and beyond that already introduced by M_A per introduced degree of freedom. Conversely, what is quantified by $\Delta RMSEA$ –beyond an unweighted mean difference in fit after additional constraints are introduced –is not entirely clear.
5) $RMSEA_D$ allows $RMSEA_A$ to be decomposed into two components: the contribution to misfit by the df_B degrees of freedom associated with M_B, and the df_D degrees of freedom introduced by M_A.
6) Unlike $\Delta RMSEA$, $RMSEA_D$ is not redundant with $RMSEA_A$ when M_B is correctly specified.
7) $\Delta RMSEA$ has low power for high df_B, low df_D comparisons, which are common in evaluations of MI.

Given the discussed issues with $\Delta GFIs$, and the utility and ease in interpretation of $RMSEA_D$, we encourage researchers looking to supplement chi-square difference tests in analyses of MI to choose $RMSEA_D$ over $\Delta GFIs$.

4.3.2. Extension to Other Fit Indices

While this manuscript focuses on extending *RMSEA* to the comparison of nested models, other GFIs based on minimized fit function values can also be extended for the purposes of nested model comparison. Consider the CFI, defined as $CFI = 1 - \frac{\hat{F}_{ML}}{\hat{F}_{ML,null}}$ (Bentler, 1990), which gives 1 minus the ratio of model misfit to independence model misfit. For the purposes of nested model comparison, one could propose $CFI_D = 1 - \frac{\hat{F}_{ML,D}}{\hat{F}_{ML,null}}$, which defines misfit of the constraints introduced by M_A as 1 minus the ratio of newly introduced misfit to independence model misfit. Note, however, that

$$
\begin{aligned}
\Delta CFI &= (1 - \frac{\hat{F}_{ML,A}}{\hat{F}_{ML,null}}) - (1 - \frac{\hat{F}_{ML,B}}{\hat{F}_{ML,null}}) \\
&= \frac{\hat{F}_{ML,B} - \hat{F}_{ML,A}}{\hat{F}_{ML,null}} \\
&= CFI_D - 1
\end{aligned}
$$

. Thus, CFI_D as defined here is a trivial transformation of ΔCFI, so critical values for model rejection by ΔCFI can be converted to the original CFI metric simply by adding 1. This perfect correlation between ΔCFI and our hypothetical CFI_D is consistent with observations made by Chen (2007), who found ΔCFI to be the most useful ΔGFI.

4.3.3. Future Research and Limitations

While it is clear that $RMSEA_D$ is a more meaningful population value than $\Delta RMSEA$ when the assumption of baseline model fit is met, in practice, difference tests will typically involve comparisons for which some degree of misfit is present. Consider that when testing measurement invariance, the difference test comparing the configural invariance model to the metric invariance model follows a test of configural invariance for which model fit has been deemed not necessarily perfect, but acceptable. Any model retained on the grounds that it is acceptable, rather than perfectly correct, will be a misspecified baseline in further comparisons. Monte Carlo research is necessary to evaluate the sensitivity of $RMSEA_D$ to violations of measurement invariance when baseline models are slightly misspecified. Additional limitations to this research include the fact that the utility of $RMSEA_D$ is primarily assessed relative to

$DGFIs$, and no other methods of assessing model fit. Future research should compare the behavior of $RMSEA_D$ to other promising methods of assessing fit of nested MI models, such as equivalence testing (Counsell, Cribbie & Flora, 2020; Yuan and Chan, 2016). Further limitations to this research are discussed in Chapter 5.

Chapter 5: General Discussion

5.1 Summary of Dissertation Topics

Broadly, the overall topic of this dissertation can be described as new methods for evaluating measurement invariance within a factor analytic context that relax some of the overly restrictive and unrealistic assumptions of the multiple-group confirmatory factor analytic framework that applied MI researchers are used to. More specifically, all studied methods relax the assumption that psychometric instruments are only invariant across groups when the null hypothesis of perfectly invariant measurement parameters for all items is retained. In Chapter 2, I introduce a structural equation modelling approach to evaluating MI of observed composite scores on psychometric instruments that does not assume invariance of each individual item. In Chapter 3, I propose a modification to Millsap and Kwok's (impact of violations of MI on classification accuracy in two groups when only partial scalar invariance can be retained and observed composite scores are used for selection/diagnosis. The proposed modification guides researchers towards critical values for selection that ensure between-group differences on their preferred measures of classification accuracy are within an acceptable range. In Chapter 4, I introduce an alternative conceptualization of how goodness-of-fit indices can be used in the analysis of MI. Specifically, I propose a method of quantifying change in model fit as measured by the RMSEA that maintains the original RMSEA metric, allowing researchers interpret the magnitude of loss of fit using familiar benchmarks (e.g., Hu and Bentler, 1999), while also allowing the computation of confidence intervals and tests of close and not-close fit (Browne & Cudeck, 1993; MacCallum, Browne & Sugawara, 1996; MacCallum, Browne & Cai, 2006).

5.2. Summary of Research Project Results

5.2.1. Chapter 2

In Chapter 2, I demonstrate that the definitions of measurement invariance given in Equation (1.1) and Equation (1.2) can be satisfied for observed composite scores when measurement parameter totals are invariant across groups. If observed composite scores are thought of as single-item tests, the definition of strict invariance given in Equation (1.1) becomes Equation (1.11), which is shown to hold when $l_g^* = {}^*$, $t_g^* = {}^*$, and $y_g^* = y^*$ for all g, where $l_g^* = w' L_g$, $t_g^* = w'{}_g$, and $y_g^* = w' Y w$. Further, the more relaxed definition of scalar invariance given in Equation (1.2) becomes Equation (1.12) for observed composites, and is satisfied when $l_g^* = {}^*$ and $t_g^* = {}^*$. I refer to the definitions of MI given in Equation (1.11) and Equation (1.12) as "scale-level" measurement invariance, and the traditional conceptualization of MI which places equality constraints on individual measurement parameters as "item-level" measurement invariance.

A series of power analyses is conducted to illustrate how the sensitivity of tests of scale-level MI to violations of MI varies in a manner that is consistent with the magnitude of the impact of violations of MI on observed composite score use. Conversely, the sensitivity of tests of item-level MI to violations of MI is often unrelated or even negatively correlated with the magnitude of the impact of violations of MI on observed composite use. In simulation 2, population models with 1 factor and 8 indicators have between 2 and 7 loadings, intercepts, or residual variances which are .1 higher in the focal group than in the reference group. When scale-level metric invariance constraints are fit to populations with non-invariant loadings, power to detect violations of MI increases as a function of between-group differences on total loadings, such that power is greatest when $\lambda^*_F - \lambda^*_R = .7$, and least when $\lambda^*_F - \lambda^*_R = .2$. Counterintuitively, this pattern is not observed when item-level constraints are fit to population models, where power is greatest when $\lambda^*_F - \lambda^*_R = .4$, and decreases for larger and smaller between-group differences on total loadings. A similar pattern of results is observed when scale-level and item-level MI constraints are fit to populations for which violations of scalar invariance are present. When violations of strict invariance are present in the population, power of tests of both scale-level and item-level strict invariance increases with the magnitude of $\psi^*_F - \psi^*_R$. This pattern of results has been previously documented by Chen (2007), who noted that it is only present when violations of MI are uniform, meaning non-invariant loadings or intercepts are always greater in the same group. I believe this phenomenon occurs because when more than half of loadings or intercepts are greater in one group, while remaining parameters are invariant, item-level measurement invariance constraints may lead to higher estimated latent variances or latent means in the group with higher loadings or intercepts, respectively. For example, if $\Lambda'_R = [.7,.7,.7,.7,.7,.7,.7,.7]$, $\Phi_R = 1$, $\Lambda'_F = [.8,.8,.8,.8,.8,.8,.8,.8]$, and $\Phi_F = 1$, between-group equality constraints on factor loadings will fit well, as model parameters in the focal group can be estimated as $\hat{\Lambda}'_F = [.7,.7,.7,.7,.7,.7,.7,.7]$ and $\hat{\Phi}_F = 1.306$, perfectly reproducing the population covariance matrix in the focal group.

In simulation 3, population models with 1 factor and 8 indicators have between 1 and 6 loadings, intercepts, or residual variances which are greater in the focal group, but in such a way that $\lambda^*_F - \lambda^*_R = .6$ for all populations where violations of metric invariance are present, $\tau^*_F - \tau^*_R = .6$ for all populations where violations of scalar invariance are present, and $\psi^*_F - \psi^*_R = .6$ for all populations where violations of strict invariance are present. Thus, as the number of non-invariant parameters decreases, the magnitude of between-group differences on individual measurement parameters increases, while between-group differences on measurement parameter totals are held constant across populations. Desirably, power of tests of scale-level MI is independent of the number of non-invariant parameters when between-group differences on measurement parameter totals are held constant, while power of tests of item-level MI is greatest when between-group differences on measurement parameter totals is due to a single highly non-invariant parameter, and lowest when total non-invariance is due to 6 slightly non-invariant parameters.

In simulation 4, multiple-group population models with 1 factor and between 4 and 14 indicators have 3 loadings, intercepts, or residual variances which are .2 higher in the focal group. Thus, between-group differences on measurement parameter totals and the number of non-invariant parameters are held constant across populations, while the number of perfectly invariant items is varied. Power of tests of scale-level MI is observed to decrease as scale-length increases, while power of tests of item-level MI is observed to increase as scale-length increases. I believe that the former pattern of sensitivity is more desirable, as power to detect violations of scale-level metric and scalar MI decreases with the proportion of observed score variability due to violations of MI.

I believe that for many researchers, the traditional item-level MGCFA approach to evaluating MI is used not out of preference, but because it is the most familiar, and because it is sufficiently restrictive to permit the inferences they hope to draw. My proposed “scale-level” MI models are useful for those applied researchers who primarily assess MI to ensure between-group comparability of observed composite scores, such as those using psychometric instruments for clinical/diagnostic purposes, or those using them in organizational settings as a hiring tool.

For researchers who cannot secure full item-level strong or strict invariance of their measures, but wish to be able to make between-group comparisons on observed composite scores, assessing scale-level invariance is more useful and appropriate as a follow-up test than is assessing partial measurement invariance, as the most useful conclusion a researcher can draw about observed composite use from partial measurement invariance is identifying which items could be removed to create a shortened, measurement invariant test (Steinmetz, 2013). Testing scale-level MI may allow applied researchers to make valid between-group comparisons on observed composite scores without the need of introducing multiple shortened versions of the same measure to the literature.

5.2.2. Chapter 3

In Chapter 3, I introduce a modification to M
quantifying the impact of violations of measurement invariance on classification accuracy when only partial scalar invariance can be retained for a measure, and observed composite scores are used for diagnosis/selection. Millsap and Kwo
measures of classification accuracy with respect to a specified critical percentile for selection in 2 groups, requiring only latent and observed means and variances as input, as well as a population mixing proportion, the proportion of the pooled population accounted for by the reference group. My modification to this method, rather than requiring a single candidate critical value be specified, presents estimated classification accuracy ratios for all percentiles as plots, allowing applied researchers to identify critical values for selection in multiple groups with acceptably small between-group differences on their preferred measures of classification accuracy.

To illustrate the utility of the proposed method, a series of classification accuracy ratio plots are generated from a variety of multiple-group models for which violations of MI are present. These population models include one in which factor loadings are higher in the focal group, one in which indicator intercepts are higher in the focal group, one in which residual variances are higher in the focal group, one in which both factor loadings and indicator intercepts are higher in the focal, and one in which factor loadings are higher in the focal group while indicator intercepts are higher in the reference group. For each of these multiple group populations, a series of $\log_{10}$ classification accuracy ratio plots are generated for four common measures of classification accuracy: sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The decision to place classification accuracy ratios on a $\log_{10}$ scale ensures that violations of MI do not differentially impact classification accuracy ratio plots depending on whether violations of MI favor the reference group or focal group. For the first three population models, classification accuracy ratio plots illustrate the differential impact that violations of MI on a single set of measurement parameters have on each of the aforementioned measures of classification accuracy. When violations of metric invariance are present, the group with higher loadings will have greater sensitivity and specificity, while the group with lower loadings will have greater PPV and NPV. When violations of scalar invariance are present, the group with higher intercepts will have greater sensitivity and NPV, while the group with lower intercepts will have greater specificity and PPV. When violations of strict invariance are present, the group with lower residual variances will have greater positive predictive value and negative predictive value, while sensitivity and specificity ratios appear largely unaffected by violations of strict invariance. For the final two population models, classification accuracy ratio plots help illustrate how violations of MI on multiple sets of parameters can have additive or compensatory effects on classification accuracy ratios, depending on which measures of classification accuracy are of interest. When violations of both metric and scalar invariance are present, and favor the same group, there are additive effects on sensitivity and positive predictive value, and compensatory effects on specificity and negative predictive value. Specifically, sensitivity is much greater in the group with higher measurement parameters, while PPV is much greater in the group with lower measurement parameters. Conversely, specificity and NPV ratios are more balanced in this population than in those where violations of MI are present on loadings only or intercepts only. When violations of both metric and scalar invariance are present such that one group has higher loadings and the other group has higher intercepts, there are additive effects on specificity and NPV, and compensatory effects on sensitivity and PPV. Specifically, specificity is much greater in the group with higher loadings, while NPV is much greater in the group with higher intercepts. Conversely, sensitivity and PPV ratios are more balanced in this population than in those where violations of MI are present on loadings only or intercepts only. Note that when violations of MI are present on multiple sets of measurement parameters, additive and compensatory effects are consistent with what is observed when violations of MI are present on a single set of measurement parameters.

5.2.3. Chapter 4

In Chapter 4, I introduce an alternative conceptualization of how to use the RMSEA for the purposes of nested measurement model comparison. Rather than using $\Delta GFIs$, which compute misfit in terms of absolute differences on estimated GFIs for a pair of nested MI models, I propose transforming between-group differences on minimized fit function values and degrees of freedom for a pair of nested models to the RMSEA metric, allowing fit to be interpreted using the same criteria applied researchers may be used to (e.g., Hu & Bentler, 1999, etc.). I refer to this modified RMSEA as RMSEA$_D$. By maintaining the original scale and non-central chi-square distribution of the RMSEA, one can also construct confidence intervals around RMSEA$_D$ estimates, as well as conduct RMSEA$_D$-based tests of small or large change in fit using the methodology applied researchers will also be familiar with (e.g., Browne and Cudeck, 1993). Other advantages of RMSEA$_D$ over $\Delta RMSEA$, illustrated in Chapter 4, include the fact that RMSEA$_D$ is not redundant with the single-model RMSEA when the baseline model is correctly specified, and the fact that RMSEA$_D$ is more sensitive to model misspecifications than $\Delta RMSEA$ when baseline model degrees of freedom are high, but df_D is small, which is often the case in analyses of MI.

5.3. Consideration: Partial Measurement Invariance and Choice of Reference Indicator for Latent Variable Identification

When fitting any structural equation model that involves latent variables, it is an essential step to identify latent variances by either fixing a factor loading associated with that latent variable to 1, or fixing the latent variance itself to 1. When models are fit to data from a single group, the choice of parameter constrained to 1 for latent variance identification often has no bearing on model fit or interpretation, and is, therefore, generally a matter of personal preference. It is not unreasonable for applied researchers to assume that this rule of thumb should extend to choice of reference indicator in analyses of MI within the MGCFA framework, however, that is not the case. When constraining a parameter to 1 while fitting a model in multiple groups, one is implicitly assuming that the parameter is invariant across groups. When this assumption is violated, it can lead to incorrect estimates on other parameters, invalidating the remainder of the assessment of MI (Cheung & Rensvold, 1999). Therefore, choice of reference indicator is a critical step in the assessment of MI that may, nevertheless, often be ignored in practice. Section

5.3 of my dissertation identifies a variety of candidate methods which may be useful for the purposes of specifying partial measurement invariance models with appropriate reference indicators, or for determining the most appropriate reference indicator to select when analyzing MI within a factor analytic context, depending on the analysis of interest. Subsequently, the relevance of these methods to those proposed in Chapter 2, Chapter 3 and Chapter 4 is discussed.

5.3.1. Cheung and Rensvold (1999)

Cheung and Rensvold (1999) proposed an ext
(1989) method of specification search for partial MI models which produces stronger evidence of invariance for individual measurement parameters. In addition to testing invariance of every
individual measurement parameter, Cheung and
individual measurement parameter for each possible choice of reference indicator. The authors argue that only the largest subset of items that are all mutually invariant on the parameter of interest with respect to each other being used as the reference indicator should be retained in the final partial MI model.

To illustrate with a simple example, consider a 1-factor model with 4 indicators that is being evaluated for partial metric invariance across two groups. Hypothetically, let us assume that when y_1 is the reference indicator, y_3 and y_4 are found to have invariant loadings, while the null hypothesis of invariant loadings cannot be retained for y_2. Further, let us assume that when y_2 is the reference indicator, y_4 is found to have an invariant loading, while y_1 and y_3 are not. When y_3 is the reference indicator, y_1 and y_4 are found to have invariant loadings, while y_2 is not. When y_4 is the reference indicator, y_1, y_2 and y_3 are all found to have non-invariant loadings. Thus, if y_4 had been the original choice of reference indicator, this hypothetical test would satisfy full item-level metric invariance. By Cheung a

results imply partial metric invariance of y_1, y_3, and y_4, as any given pair of items within this subset will always satisfy metric invariance if the other is used as the reference indicator. While this procedure may become time consuming for long scales, it provides stronger evidence for a given partial invariance model than the method of Byrne, Shavelson and Muthen (1989).

5.3.2. Alignment Method

The alignment method (Asparouhov & Muthén, 2014) is a technique which searches for an optimal partial measurement invariance model upon retention of the configural invariance model. Specifically, the method introduces a constraint to maximize the number of invariant, or near invariant parameters, while maintaining the same fit as the configural model. This constraint allows latent means and covariances to be estimated without the introduction of constraints that decrease model fit beyond that of the original configural invariance. The authors argue that this technique is analogous to rotation in exploratory factor analysis, where the best fitting model is identified and then adjusted to maximize interpretability. The primary intention of this method is to produce unbiased latent mean and covariance estimates without requiring additional constraints be placed on the configural invariance model, making it ideal for designs with large numbers of groups, to which traditional MGCFA constraints are overly sensitive. The authors note that the method produces asymptotically unbiased parameter estimates in two group designs.

The first stage of the alignment method fits the configural invariance model, where $\kappa_g = 0$, and $\Phi_g = 1$ in all groups. The variances and means of indicators can thus be expressed as $VAR(y_{pg}) = \lambda_{pg}^2 \Phi_g = \lambda_{pg,0}^2$ and $E(y_{pg}) = \alpha_{pg} + \lambda_{pg}\kappa_g = \alpha_{pg,0}$. For every possible pair of (κ_g, Φ_g), there exists a set of α_{pg} and λ_{pg} with equal likelihood to $\alpha_{pg,0}$ and $\lambda_{pg,0}$ when

$(\kappa_g, \Phi_g) = (0,1)$. The alignment method seeks to identify the set of latent mean and variance estimates that minimizes the amount of non-invariance using a *simplicity function* (Jennrich, 2006), $F = \sum_p \sum_{g_1 < g_2} w_{g_1,g_2} f(\lambda_{pg_1} - \lambda_{pg_2}) + \sum_p \sum_{g_1 < g_2} w_{g_1,g_2} f(\nu_{pg_1} - \nu_{pg_2})$, where g_1 and g_2 are the indices of a pair of groups. The simplicity function favors a solution with many invariant and few highly biased parameters over solutions with many moderately biased parameters.

The structure of the simplicity function implies that every combination of item and pair of groups makes a unique contribution to overall non-invariance. This implies that the amount of non-invariance that can be attributed to each parameter, and by extension each item, in the model can be quantified, allowing the researcher to use the alignment metho

invariant” items. The alignment method can, t

reference indicator to constrain for identification when using other methods of evaluating MI. One criticism of the method (Muthén & Asparouhov, 2013b) is that it assumes alignment is true in the population, which is not necessarily the case. Further, because the method is identical in fit to the configural invariance model, the hypothesis of alignment is not actually testable.

5.3.3. Bayesian Structural Equation Modelling

Muthén and Asparouhov (2012) proposed integrating Bayesian logic into structural equation modelling (BSEM), by replacing specification of exact zeroes with approximate zeroes: small-variance, zero-mean priors. This methodology eliminates the necessity of potentially time-consuming specification searches to identify problematic constraints, as which approximate zeroes are incorrect will be evident in their posterior distributions. The authors note that BSEM generally produces unbiased parameter estimates, and outperforms maximum-likelihood estimation in terms of parameter estimation when models are misspecified. They also note that

BSEM is particularly useful in small samples, where frequentist analyses are prone to failures of convergence and Heywood cases (Kline, 2016).

Muthén and Asparouhov (2013a) recommend using BSEM approximate equality constraints on parameter estimates across groups to evaluate *approximate measurement invariance*. BSEM is particularly useful in the context of testing measurement invariance, as non-invariant parameters can be identified without necessitating complex specification searches, and structural parameters can still be accurately estimated without relaxing approximate invariance constraints on measurement parameters. The authors recommend using BSEM to identify non-invariant parameters, and subsequently releasing invariance constraints to produce the best structural parameter estimates, a procedure which is logically equivalent to, but less time-consuming/subjective than, identifying the best partial invariance model using modification indices (Millsap & Yoon, 2007) or procedures based on iteratively introducing constraints on individual parameters (eg., Byrne, Shavelson and Muthén, 1989; Cheung & Rensvold, 1999). Because the magnitude of violations of MI are evident in the posterior distributions of between-group differences on measurement parameters, BSEM may be a good candidate method for identifying parameters that violate MI to the smallest degree, and by extension, identifying the most appropriate reference indicator.

5.3.4. Choice of Reference Indicator and Proposed Methods

While the methods discussed above are primarily used to specify partial measurement invariance models, they are also useful for selecting reference indicators outside of the context of specification searches. The scale-level MI models introduced in Chapter 2 place between-group equality constraints on parameter totals, but still require that a single loading be fixed to invariance for the purposes of latent variable identification. Because violating the assumption that the reference indicator is invariant across groups can bias parameter estimates (Cheung & Rensvold, 1999), it is particularly important to ensure that an invariant item is used for latent variable identification, as each degree of scale-level invariance essentially only has 1 estimated, measurement parameter and introducing bias to these estimates could theoretically produce highly misleading results.

With respect to Chapter 3, the partial invariance methods discussed above allow researchers to identify a partially invariant subset of items with greater certainty than the method of Byrne, Shavelson and Muthen (1989), allowing researchers to be more confident that parameter estimates are unbiased. Because the method of Millsap and Kwok (2004) takes measurement parameter totals as input, the absolute number of invariant it
partial measurement invariance model is immaterial, and so, the fact that methods like Cheung and Rensvold's (1999) leave researchers with
problematic –so long as the necessary minimum of two invariant intercepts needed to identify latent means and satisfy the definition of partial scalar invariance is met.

The methodology proposed in Chapter 4 also has implications for partial measurement invariance model specification searches, and for choice of reference indicator. The method of Cheung and Rensvold (1999) is limited in that because invariance is evaluated via binary tests of significance, there is no indication of which item in the retained partially invariant subset is the most appropriate to use as the reference indicator. I propose that the methodology proposed in Chapter 4 can be used in conjunction with the method of Cheung and Rensvold (1999) to identify the most appropriate reference indicator among the invariant subset of items in addition to helping identify the appropriate partial metric invariance model. Rather than examining a binary significance test of MI for each indicator under each choice of reference indicator, the

RMSEA_D can be used to explicitly quantify the magnitude of violations of MI for each indicator under each choice of reference indicator. One could, thus, compute an average RMSEA_D for each indicator across all possible reference indicators, and select that which introduces the least misfit on average as the most appropriate reference indicator**.**

5.4. Limitations

5.4.1. Limitations: Number of Groups and Dimensionality

It should be noted that all simulations and illustrations given in Chapter 2 and Chapter 3 make use of unidimensional models and two-group populations, making conclusions and recommendations primarily suited to researchers evaluating the invariance of unidimensional psychometric instruments across two groups. The methodology proposed in Chapter 2 should be applicable to analyses of MI across more than two groups, so long as the SEM software used has support for the kinds of model constraints necessary to evaluate the hypotheses of interest. Because multiple-group populations with more than two groups are not studied in Chapter 2, the possibility of convergence issues when scale-level measurement invariance constraints are applied to data from more than two groups is unknown. The methodology proposed in Chapter 3 can theoretically be extended to more than two groups, however, software that can compute critical values for selection given a specified percentile on the joint distribution of factor and observed scores from more than two groups does not presently exist.

The methodologies proposed in Chapter 2 and Chapter 3 can be applied to multidimensional measures so long as each factor is studied separately. When evaluating invariance of observed composite scores using linear confirmatory factor analysis, we hope to conclude that participants who are equal on true scores are likely to be equal on observed composite scores, and that participants who are equal on observed composite scores are likely to be equal on true scores. When observed composite scores are computed across multiple factors, observed scores and between-group differences on observed scores become difficult to interpret, as any given observed composite score could be reflective of infinitely many combinations of factor scores across the multiple latent variables (Millsap & Kwok, 2004).

5.4.2 Limitations: Extreme Violations of Item-Level Measurement Invariance That do not Violate Scale-Level Measurement Invariance

While the definitions of measurement invariance given in Equation (1.1) and Equation (1.2) are satisfied for observed composites when Equation (1.11) and Equation (1.12) hold, there are hypothetical multiple-group populations which satisfy scale-level invariance which may, nevertheless, provoke skepticism towards the face validity of between-group comparisons on observed composite scores. Consider a hypothetical multiple-group population with the following population parameters for a given psychometric instrument with 1 factor and 7 indicators: $\Lambda'_R = [.7,.7,0,.7,0,.7,0]$, $\Lambda'_F = [.7,0,.7,0,.7,0,.7]$, $\tau' = [0,0,0,0,0,0,0]$, $diag(\Psi) = [.51,.51,.51,.51,.51,.51,.51]$, $\kappa = 0$, and $\Phi = 1$. Based on the definitions of scale-level invariance discussed in Chapter 2 and Chapter 3, population parameters for the scale-level invariance model would be $\lambda^*_R = \lambda^*_F = 2.8$, $\tau^* = 0$, $\psi^* = 3.57$, $\kappa = 0$, and $\Phi = 1$. Thus, the present model satisfies the definition of scale-level strict invariance given in Equation (1.11): the sampling distribution of observed composite scores given unobservable true scores is independent of group membership. If we examine the item-level measurement parameters in the population, however, we have obvious reason to be skeptical of this conclusion. While measurement parameter totals are invariant across groups, there is only a single item with a non-zero factor loading in both groups. By retaining the null hypothesis of scale-level strict invariance, one is implicitly arguing that items 1, 2, 4, and 6 are measuring the same construct in the reference group as 1, 3, 5, and 7 are in the focal group.

One means of avoiding this issue would be to necessitate that all studied items adequately load on the latent variable of interest when the configural invariance model is fit to data. In factor analytic research, usually a minimum loading of .3-.4 is considered necessary for an item to be retained. While the configural invariance model would fit in this population, I believe this pattern of population factor loadings can be thought of as a sort of violation of configural invariance. It should be made clear that the purpose of the proposed scale-level definitions of MI is to allow items with lower loadings or intercepts in one group to be compensated for by items with higher loadings or intercepts in that group, and not for items that do not load at all in one group to be compensated for by items that do not load at all in the other group. Such a paradigm could lead to the development of instruments with different items for different segments of the general population. Researchers could theoretically propose an instrument with 20 unrelated items, and subsequently remove items from certain groups until estimated measurement parameter totals are consistent with scale-level MI, possibly capitalizing on chance.

In addition to ensuring that items have adequate loadings in all groups, researchers interested in evaluating MI for observed composite scores should also verify that composite reliability (Raykov, 1997) is adequate for studied groups before moving on to analysis of measurement invariance. Due to the existence of multiple items with loadings of zero in both groups, composite reliability in the above example is estimated to be $r = .68$ in both groups, which would generally be considered poor. As is the case with evaluating traditional item-level

MI, it is important not to forget about other degrees of construct validation prior to evaluating MI.

5.4.3. Limitations: Differential Attenuation Due to Unreliability

This dissertation has primarily discussed measurement invariance in terms of the impact violations of measurement invariance might have on observed composite scores when they are used for the purposes of comparing individual observed scores or means across groups. Such comparisons are only a subset of the ways observed composite scores might be used in applied behavioral research. In addition to comparing observed scores across populations, researchers might also use observed composites scores on psychometric instruments in correlational research, or as predictors in multiple regression. It is not necessarily the case that population correlations between $y_{comp,ig}$ and some external criterion, x_{ig}, will be equal across groups when correlations between ξ_{ig} and x_{ig} are equal across groups, even when scale-level strict invariance is satisfied. This is because differential attenuation of correlations due to unreliability (Spearman, 1904) across groups may still be present when scale-level strict invariance holds.

Reliability of observed composite scores can be computed as $\rho_{\xi y_{comp},g} = \dfrac{(\lambda_g^*)^2\,\Phi_g}{(\lambda_g^*)^2\,\Phi_g + \psi_g^*}$ (Raykov, 1997), where $\rho_{\xi y_{comp},g}$ is the reliability of y_{comp} as an indicator of ξ in group *g*, λ_g^* is the sum of factor loadings in group *g*, such that $\lambda_g^* = w'\Lambda_g$, ψ_g^* is the sum of residual variances in group *g*, such that $\psi_g^* = w'\Psi_g w$, and Φ_g is the latent variance in group *g*. If scale-level strict invariance holds, between-group differences in composite reliability may still be present if $\Phi_g \neq |$ for all *g*. Consider a multiple-group population where scale-level strict invariance holds, and where $\rho_{x\xi,g} = .7$ for all *g*, where *x* is an observed measure with perfect reliability. In this multiple-group population, let $l^* = 5.6$, $y^* = 4.08$, $F_R = 1$, and $F_F = 1.5$. Thus, $r_{xy_{comp},R} = .885$ and $r_{xy_{comp},F} = .920$. The relationship between $r_{xx,g}$ and $r_{y_{comp}x,g}$ can be expressed as $r_{y_{comp}x,g} = r_{xx,g}\sqrt{r_{yy_{comp},g}}$ (Spearman, 1904), resulting in population correlations of $r_{y_{comp}x,R} = .659$ and $r_{y_{comp}x,F} = .671$. This difference in correlations is small, but could still theoretically lead to the detection of a spurious group-by-y_{comp} interaction in multiple regression analysis with a sufficiently large sample size.

One method of confirming invariant attenuation due to unreliability would be to constrain latent variances to 1 for identification instead of a factor loading. Because the traditional MGCFA approach to evaluating measurement invariance involves placing between-group equality constraints on individual parameter estimates, researchers typically choose to constrain a factor loading to 1 for latent variable identification, rather than the latent variance itself, as the parameter constrained for identification is assumed to be invariant across groups. In the case of evaluating scale-level MI using the methodology proposed in Chapter 2, no specific parameters are assumed to be invariant, and thus, researchers are free to choose whether to constrain a loading or latent variance for latent variable identification. The simultaneous invariance of l^*, t^*, y^* and F across groups necessarily confirms strict invariance as defined in Equation (1.11) while also ensuring that between-group differences in regression coefficients for observed composite scores are not due to differential attenuation due to unreliability. A limitation of this approach is that it is assumed that latent variances are equal in the population, which is not necessarily true.

Alternatively, if one wishes to confirm invariance of measurement and reliability while avoiding the potentially unrealistic assumption of invariant latent variances, one could replace the test of scale-level strict invariance with a test of invariant composite reliabilities. To clarify, the following constraint: $\frac{(\lambda^*)^2 \Phi_R}{(\lambda^*)^2 \Phi_R + \theta_R^*} = \frac{(\lambda^*)^2 \Phi_F}{(\lambda^*)^2 \Phi_F + \theta_F^*}$. would ensure invariance of composite reliabilities without assuming invariance of latent variances. Such a constraint necessarily needs to replace the test of strict invariance, as equal composite reliabilities in the presence of unequal latent variances requires unequal residual variance totals. That said, an instrument for which this constraint holds in addition to scale-level scalar invariance satisfies Equation (1.12) while also confirming invariance of reliabilities without placing between-group equality constraints on latent variances.

5.4.4. Limitations: The Issue of Baseline Model Fit and Chi-Square Difference Tests

A limitation of all methods which make use of sequential nested model comparisons – including methods proposed in Chapter 2 and Chapter 3, and to some degree Chapter 4 –is the assumption of a correctly specified baseline model. As discussed by Steiger, Shapiro, and Browne (1985), the chi-square difference test statistic, *D*, asymptotically follows a central chi-square distribution with df_D degrees of freedom when both the baseline and constrained models – M_B and M_A, respectively –are correctly specified. Further, if M_B is correctly specified but M_A is not, *D* asymptotically follows a non-central chi-square distribution with df_D degrees of freedom and non-centrality parameter λ_{nonc}. However, if M_B is incorrectly specified, but constraints imposed by M_A above and beyond those already imposed by M_B are correctly specified, the asymptotic sampling distribution of *D* is not a central chi-square distribution with

df_D degrees of freedom, and thus, using the 95th percentile of that distribution to determine whether observed Deis’ sufficiently large for rejection of the null hypothesis of equal model fit is technically incorrect. It is for this reason that it is imperative to evaluate fit of the configural invariance model before evaluating metric invariance, and it is imperative to confirm metric invariance before testing scalar invariance, and so on.

Within the null-hypothesis significance testing framework, researchers do not confirm that the null hypothesis is true when it is retained, but rather, fail to find sufficient evidence suggesting that the null hypothesis is true. Thus, a null hypothesis may be retained because it is true, or it could be retained because the actual effect is small enough to avoid detection given model degrees of freedom and sample size. While this trend is arguably desirable when evaluating fit of a single model, as we do not want to reject a model that competently explains observed means and covariances over trivial misspecifications, it does, to some degree, invalidate sequential chi-square difference tests. A 2016 paper by Yuan and Chan noted that because $E(T) = df + l_{nonc}$ when a fitted model is misspecified, one can use the critical value for rejection of the null hypothesis of exact model fit, T_{crit}, to define an alternate non-central chi-square distribution for which roughly 50% of the density exceeds this critical value, and by extension, an alternate distribution that would lead to retention of the null hypothesis in roughly 50% of samples. For example, the 95th percentile of a central chi-square distribution with 10 degrees of freedom is $T_{crit} = 18.31$, meaning that if the actual population distribution of *T* given a specified model and sample size is a non-central chi-square distribution with 10 degrees of freedom and $l_{nonc} = 8.31$, the null hypothesis would be retained in roughly half of all samples. This non-centrality parameter reflects an RMSEA of .041 when N=500, an RMSEA of .064

when N=200, and an RMSEA of .092 when N=100. Yuan and Chan (2016) argue that because sequential chi-square difference tests assume a correctly specified baseline model, the significance testing process should be one which can theoretically confirm the null hypothesis, and not simply fail to reject the null hypothesis, such as via equivalence testing, or tests of the null hypothesis of not close fit (Browne & Cudeck, 1993; MacCallum, Browne & Sugawara, 1996; MacCallum, Browne & Cai, 2006). Software for conducting equivalence tests in R is provided by the authors (Yuan & Chan, 2016).

While baseline model misspecification does threaten the validity of chi-square difference tests, it theoretically should not threaten the interpretation of RMSEA$_D$, as defined in Chapter 4. Independent of whether M_B is correctly specified, RMSEA$_D$ still gives the average contribution to $\hat{F}_{ML,B}$ by each of the df_D constraints imposed on M_A above and beyond the df_B constraints already present under M_B. However, because the methodology of Browne and Cudeck (1993) assumes that the distribution of the RMSEA is a transformation of an estimated non-central chi-square distribution, attempts to construct a confidence interval around RMSEA$_D$, or perform an RMSEA$_D$-based test of close or not-close fit, assume a correctly specified baseline model. as *D* does not asymptotically follow a non-central chi-square distribution with non-centrality parameter $\lambda_{nonc,D} = \lambda_{nonc,A} - \lambda_{nonc,B}$ when M_B is misspecified (Steiger, Shapiro, and Browne, 1985). MacCallum, Browne, and Cai (2006) noted that *D* should approximately follow a non-central chi-square distribution with df_D degrees of freedom and non-centrality parameter $\lambda_{nonc,D}$ when neither model is badly misspecified. According to the authors, a model is badly misspecified when lack of model fit due to misspecification is greater than that due to sampling error. Because equivalence testing seeks to reject the null hypothesis that misfit in the population is greater than some acceptably small value, and by extension, essentially seeks to confirm misfit is smaller than this value, it may be appropriate to estimate a confidence interval around RMSEA_D when it is used to supplement equivalence tests. Further research is necessary to evaluate what degree of baseline model misspecification is necessary before estimated confidence intervals around RMSEA_D estimates poorly approximate the true sampling distribution of RMSEA_D in the population.

5.5 Recommended Workflow for Evaluating Measurement Invariance of Observed Composites

The following section offers a recommended series of analyses to perform when evaluating measurement invariance of observed composite scores within the factor analytic framework. These recommendations take into account the methods proposed in Chapter 2, Chapter 3 and Chapter 4, as well as the considerations and limitations discussed above. The proposed series of analyses is inspired by the work of Vandenberg and Lance (2000), Steenkamp and Baumgartner (1998), and Van de Schoot, Lugtig and Hox (2012), among others, who have also proposed series of MGCFA models to fit to data when evaluating psychometric instruments for MI. Unlike this previous research, the following recommendations are oriented towards researchers who wish to make between-group comparisons on observed composite scores. Those who are primarily interested in evaluating MI for the purposes of making between-group comparisons on latent variables should follow existing recommendations made by Vandenberg and Lance (2000), Steenkamp and Baumgartner (1998) or Van de Schoot, Lugtig and Hox (2012).

The first stage of one’s analysis of MI sh

fits the same single-group CFA model separately to sample data collected from each group being evaluated for measurement invariance. The chi-square test statistic, *T*, and model degrees of freedom for this analysis will always be equal to the sum of test statistics and degrees of freedom associated with each group had the single-group CFA been fit to each group in separate analyses. Prior to retaining configural invariance for a given measure, I recommend using one of the partial measurement invariance techniques discussed in section 5.3 for the purposes of selecting the most appropriate factor loading to constrain to 1 –and by extension, invariance –for the purposes of latent variable identification. While more quantitatively adept researchers may prefer computationally intensive methods such as Bayesian SEM or the alignment method, I recommend using the method of Cheung and Rensvold (1999) supplemented with RMSEA_D totals, which allows researchers to identify the most invariant factor loading to constrain for the purposes of latent variable identification, while also allowing them to work within the traditional MGCFA framework they may already be familiar with. It should be noted that while these recommendations are primarily oriented towards researchers who wish to make use of observed composite scores, choice of reference indicator is still of critical importance when analyses of MI are used for the purposes of validating between-group comparisons involving latent variables.

For researchers who are primarily interested in confirming measurement invariance of observed composite scores, it is imperative that the configural invariance model indicates that all studied items have adequately large factor loadings and that observed composite scores have adequate composite reliability in each group. If these assumptions are not met in practice, one plausibly runs the risk of concluding measurement is invariant across groups for a measure which essentially presents different items to different groups, which may provoke skepticism towards the face validity of any between-group comparisons made using observed composite scores on that measure.

Upon retention of configural invariance, researchers interested in evaluating measurement invariance of observed composite scores should subsequently evaluate scale-level metric invariance, the invariance of factor loading totals across groups, such that $/_{g}^{*} = {}^{*}/$ for all *g*. While retention of scale-level metric invariance on its own is insufficient to permit any between-group comparisons involving observed composite scores, it is a necessary first step before evaluating higher degrees of scale-level MI. If scale-level metric invariance cannot be retained, between-group comparisons on observed composite scores, means, or correlations are not warranted, and further invariance testing is not recommended. For researchers who wish to use observed composite scores for the purposes of selection/diagnosis using the method of Millsap and Kwok (2004), failure to confirm scale-level metric invariance can be followed with an evaluation of partial metric invariance, as the method of Millsap and Kwok (2004) does not need to confirm invariance of factor loading totals to be useful. As discussed in Chapter 3, between-group differences on measures of classification accuracy due to violations of metric invariance may still be compensated for by violations of measurement invariance on intercepts, depending on the direction of violations of measurement invariance, and which measures of classification accuracy are of interest. A minimum of 2 invariant loadings are necessary to confirm partial metric invariance (Byrne, Shavelson & Muthen, 1989).

Upon retention of scale-level metric invariance, researchers interested in evaluating measurement invariance of observed composite scores should subsequently evaluate scale-level scalar invariance, the invariance of indicator intercepts across groups, such that $t_{g}^{*} = {}^{*}{,}$ for all *g*. Retention of scale-level scalar invariance confirms the independence given in Equation (1.12), and thus, ensures the comparability of observed composite scores and means across groups. It should be restated that when I refer to the between-group comparability of observed composite scores across groups, I am saying that the most likely state of affairs for test respondents who are equal on X is that they are equal on y_{comp}, and that the most likely state of affairs for test respondents who are equal on y_{comp} is that they are equal on X. Further, for test respondents who are unequal on X, such that $X_{i,R} > X_{j,F}$, the most likely state of affairs is that $y_{comp,i,R} > y_{comp,j,F}$, and for test respondents who are unequal on y_{comp}, such that $y_{comp,i,R} > y_{comp,j,F}$, the most likely state of affairs is that $X_{i,R} > X_{j,F}$. Participants who are equal on unobservable true scores are expected to be equal on observed scores, and participants who are equal on observed scores are expected to be equal on unobservable true scores. Any violations of these expectations when scale-level scalar invariance holds are due to measurement error, and not between-group differences in measurement. If scale-level scalar invariance cannot be confirmed, between-group comparisons on observed composite scores or means are not justified. For those interested in using observed composite scores for the purposes of selection/diagnosis using the method of Millsap and Kwok (2004), failure to confirm scale-level scalar invariance may be followed with an evaluation of partial scalar invariance. A minimum of 2 invariant intercepts are necessary to confirm partial scalar invariance and to identify latent means, at which point the methodology discussed in Chapter 3 can be used to compute classification accuracy ratio plots. Further research is necessary to confirm the appropriateness of using scale-level measurement invariance and partial measurement invariance constraints in concert to specify a model for the purposes of applying the methodology discussed in Chapter 3.

Upon retention of scale-level scalar invariance, researchers interested in evaluating measurement invariance of observed composite scores have a few options for follow-up analyses, depending on which questions they hope to answer using their measures. If researchers wish to satisfy the relaxed definition of strict invariance for observed composite scores - given in Equation (1.1) and relaxed in Equation (1.11) –it becomes necessary to evaluate scale-level strict invariance, the invariance of residual variances across groups, such that $y^*_g = \dot{y}^*$ for all g. For researchers who wish to use observed composite scores in correlational analysis, the invariance of residual variances becomes less important than the invariance of composite reliability. As discussed in section 5.4.3, full strict invariance does not imply invariance of composite reliability unless latent variances are also equal across groups. When composite reliability is unequal across groups, spurious between-group differences on correlations or regression coefficients may be detected because of differential attenuation due to unreliability. Researchers who wish to compare correlations or regression coefficients across groups should, thus, supplement the test of strict invariance with one of invariant composite reliabilities. This can be done by either constraining latent variances to equality in a model where scale-level strict invariance holds, or by releasing the scale-level strict invariance constraint and replacing it with a between-group equality constraint on composite reliability. It should be restated that both of these models have limitations: the former assumes equality of latent variances, which may not be true in the population, and the latter necessarily violates scale-level strict invariance, as in order for between-group invariance of composite reliability to exist in the presence of unequal latent variances, residual variance totals must also be unequal across groups, violating Equation (1.11).

That said, for many researchers' purposes, re Equation (1.12) is often sufficient.

5.6 Conclusion

The methodologies proposed in Chapter 2, Chapter 3, and Chapter 4 greatly facilitate evaluations of measurement invariance when observed composite scores, rather than individual item scores, are of primary interest. I illustrate that full item-level measurement invariance is not necessarily required to justify between-group comparisons involving observed composite scores, as the impact of violations of item-level MI on observed composite use may be compensated for by other violations of MI in the aggregate. Analysis recommendations are provided depending on whether researchers are interested in making between-group comparisons on observed composite scores, means, correlations, or classification accuracy. I also provide an alternative approach to comparing fit of nested measurement invariance models which avoids some of the pitfalls associated with chi-square difference tests and $\Delta GFIs$.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modelling, 21*(4), 495-508. doi:10.1080/10705511.2014.919210

Amemiya, Y., & Anderson, T. W. (1990). Asymptotic Chi-Square Tests For A Large Class of Factor Analysis Models. *The Annals of Statistics, 18*(3), 1453 –1463.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155 –173. https://doi-org.ezproxy.library.ubc.ca/10.1007/BF02294170

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M. (2006). *EQS 6 Structural Equations Program Manual.* Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588 –606. https://doi-org.ezproxy.library.ubc.ca/10.1037/0033-2909.88.3.588

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research, 21*, 205-229.

Brace, J. C., & Savalei, V. (2016). Type I Error Rates and Power of Several Versions of Scaled Chi-Square Difference Tests in Investigations of Measurement Invariance. *Psychological Methods*, doi:10.1037/met0000097

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136 –162). Newbury Park, CA: Sage.

Browne, M. W., & du Toit, S. H. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research, 27*(2), 269 –300. https://doi-org.ezproxy.library.ubc.ca/10.1207/s15327906mbr2702_13

Browne, M. W., & Mels, G. (1990). *RAMONA user's guide*. Unpublished report, Department of Psychology, Ohio State University.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. doi:10.1037/0033-2909.105.3.456

Carter, N. M., & Pérez, E. O. (2016). Race and nation: How racial hierarchy shapes national attachments. *Political Psychology, 37*(4), 497 –513. https://doi.org/10.1111/pops.12270

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modelling, 14*(3), 464-504. doi:10.1080/10705510701301834

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462 –494. https://doi-org.ezproxy.library.ubc.ca/10.1177/0049124108314720

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*(3), 170 –175. https://doi-org.ezproxy.library.ubc.ca/10.1111/j.1467-9280.1995.tb00327.x

Chen, F. F., & West, S. G. (2008). Measuring individualism and collectivism: The importance of considering differential components, reference groups, and measurement invariance. *Journal of Research in Personality, 42*(2), 259 –294. https://doi-org.ezproxy.library.ubc.ca/10.1016/j.jrp.2007.05.006

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods, 15*(2), 167 –198. https://doi-org.ezproxy.library.ubc.ca/10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal Of Management, 25*(1), 1-27. doi:10.1177/014920639902500101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling, 9*(2), 233-255. doi:10.1207/S15328007SEM0902_5

Chou, C., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal Of Mathematical And Statistical Psychology, 44*(2), 347-357. doi:10.1111/j.2044-8317.1991.tb00966.x

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. *Multivariate Behavioral Research, 55*(2), 312 –328. https://doi-org.ezproxy.library.ubc.ca/10.1080/00273171.2019.1633617

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29. doi:10.1037/1082-989X.1.1.16

Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. *Journal of Applied Psychology, 69*(3), 498–508. https://doi-org.ezproxy.library.ubc.ca/10.1037/0021-9010.69.3.498

Edwards, M. C. (2013). Purple unicorns, true models, and other things I've never seen. *Measurement: Interdisciplinary Research And Perspectives, 11*(3), 107-111. doi:10.1080/15366367.2013.835178

Everson, H., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51*, 243–251.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1975). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal Of Applied Behavioral Science, 12*(2), 133-157. doi:10.1177/002188637601200201

Gresham, F. M., MacMillan, D. L., & Bocian, K. (1996). “Behavioral earthqu frequency, salient behavioral events that differentiate students at-risk for behavioral disorders. *Behavioral Disorders, 21*(4), 277–292.

Heine, S. J., Lehman, D. R., Peng, K., s-cultural & Gree comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology, 82*(6), 903–918. https://doi-org.ezproxy.library.ubc.ca/10.1037/0022-3514.82.6.903

Holst, E., Jorgensen, K. B., & Natolski, I. (2001). *BIVAR (Version 1.c)*. Copenhagen, Denmark: Arbejclsmijeinstuttet.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3–4), 117–144. https://doi-org.ezproxy.library.ubc.ca/10.1080/03610739208253916

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invarient: A practical scientist's look at the real concept o *Southern Psychologist, 1*(4), 179–188.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling, 6*(1), 1–55. https://doi-org.ezproxy.library.ubc.ca/10.1080/10705519909540118

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted?. *Psychological Bulletin, 112*(2), 351-362. doi:10.1037/0033-2909.112.2.351

Hukkelberg, S. (2016). The Eyberg Child Behavior Inventory: Factorial invariance in problem behaviors across gender and age. *Scandinavian Journal of Psychology, 57*(4), 298–304. https://doi.org/10.1111/sjop.12290

Innstrand, S. T. (2016). Occupational differences in work engagement: A longitudinal study among eight occupational groups in Norway. *Scandinavian Journal of Psychology, 57*(4), 338–349. https://doi.org/10.1111/sjop.12298

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. https://doi-org.ezproxy.library.ubc.ca/10.1007/BF02291366

Joreskog, K. G., & Sorbom, D. (1996). LISREL 8: User's reference gui Mooresville: Scientific Software.

Head, D., Allison, S., Lucena, N., Hassenstab, J., & Morris, J. C. (2017). Latent structure of cognitive performance in the Adult Children Study. *Journal Of Clinical And Experimental Neuropsychology, 39*(7), 621-635. doi:10.1080/13803395.2016.1252725

Kenny, D. A. (2015). *Measuring model fit.* Retrieved from http://davidakenny.net/cm/fit.htm

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486-507. doi:10.1177/0049124114543236

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modelling.* Guilford Press: New York, NY.

Lai, M. H. C., Kwok, O., Yoon, M., & Hsiao, Y. (2017) Understanding the impact of partial factorial invariance on selection accuracy: An R Script, *Structural Equation Modelling*, 24(5), 783-799, doi: 10.1080/10705511.2017.1318703

Levant, R. F., Alto, K. M., McKelvey, D. K., Richmond, K. A., & McDermott, R. C. (2017). Variance Composition, Measurement Invariance by Gender, and Construct Validity of the Femininity Ideology Scale-Short Form. *Journal Of Counseling Psychology*, doi:10.1037/cou0000230

Li, L., & Bentler, P. M. (2011). Quantified choice of root-mean-square errors of approximation for evaluation and power analysis of small differences between structural equation models. *Psychological Methods, 16*(2), 116 –126. https://doi-org.ezproxy.library.ubc.ca/10.1037/a0022657

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modelling, 13*(1), 59 –72. https://doi-org.ezproxy.library.ubc.ca/10.1207/s15328007sem1301_3

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods, 11*(1), 19-35. doi:10.1037/1082-989X.11.1.19

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modelling. *Psychological Methods, 1*(2), 130-149. doi:10.1037/1082-989X.1.2.130

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224 –253. https://doi-org.ezproxy.library.ubc.ca/10.1037/0033-295X.98.2.224

McCuish, E. C., Mathesius, J. R., Lussier, P., & Corrado, R. R. (2017). The Cross-Cultural Generalizability of the Psychopathy Checklist: Youth Version for Adjudicated Indigenous Youth. *Psychological Assessment*, doi:10.1037/pas0000468

McDermott, R. C., Levant, R. F., Hammer, J. H., Hall, R. J., McKelvey, D. K., & Jones, Z. (2017). Further Examination of the Factor Structure of the Male Role Norms Inventory-Short Form (MRNI-SF): Measurement Considerations for Women, Men of Color, and Gay Men. *Journal Of Counseling Psychology*, doi:10.1037/cou0000225

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64 –82. https://doi-org.ezproxy.library.ubc.ca/10.1037/1082-989X.7.1.64

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Statistics, 13*, 127 –143.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modelling, 14*(4), 611 –635. https://doi-org.ezproxy.library.ubc.ca/10.1080/10705510701575461

Meade, A. W., & Kroustalis, C. M. (2006). Problems With Item Parceling for Confirmatory Factor Analytic Tests of Measurement Invariance. *Organizational Research Methods, 9*(3), 369 –403. https://doi-org.ezproxy.library.ubc.ca/10.1177/1094428105283384

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing Measurement Equivalence and Invariance in Longitudinal Data With Item Response Theory. *International Journal Of Testing, 5*(3), 279-300. doi:10.1207/s15327574ijt0503_6

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi:10.1007/BF02294825

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289-311. doi:10.1007/BF02294510

Metselaar, E. E., Walet, H. J., Cozijnsen, A. J., & de Pádua, M. (1996). Sociale determinanten van innovatief gedrag = Social determinants of innovative behaviour. *Gedrag En Organisatie, 9*(1), 38 –52.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248-260. doi:10.1037/1082-989X.2.3.248

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473. doi:10.1007/s11336-007-9039-7

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

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26*(3), 479-497. doi:10.1207/s15327906mbr2603_6

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal Of Applied Psychology, 73*(3), 574-584. doi:10.1037/0021-9010.73.3.574

Millsap, R. E., & Kwok, O. (2004). Evaluating the Impact of Partial Factorial Invariance on Selection in Two Populations. *Psychological Methods, 9*(1), 93-115. doi:10.1037/1082-989X.9.1.93

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modelling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313-335. doi:10.1037/a0026802

Muthén, B. & Asparouhov, T. (2013a). BSEM measurement invariance analysis. *Mplus Web Notes: No. 17.* January 11, 2013.

Muthén, B. & Asparouhov, T. (2013b). New methods for the study of measurement invariance with many groups. *Mplus Web Notes: No. 18.* October 1, 2013.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User’s Guide. Eighth Los Angeles, CA: Muthén & Muthén

Nevitt, J., & Hancock, G. R. (2004). Evaluating Small Sample Approaches for Model Test Statistics in Structural Equation Modelling. *Multivariate Behavioral Research, 39*(3), 439-478. doi:10.1207/S15327906MBR3903_3

O'Boyle, E. H., Jr., & Williams, L. J. (2011). Decomposing model fit: Measurement vs theory in organizational research using latent variables. *Journal of Applied Psychology, 96*(1), 1–12. https://doi-org.ezproxy.library.ubc.ca/10.1037/a0020539

R Core Team, 2016. *R: A Language and Environment for Statistical Computing*, Vienna, Austria. Available at: https://www.R-project.org/.

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

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184. https://doi-org.ezproxy.library.ubc.ca/10.1177/01466216970212006

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modelling. *Journal of Statistical Software, 48*(2), 1-36. doi:10.18637/jss.v048.i02

Satorra, A. (2000). Scaled and adjusted restricted tests in multisample analysis of moment structures. In D.D.H. Heijmails, D.S.G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233-247). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *ASA Proceedings of the Business and Economic Section, 52*, 308-313.

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye && C.C. Clogg (Eds.) *Latent variable analysis: applications for developmental research* (pp. 399-419). Thousand Oaks: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514. doi:10.1007/BF02296192

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243-248. doi:10.1007/s11336-009-9135-y

Savalei, V. (2014). Understanding robust corrections in structural equation modelling. *Structural Equation Modelling, 21*(1), 149-160. doi:10.1080/10705511.2013.824793

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi-org.ezproxy.library.ubc.ca/10.2307/1412159

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, W. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Preliminary professional manual for the Test Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Steenkamp, J.E.M., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research, 25*, 78-90. doi:10.1086/209528

Steiger, J. H. (1989). *Causal modelling: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173–180. https://doi-org.ezproxy.library.ubc.ca/10.1207/s15327906mbr2502_4

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modelling, 5*, 411–419.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors*. Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika, 50*(3), 253-263. doi:10.1007/BF02294104

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough?. *Methodology: European Journal Of Research Methods For The Behavioral And Social Sciences, 9*(1), 1-12. doi:10.1027/1614-2241/a000049

Stevens, A. K., Blanchard, B. E., Shi, M., & Littlefield, A. K. (2018). Testing measurement invariance of the UPPS-P Impulsive Behavior Scale in Hispanic/Latino and non-Hispanic/Latino college students. *Psychological Assessment, 30*(2), 280 –285. https://doi-org.ezproxy.library.ubc.ca/10.1037/pas0000494

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press: Chicago.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38(1), 1 –10. https://doi-org.ezproxy.library.ubc.ca/10.1007/BF02291170

Tummers, L. G., & Van de Walle, S. (2012). Ex

implement diagnosis related groups: (No) benefits for society, patients and professionals. *Health Policy, 108*(2 –3), 158 –166. https://doi-org.ezproxy.library.ubc.ca/10.1016/j.healthpol.2012.08.024

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal Of Developmental Psychology, 9*(4), 486-492. doi:10.1080/17405629.2012.686740 van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*. https://doi-org.ezproxy.library.ubc.ca/10.3389/fpsyg.2013.00770

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi-org.ezproxy.library.ubc.ca/10.1080/17405629.2012.686740

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69. doi:10.1177/109442810031002

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modelling, 14*(3), 435-463. doi:10.1080/10705510701301677

Yuan, K.-H., & Bentler, P. M. (2004). On Chi-Square Difference and z Tests in Mean and Covariance Structure Analysis When the Base Model Is Misspecified. *Educational and Psychological Measurement, 64*(5), 737–757. https://doi-org.ezproxy.library.ubc.ca/10.1177/0013164404264853

Yuan, K., & Chan, W. (2016). Measurement Invariance via Multigroup SEM: Issues and Solutions With Chi-Square-Difference Tests. *Psychological Methods, 21*(3), 405-426. doi:10.1037/met0000080

Appendix A: Sample R Code for Implementing Scale-Level Measurement Invariance Constraints on Parameter Totals

```
library(lavaan)

base.mod<-"f1=~ c(1,1)*y1 + y2 + y3 + y4"

loadings.mod<-"f1=~ c(1,1)*y1 + c(a2,a22)*y2 + c(a3,a32)*y3 + c(a4,a42)*y4
#constraint on total loadings
a22== a2 + a3 + a4 + -a32 + -a42
"

intercepts.mod<-"f1=~ c(1,1)*y1 + c(a2,a22)*y2 + c(a3,a32)*y3 + c(a4,a42)*y4
y1 ~ c(b1,b1)*1
y2 ~ c(b2,b22)*1
y3 ~ c(b3,b32)*1
y4 ~ c(b4,b42)*1
f1 ~ c(0,NA)*1
#constraint on total loadings
a22== a2 + a3 + a4 + -a32 + -a42
#constraint on total intercepts
b22== b2 + b3 + b4 + -b32 + -b42
"

residuals.mod<-"f1=~ c(1,1)*y1 + c(a2,a22)*y2 + c(a3,a32)*y3 + c(a4,a42)*y4
y1 ~ c(b1,b1)*1
y2 ~ c(b2,b22)*1
y3 ~ c(b3,b32)*1
y4 ~ c(b4,b42)*1
y1 ~~ c(c1,c12)*y1
y2 ~~ c(c2,c22)*y2
y3 ~~ c(c3,c32)*y3
y4 ~~ c(c4,c42)*y3
f1 ~ c(0,NA)*1
#constraint on total loadings
a22== a2 + a3 + a4 + -a32 + -a42
#constraint on total intercepts
b22== b2 + b3 + b4 + -b32 + -b42
#constraint on total residuals
c12== c1 + c2 + c3 + c4 + -c22 + -c32 + -c42
"


run0<-cfa(base.mod, data=data, group="group")
run1<-cfa(loadings.mod, data=data, group="group")
run2<-cfa(intercepts.mod, data=data, group="group")
run3<-cfa(residuals.mod, data=data, group="group")


lavTestLRT(run0, run1, run2, run3)
```

Appendix B: Sample MPlus Code for Implementing Scale-Level Measurement Invariance

Constraints on Parameter Totals

Configural Invariance

```
TITLE: Configural Invariance Model
DATA: FILE IS Tummers.dat;
VARIABLE: NAMES ARE id group y1-y4 mean;
    USEVARIABLES group y1-y4;
    MISSING IS ALL (-9999);
    GROUPING IS group (1=g1 0=g2);
MODEL: f1 BY y1-y4;
    y1 ON f1@1;
    [f1@0];
MODEL g2: f1 BY y1-y4;
    y1 ON f1@1;
    [y1-y4];
```

Scale-level Constraints on Loadings

```
TITLE: Metric (Weak) Invariance Model (Scale-level Constraints)
DATA: FILE IS Tummers.dat;
VARIABLE: NAMES ARE id group y1-y4 mean;
    USEVARIABLES group y1-y4;
    MISSING IS ALL (-9999);
    GROUPING IS group (1=g1 0=g2);
MODEL: f1 BY y1-y4 (a11-a14);
    y1 ON f1@1;
    [f1@0];
MODEL g2: f1 BY y1-y4 (a21-a24);
    y1 ON f1@1;
    [y1-y4];
MODEL CONSTRAINT:
    a12 = a22 + a23 + a24 - a13 - a14;
```

Scale-level Constraints on Loadings and Intercepts

```
TITLE: Scalar (Strong) Invariance Model (Scale-level Constraints)
DATA: FILE IS Tummers.dat;
VARIABLE: NAMES ARE id group y1-y4 mean;
    USEVARIABLES group y1-y4;
    MISSING IS ALL (-9999);
    GROUPING IS group (1=g1 0=g2);
MODEL: f1 BY y1-y4 (a11-a14);
    y1 ON f1@1;
    [y1-y4] (b11-b14);
    [f1@0];
MODEL g2: f1 BY y1-y4 (a21-a24);
    y1 ON f1@1;
    [y1-y4] (b21-b24);
    [f1];
MODEL CONSTRAINT:
    a12 = a22 + a23 + a24 - a13 - a14;
    b11 = b21;
    b12 = b22 + b23 + b24 - b13 - b14;
```

Scale-Level Constraints on Loadings, Intercepts and Residuals

```
TITLE: Strict Invariance Model (Test-Level Constraints)
DATA: FILE IS Tummers.dat;
VARIABLE: NAMES ARE id group y1-y4 mean;
    USEVARIABLES group y1-y4;
    MISSING IS ALL (-9999);
    GROUPING IS group (1=g1 0=g2);
MODEL: f1 BY y1-y4 (a11-a14);
    y1 ON f1@1;
    [y1-y4] (b11-b14);
    [f1@0];
    y1 (c11);
    y2 (c12);
    y3 (c13);
    y4 (c14);
MODEL g2: f1 BY y1-y4 (a21-a24);
    y1 ON f1@1;
    [y1-y4] (b21-b24);
    [f1];
    y1 (c21);
    y2 (c22);
    y3 (c23);
    y4 (c24);
MODEL CONSTRAINT:
    a12 = a22 + a23 + a24 - a13 - a14;
    b11 = b21;
    b12 = b22 + b23 + b24 - b13 - b14;
    c11 = c21 + c22 + c23 + c24 - c12 - c13 - c14;
SAVEDATA: RESULTS ARE strictTLout.dat;
```

Item-Level Constraints on Loadings

```
TITLE: Metric (Weak) Invariance Model (Item-Level Constraints)
DATA: FILE IS Tummers.dat;
VARIABLE: NAMES ARE id group y1-y4 mean;
    USEVARIABLES group y1-y4;
    MISSING IS ALL (-9999);
    GROUPING IS group (1=g1 0=g2);
MODEL: f1 BY y1-y4;
    y1 ON f1@1;
    [f1@0];
MODEL g2: y1 ON f1@1;
    [y1-y4];
```

Item-Level Constraints on Loadings and Intercepts

```
TITLE: Scalar (Strong) Invariance Model (Item-Level Constraints)
```

```
DATA: FILE IS Tummers.dat;
VARIABLE: NAMES ARE id group y1-y4 mean;
    USEVARIABLES group y1-y4;
    MISSING IS ALL (-9999);
    GROUPING IS group (1=g1 0=g2);
MODEL: f1 BY y1-y4;
    y1 ON f1@1;
    [f1@0];
MODEL g2: y1 ON f1@1;
    [f1];
```

Item-Level Constraints on Loadings, Intercepts and Residuals

```
TITLE: Strict Invariance Model (Item-Level Constraints)
DATA: FILE IS Tummers.dat;
VARIABLE: NAMES ARE id group y1-y4 mean;
    USEVARIABLES group y1-y4;
    MISSING IS ALL (-9999);
    GROUPING IS group (1=g1 0=g2);
MODEL: f1 BY y1-y4;
    y1 ON f1@1;
    [f1@0];
    y1 (1);
    y2 (2);
    y3 (3);
    y4 (4);
MODEL g2: y1 ON f1@1;
    [f1];
SAVEDATA: RESULTS ARE strictout.dat;
```

Appendix C: Lai, Kwok, Yoon & Hsiao's (2017) Software for Executing Millsap and Kwok's (2004) Procedure.

```
pnormmix <- function(q, mean1 = 0, sd1 = 1, mean2 = 0, sd2 = 1, pmix1 = 0.5,
            lower.tail = TRUE) {
 # Distribution function (pdf) of a mixture of two normal distributions.
 #
 # Args:
 # q: vector of quantiles.
 # mean1: mean of first normal distribution.
 # sd1: standard deviation of first normal distribution.
 # mean2: mean of second normal distribution.
 # sd2: standard deviation of second normal distribution.
 # pmix1: mixing proportion for the first distribution. Should be a
 # number in the range (0, 1).
 # lower.tail: logical; if TRUE(default), probabilities are P[X <= x];
 # otherwise, P[X > x].
 #
 # Returns:
 # Cumulative probability of q or 1 minus it on the mixture normal
 # distribution.
 # Error handling
 stopifnot(pmix1 > 0, pmix1 < 1)
 as.vector(c(pmix1, 1 - pmix1) %*%
        sapply(q, pnorm, mean = c(mean1, mean2), sd = c(sd1, sd2),
            lower.tail = lower.tail))
}

qnormmix <- function(p, mean1 = 0, sd1 = 1, mean2 = 0, sd2 = 1, pmix1 = 0.5,
            lower.tail = TRUE) {
 # Quantile function of a mixture of two normal distributions.
 #
 # Args:
 # p: vector of probabilities.
 # mean1: mean of first normal distribution.
 # sd1: standard deviation of first normal distribution.
 # mean2: mean of second normal distribution.
 # sd2: standard deviation of second normal distribution.
 # pmix1: mixing proportion for the first distribution. Should be a
 # number in the range (0, 1).
 # lower.tail: logical; if TRUE(default), probabilities are P[X <= x];
 # otherwise, P[X > x].
 #
 # Returns:
 # Quantile corresponding to p or 1 - p on the mixture normal
 # distribution.
```

```
  # Error handling
  stopifnot(pmix1 > 0, pmix1 < 1, p >= 0, p <= 1)
  f <- function(x) (pnormmix(x, mean1, sd1, mean2, sd2, pmix1,
                              lower.tail) - p)^2
  start <- as.vector(c(pmix1, 1 - pmix1) %*%
                       sapply(p, qnorm, c(mean1, mean2), c(sd1, sd2),
                              lower.tail = lower.tail))
  nlminb(start, f)$par
}

.bvnorm_kernel <- function(x, y, mu_x = 0, mu_y = 0, sd_x = 1, sd_y = 1,
                           cov_xy = 0) {
  # Helper funcction for computing the kernel for bivariate normal density
  cor <- cov_xy / sd_x / sd_y
  numer <- (x - mu_x)^2 / sd_x^2 + (y - mu_y)^2 / sd_y^2 -
    2 * cor * (x - mu_x) * (y - mu_y) / sd_x / sd_y
  numer / (1 - cor^2)
}

contour_bvnorm <- function(mean1 = 0, sd1 = 1, mean2 = 0, sd2 = 1,
                           cor12 = 0, cov12 = NULL,
                           density = .95, length_out = 101,
                           bty = "L",...) {
  # Plot contour for a bivariate normal distribution
  #
  # Args:
  # mean1: mean of first normal distribution (on x-axis).
  # sd1: standard deviation of first normal distribution.
  # mean2: mean of second normal distribution (on y-axis).
  # sd2: standard deviation of second normal distribution.
  # cor12: correlation in the bivariate normal.
  # cov12: covariance in the bivariate normal. If not input, compute the
  # covariance using the correlation and the standard deviations.
  # density: density level, i.e., probability enclosed by the ellipse.
  # length_out: number of values on the x-axis and on the y-axis to be
  # evaluated; default to 101.
  # bty: argument passed to the `contour` function.
  # ...: other arguments passed to the `countour` fun
  #
  # Returns:
  # a plot showing the contour of the bivariate normal distribution on
  # a two-dimensional space.
  # Error handling
  stopifnot(cor12 >= -1, cor12 <= 1)
  if (is.null(cov12)) cov12 <- cor12 * sd1 * sd2s
  x_seq <- mean1 + seq(-3, 3, length.out = length_out) * sd1
  y_seq <- mean2 + seq(-3, 3, length.out = length_out) * sd2
  z <- outer(x_seq, y_seq, .bvnorm_kernel, mu_x = mean1, mu_y = mean2,
          sd_x = sd1, sd_y = sd2, cov_xy = cov12)
  contour(x_seq, y_seq, z, levels = qchisq(density, 2), drawlabels = FALSE,
        bty = bty, ...)
}

.partit_bvnorm <- function(cut1, cut2, mean1 = 0, sd1 = 1, mean2 = 0, sd2 = 1,
                        cor12 = 0, cov12 = cor12 * sd1 * sd2) {
  # Helper function for computing summary statistics from a selection approach
  Sigma <- matrix(c(sd1^2, cov12, cov12, sd2^2), nrow = 2)
  C <- pmnorm(c(cut1, cut2), c(mean1, mean2), Sigma)
  B <- pnorm(cut1, mean1, sd1) - C
  D <- pnorm(cut2, mean2, sd2) - C
  A <- 1 - B - C - D
  propsel <- A + B
  success_ratio <- A / propsel
  sensitivity <- A / (A + D)
  specificity <- C / (C + B)
  c(A, B, C, D, propsel, success_ratio, sensitivity, specificity)
}

PartInv <- function(propsel, cut_z = NULL, kappa_r, kappa_f = kappa_r,
                  phi_r, phi_f = phi_r, lambda_r, lambda_f = lambda_r,
                  Theta_r, Theta_f = Theta_r, tau_r, tau_f = tau_r,
                  pmix_ref = 0.5, plot_contour = TRUE, ...) {
  # Evaluate partial measurement invariance using Millsap & Kwok's (2004)
  # approach
  #
  # Args:
  # propsel: proportion of selection. If missing, computed using `cut_z`.
  # cut_z: prespecified cutoff score on the observed composite. This argument
  # is ignored when `propsel` has input.
  # kappa_r: latent factor mean for the reference group.
  # kappa_f: (optional) latent factor mean for the focal group;
  # if no input, set equal to kappa_r.
  # phi_r: latent factor variance for the reference group.
  # phi_f: (optional) latent factor variance for the focal group;
  # if no input, set equal to phi_r.
  # lambda_r: a vector of factor loadings for the reference group.
  # lambda_f: (optional) a vector of factor loadings for the focal group;
  # if no input, set equal to lambda_r.
  # tau_r: a vector of measurement intercepts for the reference group.
  # tau_f: (optional) a vector of measurement intercepts for the focal group;
  # if no input, set equal to tau_r.
  # Theta_r: a matrix of the unique factor variances and covariances
  # for the reference group.
  # Theta_f: (optional) a matrix of the unique factor variances and
  # covariances for the focal group;
  # if no input, set equal to Theta_r.
  # pmix_ref: Proportion of the reference group;
  # default to 0.5 (i.e., two populations have equal size).
  # plot_contour: logical; whether the contour of the two populations
  # should be plotted; default to TRUE.
  #    other arguments passed to the `countour` funcction.
  #
  # Returns:
  # a list of four elements and a plot if plot_contour == TRUE:
  # - propsel: echo the same argument as input.
  # - cutpt_xi: cut point on the latent scale (xi).
  # - cutpt_z: cut point on the observed scale (Z).
  # - summary: A 8 x 2 table, with columns representing the reference
  # and the focal groups, and the rows represent
  # probabilities of true positive (A), false positive (B),
  # true negative (C), false negative (D); proportion selected,
  # success ratio, sensitivity, and specificity.
  # Error handling
  stopifnot(length(kappa_r) == 1, length(kappa_f) == 1, length(phi_r) == 1,
            length(phi_f) == 1)
  #if (length(Theta_r) == length(lambda_r)) Theta_r <- diag(Theta_r) #commenting these 2 lines out
  #if (length(Theta_f) == length(lambda_f)) Theta_f <- diag(Theta_f) #were causing problems
  library(mnormt) # load `mnormt` package
  mean_zr <- sum(tau_r) + sum(lambda_r) * kappa_r
  mean_zf <- sum(tau_f) + sum(lambda_f) * kappa_f
  sd_zr <- sqrt(sum(lambda_r)^2 * phi_r + sum(Theta_r))
  sd_zf <- sqrt(sum(lambda_f)^2 * phi_f + sum(Theta_f))
  cov_z_xir <- sum(lambda_r) * phi_r
  cov_z_xif <- sum(lambda_f) * phi_f
  sd_xir <- sqrt(phi_r)
  sd_xif <- sqrt(phi_f)
  if (!missing(propsel)) {
    if (!is.null(cut_z)) {
      warning("Input to `cut_z` is ignored.")
    }
    cut_z <- qnormmix(propsel, mean_zr, sd_zr, mean_zf, sd_zf,
                      pmix_ref, lower.tail = FALSE)
  } else if (!is.null(cut_z) & missing(propsel)) {
    propsel <- pnormmix(cut_z, mean_zr, sd_zr, mean_zf, sd_zf,
                        pmix_ref, lower.tail = FALSE)
  }
  cut_xi <- qnormmix(propsel, kappa_r, sd_xir, kappa_f, sd_xif,
                     pmix_ref, lower.tail = FALSE)
  partit_1 <- .partit_bvnorm(cut_xi, cut_z, kappa_r, sd_xir, mean_zr, sd_zr,
                             cov12 = cov_z_xir)
  partit_2 <- .partit_bvnorm(cut_xi, cut_z, kappa_f, sd_xif, mean_zf, sd_zf,
                             cov12 = cov_z_xif)
  dat <- data.frame("Reference" = partit_1, "Focal" = partit_2,
                  row.names = c("A (true positive)", "B (false positive)",
                                  "C (true negative)", "D (false negative)",
                                  "Proportion selected", "Success ratio",
                                  "Sensitivity", "Specificity"))
  p <- NULL
  if (plot_contour) {
    x_lim <- range(c(kappa_r + c(-3, 3) * sd_xir,
                     kappa_f + c(-3, 3) * sd_xif))
    y_lim <- range(c(mean_zr + c(-3, 3) * sd_zr,
                     mean_zf + c(-3, 3) * sd_zf))
    contour_bvnorm(kappa_r, sd_xir, mean_zr, sd_zr, cov12 = cov_z_xir,
                   xlab = bquote("Latent Score" ~ (xi)),
                   ylab = bquote("Observed Composite" ~ (italic(Z))),
                   lwd = 2, col = "red", xlim = x_lim, ylim = y_lim,
                   ...)
    contour_bvnorm(kappa_f, sd_xif, mean_zf, sd_zf, cov12 = cov_z_xif,
                   add = TRUE, lty = "dashed", lwd = 2, col = "blue",
                   ...)
    legend("topleft", c("Reference group", "Focal group"),
           lty = c("solid", "dashed"), col = c("red", "blue"))
    abline(h = cut_z, v = cut_xi)
    x_cord <- rep(cut_xi + c(.25, -.25) * sd_xir, 2)
    y_cord <- rep(cut_z + c(.25, -.25) * sd_zr, each = 2)
    text(x_cord, y_cord, c("A", "B", "D", "C"))
    p <- recordPlot()
  }
  list(propsel = propsel, cutpt_xi = cut_xi, cutpt_z = cut_z,
       summary = round(dat, 3), p = p)
}
```

Appendix D: Replicating Millsap and Kwok's (2004) Test Anxiety Inventory Data

```
#Original Estimates
PartInv(.10, kappa_r=0, kappa_f=-.126, phi_r=.544, phi_f=.477,
      lambda_r = 7.279, lambda_f=7.679, tau_r=16.621, tau_f=16.581, Theta_r = 4.329, Theta_f = 3.405,
      pmix_ref = .5)

#Measurement Invariance Assumed
PartInv(.10, kappa_r=0, kappa_f=-.126, phi_r=.544, phi_f=.477,
      lambda_r = 7.454, lambda_f=7.454, tau_r=16.604, tau_f=16.604, Theta_r = 3.925, Theta_f = 3.925,
      pmix_ref = .5)
```

Appendix E: Software for Executing the Modified Version of 4)Millsap Method

```
###helper function to create table of proportions for method
prop.getter<-function(kappa.r, kappa.f, phi.r, phi.f,
                   lambda.r, lambda.f, tau.r, tau.f, Theta.r, Theta.f,
                   pmix.ref){

  proportions<-adply(seq(.01,.99,by=.01),1,function(x){

    j<-1-x

    analysis<-PartInv(propsel=j, plot_contour = FALSE,
            kappa_r=kappa.r,kappa_f=kappa.f,phi_r=phi.r, phi_f=phi.f,
            lambda_r=lambda.r, lambda_f=lambda.f, tau_r=tau.r,tau_f=tau.f,Theta_r=Theta.r,
Theta_f=Theta.f,
            pmix_ref=pmix.ref)

    props.r<-analysis$summary[1:4,"Reference"]
    props.f<-analysis$summary[1:4,"Focal"]
    row.out<-c(x,props.r,props.f)
    names(row.out)<-c("percentile","TP.r","FP.r","TN.r","FN.r","TP.f","FP.f","TN.f","FN.f")
    row.out
  })
  proportions
}

#general function for modified method
MillsapKwok2<-function(kappa.r, kappa.f, phi.r, phi.f,
                   lambda.r, lambda.f, tau.r, tau.f, Theta.r, Theta.f,
                   pmix.ref,method="Sensitivity",plot="Ratio"){

  prop.table<-prop.getter(kappa.r, kappa.f, phi.r, phi.f,
                     lambda.r, lambda.f, tau.r, tau.f, Theta.r, Theta.f,
                     pmix.ref)

  if(method=="Sensitivity"){
    ratios<-adply(1:99,1,function(x){
      arow<-prop.table[x,]
      Sens.r<-arow$TP.r/(arow$TP.r + arow$FN.r)
      Sens.f<-arow$TP.f/(arow$TP.f + arow$FN.f)
      out<-c(acc.r=Sens.r,acc.f=Sens.f, ratio=Sens.r/Sens.f)
      out
    })
  }

  if(method=="Specificity"){
    ratios<-adply(1:99,1,function(x){
```

```
    arow<-prop.table[x,]
    Spec.r<-arow$TN.r/(arow$TN.r + arow$FP.r)
    Spec.f<-arow$TN.f/(arow$TN.f + arow$FP.f)
    out<-c(acc.r=Spec.r,acc.f=Spec.f, ratio=Spec.r/Spec.f)
    out
  })
}

if(method=="Accuracy"){
  ratios<-adply(1:99,1,function(x){
    arow<-prop.table[x,]
    Acc.r<-(arow$TN.r+arow$TP.r)
    Acc.f<-(arow$TN.f+arow$TP.f)
    out<-c(acc.r=Acc.r,acc.f=Acc.f, ratio=Acc.r/Acc.f)
    out
  })
}

#false discovery rate
if(method=="FDR"){
  ratios<-adply(1:99,1,function(x){
    arow<-prop.table[x,]
    FDR.r<-arow$FP.r/(arow$TP.r + arow$FP.r)
    FDR.f<-arow$FP.f/(arow$TP.f + arow$FP.f)
    out<-c(acc.r=FDR.r,acc.f=FDR.f, ratio=FDR.r/FDR.f)
    out
  })
}

#positive predicted value
if(method=="PPV"){
  ratios<-adply(1:99,1,function(x){
    arow<-prop.table[x,]
    PPV.r<-arow$TP.r/(arow$TP.r + arow$FP.r)
    PPV.f<-arow$TP.f/(arow$TP.f + arow$FP.f)
    out<-c(acc.r=PPV.r,acc.f=PPV.f, ratio=PPV.r/PPV.f)
    out
  })
}

#negative predicted value
if(method=="NPV"){
  ratios<-adply(1:99,1,function(x){
    arow<-prop.table[x,]
    NPV.r<-arow$TN.r/(arow$TN.r + arow$FN.r)
    NPV.f<-arow$TN.f/(arow$TN.f + arow$FN.f)
    out<-c(acc.r=NPV.r,acc.f=NPV.f, ratio=NPV.r/NPV.f)
    out
    })
  }

  #false ommisson rate
  if(method=="FOR"){
    ratios<-adply(1:99,1,function(x){
      arow<-prop.table[x,]
      FOR.r<-arow$FN.r/(arow$TN.r + arow$FN.r)
      FOR.f<-arow$FN.f/(arow$TN.f + arow$FN.f)
      out<-c(acc.r=FOR.r,acc.f=FOR.f, ratio=FOR.r/FOR.f)
      out
    })
  }

  #false positive rate
  if(method=="FPR"){
    ratios<-adply(1:99,1,function(x){
      arow<-prop.table[x,]
      FPR.r<-arow$FP.r/(arow$TN.r + arow$FP.r)
      FPR.f<-arow$FP.f/(arow$TN.f + arow$FP.f)
      out<-c(acc.r=FPR.r,acc.f=FPR.f, ratio=FPR.r/FPR.f)
      out
    })
  }

  if(method=="FNR"){
    ratios<-adply(1:99,1,function(x){
      arow<-prop.table[x,]
      FNR.r<-arow$FN.r/(arow$TP.r + arow$FN.r)
      FNR.f<-arow$FN.f/(arow$TP.f + arow$FN.f)
      out<-c(acc.r=FNR.r,acc.f=FNR.f, ratio=FNR.r/FNR.f)
      out
    })
  }

  ### post-processing
  if(plot=="Ratio"){
    plot(x=1:99,y=log10(ratios$ratio),xlab="Percentile",ylab="log10(ratio)",
         ylim=c(log10(.5),log10(2)),type="l",main=method)
  }

  if(plot=="Identity"){
    plot(x=1:99,y=ratios$acc.r, xlab="Percentile", ylab=method, ylim=c(0,1),
         type="l", main=method, col="red")
    lines(x=1:99,y=ratios$acc.f,col="blue")
  }
  if(plot=="None"){return(ratios)}
}
```

Appendix F: Sample Code for Generating Multiple Classification Accuracy Ratio Plots on one Figure

```
#Multiple Conditions per plot, 1 plot per measure of accuracy.
#sensitivity
#condition 1 data
sens.1<-MillsapKwok2(0,0,1,1,
            4.2,4.8,0,0,3.06,3.06,
            .8,"Sensitivity","None")
#condition 2 data
sens.2<-MillsapKwok2(0,0,1,1,
            4.2,4.2,0,.6,3.06,3.06,
            .8,"Sensitivity","None")
#condition 3 data
sens.3<-MillsapKwok2(0,0,1,1,
            4.2,4.2,0,0,3.06,4.86,
            .8,"Sensitivity","None")
#condition 4 data
sens.4<-MillsapKwok2(0,0,1,1,
            4.2,4.8,0,.6,3.06,3.06,
            .8,"Sensitivity","None")
#condition 5 data
sens.5<-MillsapKwok2(0,0,1,1,
            4.2,4.8,0,-.6,3.06,3.06,
            .8,"Sensitivity","None")

#Condition 1 plot (base plot)
plot(x=1:99,y=log10(sens.1$ratio),type="l",ylim=c(log10(.5),log10(2)),
   main="Sensitivity (pmix=.8)",xlab="Percentile",ylab="log10(Sensitivity Ratio)")
points(x=c(1,10,20,30,40,50,60,70,80,90,99),
    y=log10(sens.1$ratio)[c(1,10,20,30,40,50,60,70,80,90,99)],
    pch=0) #square

#condition 2 plot
lines(1:99,y=log10(sens.2$ratio))
points(x=c(1,10,20,30,40,50,60,70,80,90,99),
    y=log10(sens.2$ratio)[c(1,10,20,30,40,50,60,70,80,90,99)],
    pch=1) #circle

#condition 3 plot
lines(1:99,y=log10(sens.3$ratio))
points(x=c(1,10,20,30,40,50,60,70,80,90,99),
    y=log10(sens.3$ratio)[c(1,10,20,30,40,50,60,70,80,90,99)],
    pch=2) #triangle

#condition 4 plot
lines(1:99,y=log10(sens.4$ratio))
points(x=c(1,10,20,30,40,50,60,70,80,90,99),
    y=log10(sens.4$ratio)[c(1,10,20,30,40,50,60,70,80,90,99)],
    pch=3) #+

#condition 5 plot
lines(1:99,y=log10(sens.5$ratio))
points(x=c(1,10,20,30,40,50,60,70,80,90,99),
    y=log10(sens.5$ratio)[c(1,10,20,30,40,50,60,70,80,90,99)],
    pch=4) #x

for(i in seq(-.3,.3,by=.05)){abline(h=i,lty="dotted")}
```

Appendix G: Sample Code for Simulating Finite Sample Estimates Of Classification Accuracy Ratio Plots

```
#defining the partial measurement invariance model being fit to each simulated dataset.
ana.model.c1<-"f1=~c(a1,a1)*y1+y2+c(a2,a2)*y3+y4+c(a3,a3)*y5+y6
y1~c(b1,b1)*1
y3~c(b2,b2)*1
y5~c(b3,b3)*1

y1~~c(c1,c1)*y1
y3~~c(c1,c1)*y3
y5~~c(c1,c1)*y5

f1~c(0,NA)*1"

#generates data, fits models, retrieves log10(sensitivity) estimates for each replication and percentile
temp<-adply(1:100,1,function(x){
  N1<-200
  N2<-200

  #datagen
  f.r<-rnorm(N1,0,1)
  f.f<-rnorm(N2,0,1)

  y1<-.7*f.r+rnorm(N1,0,sqrt(.21))
  y2<-.7*f.r+rnorm(N1,0,sqrt(.21))
  y3<-.7*f.r+rnorm(N1,0,sqrt(.21))
  y4<-.7*f.r+rnorm(N1,0,sqrt(.21))
  y5<-.7*f.r+rnorm(N1,0,sqrt(.21))
  y6<-.7*f.r+rnorm(N1,0,sqrt(.21))

  g1<-data.frame(y1,y2,y3,y4,y5,y6,group=1)

  y1<-.7*f.f+rnorm(N2,0,sqrt(.21))
  y2<-.9*f.f+rnorm(N2,0,sqrt(.21))+.2
  y3<-.7*f.f+rnorm(N2,0,sqrt(.21))
  y4<-.9*f.f+rnorm(N2,0,sqrt(.21))+.2
  y5<-.7*f.f+rnorm(N2,0,sqrt(.21))
  y6<-.9*f.f+rnorm(N2,0,sqrt(.21))+.2

  g2<-data.frame(y1,y2,y3,y4,y5,y6,group=2)

  fulldat<-rbind(g1,g2)

  #model fitting
  run0<-sem(ana.model.c1,fulldat,group="group")

  #matrix extraction
  lambda.r<-sum(inspect(run0,"est")$`1`$lambda)
  tau.r<-sum(inspect(run0,"est")$`1`$nu)
  theta.r<-sum(diag((inspect(run0,"est")$`1`$theta)))
  kappa.r<-(inspect(run0,"est")$`1`$alpha)
  phi.r<-(inspect(run0,"est")$`1`$psi)

  lambda.f<-sum(inspect(run0,"est")$`2`$lambda)
  tau.f<-sum(inspect(run0,"est")$`2`$nu)
  theta.f<-sum(diag((inspect(run0,"est")$`2`$theta)))
  kappa.f<-(inspect(run0,"est")$`2`$alpha)
  phi.f<-(inspect(run0,"est")$`2`$psi)


  log10(MillsapKwok(kappa.r=kappa.r, kappa.f=kappa.f, phi.r=phi.r, phi.f=phi.f,
                lambda.r=lambda.r, lambda.f=lambda.f, tau.r=tau.r, tau.f=tau.f, Theta.r=theta.r,
Theta.f=theta.f,
                pmix.ref=.5, method="Sensitivity",plot=FALSE)$ratio)
})

temp<-temp[,-1]

#plotting upper bound of CI
plot(1:99,sapply(temp,function(x){
  quantile(x,.975)
}),ylim=c(log10(.5),log10(2)),type="l", main="Sensitivity", ylab="log10(Sensitivity Ratio)",xlab="percentile")

#plotting lower bound of CI
lines(1:99,sapply(temp,function(x){
  quantile(x,.025)
}),ylim=c(log10(.5),log10(2)),type="l")

#plotting population values
popline<-log10(MillsapKwok(kappa.r=0, kappa.f=0, phi.r=1, phi.f=1,
                      lambda.r=4.2, lambda.f=4.8, tau.r=0, tau.f=.6, Theta.r=3.06, Theta.f=3.06,
                      pmix.ref=.5, method="Sensitivity",plot=FALSE)$ratio)

lines(1:99,popline, col="red",lwd=1)

for(i in seq(-.3,3,by=.05)){abline(h=i,lty="dotted")}
```

Appendix H: Sample Code Applying Software to Evers

```
#set up 2x2 panel
par(mfrow=c(2,2))

#sensitivity
MillsapKwok2(kappa.r=0, kappa.f=-.126, phi.r=.544, phi.f=.477,
            lambda.r=7.279, lambda.f=7.679, tau.r=16.621, tau.f=16.581, Theta.r=4.329, Theta.f=3.405,
            pmix.ref=.5,method="Sensitivity",plot="Ratio") #plot="Identity" gives group estimates
for(i in seq(-.3,3,by=.05)){abline(h=i,lty="dotted")}

#specificity
MillsapKwok2(kappa.r=0, kappa.f=-.126, phi.r=.544, phi.f=.477,
         lambda.r=7.279, lambda.f=7.679, tau.r=16.621, tau.f=16.581, Theta.r=4.329, Theta.f=3.405,
         pmix.ref=.5,method="Specificity",plot="Ratio")
for(i in seq(-.3,3,by=.05)){abline(h=i,lty="dotted")}

#PPV
MillsapKwok2(kappa.r=0, kappa.f=-.126, phi.r=.544, phi.f=.477,
         lambda.r=7.279, lambda.f=7.679, tau.r=16.621, tau.f=16.581, Theta.r=4.329, Theta.f=3.405,
         pmix.ref=.5,method="PPV",plot="Ratio")
for(i in seq(-.3,3,by=.05)){abline(h=i,lty="dotted")}

#NPV
MillsapKwok2(kappa.r=0, kappa.f=-.126, phi.r=.544, phi.f=.477,
         lambda.r=7.279, lambda.f=7.679, tau.r=16.621, tau.f=16.581, Theta.r=4.329, Theta.f=3.405,
         pmix.ref=.5,method="NPV",plot="Ratio")
for(i in seq(-.3,3,by=.05)){abline(h=i,lty="dotted")}
```

Appendix I: Sample Bollen-Stine Bootstrap for Estimating Variability of Classification Accuracy Ratio Plots Applied to Everson et al.'s (1991) Data

```
#final partial measurement invariance model
ana.model.real<-"f1=~c(a1,a1)*y1+c(a2,a2)*y2+y3+y4+c(a5,a5)*y5+c(a6,a6)*y6+c(a7,a7)*y7+c(a8,a8)*y8
y1~c(b1,b1)*1
y2~c(b2,b2)*1
y5~c(b5,b5)*1
y6~c(b6,b6)*1
y7~c(b7,b7)*1
y8~c(b8,b8)*1

f1~c(0,NA)*1"

for(j in c("Sensitivity","Specificity","PPV","NPV")){
  temp<-adply(1:100,1,function(x){
    N1<-282
    N2<-219

    #datagen
    f.r<-rnorm(N1,0,sqrt(.544))
    f.f<-rnorm(N2,-.126,sqrt(.477))

    #female
    y1<-.836*f.r+rnorm(N1,0,sqrt(.517))+2.114
    y2<-1.00*f.r+rnorm(N1,0,sqrt(.523))+2.064
    y3<-.904*f.r+rnorm(N1,0,sqrt(.631))+1.901
    y4<-.808*f.r+rnorm(N1,0,sqrt(.585))+2.004
    y5<-.903*f.r+rnorm(N1,0,sqrt(.481))+2.144
    y6<-.960*f.r+rnorm(N1,0,sqrt(.469))+1.985
    y7<-.934*f.r+rnorm(N1,0,sqrt(.551))+2.179
    y8<-.934*f.r+rnorm(N1,0,sqrt(.572))+2.230

    g1<-data.frame(y1,y2,y3,y4,y5,y6,y7,y8,group=1)

    #male
    y1<-.836*f.r +rnorm(N1,0,sqrt(.514))+2.114
    y2<-1.00*f.r +rnorm(N1,0,sqrt(.407))+2.064
    y3<-1.111*f.r+rnorm(N1,0,sqrt(.371))+1.880
    y4<-1.001*f.r+rnorm(N1,0,sqrt(.475))+1.985
    y5<-.903*f.r +rnorm(N1,0,sqrt(.392))+2.144
    y6<-.960*f.r +rnorm(N1,0,sqrt(.335))+1.985
    y7<-.934*f.r +rnorm(N1,0,sqrt(.454))+2.179
    y8<-.934*f.r +rnorm(N1,0,sqrt(.457))+2.230

    g2<-data.frame(y1,y2,y3,y4,y5,y6,y7,y8,group=2)
    fulldat<-rbind(g1,g2)

    #model fitting
    run0<-sem(ana.model.real,fulldat,group="group")

    #matrix extraction
    lambda.r<-sum(inspect(run0,"est")$`1`$lambda)
    tau.r<-sum(inspect(run0,"est")$`1`$nu)
    theta.r<-sum(diag((inspect(run0,"est")$`1`$theta)))
    kappa.r<-(inspect(run0,"est")$`1`$alpha)
    phi.r<-(inspect(run0,"est")$`1`$psi)

    lambda.f<-sum(inspect(run0,"est")$`2`$lambda)
    tau.f<-sum(inspect(run0,"est")$`2`$nu)
    theta.f<-sum(diag((inspect(run0,"est")$`2`$theta)))
    kappa.f<-(inspect(run0,"est")$`2`$alpha)
    phi.f<-(inspect(run0,"est")$`2`$psi)


    log10(MillsapKwok(kappa.r=kappa.r, kappa.f=kappa.f, phi.r=phi.r, phi.f=phi.f,
                      lambda.r=lambda.r, lambda.f=lambda.f, tau.r=tau.r, tau.f=tau.f, Theta.r=theta.r,
Theta.f=theta.f,
                      pmix.ref=.5, method=j,plot=FALSE)$ratio)
  })
  temp<-temp[,-1]


  plot(1:99,sapply(temp,function(x){
    quantile(x,.975)
  }),ylim=c(log10(.5),log10(2)),type="l", main=j, ylab=paste0("log10(",j, " Ratio)"),xlab="percentile")

  lines(1:99,sapply(temp,function(x){
    quantile(x,.025)
  }),ylim=c(log10(.5),log10(2)),type="l")

  popline<-log10(MillsapKwok(kappa.r=0, kappa.f=-.126, phi.r=.544, phi.f=.477,
                             lambda.r = 7.279, lambda.f=7.679, tau.r=16.621, tau.f=16.581, Theta.r = 4.329, Theta.f =
3.405,
                             pmix.ref = .5, method=j,plot=FALSE)$ratio)

  lines(1:99,popline, col="red",lwd=1)

  for(i in seq(-.3,3,by=.05)){abline(h=i,lty="dotted")}
}
```

Appendix J: RMSEA Confidence Intervals and Tests of Close Fit

RMSEA's popularity is, in part, motivated

its finite sample estimate can easily be obtained. The 90% confidence interval about RMSEA is computed as follows: first, the 90% confidence interval about the non-centrality parameter, $\hat{l}_{ML,N}$, is computed. The confidence interval is given as $[\ \hat{l}_{ML,.05},\ \hat{l}_{ML,.95}\ ]$, where $\hat{l}_{ML,.05}$ is the non-centrality parameter for the non-central chi-square distribution with *df* degrees of freedom for which T_{ML} is the 95th percentile, and $\hat{l}_{ML,.95}$ is the non-centrality parameter for the non-central chi-square distribution with *df* degrees of freedom T_{ML} for whi is the 5th percentile. The confidence interval about the non-centrality parameter can then be converted into a confidence interval about the RMSEA by converting it to the RMSEA metric:

$$RMSEA_{.05} = \sqrt{\frac{G*\ \hat{l}_{ML,.05}}{df\,(N-G)}},\ RMSEA_{.95} = \sqrt{\frac{G*\ \hat{l}_{ML,.95}}{df\,(N-G)}}$$ (Browne & Cudeck, 1993).

Because a given RMSEA value can easily be converted to the non-centrality parameter for a chi-square distribution –given *N*, *G*, and *df* – it can also be used to sp and not-close fit (Browne & Cudeck, 1993)." Rather than testing the perfect fit, tests of close fit allow the researcher to test H_0 *: RMSEA* ¢c, where *c* is a critical RMSEA deemed the largest acceptable value for model retention. For a specified *c*, the reference chi-square distribution against which one tests their observed *T* has non-centrality parameter $\hat{l}_{ML} = \frac{(N-G)*df*c^2}{G}$. If *T* falls above the 95th percentile of $T \sim c^2(df,\ \hat{l}_{ML})$, the researcher rejects the null hypothesis of close fit, and retains $H_1 : RMSEA > c$, implying inadequate fit. This method can also be used to specify tests of not-close fit, which test the null hypothesis H_0 *: RMSEA* ²c, where *c* is a critical RMSEA deemed the smallest unacceptably large value. If

T falls below the 5th percentile of the reference distribution, the researcher rejects the null hypothesis of not-close fit and retains $H_1 : RMSEA < c$, implying adequate fit. Note that in the case of tests of close fit, rejection of the null hypothesis implies rejection of one's model of interest, while rejection of the null hypothesis in a test of not-close fit implies re model of interest.

Appendix K: RMSEA_D Confidence Intervals and Tests of Close Fit

The 90% confidence interval about $RMSEA_{D,ML,N}$ can be computed as follows: first, the 90% confidence interval about the non-centrality parameter, $\hat{\lambda}_{D,ML}$, is computed. The confidence interval is given as $[\ \hat{\lambda}_{D,ML,.05},\ \hat{\lambda}_{D,ML,.95}\ ]$, where $\hat{\lambda}_{D,ML,.05}$ is the non-centrality parameter for the non-central chi-square distribution with df_D degrees of freedom for which observed D_{ML} is the 95th percentile, and $\hat{\lambda}_{D,ML,.95}$ is the non-centrality parameter for the non-central chi-square distribution with df_D degrees of freedom for which D_{ML} is the 5th percentile. The confidence interval about the non-centrality parameter can be converted into a confidence interval about $RMSEA_{D,ML,N}$ by converting it to the RMSEA metric:

$$RMSEA_{D,ML,N,.05} = \sqrt{\frac{G*\ \hat{\lambda}_{D,ML,.05}}{df_D(N-G)}},\ RMSEA_{D,ML,N,.95} = \sqrt{\frac{G*\ \hat{\lambda}_{D,ML,.95}}{df_D(N-G)}}\ .$$

$RMSEA_D$ can also be used to specify tests of close fit and tests of not-close fit, similar to those proposed by Browne and Cudeck (1993). The non-centrality parameter for the reference distribution, $D \sim \chi^2(df_D,\ \hat{\lambda}_{D,ML})$, is computed as $\hat{\lambda}_{D,ML} = \dfrac{c^2 * df_D * (N-G)}{G}$, where c is a specified $RMSEA_D$ deemed the largest acceptable value. The null hypothesis, $H_0 : RMSEA_D \le c$, is tested by comparing observed D_{ML} to the reference distribution, $D_{ML} \sim \chi^2(df_D,\ \hat{\lambda}_{D,ML})$. If D_{ML} falls above the 95th percentile, the null hypothesis of close fit is rejected, and $H_1 : RMSEA > c$ is retained. This method can also be used to evaluate the hypothesis of not-close fit, $H_0 : RMSEA \ge c$, where c is a specified $RMSEA_D$ deemed the largest unacceptable value. If one observed D_{ML} falls below the 5th percentile of the reference distribution

$D_{ML} \sim c^2(df_D, J_{D,ML})$, the null hypothesis of not-close fit is rejected, and $H_1 : RMSEA < c$ is retained. It is worth noting that $RMSEA_D$ -based tests of close and not-close fit are not identical to the tests of small change in fit and tests of not-small change in fit proposed by MacCallum, Browne, & Cai (2006), which involve specification of a baseline model *RMSEA* and a restricted model *RMSEA*.