ABSTRACT

Title of Document:	TESTING FOR DIFFERENTIALLY FUNCTIONING INDICATORS USING MIXTURES OF CONFIRMATORY FACTOR ANALYSIS MODELS
	Heather Marie Mann, Doctor of Philosophy, 2009
Directed By:	Professor Gregory R. Hancock, Department of Measurement, Statistics & Evaluation

Heterogeneity in measurement model parameters across known groups can be modeled and tested using multigroup confirmatory factor analysis (CFA). When it is not reasonable to assume that parameters are homogeneous for all observations in a manifest group, mixture CFA models are appropriate. Mixture CFA models can add theoretically important unmeasured characteristics to capture heterogeneity and have the potential to be used to test measurement invariance. The current study investigated the ability of mixture CFA models to identify differences in factor loadings across latent classes when there is no mean separation in both the latent and measured variables. Using simulated data from models with known parameters, parameter recovery, classification accuracy, and the power of the likelihood-ratio test were evaluated as impacted by model complexity, sample size, latent class proportions, magnitude of factor loading differences, percentage of noninvariant factor loadings, and pattern of noninvariant factor loadings. Results suggested that

mixture CFA models may be a viable option for testing the invariance of measurement model parameters, but without impact and differences in measurement intercepts, larger sample sizes, more noninvariant factor loadings, and larger amounts of heterogeneity are needed to distinguish different latent classes and successfully estimate their parameters.

TESTING FOR DIFFERENTIALLY FUNCTIONING INDICATORS USING MIXTURES OF CONFIRMATORY FACTOR ANALYSIS MODELS

By

Heather Marie Mann

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Gregory R. Hancock, Chair
Professor Jeffrey R. Harring
Professor Robert W. Lissitz
Professor Robert J. Mislevy
Professor Ellin K. Scholnick

Dedication

To Alex, for your love and friendship, and for going through this together with me.

To Mom and Dad, for your support of all my education and for encouraging me to work hard.

Acknowledgements

Thank you, Dr. Gregory R. Hancock, for being an outstanding academic and dissertation advisor. You were a great influence on the way I conduct and write about my research. You helped to prepare me with the skills, knowledge, and confidence needed to become an independent researcher. Your guidance up to and throughout the dissertation process made this research something from which I have learned so much and of which I can be proud.

I am thankful for the helpful advice from Dr. Robert J. Mislevy, Dr. Jeffrey R. Harring, Dr. Robert W. Lissitz, and Dr. Ellin K. Scholnick. I am grateful for those who aided in the timely completion of this research, in particular Payal Patel for loaning me an extra computer, Alex Buzick for being patient and supportive, and my parents, Howard and Nancy Mann, for motivating me to push the limits of my capabilities. Thank you, Daisy Wise Rutstein, for our collaborations in research and learning and also for your friendship.

This research was supported in part by the Graduate School via an Ann G. Wylie Dissertation Fellowship.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In confirmatory factor analysis (CFA) modeling, it is assumed that the same latent variables are being measured by an instrument for all individuals in a population. Traditional CFA assumes a linear relationship between continuous latent variables and continuous observable variables, implying that variability in values of measured variables across individuals can be explained by individual differences in the values of the underlying latent variables. When relationships between latent variables and their manifest indicators are moderated by group membership, these assumptions are violated and methods to account for heterogeneity are required.

Modeling and testing for heterogeneity in the relationship between latent variables and their continuous measured indicator variables in cross-sectional CFA models, herein referred to as differential indicator functioning (CFA-DIF) is of interest when the assumption of a homogeneous measurement scale may not be tenable. Testing for CFA-DIF is both practically and theoretically important to ensure that inferences drawn are about the intended population and that cross-population comparisons are accurate and valid. Validity can be substantially comprised if a measurement instrument is assumed to be invariant across populations that are being compared at the structural level when it truly is not or when inferences about a homogeneous population are drawn from a single-population model using data that is truly derived from two or more populations with noninvariant parameters.

When theory suggests that heterogeneity across populations may exist in a measurement instrument, different choices of modeling techniques are available, depending on whether or not characteristics that determine population membership

are observable. When population membership is known a priori, multigroup CFA can be used to test for CFA-DIF. When the source of heterogeneity is unobserved and population membership is unknown a priori, finite mixture modeling can be combined with CFA to model and test for CFA-DIF across latent classes.

The majority of applications to date assume that observed moderator variables are sufficient to separate data into multiple homogeneous groups, making multigroup CFA the most popular model for use in testing measurement invariance with continuous data. However, if manifest characteristics are insufficient proxies for underlying characteristics that define population membership, hypothesis tests may favor invariance with respect to the manifest groups while true differences across populations would remain undetected. Allowing parameters to vary across latent classes may result in better accuracy of parameter estimates and validity of inferences about heterogeneity when true moderating characteristics have not been observed and do not overlap well with available manifest grouping variables.

Given that latent classes may be more appropriate than manifest groups to use in defining population membership, the current study examined the feasibility of mixture CFA to model and test for CFA-DIF. Differentially functioning continuous indicator variables, like dichotomous/polytomous items, are a concern in many areas in the social and behavioral sciences, for example cross-cultural research and scale development. There is extensive methodological and applied literature on using item response theory (IRT) and CFA to model heterogeneous discrete item responses across populations with both observed and unobserved membership, using multigroup CFA to model heterogeneous continuous or discrete indicator variables across

observed populations, and using mixture CFA to model heterogeneous latent means across unobserved populations. However, little is known about how well mixture CFA can successfully model and test for heterogeneous measurement model parameters with continuous indicators across populations with unobserved population membership.

There is a need for research on the ability of mixture CFA to accurately estimate heterogeneous measurement model parameters and to test their invariance, in particular because mixture CFA models are more complex than multigroup CFA models. The advent of mixture CFA modeling occurred a little over a decade ago, yet the method has yet to achieve mainstream use in empirical research, likely because of the added complexity of modeling and estimation. The intent of the current study is to increase the consideration of latent classes when conducting invariance testing in a CFA framework by examining the viability of mixture CFA for modeling heterogeneity and conducting invariance tests on measurement model parameters using cross-sectional continuous data from multiple populations.

The current study evaluated the performance of mixture CFA for modeling and testing for CFA-DIF with data simulated under a variety of conditions that may occur in practice. For mixture CFA modeling to become popular for the purpose of testing measurement invariance, it has to be accessible to applied researchers. As such, the current study considered the capabilities of maximum-likelihood estimation using existing statistical software. There are several software programs available with the capacity to estimate mixture CFA model parameters with maximum-likelihood methods (Mplus, Muthén & Muthén, 2007; Mx, Neale, Boker, Xie, & Maes, 2000;

and Latent Gold, Vermunt & Magidson, 2000). Estimation in this study was performed in Mplus 5.1 (Muthén & Muthén, 2007).

The following chapter presents a review of existing research, providing a context and theoretical framework for testing CFA-DIF across latent classes and motivation for the current study. Chapter 3 details the design of the current study, including methods of estimation and analysis. Results are reported and described in Chapter 4. Chapter 5 comprises a discussion of results and limitations of the current study, recommendations for practitioners, and potential research extensions.

Chapter 2: Review of Literature

This chapter details existing research related to mixture CFA modeling and CFA-DIF. The review includes theoretical underpinnings of mixture modeling, substantive foundations of heterogeneous measurement model parameters, and statistical tests for invariance. Related methodological and empirical studies are also described, setting up a framework for modeling heterogeneity across latent classes and testing for CFA-DIF.

Theoretical background for mixture CFA modeling

The following sections describe general mixture modeling, major types of mixture models and exemplary applications. The mixture CFA model is introduced along with details about maximum-likelihood estimation in general and specifics about estimation in Mplus 5.1 (Muthén & Muthén, 2007). Special problems that can occur with mixture CFA models are also outlined.

Finite mixture distributions

A composite of several distributions in which component membership is unobserved for each observation is called a finite mixture distribution. Generally, finite mixture distributions provide a way to model the density of complex distributions and also can be used when it is not reasonable to assume that all observations arise from a homogeneous population. In the former case, when there is one population with a distribution that is dispersed or multi-modal, mixtures of two or more densities can approximate the distribution. In the latter case, there is more than one unobserved population and the intent is to infer or even form homogeneous

clusters of individuals. Pioneering research involving mixtures of distributions includes work by Karl Pearson (1894), who fit a mixture of two univariate normal distributions with different means and variances to measurements of the ratio of forehead to body length of crabs from the Bay of Naples to infer that the crabs had evolved into two separate species (see, for example, Cowles, 2000).

When modeling population heterogeneity using finite mixture distributions, it is typically assumed that data come from a mixture of two or more distributions from the same parametric family with parameters that are allowed to differ across components, or latent classes (e.g., Aitkin & Rubin, 1985; see also, McLachlan & Peel, 2000). The general form of a mixture of *C* densities can be specified as

$$f(y|\pi, \boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c \, f_c(y|\boldsymbol{\theta}_c),$$

where $f_c(y|\theta_c)$ is the probability density function (pdf) for class *c*, *c* = 1, …, *C*, with weight or class proportion $0 \leq \pi_c \leq 1$, and an unknown class-specific parameter vector, $\boldsymbol{\theta}_c$. The class-specific pdfs can take on a variety of forms to represent different types of modeling for different purposes, including density estimation, clustering, and random-effects modeling. The following are examples of major types of mixture distributions and their general uses.

Mixtures of univariate distributions, for example normal, with class-specific pdf

$$f_c(y|\mu_c, \sigma_c^2) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(y-\mu_c)^2}{2\sigma_c^2}},$$

where μ_c and σ_c^2 are the class-specific mean and variance, are often used to account for outliers or to estimate the density of complicated distributions (Magidson &

Vermunt, 2004). Mixtures of regression models, also known as latent class regression models, are used to model random-effects regression coefficients (Wedel & DeSarbo, 1994). For example, a latent class normal linear regression model has a class-specific pdf

$$f_c(y|\boldsymbol{\beta}_c'\mathbf{x}, \sigma_c^2) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(y-\boldsymbol{\beta}_c'\mathbf{x})^2}{2\sigma_c^2}},$$

with class-specific residual variance, σ_c^2, and matrix of class-specific regression coefficients, $\boldsymbol{\beta}_c$. Latent class analysis models (Lazarsfeld & Henry, 1968), used, for example, in educational research for response scaling or measuring rater agreement (Bergan, 1983), can be formulated in terms of mixtures of product multivariate Bernoulli probability densities with class-specific pdf

$$f_c(y|\theta_c) = \prod_{p=1}^{P} \theta_{cp}^{y_p} (1-\theta_{cp})^{(1-y_p)},$$

where $\boldsymbol{\theta}_c' = [\theta_{c1}, \ldots, \theta_{cP}]$ is a vector containing the probabilities that variables in the *c*th latent class equal unity (e.g., Keren & Lewis, 1993; McLachlan & Peel, 2000). Mixtures of item response theory (IRT) models, for example the mixed Rasch model (Rost, 1990), with class-specific pdf

$$f_c(y|\theta_j, \beta_i) = \prod_{j=1}^{J} \exp(\theta_{jc} + \beta_{ic})\, y_{ijc} / [1 + \exp(\theta_{jc} + \beta_{ic})],$$

where θ_{jc} is a class-specific, person-specific ability parameter, and β_{ic} is a class-specific item difficulty parameter, allow item parameters to differ across classes to reflect examinees who differ in their item responses due to item bias or secondary

latent traits. Mixtures of multivariate processes such as multivariate normal, with class-specific pdf

$$f_c(y|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_c|^{1/2}} e^{-\frac{1}{2}(y-\boldsymbol{\mu}_c)'\boldsymbol{\Sigma}^{-1}(y-\boldsymbol{\mu}_c)},$$

where $\boldsymbol{\mu}_c$ is a vector of class-specific means, and $\boldsymbol{\Sigma}_c$ is a class-specific variance-covariance matrix, are used for density estimation and clustering. Mixture structural equation models, within which mixture factor analysis models are nested, are derived from the multivariate normal density by imposing covariance (and mean) structure through restrictions on parameter matrices.

Mixture CFA model

The following equations define the general form of the finite mixture CFA model with continuous indicator variables for cross-sectional data. The measurement model with *p* indicators of *m* exogenous latent variables for observation *i* in latent class *c* is

$$\mathbf{X}_{ic} = \boldsymbol{\tau}_c + \boldsymbol{\Lambda}_c\boldsymbol{\xi}_{ic} + \boldsymbol{\delta}_{ic},$$

where i = 1 to N_c, $\boldsymbol{\tau}_c$ is a $p \times 1$ vector of measurement intercepts, $\boldsymbol{\Lambda}_c$ is a $p \times m$ matrix of factor loadings, $\boldsymbol{\xi}_{ic}$ is an $m \times 1$ vector of factor scores (i.e. values for individuals on a latent continuum), and $\boldsymbol{\delta}_{ic}$ is a $p \times 1$ vector of residuals. $\mathbf{X}_{ic}$ has an associated mean vector, $\boldsymbol{\mu}_{xc} = \boldsymbol{\tau}_c + \boldsymbol{\Lambda}_c\boldsymbol{\kappa}_c$, where $\boldsymbol{\kappa}_c$ is an $m \times 1$ vector of factor means, and covariance matrix $\boldsymbol{\Sigma}_c = \boldsymbol{\Lambda}_c\boldsymbol{\Phi}_c\boldsymbol{\Lambda}_c' + \boldsymbol{\Theta}_c$, where $\boldsymbol{\Phi}_c$ is an $m \times m$ factor variance/covariance matrix, and $\boldsymbol{\Theta}_c$ is a $p \times p$ residual variance/covariance matrix.

Assuming that $\boldsymbol{\xi}_{ic} \sim iid\ MVN(\boldsymbol{\kappa}_c, \boldsymbol{\Phi}_c)$ and $\boldsymbol{\delta}_{ic} \sim iid\ MVN(0, \boldsymbol{\Theta}_c)$, the finite mixture CFA density function is

$$f(x_i|\pi, \boldsymbol{\mu}_x, \boldsymbol{\Sigma}) = \sum_{c=1}^{C} \pi_c \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_c|^{1/2}} e^{-\frac{1}{2}(x_i - \boldsymbol{\mu}_{xc})' \boldsymbol{\Sigma}_c^{-1} (x_i - \boldsymbol{\mu}_{xc})}.$$

In mixture CFA modeling, it is typically assumed that each within-class model has the same structure across classes. Specifically, the number of factors is usually the same across classes, as well as the location of fixed and free parameters (e.g., Yung, 1997). Equivalent forms and numbers of factors help to facilitate comparisons of parameters across classes. Once the factor structure has been appropriately specified, individual parameters can either be class-specific or class-invariant. That is, heterogeneity can be captured in measurement model parameters (factor loadings, measurement intercepts, and/or residual variances) or in factor variances/covariances and means. Misspecification of the within-class model and violations of CFA model assumptions including linearity and normality of the continuous indicator variables may lead one to choose a model with additional spurious latent classes (Bauer & Curran, 2004).

Maximum likelihood estimation of mixture CFA models via the EM algorithm

The likelihood function for a mixture of CFA models can be written as

$$L = \prod_{i=1}^{n} f(x_i|\pi, \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \sum_{c=1}^{C} \pi_c \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_c|^{1/2}} e^{-\frac{1}{2}(x_i - \boldsymbol{\mu}_{xc})' \boldsymbol{\Sigma}_c^{-1} (x_i - \boldsymbol{\mu}_{xc})}$$

where $\boldsymbol{\mu}_{xc}$ and $\boldsymbol{\Sigma}_c$ are the previously given functions of $\boldsymbol{\Lambda}_c$, $\boldsymbol{\tau}_c$, and $\boldsymbol{\psi}_c$,. In typical maximum-likelihood estimation, the log-likelihood is maximized with respect to

model parameters to obtain estimates. However, because mixture CFA models contain latent variable values and latent class memberships that are both unobserved, there is no closed-form solution for the parameter estimates. Thus, the expectation-maximization (EM) algorithm or some modification of it needs to be used to obtain maximum-likelihood estimates (McLachlan & Krishnan, 2008). Other estimation methods include maximum-likelihood estimation with an approximate scoring algorithm described by Yung (1997), modifications of the EM algorithm, for example an alternating expectation-conditional maximization algorithm (Meng & van Dyk, 1997; McLachlan & Peel, 2000), and Markov chain Monte Carlo (Bayesian maximum posterior) estimation (Lee, 2007). Mplus 5.1 (Muthén & Muthén, 2007) uses maximum-likelihood estimation via the EM algorithm; a description of the estimation in general for latent variable mixture modeling is described in the Mplus 5.1 technical appendices (Muthén, 1998-2004) with more details provided in Muthén and Shedden (1999). The algorithm is described in detail specifically for mixture CFA models in Yung (1997).

The EM algorithm was proposed by Dempster, Laird, and Rubin (1977) to obtain maximum-likelihood parameter estimates in the presence of missing data. Maximum-likelihood estimation of traditional single population CFA models requires the EM algorithm with latent variable values treated as missing data (McLachlan & Krishnan, 2008; Rubin & Thayer, 1982). For mixture CFA models, latent class memberships are treated as missing in addition to latent variable values (e.g., McLachlan & Peel, 2000).

Letting c_{ic} be a dummy variable equal to 1 if individual *i* is in class *c* and 0 otherwise, and assuming the vector $c_i = (c_{i1},\ldots, c_{iC})'$ has a multinomial distribution with its values and the values of ξ for all examinees known, the complete-data log-likelihood for *C* mixtures of CFA models can be written as

$$\log L_c = \sum_{n=1}^{n}\sum_{c=1}^{C} c_{ic}\ln\left\{\pi_c \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_c|^{1/2}} e^{-\frac{1}{2}(x_i-\boldsymbol{\mu}_{xc})'\boldsymbol{\Sigma}_c^{-1}(x_i-\boldsymbol{\mu}_{xc})}\right\}.$$

The EM algorithm can be used to form and maximize the complete-data log-likelihood iteratively starting with the E-step. The first iteration of the E-step starts with initial values for π_c and the parameters in $\boldsymbol{\mu}_{xc}$ and $\boldsymbol{\Sigma}_c$ (namely $\boldsymbol{\Lambda}_c$, $\boldsymbol{\tau}_c$, $\boldsymbol{\psi}_c$, and $\boldsymbol{\phi}_c$) as estimates in the complete log-likelihood function, with the sufficient statistics for the missing latent variable values and latent class memberships replaced by their expectations conditional on the data and the parameters. Subsequent iterations use estimates of φ_c and the parameters in $\boldsymbol{\mu}_{xc}$ and $\boldsymbol{\Sigma}_c$ from the previous M-step.

In the E-step, the conditional expectation of the probability of class membership is computed given the data and estimates of φ_c and the parameters in $\boldsymbol{\mu}_{xc}$ and $\boldsymbol{\Sigma}_c$ from the previous M-step followed by the calculation of the conditional expectation of cross-products of latent variable values and observables appearing in the complete-data log-likelihood given the estimated probabilities of class membership. By Bayes' theorem, the posterior probability that observation *i* is a member of class *c* is given by the formula

$$P_{ic} = \mathrm{E}(c_{ic}|x_i, \varphi_c, \mu_c, \Sigma_c) = P(c_{ic}=1|x_i) = \frac{\pi_c f(x_i|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)}{\sum_{c=1}^{C}\pi_c f(x_i|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)}.$$

The estimate of the proportion of individuals in class *C* can be computed as the average of posterior probabilities:

$$\pi_c = \frac{1}{n}\sum_{i=1}^{n} P_{ic}.$$

Given the estimated posterior probabilities of class membership and the observed data, *X*, the expectations for each class's means and sums of squares and cross-products (SSCP) matrices, joint with *X* conditional on the estimates of the parameters in $\boldsymbol{\mu}_{xc}$ and $\boldsymbol{\Sigma}_c$, are obtained.

In the M-step, the posterior probabilities of class membership and estimated means and SSCP matrices of factor values from the E-step are used to form the complete-data log-likelihood. To maximize the complete-data log-likelihood with respect to the parameters in $\boldsymbol{\mu}_{xc}$ and $\boldsymbol{\Sigma}_c$, the expected values of interactions between posterior probabilities of class membership and the means and covariance of factors and measured variables can then be computed, resulting in estimates of factor scores and their conditional covariances (Muthén & Shedden, 1999; Yung, 1997). General formulas for mixtures of latent variable models with covariates are described in Muthén and Shedden (1999) and the Mplus 5.1 technical appendices (Muthén, 1998-2004). These formulas are simplified for mixture CFA models without covariates and a description under the assumption that residual covariance matrices are constrained equal across classes can be found in Yung (1997). When residual covariances are allowed to vary across classes, modifications need to be made to the EM algorithm described by Yung (1997) because closed-form solutions do not exist in the M-step. There are a number of modified EM algorithms that have been described (see, for

example, Dolan & van der Maas, 1998; McLachlan & Peel, 2000). Muthén and Shedden (1999) stated that closed-form solutions in the M-step can be achieved by using one or two Newton-Raphson steps (see McLachlan & Krishnan, 2008 for technical details).

The EM algorithm along with any necessary modifications provides estimates of $\boldsymbol{\Lambda}_c$, $\boldsymbol{\tau}_c$, $\boldsymbol{\psi}_c$, and $\boldsymbol{\phi}_c$, c = 1, …, *C* which are then used in the next E-step to estimate another set of latent class proportions, factor scores, and conditional variances. The algorithm continues iterating until the parameter estimates stabilize, as measured by a sufficiently small change in the observed data log-likelihood between iterations. By default, Mplus 5.1 (Muthén & Muthén, 2007) supplements the steps in the EM algorithm with quasi-Newton or Fisher scoring steps (see McLachlan & Krishnan, 2008 for technical details) when needed to accelerate convergence (Muthén & Muthén, 1998-2007).

Nonconvergence and local maxima

Failure to converge to a stable solution within a given number of iterations or converging to a local maximum are common problems when estimating any type of mixture model using maximum-likelihood. In univariate and multivariate normal mixtures with heterogeneous covariance matrices, the likelihood function is unbounded, which frequently can lead to nonconvergence or convergence to one of multiple local maxima that likely exist rather than a global solution (McLachlan & Peel, 2000). Thus, while nonconvergence has been a nonissue in previous methodological studies that have estimated multigroup CFA with observed group membership and continuous indicator variables using maximum-likelihood estimation

(e.g., French & Finch, 2006; Meade & Bauer, 2007), failure to obtain a global solution is a concern when studying and using mixture CFA models.

Empirical studies have illustrated the potential for nonconvergence when estimating mixture CFA model parameters under various conditions. For example, Lubke and Muthén (2007) provided evidence that suggests that when estimating mixture latent means models, convergence rates may be lower when factor loadings are free to vary across classes and when class separation is low. Specifically, using 100 replications, they found a 10% rate of nonconvergence for a two-class, two-factor model with noninvariant factor loadings and low class separation. Gagné (2004) found lower convergence rates for completely invariant models versus models with either noninvariant measurement intercepts or noninvariant factor loadings.

As is the case when estimating mixture models in general, using multiple starting values during maximization is typically necessary to avoid local maxima in mixture CFA models (Vermunt & Magidson, 2005). The default for estimation of latent variable mixture models in Mplus 5.1 (Muthén & Muthén, 2007) is 10 random sets of start values with two of the solutions with the highest log-likelihood from applying the EM algorithm chosen to be iterated until convergence (Muthén & Muthén, 1998-2007). If the highest log-likelihood value is not replicated in two or more final stage solutions, a local solution is a possibility and a warning message will appear in the output.

The complexity of mixture CFA models relative to multigroup CFA models increases the potential for unsuccessful estimation of or incorrect inferences about heterogeneous measurement model parameters if data have not been collected under

favorable conditions. There are costs and benefits of using mixture CFA instead of multigroup CFA, both substantive and methodological. The current study, together with existing work, helps to move towards a further understanding of conditions that impact the success of mixture CFA models when used to model and test for CFA-DIF.

Substantive foundations of CFA-DIF

The following sections provide a context for modeling CFA-DIF, which occurs when factor loadings and/or measurement intercepts are heterogeneous. Latent variable models for measurement heterogeneity and their purposes are described. Sources of heterogeneity and interpretations of noninvariant factor loadings and measurement intercepts are then presented.

Latent variable models for heterogeneity

Mixture CFA models are a combination of traditional CFA with continuous measured and latent variables and latent class analysis with categorical latent variables that explain heterogeneity across multiple unobserved populations. Muthén (2008) has divided mixture CFA models into two subcategories based on whether or not measurement equivalence is of substantive interest. In one branch described by Muthén, latent classes are hypothesized to impact the factor mean while tests of invariance or partial invariance of factor loadings and measurement intercepts are of secondary concern and are modeled only to improve estimation or to ensure latent mean comparisons are meaningful. In the second branch, the intent is to model the impact of latent classes on factor loadings and measurement intercepts and possibly to

cluster individuals without further substantive interest in latent means. Testing for CFA-DIF using mixtures of CFA models falls into the latter category.

Examples of heterogeneous latent variable models with the primary focus on latent mean differences include multigroup mean and covariance structure models (e.g., Sörbom, 1974), MIMIC models (Muthén, 1989) and mixtures of CFA models for cross-sectional data (e.g., Gagné, 2004; Lubke & Muthén, 2005; Lubke & Muthén, 2007; Yung, 1997), longitudinal data (Muthén, 2004; Muthén & Shedden, 1999), and hierarchical data (e.g., Ansari, Jedidi, & Dube, 2002; Longford & Muthén, 1992; Muthén, 1989; Allua, Stapleton, & Beretvas, 2008). The goal when using such models is to make inferences about latent means across populations. A key assumption is that the measures are invariant across populations, which allows one to infer that differences in the observed variables are due to differences in the latent variables. In practice, measurement model parameters are either assumed invariant or their invariance is verified through statistical tests (described below).

Examples of heterogeneous latent variable models with the primary focus on measurement equivalence include multigroup CFA for cross-sectional data (e.g., Jöreskog, 1971), longitudinal data (Bollen & Curran, 2006) and dichotomous data (e.g., Muthén and Cristoffersson, 1981), mixture CFA for cross-sectional data (e.g., Yung, 1997) and longitudinal data (Bollen & Curran, 2006), multilevel (hierarchical, or random coefficient) structural equation modeling (e.g., Jedidi, Jagpal, & DeSarbo, 1997) and mixtures of IRT models (e.g., Mislevy & Verhelst, 1990; Mislevy & Wilson, 1996; Rost, 1990). Inferences pertaining to factor loading and measurement intercept heterogeneity are of interest when using these models.

Heterogeneous factor loadings

Each factor loading in a CFA model, the slope in the regression of a measured variable on a latent variable, is assumed to be the same for all members in a population. When a single-population CFA model is applied to sample data, it is assumed that all observations in the sample come from the same population with one set of parameters. Multigroup and mixture CFA models assume that a sample comprises individuals from more than one population, and allow for parameters, including factor loadings, to differ across populations.

Modeling heterogeneous factor loadings has been popular in both methodological and applied research using multigroup CFA models. In 2000, Vandenberg and Lance reviewed 14 methodological studies and 67 applications of invariance testing using multigroup CFA models. They found that invariance testing on the factor loadings was discussed more than the invariance of any other parameter in the multigroup methodological literature. They identified all but one of the reviewed applied studies as ones that tested the invariance of factor loadings using a multigroup CFA model with only 14 testing differences in latent means.

The statistical evolution of modeling heterogeneous factor loadings emerged from concerns within an exploratory factor analysis framework about sample selection causing noninvariance of factor loadings across groups taking mental tests (e.g., Thomson, 1939; Thurstone, 1947; see Cudeck & MacCallum, 2007, for a more extensive historical summary). Thomson (1939) thought that differences in factor loadings exist due to selection which would deem exploratory factor analysis unusable for group comparisons of factor scores.

Those researching the impact of sample selection on factor loading invariance held the perspective that samples were taken from subpopulations derived from a common larger theoretical population. Later methodological work showed that rotating factors can ensure that factor loadings are invariant across samples from the same theoretical parent population (e.g., Ahmavaara, 1954; Meredith, 1964). Group comparisons using exploratory factor analysis still relied on the strong assumption that groups were all from subpopulations selected from the same common population until Jöreskog (1971) posited a CFA model for multiple known populations with the possibility of group-specific covariance structure parameters (i.e., factor loadings, factor variances/covariances, and residual variances). Group-specific measurement intercepts were also included in Jöreskog's model, though testing their invariance was not a concern. Jöreskog's specifications and extensions are currently used to model and test for heterogeneous factor loadings across multiple populations with known group membership.

In educational and psychological research, differences in factor loadings are substantively interpreted as qualitative differences between groups of subjects or that indicators are measuring different hypothetical constructs for different populations. Early substantive interest in qualitative differences in factor loadings across populations includes work by French (1965) who hypothesized that factor loadings are different for different test-taking populations due to differences in problem-solving skills and abilities. Results from applying exploratory factor analytic models to a battery of tests and dividing subjects into sets of pairs based on problem-solving skills supported his hypothesis that different approaches to solving problems do not

change the factorial composition of a test (that is, the test is measuring the same latent attributes for all individuals), but instead indicates qualitative differences among examinee groups that impact the factor-indicator relationship.

When dichotomous or polytomous indicators are obtained from an educational or psychological test and modeled with factor analysis, heterogeneous factor loadings represent measurement bias (for a review of measurement bias, see Millsap and Everson, 1993). For example, Drasgow (1984), motivated by the black-white achievement test score gap, argued that test bias can occur because of measurement nonequivalence, which he defined in terms of the relationship between observed test scores and latent ability. Jensen (1985) used exploratory factor analysis to correlate loadings on Spearman's *g* with group differences in test scores. A volume in *Multivariate Behavioral Research* (Issue 2, Volume 27, 1992) is almost fully devoted to group differences, exploratory factor analysis, and the role of factor loadings in the study of group differences, based on a posthumous reproduction of Guttman's (1992) critique of Jensen (1985). Following many substantive and statistical debates, Lubke, Dolan, and Kelderman (2001) showed that CFA is a superior method for comparing groups relative to exploratory factor analysis for dichotomous/polytomous items.

More recently in the test theory literature, the term measurement bias has been reserved for situations where nonequivalence of measurement scales is due to secondary characteristics that are unrelated to the measurement instrument. The general use of the term measurement bias has been replaced with the more statistically neutral terms differential item functioning (DIF) and differential test functioning, which describe heterogeneity that can be explained by either

substantively relevant or irrelevant secondary characteristics. Once DIF is established, further investigation of an item using contextual and theoretical information can reveal whether bias exists (Camilli & Shepard, 1994).

In the domain of survey research, where substantive interest lies in opinions and attitudes, heterogeneous factor loadings may be associated with group differences in response styles. For example, Cheung and Rensvold (2000) discussed how one group of respondents may follow an extreme response style (ERS), where their responses are either at the high or the low end of a psychological scale. Their factor loadings will differ from respondents who are not following ERS and whose responses are in the middle of the scale. Statistically, when respondents are at one of the extremes of the measured variable scale for all levels of the latent variable, the range of values is restricted, leading to a lower correlation between observed and latent variables and thus, lower factor loadings, compared to a group of individuals who are using the entire scale.

Heterogeneous measurement intercepts

In single-group CFA modeling with no mean structure, the covariance structure is modeled alone and measured variables are assumed to be deviations from their mean, with substantive interest lying in factor loadings, the variance/covariance of the factors, and residual variances/covariances. When there is substantive interest in latent means, CFA with mean structure, which includes latent means and measurement intercepts, is modeled. When both mean and covariance structure are modeled in single-group CFA, average values of measured variables are a function of the associated latent mean and measurement intercept, $\mathrm{E}(X) = \tau + \lambda\mathrm{E}(\xi)$, assuming

that the expected value of the residuals is zero. Thus, each measurement intercept in a CFA model represents the average value of its corresponding indicator variable over individuals with a factor score equal to zero. In single-group analysis, it is assumed that individuals with the same latent variable value have the same expected value on an associated manifest variable. Differences in measurement intercepts across populations are reflected in uniform differences in average indicator values across all levels of the associated latent variable.

When substantive interest is focused only on factor loadings, the mean structure is typically ignored in traditional single-sample CFA and is sometimes included in multisample CFA. It is recommended that the mean structure be included and measurement intercept invariance tested prior to investigating the invariance of parameters in the structural portion of a model, such as latent means (Meredith, 1993). However, prior to 2000, very few studies tested differences in measurement intercepts in multigroup CFA analyses (Vandenberg & Lance, 2000). Since then, tests of measurement intercept invariance have increased substantially from 12% to 54% of those reviewed (Schmitt & Kuljanin, 2008).

When measurement intercepts are allowed to vary across populations, differences in observable variables may be due to either differences in true values of the associated latent variables or differences in measurement intercepts. Substantive interpretations of heterogeneous measurement intercepts depend on the discipline. Heterogeneous measurement intercepts can represent systematic bias on all or part of a measurement instrument (Bollen, 1989). Cheung and Rensvold (2000) defined scale displacement as when one group systematically scores higher or lower than the other

group on Likert-style items. A difference in measurement intercepts across populations is also called additive bias (Meredith, 1993) or uniform bias (Lubke, Dolan, Kelderman, & Mellenbergh, 2003).

Differences in response styles on psychological or mental tests can contribute to additive bias (Guilford, 1954). Examples of response styles from Likert-scale questionnaires measuring psychological variables such as opinions or attitudes include social desirability (e.g., Tourangeau & Smith, 1996) and acquiescence (e.g., McClendon, 1991). Social desirability occurs when survey respondents choose answers that make them appear favorably and is likely to occur on sensitive questions such as those about substance abuse. Acquiescence, also called yea-saying, occurs when respondents always answer a question positively. Acquiescence is a concern in areas such as cross-cultural research (Cheung & Rensvold, 2000).

In educational testing, an example of additive bias is rater leniency, which occurs when one group of raters consistently gives higher scores to all examinees than another group of raters. When modeling dichotomous or polytomous item responses from an assessment, heterogeneous measurement intercepts are interpreted as item difficulty differences. For example, an algebra word problem may be more difficult for examinees whose first language is not English. Heterogeneous intercepts may also represent stereotype threat; that is, for example, stigmatized groups may score lower on items than the general population (Wicherts, Dolan, & Hessen, 2005). Threshold differences may also account for differences in measurement intercepts (Vandenberg & Lance, 2000). For example, groups may perceive pain differently and

thus use the highest score on a continuous pain scale differently (Maydeu-Olivares & Coffman, 2006).

Theoretical framework for testing CFA-DIF

The following sections detail measurement invariance testing in its broader context. A formal definition of CFA-DIF is presented along with descriptions of the types of CFA-DIF that can occur. How CFA-DIF is related to its counterpart in the domain of modeling with dichotomous/polytomous item responses is also described. An overview of model selection when testing CFA-DIF is presented along with motivation for testing across latent classes instead of manifest groups.

Measurement invariance testing sequence

In order to minimize Type I and Type II errors and obtain accurate and meaningful inferences about population differences in latent means and structural parameters, manifest indicators must be measuring the same latent variables across the multiple populations being compared. Measurement invariance holds when differences in factor scores are due to differences in manifest indicator values and are not moderated by population membership. Thus, testing for and verifying the invariance of measurement model parameters, comprising measurement intercepts, factor loadings, and residual variances, typically precludes group comparisons of other mean and covariance structure parameters.

Invariance testing across populations in the context of mean and covariance structure models has mainly been discussed from a hierarchical perspective (e.g., Bollen, 1989). Jöreskog (1971) first recommended testing increasingly restrictive

hypotheses in order using multigroup CFA with no mean structure beginning with a test of the equality of covariance matrices across groups. If covariance matrices were found to be noninvariant, Jöreskog recommended testing for the same number of factors across populations. If this hypothesis remained tenable, a sequence of tests could be conducted for hypotheses about increasing restrictions on measurement model parameters. Specifically, Jöreskog recommended testing the invariance of factor loadings (assuming the same pattern of fixed and free factor loadings across populations) with all other parameters unconstrained across populations. A finding of invariant factor loadings would then lead to a stricter test of invariance in which the residual variances would be constrained equal across populations in addition to the factor loadings. If it is found that all measurement parameters are invariant, Jöreskog suggested that one would be allowed to proceed to a test of the equality of the factor variance-covariance matrix. Sörbom (1974) later introduced mean structure in the multigroup CFA model and tested hypotheses about cross-population parameter invariance in the covariance as well as the mean structure following the same steps as Jöreskog.

There has been overall agreement with Jöreskog's (1971) recommendation to ensure that the number of factors and the pattern of fixed and free factor loadings is the same in all populations before conducting any stricter tests of invariance. Given this, the strictest form of measurement invariance for covariance structure models occurs when all factor loadings and residual variances/covariances are equal across populations. For latent means models, the strongest form of measurement invariance

occurs when measurement intercepts are equal across populations in addition to the factor loadings and residual variances/covariances.

Terms for various levels within the sequence of measurement invariance tests have been coined in the literature; yet, as the review by Vandenberg and Lance (2000) noted, they have not been standardized. The least severe level is lack of *configural invariance*, a term adopted from single sample exploratory factor analysis (Thurstone, 1947) describing factor models that have the same number of factors and the same simple structure across different samples selected from the same population. Meredith (1993) extended this definition with the additional characteristic that nonzero loadings have the same sign across populations. A model that achieves configural invariance and allows for all other parameters except those necessary for identification to be freely estimated for each population can be used to compare models with restrictions on factor loadings, measurement intercepts, and residual variances/covariances.

Given configural invariance, the least severe restriction on measurement model parameters occurs when factor loadings are constrained equal across populations. If invariance is tenable for all factor loadings, *metric invariance* is said to hold (Horn & McArdle, 1992). When all factor loadings and all measurement intercepts are equal across populations, *strong factorial invariance* (Meredith, 1993) is said to exist. The most severe restriction on measurement model parameters, called *strict factorial invariance* (Meredith, 1993), holds when all factor loadings, measurement intercepts, and residual variances/covariances are equal across populations.

The types of invariance described above refer to completely invariant parameter matrices which are based on the multigroup invariance tests described by Jöreskog (1971) and Sörbom (1974) that were omnibus tests of an entire parameter matrix at a given level. Byrne, Shavelson, and Muthén (1989) proposed testing individual parameters or subsets of parameters independently conditional on finding a matrix of parameters to be noninvariant. The authors used the term *partial measurement invariance* to describe a situation in which only one factor loading beyond the referent must be equal across populations and the remaining may vary across populations. Measurement intercepts and residual variances/covariances can also be tested for cross-population invariance individually.

Despite ample discussion of levels of measurement invariance, no necessary and sufficient conditions have been established for deciding on the minimum level of noninvariance in order to achieve accurate and meaningful inferences about differences in parameters beyond the measurement model (Millsap & Kwok, 2004). For example, Rock, Werts, and Flaugher (1978) suggested that meaningful interpretations of latent mean differences cannot be made if differences in covariance matrices across groups are due to population differences in factor loadings or measurement intercepts. Bollen (1989) stated that conventionally, factor loadings should be invariant in order to test for differences in measurement intercepts and factor means. Meredith (1993) posited that all factor loadings, measurement intercepts, and residual variances should be equal across groups in order to make meaningful comparisons of latent means. On the other hand, Byrne et al. (1989) showed that establishing full measurement invariance is not necessary before testing

invariance in covariance and mean structure parameters. Hancock (2004) stated that if there are truly noninvariant factor loadings or intercepts, accurate comparisons of latent means are attainable as long as the heterogeneous measurement model parameters are located and correctly allowed to vary across populations.

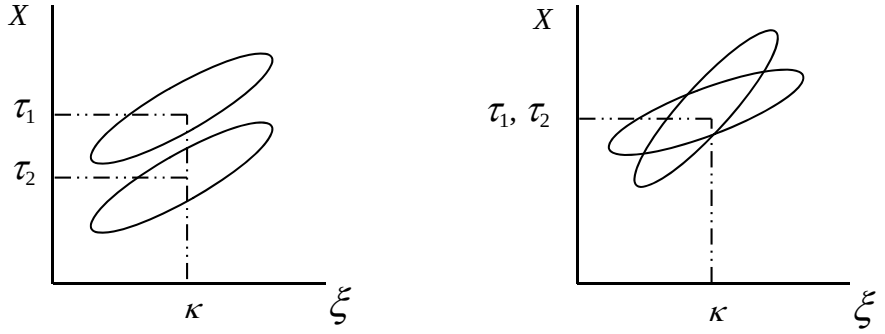
In addition to the lack of consensus on the minimum level of invariance required to ensure that the measurement model is not statistically significantly different across populations, there is also no consensus on what constitutes a small size difference in factor loadings or a small number of noninvariant factor loadings (Cudeck & MacCallum, 2007). More research on the consequences of partial measurement invariance has been called for (Schmitt & Kuljanin, 2008). Meanwhile, examples of latent mean comparisons in the presence of partial measurement invariance show that accurate inferences can be drawn. For example, Muthén and Christoffersson (1981) tested for differences in latent means using factor analysis with dichotomous indicator variables in the presence of partial measurement invariance. More recent methodological studies in the continuous indicator domain include cases of partial measurement invariance when testing for differences in latent means; for example, in the context of multigroup structured means modeling (Hancock, Lawrence, & Nevitt, 2000), mixture latent means modeling (Gagné, 2004; Lubke & Muthén, 2005), and multilevel mixture latent means modeling (Allua, Stapleton, & Beretvas, 2008).

CFA-DIF

In CFA models, factor loadings and measurement intercepts inform the relationship between latent factors and their manifest indicators. There are two ways

that indicators can function differently across populations due to the factor-indicator relationship. A simplified conceptualization of CFA-DIF with one factor and one indicator is shown in Figure 1. The figure depicts representations in two-dimensional space of the relationship between values of a manifest indicator variable, *X*, and values of a latent variable, ξ, for two populations by plotting the confidence ellipses that correspond to a scatterplot in ξ-*X* space. The factor loadings for each population are represented by the slope of the major axis of the ellipse for each population and the measurement intercepts are labeled τ_1 and τ_2.

Figure 1: Graphical representation of a factor-indicator relationship



(A)
Uniform CFA-DIF
Equal factor loadings, different measurement intercepts

(B)
Nonuniform CFA-DIF
Different factor loadings, same (or different) measurement intercepts

Figure 1(A) illustrates the indicator-factor relationship for two populations with identical factor loadings and different measurement intercepts. When the factor loadings are equal but the intercepts are different across populations, individuals in one population with the same level of ξ as individuals in the other population have higher values on *X* across the continuum of the latent variable. This type of noninvariance involving only the measurement intercepts is herein referred to as

uniform CFA-DIF. Uniform CFA-DIF is equivalent to a lack of scalar invariance (Meredith, 1993).

Figure 1(B) displays an indicator-factor relationship where the factor loadings differ across populations while, for example, the measurement intercepts are equal. In this case, one population has higher values on *X* than the other population for a given ξ when ξ is low and lower values on *X* for a given ξ when ξ is high, herein referred to as *nonuniform* CFA-DIF. Nonuniform CFA-DIF occurs when factor loadings are noninvariant across populations regardless of the invariance status of measurement intercepts and corresponds to a lack of metric invariance (Horn & McArdle, 1992).

There are two main purposes for modeling heterogeneity within a CFA framework and testing for CFA-DIF. A heterogeneous indicator-latent variable relationship may be modeled for substantive reasons, such as when theory suggests that a construct differs across populations. For example, scale items may function differently across ethnic groups due to secondary attributes in addition to the primary construct. Alternatively, a researcher addressing questions of moderation in structural relations or latent means must first verify the cross-population equality of parameters residing in the CFA portion of the model. For example, if latent reading self-concept scores for boys are on a different scale than for girls, mean levels of self-concept cannot be compared across gender. In the former case, the intent is to find evidence of CFA-DIF to support a hypothesis. In the latter case, the intent is to find sufficient statistical support for a lack of CFA-DIF so that inferences about structural parameters or latent means will be accurate and meaningful.

The choice of testing for uniform or nonuniform CFA-DIF depends on modeling purposes. Uniform CFA-DIF is typically tested in applications where latent mean or structural parameter comparisons are of primary interest. When conducting tests of uniform CFA-DIF using mean structure CFA models, researchers typically test factor loading invariance first, then subsequently test measurement intercept invariance for those intercepts associated with factor loadings that have already been deemed invariant. Modeling heterogeneous factor loadings and testing for nonuniform CFA-DIF have been the primary interest in multigroup CFA models (Meade & Bauer, 2007; Vandenberg & Lance, 2000). Typically in multigroup CFA applications, the covariance structure is important while the mean structure is ignored by assuming that the measurement intercepts have been absorbed into mean-centered indicator variables. In such studies, the invariance of all factor loadings is tested simultaneously, followed by individual tests of substantively important factor loadings if necessary.

Correspondence between CFA-DIF and IRT-DIF

While CFA assumes a linear relationship between continuous latent variable(s) and continuous observable variables, in IRT modeling, discrete item scores have a logistic relationship to a corresponding continuous latent trait. In IRT, differential *item* functioning (DIF) refers to variability in scores on a dichotomous or polytomous item that cannot be explained by subpopulation differences on the underlying constructs(s) that the item is intended to measure (Clauser & Mazor, 1998; Crocker & Algina, 1986). IRT models assume that examinees with the same value on the primary latent trait(s), θ, have the same probability of a correct response on any

item that is a measure of θ. When an item exhibits DIF, there is a second latent trait which influences the probability of a correct response on that item (Clauser & Mazor, 1998). Secondary latent trait(s) may either be latent variables that are not related to the primary latent trait(s), such as English language ability on a mathematics test, or latent variables that are relevant to the primary latent trait(s) (Embretson & Reise, 2000), such as testwiseness. Testing for DIF is motivated by both theoretical perspectives (e.g., what are the secondary latent variables and how does their presence or absence advantage or disadvantage some examinees more than others?) and statistical properties (e.g., if an item exhibits DIF, it could lead to inappropriate or incorrect inferences about the latent variable for an examinee or groups of examinees).

Conceptually, examining DIF is similar to tests of CFA-DIF in covariance and mean structure CFA models with continuous indicator variables. That is, as Embretson and Reise (2000, p. 251) stated, "analogous to measurement invariance investigations under a [covariance structure analysis] CSA paradigm, DIF is said to occur when a test item does not have the same relationship to a latent variable (or a multidimensional latent vector) across two or more groups." Factor loadings and measurement intercepts model the relationship between the latent factor and the manifest indicators in CFA models just as item difficulty and discrimination parameters in IRT model the relationship between the latent trait and the item responses. Thus, the interpretations of item discrimination parameters in IRT are roughly equivalent to factor loadings in CFA models and difficulty parameters in IRT are conceptually related to measurement intercepts in CFA models.

CFA modeling has been applied to dichotomous/polytomous scale items as an alternative to IRT models in order to detect DIF (Raju, Laffitte, & Byrne, 2002). In CFA modeling with discrete data, it is assumed that each item response is realized from an underlying continuous normal distribution (e.g., Muthen & Cristoffersson, 1981). Discrete item responses arise from threshold values that categorize the underlying continuous distribution. Each factor loading is the slope of the regression of an underlying continuous item score on a latent variable and is interpreted as item discrimination. Measurement intercepts are not directly estimated when applying CFA to discrete data; instead, item thresholds are estimated. Measurement intercepts and item thresholds are inversely related to each other substantively and are both indicative of item difficulty (Stark, Chernyshenko, & Drasgow, 2006).

Historically, different analytic strategies have been used in DIF testing in IRT and CFA modeling. The approach in IRT has been to specify a restricted baseline model and simultaneously test for differences across populations in item difficulty and discrimination parameters (Stark et al., 2006). In CFA modeling, an unrestricted baseline model is typically specified with factor loading differences tested first, followed by tests of measurement intercepts for items whose factor loadings are inferred to be invariant (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000).

Several studies have compared CFA to IRT for detecting DIF using real data examples and didactic exposition (e.g., Raju et al., 2002; Reise, Widaman, & Pugh, 1993) and simulated data (e.g., Meade & Lautenschlager, 2004; Stark et al., 2006). Raju et al., Reise et al., and Meade and Lautenschlager found different results from the two methods using the traditional approaches of each method described above.

However, Stark et al. (2006), who developed a common strategy in which increasingly constrained models are compared by restricting the factor loading and measurement intercept of one item at a time to be equivalent across populations, showed similar results from the two methods for dichotomous items and superior results for CFA over IRT for polytomous items.

Invariance testing using CFA models for continuous data

Invariance tests use fit statistics to select the best fitting model across nested models with different amounts of heterogeneity across populations. A model with fewer parameters is nested within a model with more parameters if the smaller model can be achieved by restricting one or more parameters of the larger model. When conducting invariance tests within a CFA framework with observed group membership, separate mean and covariance structures are specified for each population and goodness-of-fit can be compared across models with varying cross-population constraints on parameters of theoretical interest. In mixture CFA, separate mean and covariance structure models are specified for each population and models with different class-specific parameters can be compared across a theoretically known number of latent classes or different numbers of latent classes (e.g., Lubke & Muthén, 2005).

When population membership is observed or the number of latent classes is known, invariance testing involves only cross-population constraints. Comparing goodness-of-fit between alternative configurations of cross-population constraints requires estimation of a null hypothesized model with all parameters of theoretical interest constrained equal across populations while all other parameters except those

needed for identification are free to vary. An alternative model also is estimated with all parameters of theoretical interest free to vary across populations in order to compare goodness-of-fit across the models. The most commonly used method of invariance testing in multigroup CFA is the likelihood-ratio test (Vandenberg & Lance, 2000) which allows for goodness-of-fit comparisons of nested models and can also be used in mixture CFA when the number of latent classes is known.

The likelihood-ratio test statistic, also called the χ^2 difference statistic, is defined as $\chi^2_{difference} = -2(\log L_C - \log L_U)$, where L_U is the likelihood value for a model with parameters of theoretical interest free or unconstrained across populations and L_C is the likelihood value for a model with parameters of theoretical interest constrained to be equal across populations. Parameters needed for model identification and parameters not of theoretical interest can be fixed to a constant or constrained equal across populations identically in both the constrained and unconstrained models. A statistically significant $\chi^2_{difference}$ with degrees of freedom equal to the number of additional constrained parameters is evidence in favor of heterogeneity.

A criticism of the likelihood-ratio test is that the likelihood depends on sample size such that larger samples are more likely to lead one to infer that parameters are heterogeneous, even if the difference may not be practically significant. Alternative goodness-of-fit indices that do not depend as much on sample size have also been proposed in the multigroup CFA literature. Cheung and Rensvold (2002) compared the performance of 20 change in goodness-of-fit indices for testing measurement invariance using multigroup CFA models. They found that comparing models using

the comparative fit index (CFI; Bentler, 1990), gamma hat (Steiger, 1989), and McDonald's Noncentrality Index (McDonald, 1989) were superior to the others in terms of robustness and were shown to be unaffected by model complexity. However, these fit indices do not have criteria tied to statistical significance. French and Finch (2006) found the χ^2 difference statistic to perform at least as well as the change in CFI. Chen (2007) conducted two simulation studies to compare the three fit indices from Cheung and Rensvold in addition to two others, root mean square error of approximation and standardized root mean square residual. Results showed CFI to perform the best while all indices were found to be affected by factors including pattern of noninvariance, sample size, ratio of sample size, and model complexity.

Invariance testing within a mixture modeling framework when the number of latent classes is theoretically unknown poses challenges because model selection involves both choosing the number of latent classes and choosing between different configurations of cross-population invariance constraints. Models with different numbers of latent classes are in fact nested; however, the likelihood-ratio statistic for comparing models with different numbers of latent classes does not follow the theoretical χ^2 distribution and as such, the models cannot be compared with the likelihood-ratio test (see Dayton, 1998, P. 17-18, or Nylund, Asparouhov, & Muthén, 2007, P. 534, for a conceptual explanation; for technical details, see McLachlan & Peel, 2000, P. 185-186).

Fit statistics that are commonly used in latent class analysis to test for the number of latent classes include the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the consistent Akaike Information Criterion

(CAIC) and the sample-size adjusted BIC (see, for example, Henson, Reise, & Kim, 2007). Goodness-of-fit statistics for deciding on the number of latent classes in structured means mixture modeling and growth mixture modeling have been studied empirically by several authors (e.g., Henson et al.; Nylund et al., 2007; Tofighi & Enders, 2008). Equivalent study would be needed for mixture models concerned with CFA-DIF once it is determined that these models perform successfully under the ideal condition that the number of latent classes is known.

In addition to choice of fit statistics for invariance testing using mixture models, there is debate about the order of tests for deciding on the number of classes and the pattern of cross-population constraints. For example, Lubke and Muthén (2005) simultaneously tested for the number of latent classes and the pattern of noninvariant parameters. They found that inferences about the number of latent classes and the pattern of noninvariance depended on the choice of fit statistics. This simultaneous approach could be criticized for its exploratory nature and potential for inflated type I error rates (i.e., overextraction of latent classes). An alternative to the simultaneous approach is a stepwise approach, which would first determine the number of latent classes and secondly conduct invariance tests on parameters of interest. As an empirical stepwise approach, Bauer and Curran (2004) advised to first estimate a saturated model with different numbers of latent classes to find the best fitting number of latent classes, then test different specifications of the within-class model with the number of latent classes previously found. A theory-based stepwise approach would choose the number of latent classes according to past research or

based on the number of possible manifest groups (Samuelson, 2005) and subsequently conduct invariance tests on parameters of interest.

Model identification in mixture CFA models

In order to estimate parameters from a CFA model with no mean structure, the covariance structure must be identified by assigning a metric to the factors, which is usually achieved by choosing a referent loading and fixing it to a constant, typically unity. When there are multiple populations, the referent loading must be fixed for each population. The choice of referent may impact inferences drawn from invariance testing using heterogeneous CFA models (Yoon & Millsap, 2007). Statistical procedures have been proposed to detect factor loadings that are truly invariant across populations to use as referents (see French & Finch, 2008 for a review and performance results from simulated data), yet the degree to which they are successful is unclear.

In CFA modeling with mean structure and multiple populations, both the covariance structure and the mean structure must be identified. This is accomplished by assigning a metric and an origin for the factors. The metric can be set in a similar way to covariance structure analysis, by fixing a referent indicator for each factor in each population. There are two statistically equivalent ways in which parameters can be fixed to set the origin of the factors in a mean structure CFA model. One way is to set the factor means of each population to zero and allow all measurement intercepts in all populations to be freely estimated. A second way is to set the factor mean of one population to zero while constraining the measurement intercept associated with the referent indicator(s) equal across populations. Note that in all models, fixing

factor loadings to a constant can be replaced by fixing factor variances to unity across all populations.

When conducting invariance tests using mixture CFA, in addition to restrictions for identification, homogeneity can potentially be imposed on parameters that are not of theoretical interest in order to model and test for heterogeneity in the desired portion of the model. This results in more parsimony which can help in achieving convergence. When testing for CFA-DIF, the focus is on heterogeneity in the relationship between latent factors and their manifest indicator variables due to differences in factor loadings and/or measurement intercepts across unknown populations. As such, if measurement intercepts are free to vary across classes to account for mean differences in indicator variable values, factor means can be set to zero in all classes (e.g., Muthén, 2008; Yung, 1997). This modeling is appropriate if substantive theory suggests that factor means should not differ across populations.

In some research areas, latent means may be expected to differ across populations. For example, in educational testing, a difference in latent means across populations is referred to as impact. Specifically, impact occurs when there are differences in skills between groups that cause differences in probabilities of item responses (Ackerman, 1992). DIF tests using IRT methods can be conducted with or without the existence of impact (e.g., Stark et al., 2006). In psychological testing with continuous observable variables, impact occurs when differences in scale scores are due to differences in construct means across populations. If true differences in factor means exist and factor means are set to zero for identification, impact will be reflected in differences in measurement intercepts.

Factor variances/covariances and residual variances/covariances either can be allowed to vary or constrained equal across latent classes in a mixture CFA model (Yung, 1997). Adding additional constraints to the model decreases the number of free parameters to estimate which can improve convergence when estimating mixture models at the expense of a decrement in accuracy of parameter estimates if parameters are truly noninvariant. This tradeoff is related to problems associated with specification searches in the single population case (Millsap & Kowk, 2004).

Latent classes vs. manifest groups

There has been growing interest in modeling heterogeneity across populations using mixture CFA models, especially for comparing latent means and in the domains of growth and hierarchical latent variable modeling. Recently, Lubke and Neale (2007) evaluated the power of mixture CFA models to detect DIF with Likert data. However, the emphasis in mixture CFA modeling with continuous data has eluded testing for CFA-DIF in part due to the widespread use of multigroup CFA models which can account for measurement model heterogeneity but require group membership to be known *a priori*. Testing for CFA-DIF with observable group membership may be problematic when it is not reasonable to assume that the indicator-latent variable relationship is the same for all observations in a manifest group, particularly when group membership is determined by broad-based categories such as gender, race, or ethnicity.

The use of observed groups out of convenience or *post hoc* consideration can result in the inability of the multigroup CFA model to detect heterogeneity when it exists. The implication is that when CFA-DIF testing is used as a precursor to testing

for heterogeneity in the structural portion of a latent variable model, results may be inaccurate if the model is deemed not to exhibit CFA-DIF and invariance does not actually hold for all observations in each population. Additionally, when testing for CFA-DIF for theoretical purposes, the consequences may be that inferences drawn from the model may be incorrect or meaningless. Mixture CFA models can ease these problems by incorporating theoretically important unmeasured characteristics to capture heterogeneity and test for CFA-DIF.

In some disciplines, manifest groups may be sufficient for modeling heterogeneity and statistical analyses can verify substantive theories about population differences. For example, sources of self-esteem are hypothesized to be different for Western and Eastern cultures implying that items on a questionnaire translated from English to Chinese may have weaker factor loadings (Chen, 2008). However, oftentimes manifest groups are used in empirical research as proxies for other more nuanced characteristics that are not included in a dataset. Insufficient time or funding may make it impractical to gather data to determine characteristics that define correct population membership, so applied researchers often use broad characteristics such as socio-economic status or race that are easier to observe as measures of other unobserved characteristics out of convenience.

The psychometric insufficiency of manifest groups has been discussed in the social and behavioral sciences, for example, when modeling data from psychological instruments (e.g., Muthén, 1989), marketing research (e.g., Moore, 1980) and test items (e.g., Mislevy & Verhelst, 1990). There may be substantive reasons why it is not tenable to assume that a manifest group is homogeneous with respect to indicator

variable performance. Latent classes can be used when important moderator variables are unavailable either because they are unobservable or a priori unknown. In educational testing, for example, latent classes can represent qualitative characteristics, such as instructional background (Muthén, 1989) or solution strategies (Mislevy & Verhelst; Rost, 1990). The use of latent classes instead of manifest groups has been recently advocated in the test theory literature when conducting DIF studies (Samuelsen, 2007) and has been shown to provide more meaningful inferences in an application (Webb, Cohen, & Schwanenflugel, 2008).

While capturing heterogeneity in CFA models originated with the use of manifest groups and tests for CFA-DIF have mostly involved observable groups, there has been a burgeoning interest in using latent classes instead of manifest groups. Reviews of invariance testing by Vandenberg and Lance (2000) and Schmitt and Kuljanin (2008) only include tests of measurement invariance across manifest groups. However, more recently, latent class CFA models have been considered in the methodological literature (Ansari, Jedidi, & Dube, 2002). Several studies that focus on latent means have looked at the viability of mixture CFA models to recover parameters (e.g., Lubke & Muthén, 2007; Lubke & Neale, 2008) and to choose the correct number of latent classes (e.g., Bauer & Curran, 2004; Tofighi & Enders, 2008).

Using manifest groups when it is not appropriate may lead to untested assumptions that CFA-DIF is nonexistent in a given empirical application if unobserved characteristics that define population membership are not considered. In the context of IRT, Samuelson (2007) showed that there will be statistical problems

such as type I error if there is not sufficient overlap between manifest groups and latent classes when manifest groups are used. Samuelson advocated for considering latent classes as the first step in modeling heterogeneity—a concept that can be further studied and applied to mixture CFA models.

Simulation studies and empirical examples

This section provides a more detailed review of existing work on heterogeneous measurement model parameters, focusing primarily on methodological aspects of prior research, including parameter values and separation across populations. Further background is provided for research on heterogeneous factor loadings in CFA models for cross-sectional, continuous data with observed group membership (multigroup CFA) and unobserved group membership (mixture CFA). Both models can be used to capture measurement model heterogeneity, of which CFA-DIF is a main source. The literature on multigroup CFA models has developed over several decades whereas the mixture CFA literature is relatively nascent. As such, research on multigroup CFA models significantly contributes to the current body of knowledge about mixture CFA models when factor loading heterogeneity is of primary interest.

In addition, relevant literature on testing for differentially functioning dichotomous/polytomous items across observed groups using IRT- and CFA-based methods and across unobserved groups using mixture IRT modeling is also reviewed. Testing for differentially functioning discrete item responses is an important part of test and scale development and has provided a fertile area for research on appropriate procedures and conditions that impact them. While modeling, estimation, and

invariance testing are different when items are dichotomous/polytomous versus continuous, research in this area has been concerned with testing for heterogeneity in measurement model parameters comparable to factor loadings and measurement intercepts and as such is an applicable resource for the study of CFA-DIF.

Multigroup CFA

As discussed above, current practices of invariance testing using multigroup CFA are grounded in work by Jöreskog (1971), Sörbom (1974), and Byrne et al. (1989). All three provided empirical examples to illustrate parameter estimation and invariance testing across multiple known populations. Both Jöreskog and Sörbom modeled responses from a battery of psychological tests measuring special ability, verbal ability, and memory administered to students in two different schools. Byrne et al. modeled scale scores for high and low academically tracked students from a questionnaire measuring four types of self-concept: general, academic, English, and mathematics. They each used the likelihood-ratio test to compare models with varying severity of invariance. Decisions about whether to reject or retain the null hypothesis of invariance of the entire parameter matrix were based strictly on statistical evidence. However, unlike the other two, once they found the hypothesis of invariant factor loading matrices untenable, Byrne et al. combined fit indices with substantive theory in order to judge which individual or subsets of cross-population constraints were to be released.

Following the introduction of such methods, reviews of methodological and applied studies (Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008) revealed the popularity of invariance testing using multigroup CFA models and provided

guidelines for best practices. In 2002, Vandenberg called for more research on issues associated with invariance testing, in particular, scientific research on conditions that impact measurement invariance testing to ensure valid inferences from multigroup CFA models. Three simulation studies in the multigroup CFA literature then followed: Meade and Lautenschlager (2004), French and Finch (2006), and Meade and Bauer (2007). All three pertain to covariance structure analysis, and did not consider mean structure in the CFA model.

Meade and Lautenschlager (2004) conducted the first published simulation study investigating the performance of multigroup CFA models in terms of the power of the omnibus test of equal covariance matrices and an omnibus test of invariant factor loadings. French and Finch (2006) expanded on the work of Meade and Lautenschlager by considering both power and type I error rates and also including fit indices in addition to the χ^2 difference statistic. Meade and Bauer (2007), seemingly concurrent with French and Finch, also conducted a simulation study as a follow-up to Meade and Lautenschlager. Meade and Bauer's study added to the analysis of multigroup CFA models by including the precision of factor loadings and their difference as an outcome measure in addition to power.

Each of the three simulation studies included conditions that were expected to be found when using real data and that allowed generalizability to a wider range of conditions. All three simulated continuous indicator values according to CFA models with two populations and observed group membership. Specifically, Meade and Lautenschlager (2004) generated data from one-factor models with 6 and 12 indicators, French and Finch (2006) simulated indicator values from two- and four-

factor models with 3 and 6 indicators per factor and Meade and Bauer (2007) generated data from models with 20 indicators measuring 3 and 6 factors.

Conditions in Meade and Lautenschlager (2004) included sample size (75, 250, and 500 per group with equal group sizes), number of items with factor loading differences (ranging from 17% to 67%), and directionality of the differences (mixed or uniform). French and Finch (2006) considered both equal and unequal group size conditions (150/150, 150/500, and 500/500) with three levels of percent of noninvariant factor loadings: 0, 17% and 33%. Factor loadings were chosen to be uniformly higher for one population and all communalities were high. Meade and Bauer (2007) had equal group sizes with total sample sizes of 200, 400, and 800. Other conditions included communalities (between 0.2 and 0.7) and which indicators were chosen to be noninvariant (high communality or low communality indicators), and directionality of factor loading differences (mixed or uniform). The study by Meade and Bauer differs from the other two by the modeling of both homogeneous as well as heterogeneous factor covariances.

Taken together, the results of the three simulation studies show the impact of a wide range of conditions that may be encountered in applications on the ability of multigroup CFA models to successfully test for the invariance of factor loadings. Meade and Lautenschlager (2004) found that large sample sizes and mixed loading differences (vs. uniform) increased the power of the omnibus likelihood-ratio test of loading invariance. Their study was conducted under ideal conditions (e.g., equal groups and high numbers of indicators per factor) and did not address accuracy of parameter estimates or partial invariance. French and Finch (2006), using the

likelihood-ratio test, the difference CFI, and a combination of the two based on Cheung and Rensvold (2002), found that the likelihood-ratio test performed better than the CFI difference test and resulted in good control of type I error and relatively high power, especially for large samples sizes and higher numbers of indicators per factor. French and Finch also looked at the power of the likelihood-ratio test to detect group differences on an individual loading after an initial finding of noninvariance from the omnibus test. They found greatly reduced power when just studying one noninvariant loading, especially as models became more complex, indicating that tests of partial measurement invariance can be problematic in multigroup CFA modelsing.

Like French and Finch (2006), Meade and Bauer (2007) used both the likelihood-ratio test and the difference CFI proposed by Cheung and Rensvold (2002) to test factor loading invariance. Meade and Bauer found that larger sample sizes and higher numbers of factors per indicator were associated with increased precision. Additionally, they found that precision was better for high communalities when there were many indicators per factor but the same when there were fewer indicators per factor. Larger sample sizes, a mixed pattern of noninvariance, and larger numbers of indicators per factor were associated with higher power in their study.

A combination of methodological expositions and examples, reviews with recommendations for best practices and several simulation studies has taken shape as a resource for helping applied researchers to make valid inferences when using multigroup CFA to test for nonuniform CFA-DIF when population membership is observed. However, when manifest groups are not appropriate, mixture CFA models

are a logical alternative. Research on mixture CFA models has mostly focused on detecting differences in latent means across populations and more studies have been methodologically oriented, rather than applied.

Mixture CFA

Applying mixture modeling to latent variable models was proposed simultaneously by several authors to handle parameter heterogeneity when population membership is not known *a priori*. Specifically, general mixture structural equation models, of which mixture CFA models are a subset, were described by Arminger and Stein (1997), Dolan and van der Mass (1998) and Jedidi, Jagpal, and Desarbo (1997). Yung (1997) proposed modeling and estimation of mixture CFA models. The proposed models were all finite mixtures—that is, the number of unknown groups was specified before estimation—and allowed for class-specific parameters to be incorporated into the covariance and mean structure portions of a latent variable model.

Yung (1997) provided illustrative examples of how heterogeneity could be modeled in both the covariance and mean structure of a mixture CFA model with parameter estimates obtained through maximum-likelihood estimation. Yung also discussed the ability of the model to handle partial invariance of the factor loadings and described the hierarchical nature of restrictions in terms of using likelihood-ratio tests but did not advocate for a particular order. Instead, Yung suggested that the parts of the model that are allowed to be heterogeneous be based on substantive theory and parsimony.

The use of mixture CFA models continued to be limited in applied studies, perhaps because of the added complexity of mixture CFA over multigroup CFA models. Lubke and Muthén (2005) provided a more didactic, application-oriented explanation of the mixture CFA model by comparing it to other heterogeneous latent variable models and illustrating different forms of heterogeneity that can be modeled. They covered topics including covariates, model selection, and the connection between concepts and terminology from the multigroup measurement invariance testing literature and mixture CFA models. The main focus of their methodological exposition and empirical example was on latent means modeling using a mixture CFA model, with factor loadings assumed invariant in many of their exemplary models.

Several simulation studies (Gagné, 2004; Lubke & Muthén, 2007; Lubke & Neale, 2006) have evaluated mixture CFA models for conditions that impact convergence and the ability to obtain correct parameter estimates. In each of these three studies, the main focus was on the heterogeneity of mean structure parameters, but all included an analysis with heterogeneous factor loadings. Lubke and Neale chose noninvariant factor loadings to be equal to 0.8 for one class and ranged from 0.49 to 0.96 for the second class. Lubke and Muthén evaluated a two-class, two-factor model with heterogeneous factor loadings that ranged from 0.6 to 0.9 within class and differed across classes by 0, 0.1, and 0.3. Gagné estimated a two-class one-factor model with factor loadings chosen to range between 0.4 and 0.8 within and between classes. When factor loadings differed between latent classes (25% differed by 0.4), only equal class proportions were considered.

Since the focus in all three studies was on latent means, all study conditions had some level of separation in the mean structure. Lubke and Neale (2006) and Lubke and Muthén (2007) both defined class separation in terms of the multivariate Mahalanobis distance between two classes, $M = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the latent variable means for classes 1 and 2 and Σ is the invariant covariance matrix. Latent means varied such that the Mahalanobis distance was either 1 or 1.5 in both studies. Gagné used two standardized differences between latent means (2 and 2.5) and three levels of intercept invariance (completely invariant, one noninvariant, two nonninvariant).

Results from Gagné (2004) showed that differences in measurement intercepts or factor loadings across classes, larger sample sizes, and larger differences in latent means across latent classes were associated with higher convergence rates and more accurate parameter estimates. Lubke and Neale (2006) found that noninvariant factor loadings and measurement intercepts were associated with an increase in correct model selection for models with latent mean separation with more restrictions leading to spurious latent classes. They showed that there is a trade-off between sample size and latent class separation when estimating mixture CFA model parameters. Lubke and Muthén (2007) found that fixed factor loadings led to higher convergence rates when fitting a latent means mixture model with the number of latent classes known. In addition, model performance was shown to improve when measurement intercepts were allowed to vary across classes with their standardized difference ranging from 0.1 to 0.2. They also found that constraining factor loadings resulted in more accurate

recovery of parameters and class assignment, especially when factor loadings differed by only a small amount across classes.

In addition to convergence and parameter estimation, challenges that can occur when using mixtures of latent variable models which can potentially apply to mixture CFA models have been discussed by several authors. Yung (1997) cautioned that restrictions placed on parameters for identification may have an impact on estimation in mixture CFA models. Bauer and Curran (2004) used examples to show that misspecification and nonnormality in latent variable mixture models may result in spurious latent classes. Lubke and Muthén (2005) pointed out that interpretation of factor mean differences across latent classes may be more difficult when partial invariance exists in the factor loadings.

Heterogeneous factor loadings have only been a secondary concern in the research on mixture CFA modeling with continuous indicators to date and there is no research that aims to evaluate the performance of mixture CFA for conducting invariance tests like one would using multigroup CFA with continuous indicators. A related area in which heterogeneous factor loadings are of primary concern is when modeling dichotomous/polytomous items with multigroup CFA, which developed because, in practice, many psychological scales are ordinal rather than continuous. Similarly, the focus on the invariance of measurement model parameters is a major concern in IRT.

Differential functioning of dichotomous/polytomous items

As described above, factor analytic techniques have been used to model heterogeneous parameters in order to explain differences in dichotomous/polytomous

item responses across multiple populations. Testing for heterogeneity in the entire covariance matrix and differential functioning of all items in a scale have been studied across both manifest and latent populations, while differential functioning of individual dichotomous/polytomous items has been studied across manifest groups. In such studies, factor loading differences have been of substantive interest more than measurement intercepts and factor means.

Modeling dichotomous item responses using CFA with parameters that differ across multiple populations with observed group membership was developed by Muthén and Cristoffersson (1981). In an illustrative example, the authors tested the invariance of measurement model parameters using data from married couples (266 males and 318 females) in Australia who were asked questions about “interpersonal relationships” and “neurotic illness”. The authors used generalized least squares to estimate a model with factor means set to zero in both groups with factor loadings and measurement intercepts free to vary across populations except for those associated with the referent items for each of the two factors. They found that the likelihood-ratio test favored the heterogeneous model over a completely homogeneous model with differences in parameter estimates across groups for factor loadings and measurement intercepts ranging from small to medium.

Muthén and Asparouhov (2006) applied mixture factor analysis to polytomous item responses from a tobacco dependence survey, with very large subsamples ranging from approximately 8,000 to 20,000 respondents. They used maximum-likelihood estimation and allowed factor loadings, measurement intercepts, and residual variances to vary across two latent classes. Factor means were set to zero for

both classes for model identification. Muthén and Asparouhov found that the two class factor mixture model provided a better fit than other models fit to the same subsamples, including a one class IRT model, a two factor CFA model, and latent class analysis models with 2 to 5 classes.

French and Finch (2006), in addition to evaluating CFA invariance testing across known groups for continuous observable variables, also considered simulated dichotomous item responses modeled with two- and four-factor CFA models and estimated with robust weighted least squares, with study conditions described in the previous section. They found that type I error rates were higher when sample size was larger and when there were fewer items, nonconvergence occurred frequently, especially with lower sample sizes, and power was low across all study conditions.

Lubke and Neale (2008) studied mixture CFA models for dichotomous and polytomous scale items and their ability to test for the invariance of factor loadings and item thresholds. Using 30 replicated datasets per condition, they estimated two-factor models with two classes and compared models with noninvariant factor loadings to models with invariant factor loadings using the BIC and sample-size adjusted BIC. They considered total sample sizes of 300, 400, and 1500. They found high power for detecting factor loading noninvariance, but it was unacceptable for detecting noninvariant item thresholds.

Of the five studies, both Muthén and Cristoffersson (1981) and Muthén and Asparouhov (2006) estimated noninvariant thresholds and factor loadings and successfully estimated parameters. French and Finch (2006) simulated data with noninvariant thresholds and found that nonconvergence rates were high. Lubke and

Neale (2008) generated data from models with noninvariant measurement intercepts and factor means to achieve separation of latent classes, with the Mahalanobis distance equal to either 1.5 or 2.0. The standardized group differences in the estimates of item thresholds from Muthén and Cristoffersson ranged from approximately 0.01 to 0.45 while the standardized differences in estimated factor loadings across the two groups ranged from approximately 0.01 to 0.15. French and Finch set measurement intercepts invariant across groups at 0.25 for model identification and chose noninvariant factor loadings to differ by 0.25 across groups, which is roughly equivalent to a difference of 0.6 within an IRT framework (French & Finch). In Muthén and Asparouhov, estimated factor loadings differed within a range of 0.06 to 0.62 across the two latent classes. Estimates of thresholds were not reported, but were described as statistically significantly different across classes.

Other methods of detecting DIF for dichotomous/polytomous items are grounded in classical test theory and IRT (see Millsap & Everson, 1993, and Mapuranga, Dorans, & Middleton, 2008, for reviews). Under test theory modeling, uniform DIF has been considered more in previous methodological research than nonuniform DIF, likely because it is cited as occurring less frequently in applications (Finch & French, 2007) and because some procedures, for example, the Mantel-Haenszel χ^2 test (Holland & Thayer, 1988), are not designed to detect nonuniform DIF (Li & Stout, 1996).

Studies investigating uniform DIF with non-IRT procedures across manifest groups (e.g., Klockars & Lee, 2008; Mazor, Clauser, & Hambleton, 1992; Narayanan & Swaminathan, 1994, 1996) have found that mean differences in ability distributions

across groups are associated with decreased power to detect uniform DIF. This is due to the fact that non-IRT methods have trouble separating differences in item difficulty from differences in item impact (Shealy & Stout, 1993). The Mantel-Haenszel procedure and modifications have been shown to have higher Type I error rates as the means of the ability distributions become more spread out, the discrimination parameter differs across groups, and item responses follow something other than the Rasch model (Fidalgo & Madeira, 2008). Simulations based on the mixed Rasch model (Samuelson, 2005) showed that when latent means differed by one standard deviation, heterogeneous item difficulty parameters could be successfully recovered while estimates of latent means were not recovered successfully for all groups.

On average, studies that consider heterogeneous factor loadings (nonuniform DIF) have found that increased differences in latent means across populations leads to decreased power to detect differences in thresholds and a decrease in power to detect small differences in factor loadings when thresholds are equal, regardless of whether or not latent means differ. In the presence of simulated nonuniform DIF, Li and Stout (1996) found that the Mantel-Haenszel procedure and SIBTEST (Shealy & Stout, 1993) have low power detect nonuniform DIF. With simulated ability distribution differences of 0, 0.5, and 1.0, power was found to further decrease when mean ability differed between populations.

Cohen and Bolt (2005) examined differential functioning of item responses from 1,000 examinees taking a college mathematics placement test across gender groups and across latent classes. Estimates from applying a latent class three-parameter logistic model to the data showed evidence of impact across genders and

latent classes; latent means differed by 0.25 standard deviations across males and females and 2.27 between latent classes. They analyzed items one at a time for population-specific difficulty, discrimination, and guessing parameters and found 5 of 32 items to exhibit uniform DIF. Finch and French (2008) found that in a test of uniform DIF, truly noninvariant factor loadings had the effect of increased type I error rates (ranging from 0.55 to 0.62 across study conditions), but the type I error rate was smaller when ability distributions were equal than when they were one standard deviation apart.

As mentioned above, several studies have compared CFA and IRT-based methods to detect differentially functioning discrete items (e.g., Meade & Lautenschlager, 2004; Raju et al., 2002; Reise et al., 1993). Most recently, Stark et al. (2006) compared CFA and IRT methods of detecting differences in factor loadings and item thresholds across two manifest groups for dichotomous and polytomous item responses. Their study design included three magnitudes of DIF: none, small and large. Small DIF was achieved by decreasing factor loadings for the second group by 0.15. Large DIF conditions were established by decreasing factor loadings in the second group by 0.4. Three types of DIF were considered: DIF due to item thresholds only, DIF due to factor loadings only, and DIF due to both factor loadings and item thresholds. They also considered two levels of impact: no impact, in which latent means were equal across populations, and moderate impact, in which one population's latent mean was half a standard deviation lower than the latent mean of the other population. With sample sizes of 500 and 1,000, their results showed a decrease in power to detect small magnitude DIF in the factor loadings when item

thresholds did not vary across populations. They also found that impact had little influence on power and type I error rates for both the IRT and CFA methods.

Summary of the current study

Invariance testing is an important part of latent variable modeling with continuous, cross-sectional data, whether for the purpose of estimating CFA model parameters or as part of a larger structural model. Heterogeneous measurement model parameters are of interest both substantively and methodologically throughout a variety of disciplines. Although most applications have used manifest grouping variables to capture heterogeneity, previous research has suggested that latent classes may be more appropriate when modeling cross-population measurement model differences. With increased computing power and methodological studies arising, latent variable mixture models are now more than ever a practical option for empirical analysis and it is of interest to assess the viability of mixture CFA for modeling heterogeneity and conducting invariance tests.

The inclusion of latent classes in a CFA model adds complexity, creating concerns about model selection, estimation, and accuracy of parameter estimates that cannot be addressed through existing research on multigroup CFA. The methodological literature on mixture CFA models with continuous indicators has thus far been concerned with modeling and testing for heterogeneity in latent means. Because modeling heterogeneity in factor loadings is a primary concern in the multigroup CFA domain, there is a need for comparable methodological research on modeling factor loading heterogeneity using mixture CFA models. In addition, the relationship between degree of differences in latent variable means, item difficulty,

and item discrimination has been a concern in the DIF literature where manifest variables are discrete and there is a need for those same concerns to be addressed when modeling and testing for CFA-DIF.

The focus of the current study was on modeling heterogeneous factor loadings and testing for nonuniform CFA-DIF, as factor loadings are the first and most commonly tested set of parameters when investigating the equivalence of measurement models across populations. Specifically, this study assessed whether mixture CFA parameters can be recovered successfully and the extent to which invariance testing on factor loadings results in accurate inferences. Data were generated from known population values to evaluate conditions that impact the ability of mixture CFA models to test for nonuniform CFA-DIF given the correct number of latent classes.

Studies have shown that substantial separation of latent classes, predominantly in the mean structure, must exist when estimating mixture CFA models to achieve convergence and successfully estimate latent means and the correct number of latent classes (e.g., Gagné, 2006; Lubke & Neale, 2006). These and related studies have also evaluated the impact of modeling noninvariant factor loadings on the ability of CFA mixture models to estimate heterogeneous latent means (e.g., Gagné, 2004; Lubke & Muthén, 2007). However, there have been no methodological studies to date that have shown evidence that separation of either latent means or measurement intercepts is a necessary condition for detecting heterogeneous factor loadings in mixture CFA models with continuous indicators.

If mixture CFA models are to be a viable alternative to or improvement over multigroup CFA models for modeling heterogeneous factor loadings, the models have to be able to perform well even when there is no mean separation in the factors and measurement intercepts. Mean separation is not a requirement in multigroup CFA modeling with observed group membership when solely testing parameters in the covariance structure. In mixture CFA modeling, a mean structure is imposed even when the focus is on heterogeneity in the covariance structure (e.g., Yung, 1997) and is currently required in modeling software such as Mplus 5.1 (Muthén & Muthén, 2007). Thus, the current study aimed to determine how well mixture CFA models can detect differences in factor loadings when factor means and measurement intercepts overlap across latent classes. The extreme case of a completely invariant mean structure was important to focus on initially since it was expected that heterogeneous factor loadings would be most difficult to detect in this situation (see Chapter 3 for a detailed explanation of hypotheses). If mixture CFA modeling can successfully determine the existence of nonuniform CFA-DIF when the location of latent classes is the same, then it is expected that mixture CFA can be used with confidence whether or not impact is hypothesized to exist. As such, the performance of mixture CFA models in this context was evaluated under a variety of study conditions. These conditions, discussed in detail in Chapter 3, were chosen based on the review of methodological and applied research to represent a broad range of research scenarios found in practice.

Chapter 3: Methodology

The goal of the current study is to identify conditions under which mixture CFA model parameters can be estimated successfully and to evaluate the impact of study conditions on the ability to detect nonuniform CFA-DIF across two populations in the presence of a completely invariant mean structure. Mixture CFA models allow researchers to form homogeneous clusters of individuals based on characteristics that may be difficult or impossible to measure quantitatively. While testing continuous indicators for CFA-DIF occurs widely in applications in which group membership is observed, testing for CFA-DIF across latent classes has been limited in practice, perhaps because previous studies with continuous indicators that have evaluated conditions that impact the performance of mixture CFA models have paid relatively little attention to factor loadings.

Design of study

True two-class one- and two-factor mixture CFA models were fit to simulated data with known population generating values in order to evaluate the accuracy of parameter estimates and the ability to detect noninvariant factor loadings. Using simulated data allowed for the analysis of a variety of conditions that may occur in applications. Fitting models with the true number of latent classes avoided potential complications that can arise when choosing the correct number of latent classes (described in Chapter 2) and focused the study on unconditional parameter estimates and their standard errors. Sample data was generated according to study conditions that were hypothesized to impact outcome measures. Study conditions were created

by varying sample size, latent class proportions, the pattern of noninvariant factor loadings, the size of nonuniform CFA-DIF, and the percentage of noninvariant factor loadings. The conditions were chosen to provide analyses comparable to methodological studies on invariance testing when group membership is known as well as to represent conditions that occur in practice in the social and behavioral sciences.

Samples sizes and latent class proportions were chosen to represent moderately small to large numbers of individuals. Total sample sizes of 200, 400, and 800 were used with equal class proportions. An unequal group size condition was also considered with latent class proportions equal to 0.25 and 0.75 for total sample sizes of 400 and 800. The smallest class studied had 100 observations. Pilot analyses suggested that smaller sample sizes were infeasible under the current study design with a completely noninvariant mean structure.

The noninvariance of factor loadings was manipulated in three ways: the pattern of noninvariance across classes was either uniform or mixed, the size of nonuniform CFA-DIF was either 0.25 or 0.40 and the percentage of noninvariant loadings was either low, medium, or high. In conditions with uniform noninvariance, the population generating values of all noninvariant class 1 factor loadings were higher than those for class 2. In mixed conditions, half of the noninvariant factor loadings were higher for class 1 and half were higher for class 2. The percentages of factor loadings chosen to be noninvariant were 25% and 50% for conditions with low or medium percentages, with the addition of high conditions with 75% of the factor loadings noninvariant for the two-factor model and 88% for the one-factor model.

For the one-factor model, there are 36 conditions with equal latent class proportions and 24 conditions with unequal latent class proportions for a total of 60 conditions. For the two-factor model, since the low percent noninvariant conditions only have one noninvariant loading, there is no mixed pattern condition, resulting in 50 conditions. In total, data was simulated from mixture CFA models under 110 study conditions.

Two levels of magnitude of factor loadings (low and high) were evaluated in pilot analyses. However, it was discovered that conditions with low factor loadings (all indicators loading either 0.3 or 0.7) required extremely long computation times and had very low rates of convergence to the global solution (approximately 10% to 12%) for all study conditions. This was expected since low factor loadings have been shown to be problematic when using maximum-likelihood estimation in the single population case (Ximénez, 2006). In practice, a model with all weak loadings for at least one population is indicative of a poorly constructed instrument or it may be that the model is misspecified for a particular set of data. Thus, conditions with low factor loadings were deemed infeasible under the current study conditions and not analyzed further.

Data generation

Two-class one- and two-factor models were used to generate eight continuous indicator variables from a multivariate normal distribution depending on class membership. These numbers of factors are consistent with previous simulation studies (e.g., Gagné, 2004; Lubke & Muthén, 2007; Lubke & Neale, 2006; Meade & Lautenschlager, 2004) and allow for the study of models both with and without a

factor covariance matrix. Both models are commonly used in practice. One-factor models are appropriate when an instrument is unidimensional or when subscales are individually estimated; multidimensional measurement instruments are represented by models with two or more factors.

The within-class models are shown in Figures 2 and 3. The eight indicator variables measure either one factor or two factors, resulting in a model with eight indicators per factor and a model with four indicators per factor. The numerical values in the figures represent parameters that were fixed for identification during estimation and are discussed more below.

Figure 2: One-factor model with eight indicator variables.

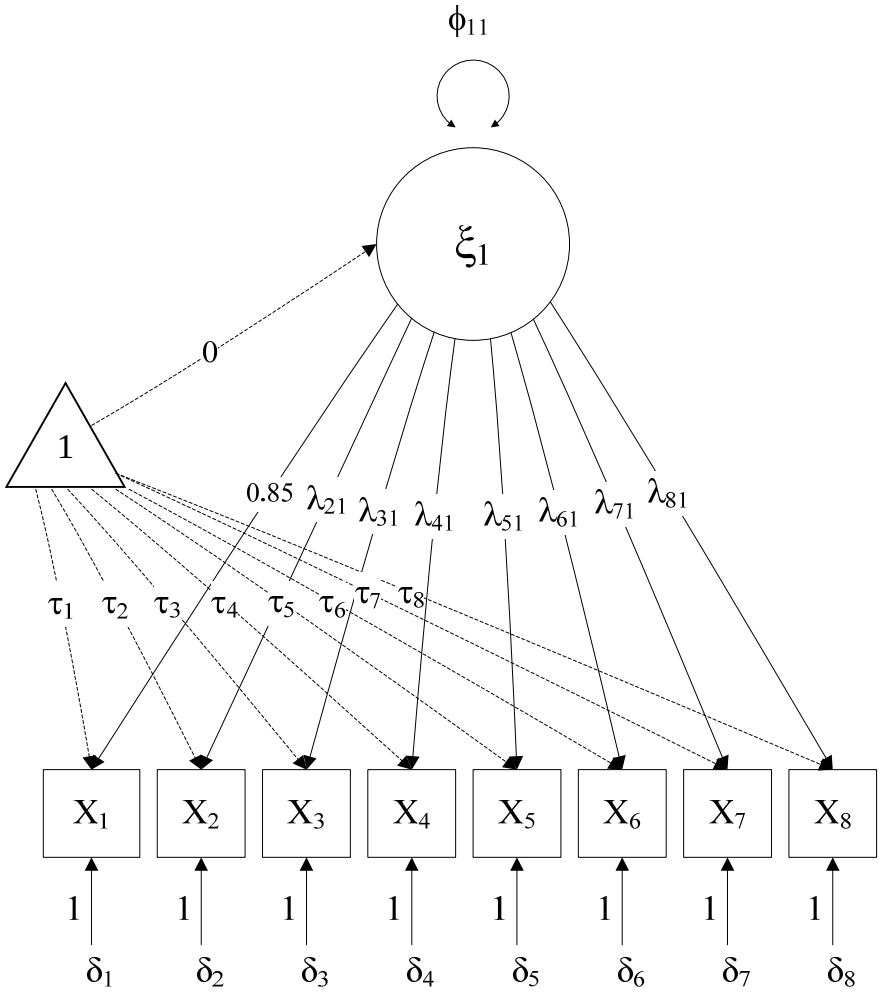
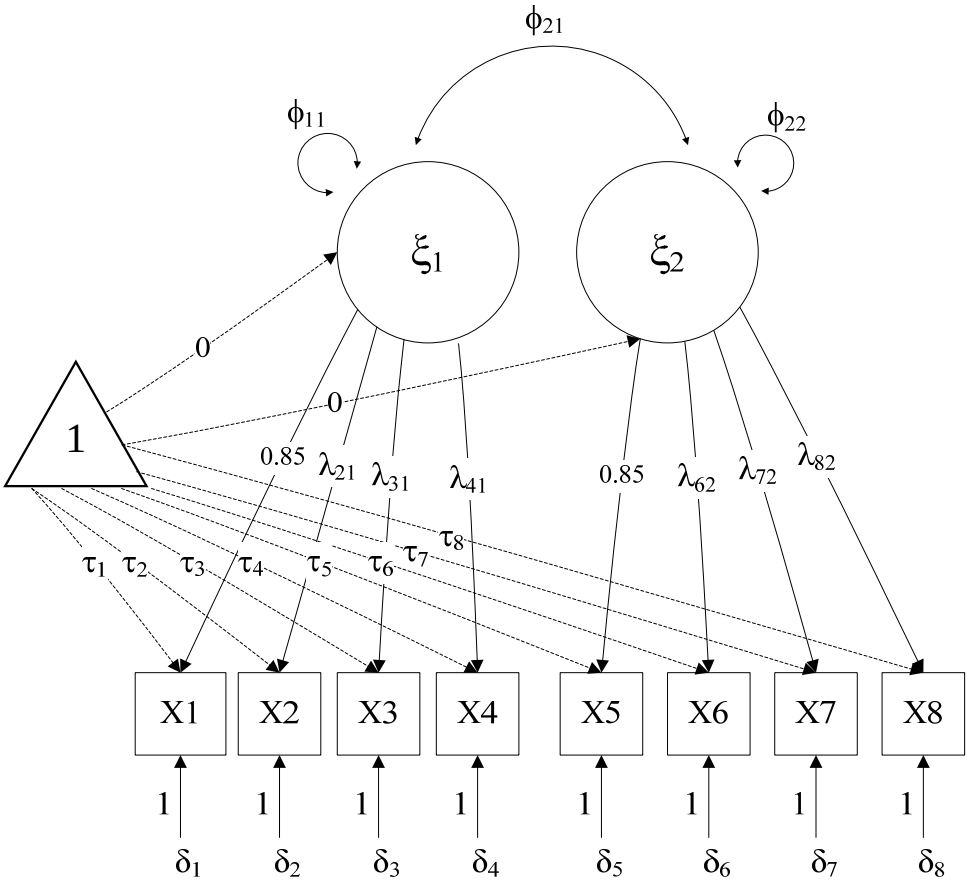


Figure 3: Two-factor model with four indicator variables per factor.



Population generating values for the factor loadings are reported in Table 1 for the one-factor model and in Table 2 for the two-factor model. Factor loadings were chosen to range from 0.45 to 0.85 and vary according to study conditions described above. A factor loading below 0.6 is indicative of a mediocre to weak relationship between a latent variable and its measured indicator (Ximénez, 2006). For conditions with unequal latent class proportions and a uniform pattern of noninvariance, the larger class had the smaller factor loadings for both models.

Table 1: Factor loading values for the one-factor model.

Size of CFA-DIF	% Noninvariant	Pattern	Class	λ_{11}	λ_{21}	λ_{31}	λ_{41}	λ_{51}	λ_{61}	λ_{71}	λ_{81}
0.40	High	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.45	0.45	0.45	0.45	0.45	0.45	0.45
		Mixed	1	0.85	0.85	0.85	0.85	0.45	0.45	0.45	0.45
			2	0.85	0.45	0.45	0.45	0.85	0.85	0.85	0.85
	Medium	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.85	0.85	0.85	0.45	0.45	0.45	0.45
		Mixed	1	0.85	0.85	0.85	0.85	0.85	0.85	0.45	0.45
			2	0.85	0.85	0.85	0.85	0.45	0.45	0.85	0.85
	Low	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.85	0.85	0.85	0.85	0.85	0.45	0.45
		Mixed	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.45
			2	0.85	0.85	0.85	0.85	0.85	0.85	0.45	0.85
0.25	High	Uniform	1	0.85	0.60	0.60	0.60	0.60	0.60	0.60	0.60
			2	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
		Mixed	1	0.85	0.60	0.60	0.60	0.60	0.85	0.85	0.85
			2	0.85	0.85	0.85	0.85	0.85	0.60	0.60	0.60
	Medium	Uniform	1	0.85	0.60	0.60	0.60	0.60	0.60	0.60	0.60
			2	0.85	0.60	0.60	0.60	0.85	0.85	0.85	0.85
		Mixed	1	0.85	0.60	0.60	0.60	0.60	0.60	0.85	0.85
			2	0.85	0.60	0.60	0.60	0.85	0.85	0.60	0.60
	Low	Uniform	1	0.85	0.60	0.60	0.60	0.60	0.60	0.60	0.60
			2	0.85	0.60	0.60	0.60	0.60	0.60	0.85	0.85
		Mixed	1	0.85	0.60	0.60	0.60	0.60	0.60	0.60	0.85
			2	0.85	0.60	0.60	0.60	0.60	0.60	0.85	0.60

Table 2: Factor loading values for the two-factor model.

Size of CFA-DIF	% Noninvariant	Pattern	Class	λ_{11}	λ_{21}	λ_{31}	λ_{41}	λ_{52}	λ_{62}	λ_{72}	λ_{82}
0.40	High	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.45	0.45	0.45	0.85	0.45	0.45	0.45
		Mixed	1	0.85	0.85	0.45	0.45	0.85	0.85	0.45	0.45
			2	0.85	0.45	0.85	0.85	0.85	0.45	0.85	0.85
	Medium	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.85	0.45	0.45	0.85	0.85	0.45	0.45
		Mixed	1	0.85	0.85	0.85	0.45	0.85	0.85	0.85	0.45
			2	0.85	0.85	0.45	0.85	0.85	0.85	0.45	0.85
	Low	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.85	0.85	0.45	0.85	0.85	0.85	0.45
0.25	High	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.60	0.60	0.60	0.85	0.60	0.60	0.60
		Mixed	1	0.85	0.85	0.60	0.60	0.85	0.85	0.60	0.60
			2	0.85	0.60	0.85	0.85	0.85	0.60	0.85	0.85
	Medium	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.85	0.60	0.60	0.85	0.85	0.60	0.60
		Mixed	1	0.85	0.85	0.85	0.60	0.85	0.85	0.85	0.60
			2	0.85	0.85	0.60	0.85	0.85	0.85	0.60	0.85
	Low	Uniform	1	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
			2	0.85	0.85	0.85	0.60	0.85	0.85	0.85	0.60

The population generating values for the remaining model parameters are shown in Table 3 for the one-factor model and in Table 4 for the two-factor model. For the one-factor model, the factor means (κ_c; c = 1, 2) and measurement intercepts ($\tau_{j,c}$; c = 1, 2, $j \in \{1 \ldots 8\}$) were set to zero in the population for both classes. Each within-class factor variance (ϕ_{11}) was set to unity. The first factor loading (λ_{11}) was set to 0.85 for both classes and was chosen as the referent loading for model identification during estimation while the remaining factor loadings vary across classes according to study conditions. The residual variances associated with each indicator variable were set to $1-\lambda_j^2$, $j \in \{1 \ldots 8\}$. Thus, all residual variances except the

one associated with the referent indicator varied across classes according to study conditions.

Table 3: Parameter generating values for the one-factor model.

Parameter	True Value	Index
$\tau_{j,c}$	0	$\forall$c, $\forall$j
κ_c	0	$\forall$c
$\phi_{11,c}$	1	$\forall$c
$\delta_{j,c}$	1- $\lambda^2_{j,c}$	$\forall$c, $\forall$j

c = 1, 2; j ∈{1…8}

Table 4: Parameter generating values for the two-factor model.

Parameter	True Value	Index
$\tau_{j,c}$	0	$\forall$c, $\forall$j
$\kappa_{1,c}$	0	$\forall$c
$\kappa_{2,c}$	0	$\forall$c
$\phi_{11,c}$	1	$\forall$c
$\phi_{22,c}$	1	$\forall$c
$\phi_{21,c}$	.5	$\forall$c
$\delta_{j,c}$	1- $\lambda^2_{j,c}$	$\forall$c, $\forall$j

c = 1, 2; j ∈{1…8}

The population generating values were chosen similarly for the two-factor model except for the addition of a referent loading associated with the second factor and the covariance between the two factors. Specifically, within class, the first factor loading for each factor (λ_{11} and λ_{52}) was fixed at 0.85 and the covariance between the two factors (ϕ_{21}) was set equal to 0.5. With these parameter values, each model was used to generate continuous indicator variables for each class distributed multivariate normal with a mean vector equal to 0 and unity variance on the diagonal of the covariance matrix.

Model estimation

Mplus 5.1 (Muthén & Muthén, 2007) was applied to the generated data and used to obtain maximum-likelihood estimates of parameters in the true two-class one- and two-factor models. Fifty sets of random starting values were used with ten iterations for each set, consistent with previous research on mixture CFA modeling (e.g., Lubke & Muthén, 2007; Lubke & Neale, 2006). The number of starting values used was higher than the default in Mplus 5.1 (Muthén & Muthén) in order to investigate local solutions more thoroughly.

For each model, the first loading associated with each factor was fixed to the true value (0.85) for each class to provide a metric for the factor. Factor means were constrained equal across classes and set to zero for model identification (Muthén, 2008; Yung, 1997). Residual variances were freely estimated. Additionally, all measurement intercepts and factor variances/covariances, although truly invariant, were freely estimated and unconstrained across classes, consistent with the hierarchy recommended in the invariance testing literature (described in Chapter 2).

In order to conduct the analyses described below, three nested models were estimated with varying degrees of severity of cross-population constraints on the factor loadings (excluding the referent factor loadings): completely unconstrained factor loadings, all factor loadings unconstrained except for one truly noninvariant factor loading, and completely constrained factor loadings. The remaining parameters were freely estimated as described above identically in all three configurations of factor loading invariance.

Replications

Mixture CFA models require more computation time than multigroup CFA models due to the addition of an extra parameter, the latent class proportion, which is treated as a missing value in the EM algorithm, as well as the propensity for mixture models to have multiple singularities in the likelihood surface. As such, studies on mixture CFA models have used few replications (e.g., 100; Lubke & Muthén, 2007; Lubke & Neale, 2006). In comparison, relative ease of estimation has allowed previous studies on CFA-DIF with manifest groups to use larger numbers of replications; for example 500 (Meade & Bauer, 2007) or 1,000 (French & Finch, 2006).

The potential for nonconvergence or convergence to a local solution implies that many more replications would be required to achieve a similar number of usable replications when evaluating a mixture CFA model with simulated data compared to a CFA model with observed population membership. Based on pilot analyses and existing work described in Chapter 2, the rates of nonconvergence or local maxima in the current study were hypothesized be higher than the aforementioned studies on mixture CFA, given a completely invariant mean structure. Therefore, datasets were generated in SAS 9.0 for each study condition and estimated in Mplus 5.1 (Muthén & Muthén, 2007) until global solutions were achieved for 500 replications, with an upper limit of 5,000 replications. Conditions under which 5,000 replications were reached without achieving 500 global solutions would then have a 90% or higher rate of unsuccessful replications. The number of replications required to achieve 500

global solutions was recorded for each study condition and the datasets with properly converged solutions were used to evaluate the outcome measures (described below).

A potential solution to intractable convergence problems is to estimate a more parsimonious model. When estimating mixture CFA models with noninvariant factor loadings and latent means, previous studies have constrained residual variances and factor variances/covariances to be equal across classes (e.g., Lubke & Muthén, 2007; Lubke & Neale, 2006). Pilot analyses indicated that constraining factor variances afforded little improvement in convergence rates. However, preliminary results revealed that global solutions were achieved at a much higher rate when residual variances were constrained equal across latent classes, at the expense of a severe decrement in the accuracy of parameter estimates due to the introduction of known misspecification into the estimation (since population values of the residual variances were set to be heterogeneous across latent classes). It was determined that constraining truly invariant residual variances to be equal across classes was not a viable option under the models and conditions in the current study. As such, both factor and residual variances were freely estimated with the expectation that parameter estimates would improve while the number of replications required to achieve 500 solutions that converged to a global maximum would increase.

Outcome measures

Modeling heterogeneous factor loadings across latent classes with continuous indicators in the presence of a completely invariant mean structure has not been previously studied; as such, the current research evaluated the viability of mixture CFA models for this purpose using several outcome measures. The outcome measures

are 1) parameter recovery; 2) classification accuracy; 3) power of the likelihood-ratio test to detect CFA-DIF in the entire factor loading matrix; and 4) power of the likelihood-ratio test to detect one truly noninvariant factor loading.

Parameter recovery

The impact of study conditions on the ability of mixture CFA to accurately estimate parameters of the completely unconstrained model was evaluated. In practice, the completely unconstrained model is typically used in the test of the null hypothesis of complete invariance of the factor loading matrix (Cheung & Rensvold, 2002). 95% confidence intervals were computed using parameter estimates and their standard errors for each of the 500 replications that achieved a global solution. The percentage of confidence intervals that contained the parameter generating value across replications was recorded for each individual parameter for each study condition and model. The coverage rates were then averaged across latent classes and parameters in each of the following matrices: factor loadings, measurement intercepts, residual variances, factor variances/covariances, and latent class proportions.

Label switching, a common problem in Bayesian analysis of latent class models in which the latent class labels switch on different Markov chain Monte Carlo iterations, does not occur when estimating a mixture CFA model via maximum-likelihood estimation with the EM algorithm (McLachlan & Peel, 2000). However, a similar problem can occur when using maximum-likelihood estimation, random starting values, and multiple replicated datasets to study parameter recovery. Specifically, since Mplus 5.1 (Muthén & Muthén, 2007) does not order latent classes,

estimated latent classes may differ across replicated datasets from the same model, resulting in the incorrect comparison of parameter generating values from one class with parameter estimates from a different class. The extent to which this has occurred in existing work and how it has been remedied is unclear. For example, Lubke and Muthén (2007) used random starting values in their analysis of parameter recovery but did not mention whether or not this type of label switching was an issue. Nylund et al. (2007) avoided this problem by specifying starting values equal to parameter generating values instead of using random starting values when measuring parameter recovery. However, this can artificially increase coverage rates by providing information that would not be available in practice, and would only give an upper bound on the accuracy of parameter estimates. In the current study, visual inspection of parameter estimates revealed this type of label switching did occur. As such, in the computation of parameter coverage rates, recovery of the total parameter matrix was computed by comparing population generating values to the estimates and also by reversing the latent classes and comparing the estimates to the population values, with the highest recovery rate of the two used as the measure of parameter recovery.

Classification accuracy

While modeling and testing for heterogeneous parameters, researchers may also be interested in using a mixture CFA model to cluster individuals or classify them into the appropriate population. It is possible that measurement model and structural parameters may be accurately estimated while estimates of individual-level parameter may or may not be accurate. If accurate estimation of CFA model

parameters can be achieved, it may be of interest to assess the ability of the model to assign individuals to their appropriate latent class.

Mplus 5.1 (Muthén & Muthén, 2007) uses the posterior probabilities of latent class membership computed during estimation through the EM algorithm (described in Chapter 2) to assign individuals to the most likely latent class. In order to compute correct class assignment in the current study, these estimated latent class memberships were compared to the population values for each individual. The differences were then averaged over all individuals for each replication, giving a value of classification accuracy for each replication and each condition. These values were then averaged across replications resulting in averages over study conditions.

In addition to correct class assignment, other measures are available that gauge the ability of the mixture CFA model to assign individuals to the correct latent class. Entropy is one such measure of classification accuracy that can be used in practice once a well-fitting model has been established. The formula for entropy can be written as

$$E_c = 1 - \frac{\sum_{i=1}^{n}\sum_{c=1}^{C}(-\hat{P}_{ic}\ln\hat{P}_{ic})}{n\ln C}$$ (Muthén, 1998-2004),

where $\hat{P}_{ic}$ is the estimated posterior probability of individual *i* belonging to latent class, c. Values of entropy range from 0 to 1, with values closer to unity indicating that individuals are assigned to a particular latent class with more certainty (Celeux & Soromenho, 1996). In the presence of a noninvariant mean structure and class-specific factor loadings, Lubke and Muthén (2007) found entropy values that ranged from 0.35 to 0.82, with the values increasing as class separation increased. Values of

entropy in the current study were expected to be lower given an invariant mean structure.

Post hoc analyses compared entropy to correct class assignment to evaluate the strength of entropy as an indicator of classification accuracy when modeling CFA-DIF in empirical applications using mixture CFA. To evaluate the potential of entropy as a predictor of classification accuracy, correct class assignment was regressed on entropy. Simple linear regression was conducted separately for each study condition.

Classification accuracy can also be determined in practice through graphical representation of the posterior probabilities of membership in a latent class. A clear delineation of latent classes would be evidenced by well-separated clusters of individuals. As an illustrative example, scatterplots of posterior probabilities of being a member of latent class 1 were created using one successfully estimated replicated dataset for select study conditions.

Testing for nonuniform CFA-DIF

An omnibus likelihood-ratio test on the entire factor loading matrix was performed first for both the one- and two-factor models, comparing a two-class model with all factor loadings constrained to be equal across classes (H_0) to a two-class model with all factor loadings free to vary across classes except for referent factor loadings (H_A). All other parameters were estimated freely except for factor means which were fixed to 0 for identification as described above for both null and alternative models. Power of the omnibus test for each study condition was computed

as the percentage of correctly rejected likelihood-ratio tests across the 500 replications with global solutions using $\alpha = 0.05$.

In addition to conducting invariance tests on the entire matrix of factor loadings, a second analysis was conducted on the power of the mixture CFA model to detect one truly noninvariant loading. This test was motivated by applications of measurement invariance testing with manifest groups which are concerned with partial measurement invariance and DIF in an IRT framework which is concerned with differential functioning of individual test items. When testing the invariance of factor loadings is of interest in applications using multigroup CFA, after the loading matrix is deemed noninvariant, the next step is often to examine individual factor loadings or sets of factor loadings to determine the source of noninvariance. To evaluate the ability of mixture CFA models to detect a cross-population difference in a single factor loading, the current study compared a model with all factor loadings except the referents free to vary (H_A) to a model with one truly noninvariant loading constrained equal across classes (H_0). The likelihood-ratio test was used and power was calculated as the percentage of replications out of 500 which lead to rejection at the $\alpha = 0.05$ level of significance.

Hypotheses

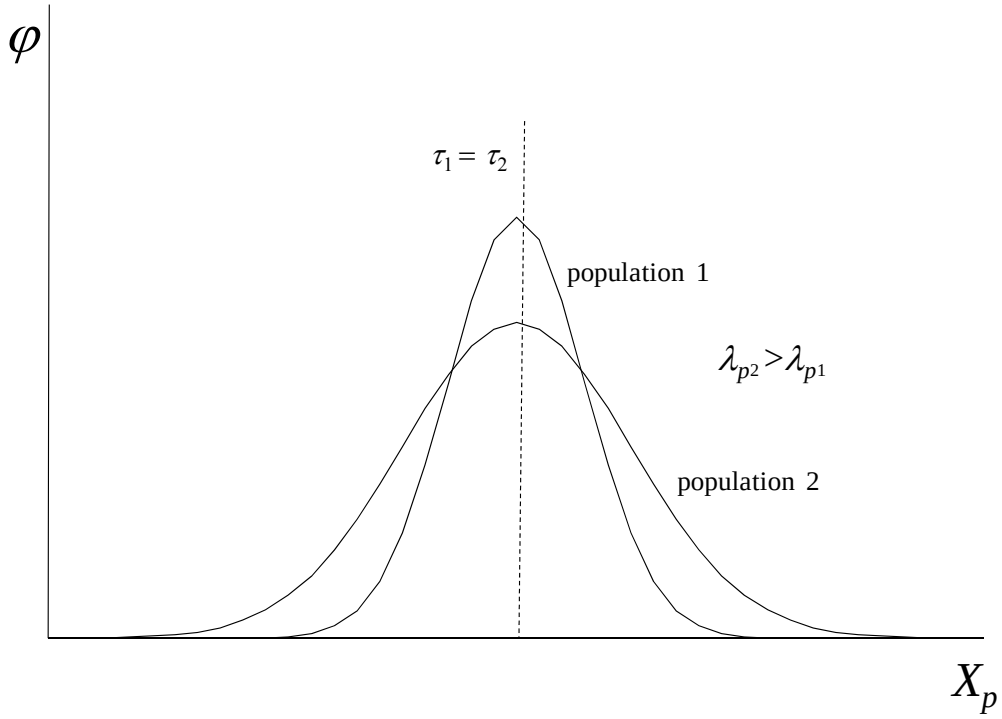
The primary focus of the current study was on the successful estimation and detection of heterogeneous factor loadings and other CFA parameters across populations with unobserved group membership in the presence of a completely invariant mean structure. An invariant mean structure is associated with a complete overlap in the location of each latent class's distribution of manifest variables,

making the aggregate distribution appear unimodal. When group membership is observed, the location of the distribution for each group does not impact the ability to obtain estimates of parameters in the covariance structure. However, in general mixture modeling, and therefore in mixture CFA modeling, because latent class memberships must be estimated, too much overlap among component distributions may lead to difficulty in differentiating between latent classes.

In the presence of heterogeneous factor loadings and a completely invariant mean structure, distinctions among latent classes and estimates of mixture model parameters are based on differences in spread, not location, since the location is equal for all latent classes. This can be conceptualized using a simplified example. Consider the equation for one of *P* indicator variables measuring a single latent variable (with individual subscripts suppressed), $X_p = \tau + \lambda_p\xi + \delta_p$. The pdfs corresponding to X_p for two populations with heterogeneous factor loadings and equal measurement intercepts and latent means are shown in Figure 4. As the diagram shows, the two distributions differ in spread but have the same location. An invariant mean structure implies that the location of the distribution of manifest variables the same for all latent classes. To see this mathematically, consider the expected value of X_p, which is $E(X_p) = E(\tau) + \lambda_p E(\xi) + E(\delta_p)$. When the mean structure is completely invariant, latent means are equal across classes. In this case, the term $\lambda_p E(\xi)$ will always equal 0 due to the need to identify the model. That is, either latent variable means in all classes are fixed to 0 or the latent mean of one class is fixed to zero, implying that the other invariant latent means are also equal to zero. As such, $E(\lambda_p\xi) = \lambda E(\xi) = 0$ for all classes, implying that differences in factor loadings do not impact the location of the

manifest variables. Further, assuming residuals are distributed with a mean of 0, then $E(X) = E(\tau)$, which is the same for all latent classes when measurement intercepts are invariant across populations.

Figure 4: Probability density function for one manifest variable for two populations with heterogeneous factor loadings and a noninvariant mean structure.



While differences in factor loadings do not impact the location of the distributions when the mean structure is invariant, heterogeneous factor loadings do impact the variability of the manifest variables. Specifically, higher factor loadings are associated with a larger spread in their associated manifest variables, *ceteris paribus*. Considering again the equation for X_p shown above and the pdfs in Figure 4, the variance of the observable variable is $\text{Var}(X_p) = \lambda_p^2\text{Var}(\xi) + \text{Var}(\delta_p)$. Holding factor and residual variances constant and equal across latent classes, the variability of a manifest variable will be higher for latent classes with higher values of associated

factor loadings. Thus, in the presence of a completely invariant mean structure and heterogeneous factor loadings, only the difference in the variability of the observable variables is available to differentiate between latent classes when estimating with maximum-likelihood methods.

Given an invariant mean structure, the following are hypotheses about the impact of manipulated study conditions on parameter recovery, classification accuracy, and power of the likelihood-ratio test.

- As sample size increases, the accuracy of parameter estimates and classification will improve and power to detect cross-population differences in factor loadings will increase.
- As latent class proportions become less equal, the accuracy of parameter estimates and classification will decrease and power will be negatively impacted by the decrease in accuracy.
- As the size of the difference in factor loadings across latent classes increases, the accuracy of parameter estimates and classification will increase and the power to detect differences in factor loadings across latent classes will increase.
- As the percentage of noninvariant factor loadings increases, the accuracy of parameter estimates and classification will increase and the power to detect noninvariant factor loadings will increase.
- A mixed pattern of noninvariant factor loadings will result in higher accuracy of parameter estimates and classification and increased

power to detect noninvariant factor loadings than a uniform pattern across latent classes.

- Entropy will be a better predictor of correct class assignment under conditions in which CFA-DIF and sample size are large.
- Plots of posterior probabilities of latent class membership will show clearer delineation of two latent classes as the size of CFA-DIF and sample size increase.
- The two-factor model has two more parameters to estimate and fewer indicators per factor than the one-factor model so it is expected to perform slightly worse than the one-factor model.

Results

The following chapter presents results separately for the one- and two-factor models. The number of replications required to achieve 500 global solutions is reported for all study conditions. Analysis of variance (ANOVA) with a 5-way [2 (latent class proportions) × 3 (sample size) × 2 (size of CFA-DIF) × 2 (pattern of factor loading noninvariance) × 3 (percentage of noninvariant factor loadings)] unbalanced design was performed to evaluate the impact of study conditions on parameter recovery, correct class assignment, and power. Results are reported for study conditions that are included in the highest significant interaction from each 5-way ANOVA.

The results show the ability of mixture CFA models to estimate heterogeneous factor loadings and test for CFA-DIF in the presence of a completely invariant mean structure. The impact of study conditions on the outcome measures provides

information about conditions under which mixture CFA models may be a feasible option for measurement invariance testing. Evaluation of mixture CFA modeling when the locations of the latent class distributions are the same offers insight into the performance of mixture CFA for modeling and testing for CFA-DIF under an extreme and undesirable modeling situation. It is expected that as the locations of latent class distributions become more separated, accuracy of estimation and testing would improve for all study conditions.

Chapter 4: Results

Results from estimating parameters from two-class one- and two-factor mixture CFA models with data generated under known study conditions are presented in four sections. Initial information about the viability of the mixture CFA model under each study condition is presented in the first section via rates of convergence to the global maximum. The remaining three sections present results based on the outcome measures: parameter recovery, classification accuracy and power to detect nonuniform CFA-DIF. Results are reported separately for the one- and two- factor models for study conditions in the highest statistically significant interaction from a 5-way ANOVA on each outcome measure.

Convergence to the global maximum

The number of replications required to achieve 500 global solutions, with an upper limit of 5,000 replications for each study condition, is reported for the one-factor model in Table 5. Counts closer to 500 indicate better rates of convergence to the global maximum. As counts increase, the feasibility of estimating a CFA mixture model under a given study condition decreases. In general, as expected, larger sample sizes, larger differences in factor loadings, a mixed pattern of noninvariance, and more noninvariant factor loadings were associated with higher rates of convergence to the global solution in the one-factor model.

Table 5: Number of replications out of 5,000 needed to achieve 500 global solutions for the one-factor model.

Latent class proportion	Size of CFA-DIF	Noninvariance of factor loadings		Number of replications out of 5000 to reach 500 global solutions		
		Percentage	Pattern	N=800	N=400	N=200
0.5/0.5	0.40	high	uniform	500	543	1107
			mixed	500	500	631
		medium	uniform	504	702	1848
			mixed	501	680	1606
		low	uniform	1011	2677	4488
			mixed	930	2493	4953
	0.25	high	uniform	562	1276	3861
			mixed	507	796	2412
		medium	uniform	1125	2632	4545
			mixed	957	2364	4206
		low	uniform	3501	4629	5000+[1]
			mixed	3235	4800	5000+[2]
0.25/0.75	0.40	high	uniform	534	1066	
			mixed	501	532	
		medium	uniform	977	2719	
			mixed	544	1024	
		low	uniform	3876	5000+[3]	
			mixed	1445	3398	
	0.25	high	uniform	1570	4604	
			mixed	600	1399	
		medium	uniform	3673	5000+[4]	
			mixed	1641	3684	
		low	uniform	5000+[5]	5000+[6]	
			mixed	4102	5000+[7]	

Total number of replications out of 5,000 with global solutions: 462[1], 468[2], 386[3], 467[4], 450[5], 340[6], 473[7]

Six of the 60 study conditions for the one-factor model had less than a 10% rate of convergence to the global solution. Data generated under these conditions was characterized by one or more of the following: fewer than 100 observations in the smallest latent class, small nonuniform CFA-DIF, few noninvariant factor loadings, and a uniform pattern of noninvariant factor loadings. Eleven of 60 conditions had a 90% or better rate of convergence to the global solution. Data generated under these

conditions was characterized by more than 200 observations in the smaller latent class, many noninvariant factor loadings, and large CFA-DIF. Conditions with 100 observations in the smaller latent class achieved acceptable rates of convergence to the global solution only when CFA-DIF was large and there were many noninvariant factor loadings following a mixed pattern of noninvariance.

For a given total sample size, conditions with unequal latent class proportions had fewer successfully converged replications than conditions with equal latent class proportions. In particular, with a total sample size of 800 and unequal latent class proportions (N = 200 for one class and N = 600 for the other), the numbers of replications needed to achieve convergence to the global solution were higher than for conditions with N = 800 and equal class proportions. When comparing conditions with the same number of observations in the smaller class, conditions with unequal latent class proportions (and therefore a higher total sample size) had higher rates of convergence to the global solution than conditions with equal latent class proportions when the pattern of noninvariance was mixed and there were many noninvariant factor loadings.

For the two-factor model, the number of replications required to achieve 500 global solutions is reported in Table 6. For all study conditions, more replications were required to achieve 500 replications with global solutions for the two-factor model than for the one-factor model. 29 out of 50 conditions did not achieve 500 global solutions within the upper limit of 5,000 replications. All conditions with small CFA-DIF and 200 observations or fewer in the smaller class had a less than 10% rate of convergence to the global maximum. This suggests that when the model is more

complex or there are fewer indicators per factor, larger sample sizes are required, especially when CFA-DIF is small. Similar to the one-factor model, conditions with a combination of large sample size, large CFA-DIF, and many noninvariant factor loadings required the fewest replications to achieve the desired number of global solutions and conditions with a mixed pattern of factor loading noninvariance had higher rates of convergence than conditions with a uniform pattern of factor loading noninvariance.

Table 6: Number of replications out of 5,000 needed to achieve 500 global solutions for the two-factor model.

Latent class proportion	Size of CFA-DIF	Noninvariance of factor loadings		Number of replications out of 5000 to reach 500 global solutions		
		Percentage	Pattern	N=800	N=400	N=200
0.5/0.5	0.40	high	uniform	743	1643	5000+[1]
			mixed	549	822	3475
		medium	uniform	794	1850	5000+[2]
			mixed	698	1487	5000+[3]
		low	uniform	2522	5000+[4]	5000+[5]
	0.25	high	uniform	1779	5000+[6]	5000+[7]
			mixed	1523	5000+[8]	5000+[9]
		medium	uniform	3938	5000+[10]	5000+[11]
			mixed	3301	5000+[12]	5000+[13]
		low	uniform	5000+[14]	5000+[15]	5000+[16]
0.25/0.75	0.40	high	uniform	1479	5000+[17]	
			mixed	703	2009	
		medium	uniform	2944	5000+[18]	
			mixed	1030	3754	
		low	uniform	5000+[19]	5000+[20]	
	0.25	high	uniform	5000+[21]	5000+[22]	
			mixed	3165	5000+[23]	
		medium	uniform	5000+[24]	5000+[25]	
			mixed	5000+[26]	5000+[27]	
		low	uniform	5000+[28]	5000+[29]	

Total number of replications out of 5,000 with global solutions: 397[1], 492[2], 435[3], 347[4], 214[5], 406[6], 220[7], 482[8], 220[9], 231[10], 140[11], 240[12], 145[13], 210[14], 188[15], 190[16], 367[17], 252[18], 206[19], 152[20], 259[21], 145[22], 250[23], 170[24], 140[25], 361[26], 210[27], 130[28], 225[29]

Taken together, the results from the one- and two-factor models suggest a trade-off among model complexity, sample size, size of CFA-DIF, and percentage and pattern of noninvariant factor loadings. When the pattern of noninvariance was uniform, the communalities in one class were lower than in the other, especially when there were many noninvariant factor loadings and large CFA-DIF. As such, larger sample sizes were required to compensate for the difficulty in estimating parameters for the class with the lower factor loadings. Conditions with few noninvariant factor

loadings posed problems for attaining global solutions, even with as many as 400 observations per class. This suggests the need for a more theory-based and parsimonious model, at least when the mean structure is completely invariant. Overall, datasets that were characterized by larger sample sizes, larger CFA-DIF, and many noninvariant factor loadings showed promise in the ability to achieve successful estimation of mixture CFA model parameters.

The following sections contain results from the impact of study conditions on parameter recovery, classification accuracy, and power. Results are based on the number of replications that achieved convergence to the global solution. As such, the results are conditional and may be biased upward for study conditions that required more than 500 replications to achieve the global solution. In addition, outcome measures may be less reliable for conditions that did not achieve 500 proper solutions because results are based on fewer replications. Since the one-factor model achieved higher rates of convergence to the global solution across all study conditions, the following sections focus more on the results of the one-factor model than the two-factor model.

Parameter recovery

Parameter coverage rates for the one-factor model are reported separately for each parameter matrix in Table 7. Percentages are reported for all study conditions since the 5-way interactions from individual ANOVAs with average parameter recovery for each matrix as the dependent variables were all statistically significant. Average parameter coverage rates for the factor loading matrix ranged from 79% to 96% across conditions. For measurement intercepts, average parameter coverage rates

ranged from 52% to 97%. Average parameter coverage rates for the remaining parameters were 68% to 95% for residual variances, 72% to 97% for factor variances, and 25% to 99% for latent class proportions.

Table 7: Parameter coverage rates, one-factor model.

Sample size	Size of CFA-DIF	Noninvariance of factor loadings		Percentage of replications out of 500 in which 95% confidence interval covers parameter									
				Equal latent classes					Unequal latent classes				
		Percentage	Pattern	λ	τ	δ	ϕ	π	λ	τ	δ	ϕ	π
N=800	0.40	high	uniform	95	95	94	96	98	96	96	94	95	87
			mixed	95	95	94	94	99	95	96	94	94	79
		medium	uniform	96	97	95	97	98	92	92	91	94	85
			mixed	95	96	94	95	97	95	95	94	95	88
		low	uniform	93	92	91	93	82	81	69	78	81	48
			mixed	92	92	91	93	84	90	87	87	92	74
	0.25	high	uniform	95	95	94	97	94	89	90	87	92	77
			mixed	96	96	95	96	97	95	96	93	95	88
		medium	uniform	92	92	90	94	82	80	71	73	84	50
			mixed	93	93	90	94	84	90	88	85	90	75
		low	uniform	82	70	76	82	36	79	62	73	79	33
			mixed	83	72	77	83	42	83	68	74	80	44
N=400	0.40	high	uniform	95	95	94	95	97	92	93	90	93	84
			mixed	95	96	93	95	98	95	95	92	94	91
		medium	uniform	93	94	92	92	90	84	78	81	86	63
			mixed	94	94	92	94	92	92	92	88	92	83
		low	uniform	85	78	80	85	56	79	61	74	77	46
			mixed	86	79	81	86	56	84	73	78	83	60
	0.25	high	uniform	92	92	87	92	82	81	72	72	81	34
			mixed	93	93	90	94	86	91	89	85	91	81
		medium	uniform	86	78	77	86	55	80	63	71	79	49
			mixed	84	79	76	85	56	82	72	73	82	61
		low	uniform	81	61	70	78	26	81	55	73	75	41
			mixed	82	59	71	78	28	81	60	71	77	43
N=200	0.40	high	uniform	92	92	87	91	86					
			mixed	94	94	90	93	94					
		medium	uniform	88	82	81	85	66					
			mixed	89	83	82	89	69					
		low	uniform	81	59	71	77	36					
			mixed	81	59	71	75	32					
	0.25	high	uniform	84	74	71	79	54					
			mixed	87	81	79	88	66					
		medium	uniform	81	58	68	74	35					
			mixed	80	63	68	76	37					
		low	uniform	81	52	70	72	26					
			mixed	80	54	69	72	25					

Note: Highest significant interaction from 5-way ANOVA*: pattern*percent*CFA-DIF size*sample size*proportion ($p<.01$) for λ, τ, δ, ϕ, and* π. For 5-way interaction: λ: $df = 2$, $F = 7.71$, $\omega^2 = 0.0003$; τ: $df = 2$, $F = 18.85$, $\omega^2 = 0.0008$; δ: $df = 2$, $F = 8.94$, $\omega^2 = 0.0003$; ϕ: $df = 2$, $F = 7.20$, $\omega^2 = 0.0003$; π: $df = 2$, $F = 5.93$, $\omega^2 = 0.0002$;

∗Percentages out of fewer than 500 replications for conditions where noted in Table 5 above.

For the one-factor model, latent class proportions had the widest range of coverage rates among the parameter matrices. Many study conditions had coverage rates lower than 50%, pointing to difficulty in distinguishing two separate latent class distributions from the unimodal aggregate distribution. For all parameter matrices, the lowest coverage rates occurred for conditions with N = 200, low CFA-DIF, and few noninvariant factor loadings. The highest coverage rates occurred for conditions with N = 800, high CFA-DIF, and a moderate to high percentage of noninvariant factor loadings. For the smallest class size of 100 with either equal or different latent class proportions, coverage rates for all parameter matrices were 90% or above when the data exhibited all of the following: all factor loadings noninvariant (except the referent) in a mixed pattern across latent classes with nonuniform CFA-DIF equal to 0.40.

Coverage rates for the factor loading matrix and the residual variance matrix were generally higher than for the other parameter matrices, as expected, since those matrices were truly noninvariant. For parameters in the factor loading matrix, as CFA-DIF became larger, the accuracy of parameter estimates was higher. A mixed pattern of noninvariant factor loadings across latent classes was also associated with higher coverage rates in general. Alternatively, coverage rates for the factor loadings were never above 90% when there were only one or two truly noninvariant factor loadings in the entire parameter matrix.

Parameter coverage rates for the two-factor model are reported in Table 8 though Table 11. The 5-way interaction of study conditions was not statistically significant for any parameter matrix, perhaps because there were more cells with

fewer than 500 replications for the two-factor model. There were several significant 4-way interactions, each presented individually in a table. In Table 8, results for all parameter matrices are aggregated over the sizes of latent class proportions. Results in Table 9 are combined across pattern of noninvariance for all parameter matrices. Coverage rates for the factor loading matrix and latent class proportion, averaged across sizes of CFA-DIF are reported in Table 10. Results for residual and factor variances are aggregated across percentage of noninvariant factor loadings in Table 11.

Table 8: Parameter coverage rates, two-factor model.

Sample size	Size of CFA-DIF	Noninvariance of factor loadings		Percentage of replications out of 500 in which 95% confidence interval covers parameter				
		Percentage	Pattern	λ	τ	δ	ϕ	π
N=800	0.40	high	uniform	96	95	94	95	92
			mixed	96	96	95	95	94
		medium	uniform	94	94	93	94	91
			mixed	95	95	94	95	90
		low	uniform	89	88	88	91	76
	0.25	high	uniform	93	92	90	94	89
			mixed	93	94	91	95	91
		medium	uniform	90	87	87	92	79
			mixed	90	89	87	92	80
		low	uniform	85	75	81	82	61
N=400	0.40	high	uniform	94	93	91	93	92
			mixed	95	95	93	94	94
		medium	uniform	92	91	91	93	91
			mixed	91	92	89	92	87
		low	uniform	87	78	83	86	67
	0.25	high	uniform	89	86	84	87	77
			mixed	90	87	85	90	79
		medium	uniform	88	77	82	85	68
			mixed	85	77	79	86	62
		low	uniform	82	66	76	78	55
N=200	0.40	high	uniform	90	87	83	88	82
			mixed	92	89	87	92	85
		medium	uniform	78	75	74	78	55
			mixed	88	83	84	88	74
		low	uniform	85	68	79	78	67
	0.25	high	uniform	88	76	81	78	75
			mixed	88	77	82	84	73
		medium	uniform	84	63	78	78	54
			mixed	83	55	69	71	62
		low	uniform	82	57	77	76	55

Note: Highest significant interaction from 5-way ANOVA: *pattern*percent*CFA-DIF size*sample size ($p<.01$) for λ, τ, δ, ϕ, and π.* For 4-way interaction: *λ*: $df = 2$, $F = 6.14$, $\omega^2 = 0.0004$; *τ*: $df = 2$, $F = 10.89$, $\omega^2 = 0.0008$; *δ*: $df = 2$, $F = 14.22$, $\omega^2 = 0.0011$; *ϕ*: $df = 2$, $F = 11.88$, $\omega^2 = 0.0008$; *π*: $df = 2$, $F = 11.05$, $\omega^2 = 0.0009$.

∗Percentages out of fewer than 500 replications for conditions where noted in Table 6 above.

Table 9: Parameter coverage rates, two-factor model, continued.

Sample size	Size of CFA-DIF	Percentage of noninvariant factor loadings	Percentage of replications out of 500 in which 95% confidence interval covers parameter									
			Equal latent classes					Unequal latent classes				
			λ	τ	δ	ϕ	π	λ	τ	δ	ϕ	π
800	0.40	high	96	95	95	95	97	96	95	94	95	88
		medium	95	96	95	96	95	94	94	92	94	87
		low	92	93	91	93	79	83	77	82	84	70
	0.25	high	94	94	93	95	91	91	91	88	93	89
		medium	91	90	89	93	79	87	84	83	90	80
		low	85	77	82	83	56	85	71	80	82	69
400	0.40	high	95	95	94	94	94	94	93	91	93	92
		medium	93	93	92	94	90	89	89	87	91	87
		low	88	79	84	86	65	85	75	79	84	79
	0.25	high	90	87	85	89	78	85	82	78	87	90
		medium	87	78	81	86	62	85	72	78	81	83
		low	85	66	80	79	50	70	67	57	73	76
200	0.40	high	91	88	85	90	84					
		medium	83	79	79	82	64					
		low	85	68	79	78	67					
	0.25	high	88	77	82	81	74					
		medium	84	59	73	74	58					
		low	82	57	77	76	55					

Note: Highest significant interaction from 5-way ANOVA: *percent*CFA-DIF size*sample size*proportion (p<.01)* for *λ, τ, δ, ϕ,* and π. For 4-way interaction: *λ*: $df = 2$, $F = 63.39$, $\omega^2 = 0.0059$; *τ*: $df = 2$, $F = 7.47$, $\omega^2 = 0.0006$; *δ*: $df = 2$, $F = 19.92$, $\omega^2 = 0.0017$; *ϕ*: $df = 2$, $F = 184.64$, $\omega^2 = 0.0155$; *π*: $df = 2$, $F = 5.45$, $\omega^2 = 0.0005$.

∗Percentages out of fewer than 500 replications for conditions where noted in Table 6 above.

Table 10: Parameter coverage rates, two-factor model, continued.

Sample size	Noninvariance of factor loadings		Percentage of replications out of 500 in which 95% confidence interval covers parameter			
			Equal latent classes		Unequal latent classes	
	Percentage	Pattern	λ	π	λ	π
800	high	uniform	95	95	93	84
		mixed	95	93	94	92
	medium	uniform	94	87	90	85
		mixed	93	87	93	83
	low	uniform	90	72	84	69
400	high	uniform	92	86	93	90
		mixed	93	87	93	92
	medium	uniform	92	84	85	88
		mixed	90	78	89	86
	low	uniform	87	59	79	78
200	high	uniform	90	81		
		mixed	92	84		
	medium	uniform	78	54		
		mixed	88	73		
	low	uniform	84	65		

Note: Highest significant interaction from 5-way ANOVA: *pattern*percent *sample size*proportion (p<.01)* for λ and π. For 4-way interaction: λ: $df = 1$, $F = 8.95$, $\omega^2 = 0.0003$; π: $df = 1$, $F = 11.31$, $\omega^2 = 0.0005$.

∗Percentages out of fewer than 500 replications for conditions where noted in Table 6 above.

Table 11: Parameter coverage rates, two-factor model, continued.

Sample size	Size of CFA-DIF	Pattern of noninvariance	Percentage of replications out of 500 in which 95% confidence interval covers parameter			
			Equal latent classes		Unequal latent classes	
			δ	ϕ	δ	ϕ
800	0.40	uniform	93	94	90	92
		mixed	95	95	94	95
	0.25	uniform	89	92	83	87
		mixed	91	94	87	93
400	0.40	uniform	90	92	87	90
		mixed	92	94	90	93
	0.25	uniform	82	85	70	78
		mixed	84	89	76	86
200	0.40	uniform	78	82		
		mixed	86	90		
	0.25	uniform	79	77		
		mixed	77	78		

Note: Highest significant interaction from 5-way ANOVA: *pattern *sample size*size of CFA-DIF*proportion (p<.01)* for δ and ϕ. For 4-way interaction: λ: $df = 1$, $F = 21.99$, $\omega^2 = 0.0009$; ϕ: $df = 1$, $F = 9.70$, $\omega^2 = 0.0003$.

∗Percentages out of fewer than 500 replications for conditions where noted in Table 6 above.

Coverage rates from the two-factor model show that the interaction of sample size, size of CFA-DIF and noninvariance of factor loading is similar to the one-factor model. Relative to the one-factor model, latent class proportions have higher coverage rates in the two-factor model in general, while interacting with sample size and size of CFA-DIF similarly. The factor variance/covariance matrix had recovery rates similar to the one-factor model, despite the addition of four parameters. Sample size had the largest effect and was a part of all significant 4-way interactions.

Classification accuracy

Correct class assignment and entropy are reported in Table 12 for the one-factor model for study conditions according to the highest interaction from a 5-way ANOVA with average correct class assignment over individuals for each replication as the dependent variable. The 5-way interaction was significant for the one-factor model so average values for all study conditions are reported. Correct class assignment ranged from 53% to 82% across conditions with equal latent class proportions and 51% to 64% across conditions with unequal latent class proportions.

Table 12: Percentage of individuals assigned to the correct latent class and entropy, one-factor model.

Latent class proportion	Size of CFA-DIF	Noninvariance of factor loadings		Correct class assignment/entropy		
		Percentage	Pattern	N=800	N=400	N=200
0.5/0.5	0.40	high	uniform	0.82/0.48	0.81/0.50	0.77/0.59
			mixed	0.81/0.47	0.80/0.49	0.78/0.55
		medium	uniform	0.76/0.37	0.73/0.44	0.68/0.60
			mixed	0.74/0.37	0.72/0.43	0.67/0.58
		low	uniform	0.64/0.37	0.59/0.56	0.56/0.73
			mixed	0.63/0.35	0.59/0.55	0.56/0.73
	0.25	high	uniform	0.74/0.37	0.71/0.47	0.64/0.66
			mixed	0.74/0.35	0.71/0.43	0.66/0.60
		medium	uniform	0.66/0.37	0.60/0.57	0.56/0.72
			mixed	0.65/0.36	0.60/0.56	0.57/0.72
		low	uniform	0.54/0.64	0.53/0.73	0.53/0.77
			mixed	0.54/0.61	0.53/0.71	0.53/0.77
0.25/0.75	0.40	high	uniform	0.63/0.57	0.63/0.61	
			mixed	0.64/0.56	0.63/0.58	
		medium	uniform	0.59/0.51	0.56/0.63	
			mixed	0.60/0.49	0.59/0.56	
		low	uniform	0.52/0.66	0.52/0.73	
			mixed	0.55/0.50	0.54/0.65	
	0.25	high	uniform	0.58/0.55	0.55/0.67	
			mixed	0.60/0.48	0.59/0.56	
		medium	uniform	0.53/0.66	0.52/0.72	
			mixed	0.55/0.53	0.54/0.64	
		low	uniform	0.51/0.77	0.52/0.76	
			mixed	0.52/0.69	0.52/0.74	

Note: Highest significant interaction from 5-way ANOVA for correct class assignment: *proportion*CFA-DIF size*percent*pattern*sample size* ($p<.01$). For 5-way interaction: $df = 2$, $F = 8.47$, $\omega^2 = 0.0001$.

Similar to the accuracy of parameter estimates, correct class assignment was highest for conditions with a large sample, large CFA-DIF, and many noninvariant factor loadings. When latent class sizes were equal, sample size had a larger impact on classification accuracy when there were only several factor loadings that were noninvariant and when CFA-DIF was small. When latent classes were unequal, conditions with a mixed pattern of noninvariant factor loadings had higher classification accuracy than when noninvariance was uniform. Under the best

conditions, that is, when data was generated with equal latent class proportions and many factor loadings exhibited large CFA-DIF, a large portion, about 20% of observations, were misclassified.

When latent class proportions were equal, under the worst conditions, classification accuracy was only slightly better than chance. That is, a 50% classification accuracy rate could be achieved simply by assigning all individuals to one latent class. With unequal latent class proportions, under the worst conditions, classification accuracy was worse than the 75% rate that could be achieved by assigning all individuals to the same latent class. Across all conditions, Cohen's kappa coefficient, defined by the equation, $\kappa = \frac{p_a - p_c}{1 - p_c}$, where p_a is the proportion of individuals correctly assigned to the appropriate latent class and p_c is the proportion of individuals that would be correctly assigned due to chance, ranged from 0.007 to 0.67, indicative of poor to moderate levels of classification accuracy.

Entropy ranged from moderate to good (0.35 - 0.77) across study conditions for the one-factor model. However, for all study conditions except for those with equal latent class proportions, N = 800, and CFA-DIF equal to 0.40, there was an inverse relationship between entropy and correct class assignment, indicating that entropy is not a good measure of classification accuracy under most study conditions when the mean structure is completely invariant. This relationship is described further after the results from the two-factor model are reported.

Correct classification and entropy are reported in Table 13 for the two-factor model. There were no significant interactions higher than the three 4-way interactions that are combined in the table. Classification accuracy was generally lower for the

two-factor model than for the one-factor model. The highest classification accuracy value in the table is merely 0.77, which was the average over conditions with a mixed or a uniform pattern of noninvariance for a sample size of 800, equal latent class proportions, large CFA-DIF, and many noninvariant factor loadings. When taking chance correct classifications into account, classification accuracy was even worse for the two-factor model relative to the one-factor model. Specifically, Cohen's kappa coefficient was low to moderate across all study conditions, ranging from .007 to 0.57. These results show that the two-factor model was not successful at classifying individuals under any study condition, likely due to the increased complexity of the model without an increase in the number of indicator variables relative to the one-factor model.

Table 13: Percentage of individuals assigned to the correct latent class and entropy, two-factor model.

Latent class proportions	Size of CFA-DIF	Pattern of noninvariance	Sample size			Pattern of noninvariance	
			800	400	200	uniform	mixed
0.5/0.5	0.4	uniform	0.71/0.37	0.69/0.46	0.66/0.57		
		mixed	0.73/0.36	0.71/0.43	0.66/0.63		
	0.25	uniform	0.64/0.39	0.60/0.51	0.58/0.64		
		mixed	0.64/0.35	0.62/0.47	0.57/0.63		
0.75/0.25	0.4	uniform	0.58/0.51	0.57/0.54			
		mixed	0.60/0.49	0.59/0.53			
	0.25	uniform	0.54/0.51	0.53/0.55			
		mixed	0.56/0.45	0.55/0.53			
		Percent noninvariant					
0.5/0.5	0.4	high	0.77/0.39	0.75/0.44	0.71/0.56	0.75/0.47	0.74/0.46
		medium	0.72/0.35	0.69/0.43	0.63/0.62	0.70/0.44	0.67/0.48
		low	0.62/0.36	0.58/0.51	0.57/0.62	0.60/0.46	
	0.25	high	0.68/0.34	0.64/0.46	0.60/0.60	0.65/0.44	0.64/0.43
		medium	0.62/0.37	0.58/0.52	0.55/0.65	0.60/0.46	0.59/0.46
		low	0.54/0.53	0.54/0.58	0.53/0.70	0.54/0.60	
0.75/0.25	0.4	high	0.61/0.51	0.60/0.53		0.60/0.53	0.61/0.51
		medium	0.58/0.47	0.57/0.53		0.57/0.49	0.58/0.50
		low	0.53/0.53	0.53/0.55		0.53/0.54	
	0.25	high	0.57/0.45	0.55/0.53		0.55/0.51	0.57/0.46
		medium	0.54/0.50	0.53/0.55		0.53/0.54	0.54/0.51
		low	0.51/0.56	0.52/0.55		0.52/0.55	

Note: Highest significant interactions from 5-way ANOVA for correct class assignment: *proportion*magnitude*pattern*sample size ($p<.01$),df =1 , F = 14.1,* $\omega^2 = 0.0001$; *proportion*percent*CFA-DIF size*sample size ($p<.01$), df =2 , F =20.25 ,* $\omega^2 = 0.0005$; *and proportion*percent*magnitude*pattern ($p<.01$), df =1 , F =6.95 ,* $\omega^2 = 0$.

Entropy values for the two-factor model ranged from 0.33 to 0.70 across all study conditions for the two-factor model, a similar range relative to the one-factor model. Aggregate values reported in Table 13 show that correct classification rates and entropy are also inversely related in the two-factor model. To further evaluate the ability of entropy to predict correct class assignment and its usefulness in practice when correct class assignment is unknown, simple linear regression was used for each condition.

Table 14 shows the regression coefficients (with standard errors in parentheses) from the equation predicting correct class assignment from entropy for the one-factor model. Only several conditions have positive coefficients, indicative of

a positive relationship between correct class assignment and entropy. These occur for conditions with equal latent class proportions, large CFA-DIF, and a combination of $N = 800$ with a medium or high percentage of noninvariant factor loadings or $N = 400$ with a high percentage of noninvariant factor loadings. While coefficients for these conditions are statistically significantly greater than 0 at the .01 level, the associated R^2 values ranged from merely 0.003 to 0.103, implying that little of the variation in correct class assignment can be explained by entropy. For all other conditions, the regression coefficients are negative, indicating an inverse relationship between correct class assignment and entropy. In these conditions, a majority of the observations are being classified into one class with relatively high certainty. Small within-class samples, few noninvariant factor loadings, small CFA-DIF, and a uniform pattern of factor loadings made differences between latent classes hard to detect and caused the data appear to be derived from a single population.

Table 14: Regression coefficients in the prediction of correct class assignment from entropy, one-factor model.

Latent class proportion	Size of CFA-DIF	Noninvariance of factor loadings		Regression coefficient		
		Percentage	Pattern	N=800	N=400	N=200
0.5/0.5	0.40	high	uniform	0.17 (0.024)	0.10 (0.029)	-0.54 (0.029)
			mixed	0.2 (0.027)	0.19 (0.029)	-0.17 (0.036)
		medium	uniform	0.03 (0.028)	-0.43 (0.019)	-0.43 (0.016)
			mixed	0.11 (0.031)	-0.36 (0.019)	-0.38 (0.017)
		low	uniform	-0.26 (0.007)	-0.23 (0.008)	-0.2 (0.011)
			mixed	-0.23 (0.006)	-0.19 (0.006)	-0.17 (0.009)
	0.25	high	uniform	-0.21 (0.029)	-0.41 (0.013)	-0.4 (0.015)
			mixed	-0.01 (0.034)	-0.38 (0.019)	-0.41 (0.016)
		medium	uniform	-0.31 (0.008)	-0.27 (0.008)	-0.19 (0.011)
			mixed	-0.27 (0.008)	-0.25 (0.008)	-0.2 (0.011)
		low	uniform	-0.13 (0.004)	-0.09 (0.005)	-0.09 (0.008)
			mixed	-0.12 (0.004)	-0.08 (0.005)	-0.08 (0.008)
0.25/0.75	0.40	high	uniform	-0.15 (0.01)	-0.15 (0.011)	
			mixed	-0.12 (0.014)	-0.15 (0.014)	
		medium	uniform	-0.17 (0.006)	-0.13 (0.008)	
			mixed	-0.15 (0.008)	-0.15 (0.008)	
		low	uniform	-0.06 (0.003)	-0.06 (0.004)	
			mixed	-0.09 (0.004)	-0.08 (0.005)	
	0.25	high	uniform	-0.16 (0.005)	-0.12 (0.006)	
			mixed	-0.15 (0.007)	-0.15 (0.008)	
		medium	uniform	-0.09 (0.003)	-0.06 (0.004)	
			mixed	-0.1 (0.004)	-0.08 (0.006)	
		low	uniform	-0.03 (0.002)	-0.03 (0.004)	
			mixed	-0.05 (0.002)	-0.04 (0.003)	

Note: All coefficients are significant at the .01 level. For 5-way interaction: $df = 2$, $F = 5.24$, $\omega^2 = 0.0002$.

Regression coefficients from the equation predicting correct classification accuracy from entropy for all study conditions for the two-factor model are reported in Table 15. There are no positive, significant coefficients. While parameter recovery results showed improved estimates of latent class proportions for the two-factor model relative to the one-factor model, it was more difficult to correctly classify individuals in the two-factor model. As such, entropy was not a useful measure of classification accuracy for any study condition for the two-factor model under a completely invariant mean structure.

Table 15: Regression coefficients in the prediction of correct class assignment from entropy, two-factor model.

Latent class proportion	Size of CFA-DIF	Noninvariance of factor loadings		Regression coefficient		
		Percentage	Pattern	N=800	N=400	N=200
0.5/0.5	0.40	high	uniform	0.04* (0.029)	-0.19 (0.031)	-0.47 (0.037)
			mixed	-0.01* (0.034)	-0.10 (0.038)	-0.40 (0.03)
		medium	uniform	-0.22 (0.024)	-0.33 (0.027)	-0.34 (0.013)
			mixed	-0.11 (0.032)	-0.32 (0.027)	-0.35 (0.025)
		low	uniform	-0.20 (0.011)	-0.18 (0.014)	-0.17 (0.023)
	0.25	high	uniform	-0.35 (0.015)	-0.32 (0.017)	-0.33 (0.04)
			mixed	-0.36 (0.015)	-0.29 (0.017)	-0.21 (0.03)
		medium	uniform	-0.28 (0.01)	-0.23 (0.018)	-0.11 (0.03)
			mixed	-0.24 (0.01)	-0.20 (0.016)	-0.14 (0.02)
		low	uniform	-0.09 (0.007)	-0.09 (0.012)	-0.05 (0.01)
0.25/0.75	0.40	high	uniform	-0.14 (0.008)	0* (0.016)	
			mixed	-0.14 (0.01)	0* (0.01)	
		medium	uniform	-0.11 (0.007)	0.03* (0.026)	
			mixed	-0.13 (0.006)	0* (0.011)	
		low	uniform	-0.06 (0.005)	-0.01* (0.021)	
	0.25	high	uniform	-0.10 (0.008)	-0.08 (0.013)	
			mixed	-0.11 (0.007)	-0.08 (0.012)	
		medium	uniform	-0.07 (0.008)	-0.05 (0.011)	
			mixed	-0.08 (0.005)	-0.05 (0.011)	
		low	uniform	-0.03 (0.005)	-0.02 (0.006)	

Note; Coefficients statistically significant (p <.01), except where noted with *.

Scatterplots of posterior probabilities of belonging to latent class 1 are shown in Figures 5, 6, and 7 using results from one replication of each study condition with equal latent class proportions. These particular replications were chosen as representatives of the results from aggregating replications for the other outcome measures described above. The results from evaluating global solutions and parameter recovery showed that estimation of the mixture CFA models was more successful as sample size and CFA-DIF became larger and as more factor loadings were noninvariant. Therefore, looking at Figures 5, 6, and 7, estimation was more successful moving down and towards the lower right in each figure and improved as sample size became larger across figures.

Plots in which most of the density falls to the extreme right indicate that most individuals are being assigned to the same single latent class, which demonstrates difficulty in distinguishing between the two latent class distributions. In such cases, values of entropy were high because most individuals were being classified in latent class 1 with almost perfect certainty (the posterior probabilities of belonging to latent class 1 were close to 1), but classification accuracy is poor because half of the individuals truly belong in latent class 2. On the other hand, scatterplots which contain clusters of individuals at both extremes represent higher classification accuracy. This can be seen in the scatterplots in the lower right quadrant of Figure 7. The clusters are more separated with increased density at both 0 and 1 and classification accuracy was around 80%. The associated entropy values were moderate due to individuals with posterior probabilities that are close to 0.5. Overall,

plots of the posterior probabilities were better indicators of classification accuracy than values of entropy.

Figure 5: Scatterplots of posterior probabilities of class membership, one-factor model, $N = 200$, equal latent class proportions.

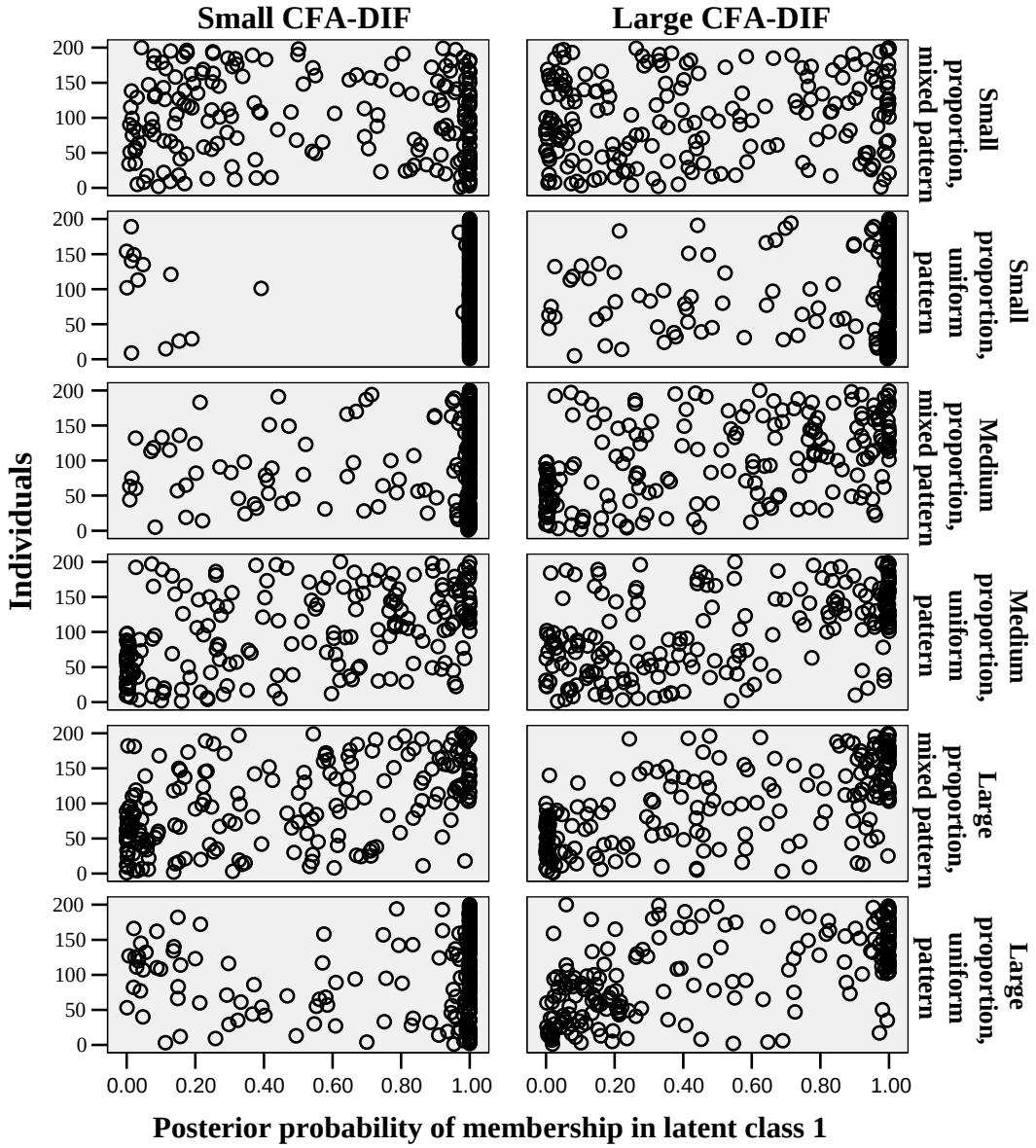


Figure 6: Scatterplots of posterior probabilities of class membership, one-factor model, $N = 400$, equal latent class proportions.

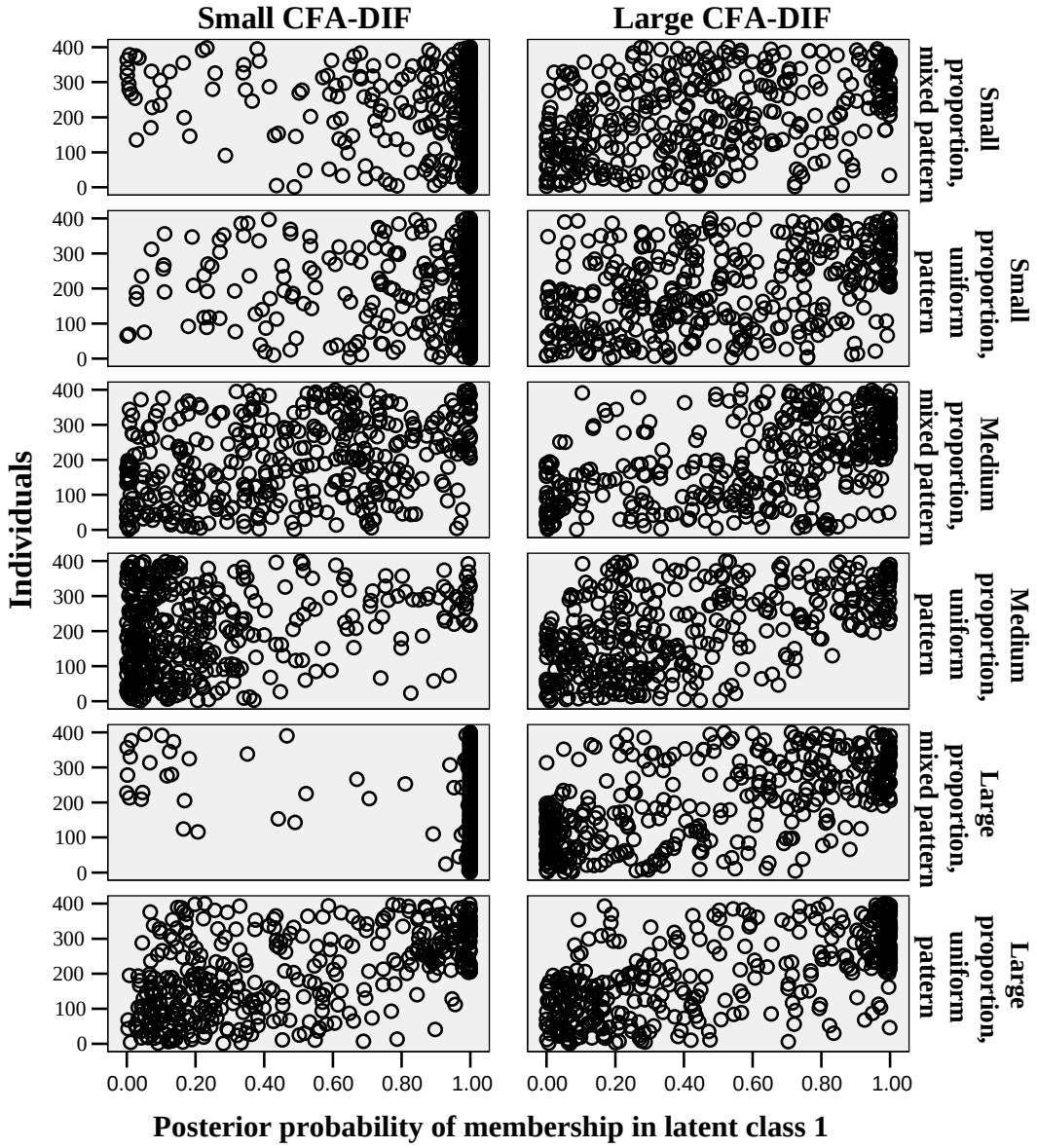
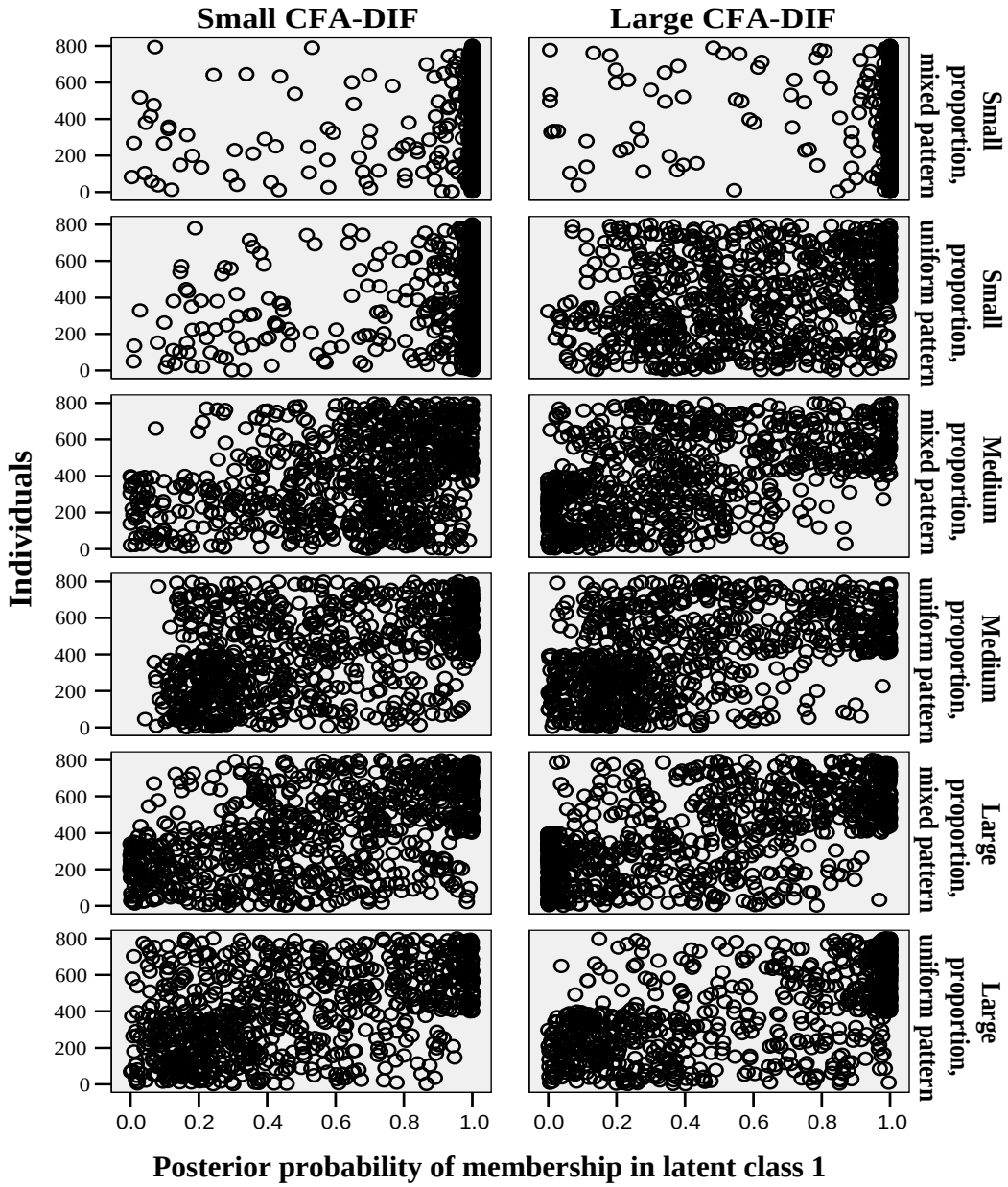


Figure 7: Scatterplots of posterior probabilities of class membership, one-factor model, *N* = 800, equal latent class proportions.



Power

When factor loadings were completely constrained across latent classes in order to estimate the null hypothesize model for the omnibus likelihood-ratio test, fewer replications converged to the global maximum than when the factor loadings

were freely estimated for some study conditions and across models. This is not surprising since noninvariant factor loadings in the constrained model were misspecified while only residual variances (both invariant and noninvariant depending on the study conditions) and parameters that were truly invariant (measurement intercepts and factor variances/covariances) were allowed to vary across classes during estimation. As such, the only source of heterogeneity in the estimated null hypothesized model was from deviations due to random sampling error and differences in the residual variances across classes, leading to more replications in which the global solution was not found.

When only one truly noninvariant factor loading was constrained equal across classes for the likelihood-ratio test for an individual factor loading, even fewer replications converged to the global solution, especially with only a few truly noninvariant factor loadings. A combination of misspecification and too many parameters to estimate likely contributed to problems in finding the global solution in the null hypothesized model.

That the null hypothesized model could not be estimated successfully under some conditions pointed to the need to estimate a one-class model instead of a completely constrained two-class model. As such, the problem of nonconvergence to the global solution under the constrained model was investigated more thoroughly in a small subset of replications by estimating a one-class model and comparing the AIC and sample-size adjusted BIC of the one-class model and the two-class unconstrained model. In every instance, these fit indices favored the unconstrained two-class model and two conclusions were drawn. First, using the likelihood-ratio test and comparing

models with the same number of latent classes may not be feasible under some conditions when the mean structure is invariant. Second, power results for the current study, reported for only those replications in which the constrained model also converged to a global solution, represent a lower bound since results do not include replications in which model selection across different numbers of latent classes would have favored the two-class model with heterogeneous factor loadings.

Power of the omnibus likelihood-ratio test for invariance of factor loadings

Power of the omnibus likelihood-ratio test is reported in Table 16 for the one-factor model and in Table 17 for the two-factor model with the total number of replications in which the constrained model converged to a global solution in parentheses. Results are reported for study conditions in the highest significant interaction from a 5-way ANOVA with a dichotomous dependent variable indicating whether or not the null hypothesis of invariance was correctly rejected for each replication.

Table 16: Power for omnibus likelihood-ratio test, one-factor model.

Latent class proportion	Size of CFA-DIF	Noninvariance of factor loadings		Percentage of correct rejections		
		Percentage	Pattern	N=800	N=400	N=200
0.5/0.5	0.40	high	uniform	99.6 (500)	89.8 (499)	64.3 (473)
			mixed	100 (497)	100 (470)	99.7 (369)
		medium	uniform	100 (484)	98.5 (407)	85.2 (324)
			mixed	100 (464)	99.7 (366)	90 (321)
		low	uniform	96.4 (358)	80.3 (290)	62.3 (265)
			mixed	97.6 (339)	81.7 (278)	63 (246)
	0.25	high	uniform	81.3 (497)	56.1 (460)	55.9 (381)
			mixed	99.8 (461)	97.6 (377)	86 (329)
		medium	uniform	93.5 (397)	73.2 (321)	56 (282)
			mixed	96.8 (380)	81.1 (307)	65 (260)
		low	uniform	74.9 (299)	59.6 (245)	57.4 (216)
			mixed	69.5 (279)	61.9 (247)	55.7 (228)
0.25/0.75	0.40	high	uniform	100 (497)	93.7 (441)	
			mixed	100 (495)	99.8 (441)	
		medium	uniform	98.1 (359)	82 (261)	
			mixed	99.8 (438)	99.1 (343)	
		low	uniform	69.9 (296)	58.7 (189)	
			mixed	93.4 (334)	78.3 (277)	
	0.25	high	uniform	77.9 (438)	60.3 (343)	
			mixed	99.8 (443)	92 (364)	
		medium	uniform	73.7 (289)	57.7 (241)	
			mixed	89.3 (337)	68.4 (288)	
		low	uniform	68.9 (219)	56.5 (170)	
			mixed	64 (286)	63.4 (191)	

Note: Highest significant interaction from 5-way ANOVA: *proportion* CFA-DIF size*percent*pattern*sample size (p<.01), df* = 2, $F = 5.24$, $\omega^2 = 0.0002$.

Table 17: Power for omnibus likelihood-ratio test, two-factor model.

Sample size	Size of CFA-DIF	Pattern of noninvariance	Latent class proportions		Percentage of noninvariant factor loadings		
			equal	unequal	high	medium	low
800	0.40	uniform	98.6 (925)	93.1 (350)	99.2 (738)	96.7 (398)	87.1 (139)
		mixed	99.8 (509)	100 (521)	100 (690)	99.7 (340)	
	0.25	uniform	74.8 (643)	46.6 (148)	71.4 (472)	75.1 (237)	42.7 (82)
		mixed	91.8 (280)	82.7 (266)	93 (344)	77.7 (202)	
400	0.40	uniform	85.5 (760)	69.0 (213)	83.1 (585)	89.1 (258)	62.3 (130)
		mixed	97.0 (305)	95.6 (342)	98.3 (407)	92.9 (240)	
	0.25	uniform	43.3 (349)	35.7 (140)	44.9 (287)	39.4 (99)	32.0 (103)
		mixed	74.9 (191)	45.8 (120)	71.3 (195)	50.9 (116)	
200	0.40	uniform	58.0 (371)		53.4 (223)	69.2 (104)	54.5 (44)
		mixed	83.1 (178)		96 (99)	67.1 (79)	
	0.25	uniform	26.7 (150)		28.6 (70)	42.9 (35)	11.1 (45)
		mixed	26.7 (75)		42.9 (35)	12.5 (40)	

Note: Highest significant interactions from 5-way ANOVA: *proportion*CFA-DIF size*pattern*sample size (p<.01), df* = 1, *F* = 23.38, ω^2 = 0.0031; and *percent*CFA-DIF size*pattern*sample size (p<.05), df* = 2, *F* = 2.85, ω^2 = 0.0002.

Across all study conditions, power ranged from 56% to 100% for the one-factor model and from 11% to 100% for the two-factor model. Power was lowest for conditions with a combination of small sample size, small CFA-DIF, and a uniform pattern of noninvariance, as expected. A mixed pattern of noninvariant factor loadings was associated with higher power relative to a uniform pattern of noninvariance except when only two factor loadings exhibited small CFA-DIF. When the pattern of noninvariance was uniform, power was considerably lower for conditions with many noninvariant factor loadings. Conditions with small CFA-DIF

and small sample size had much lower power for the two-factor model relative to the one-factor model.

For the one-factor model, conditions with unequal latent class proportions had similar power levels to conditions with equal latent class proportions when total sample size was 800, the percentage of noninvariant factor loadings was high and CFA-DIF was large. Conditions with a mixed pattern of noninvariance also had similar levels of power for both equal and unequal latent class proportions when there were many factor loadings with small CFA-DIF or a moderate percentage of factor loadings with large CFA-DIF.

Power of the likelihood-ratio test for invariance of one factor loading

Power of the likelihood-ratio test to detect one truly noninvariant factor loading is reported in Table 18 for the one factor model and in Table 19 for the two-factor model. Power rates are reported for successfully converged replications with the number of replications noted in parentheses. Power for the test of one truly invariant factor loading was much lower relative to testing the entire factor loading matrix for both models across all study conditions, but the pattern across study conditions was similar to the omnibus test (r = 0.89 for the one-factor model and r = 0.85 for the two-factor model).

Table 18: Power of test of one truly invariant factor loading, one-factor model.

Latent class proportion	Size of CFA-DIF	Noninvariance of factor loadings		Percentage of correct rejections		
		Percentage	Pattern	N=800	N=400	N=200
0.5/0.5	0.40	high	uniform	99.4 (498)	86 (485)	56.3 (368)
			mixed	99.8 (499)	95.8 (496)	76 (425)
		medium	uniform	100 (493)	87 (392)	57.7 (248)
			mixed	99.6 (494)	90.7 (407)	56.8 (257)
		low	uniform	91.3 (254)	56.9 (174)	29.7 (158)
			mixed	92.8 (279)	57.3 (178)	30.4 (161)
	0.25	high	uniform	76.2 (484)	45.4 (372)	22.5 (209)
			mixed	84.7 (491)	57.9 (418)	35 (246)
		medium	uniform	66.9 (341)	41.3 (225)	23.3 (176)
			mixed	70.5 (356)	35 (234)	25.7 (171)
		low	uniform	38.5 (195)	32.1 (165)	22.2 (126)
			mixed	37.9 (169)	25 (180)	17.2 (134)
0.25/0.75	0.40	high	uniform	99.2 (472)	80 (320)	
			mixed	97.6 (499)	79.8 (476)	
		medium	uniform	92.2 (309)	51.1 (188)	
			mixed	96.8 (466)	71.7 (350)	
		low	uniform	39.1 (128)	33.6 (119)	
			mixed	76.6 (239)	47.1 (204)	
	0.25	high	uniform	61.3 (297)	28.6 (185)	
			mixed	68.8 (462)	40.9 (320)	
		medium	uniform	42.2 (180)	19.9 (141)	
			mixed	51.4 (292)	30.3 (201)	
		low	uniform	34.2 (117)	41.7 (12)	
			mixed	40.3 (159)	24.7 (154)	

*Note: Highest significant interaction from 5-way ANOVA: proportion*CFA-DIF size*percent*pattern*sample size ($p<.01$), $df = 1$, $F = 8.73$, $\omega^2 = 0.0004$.*

Table 19: Power of test of one truly invariant factor loading, two-factor model.

Size of CFA-DIF	Noninvariance of factor loadings		Power		
	Percentage	Pattern	N=800	N=400	N=200
0.40	high	uniform	93.6 (637)	70.1 (442)	31.3 (163)
		mixed	93.6 (723)	79.1 (560)	57.1 (210)
	medium	uniform	89.9 (434)	80.6 (216)	45.4 (97)
		mixed	94.1 (657)	75.3 (421)	46.3 (121)
	low	uniform	74.4 (180)	32.6 (89)	21.6 (37)
0.25	high	uniform	53.3 (441)	32.4 (207)	7.7 (65)
		mixed	55.9 (587)	37.8 (278)	30 (50)
	medium	uniform	54.8 (263)	33.3 (66)	0 (25)
		mixed	50.3 (310)	22.7 (119)	50 (20)
	low	uniform	30.4 (79)	19.4 (67)	28.6 (35)

Note: Highest significant interaction from 5-way ANOVA: *CFA-DIF size*percent*pattern*sample size* *($p<.01$) df* = 2, F = 24.01, ω^2 = 0.0047.

For the one-factor model, power was less than 80% for 48 of the 60 study conditions. Most notably, power was very low when only two factor loadings were truly noninvariant, except for conditions with N = 800, equal latent class proportions, and large CFA-DIF. Similar to the omnibus test, conditions with a mixed pattern of noninvariance had higher power, except for when there were few noninvariant factor loadings with small CFA-DIF.

There was only one significant 4-way interaction in the 5-way ANOVA for the two-factor model. Power results from the two-factor model, aggregated over latent class proportions, show that power was 89.9 and above for conditions with a total sample size of 800 and high or medium sized CFA-DIF. The remaining results

show poor rates of convergence for the constrained model, as noted by the few number of converged solutions in parentheses, and low power rates for the two-factor model when testing the invariance of one-factor loading.

Overall, power was highest under conditions that more easily achieved global solutions and produced accurate parameter estimates, total sample size equal to 800 and many factor loadings with CFA-DIF equal to 0.40. Both the omnibus test and the test of an individual factor loading were successful under these conditions. The inability of the models to distinguish between latent classes when sample size and effect size were small and when there were few noninvariant factor loadings negatively impacted power, as expected. Power results from the omnibus test of invariance of the factor loading matrix showed promise, at least for the one-factor model, while the power of the invariance test of an individual factor loading was not acceptable for either model under conditions that achieved moderate success in estimation to a global solution and accuracy of parameter estimates

Chapter 5: Discussion

The current study provided an analysis of accuracy and power when testing for CFA-DIF with mixture CFA modeling, thereby assessing the potential of using latent classes instead of manifest groups when examining differential indicator functioning with continuous, cross-sectional data. The focus was on nonuniform CFA-DIF as invariance testing of factor loadings is theoretically and statistically important for accurate and valid inferences drawn from a measurement model. The following sections include a summary of this research, limitations of the research design, implications of the results, recommendations for applied researchers, and methodological extensions.

Summary of study

Testing for measurement invariance is important in latent variable modeling and using a CFA model for this purpose is common when data are continuous. The assumption of a homogeneous measurement instrument is often unrealistic in applications in the social and behavioral sciences and much attention is paid in particular to whether or not populations differ with respect to their factor loadings. Researchers who wish to make inferences about cross-population differences in construct representation, values of latent means, or structural relations must assess measurement invariance to ensure valid inferences.

There has been strong methodological interest in the accuracy of parameter estimates and the quality of inferences drawn from heterogeneous latent variable models. Research on multigroup CFA with maximum-likelihood estimation has

shown that invariance testing can be successful when groups are in fact known under many conditions seen in practice (French & Finch, 2006; Meade & Bauer, 2007). Omnibus tests of the entire factor loading matrix have been shown to result in accurate inferences even with small sample sizes while tests of individual factor loadings may be suspect even under ideal conditions (French & Finch). Existing work suggests that mixture CFA models provide accurate estimates of heterogeneous measurement model parameters across latent classes when there are differences in latent means whether or not there are differences in measurement intercepts, even for relatively small samples (Gagné, 2004; Lubke & Muthén, 2007).

The current simulation study investigated the viability of using mixture CFA models to conduct invariance tests on factor loadings with continuous data from two heterogeneous populations. The extreme case of an invariant mean structure was chosen to ascertain whether the use of a mixture CFA model necessitates the presence of impact or differences in measurement intercepts, which are not requirements for multigroup CFA, when testing for nonuniform CFA-DIF. Without differences in the mean structure, the locations of the latent class distributions overlap and the aggregate distribution is unimodal, making it potentially difficult to distinguish among latent classes and accurately estimate their parameters.

The current study has shown some potential situations under which mixture CFA models can successfully estimate parameters and detect heterogeneous factor loadings when the mean structure is completely invariant. Performance was generally better for the one-factor model with eight indicators relative to the two-factor model with eight indicators. In general, parameters that were truly noninvariant (factor

loadings and residual variances) were estimated most successfully while latent class proportions were least accurately estimated. Convergence and parameter coverage rates were low when there were only one or two heterogeneous factor loadings and all factor loadings (except referent loadings) were free to vary across classes. A uniform pattern of noninvariance also made it difficult to successfully estimate the mixture CFA model. The mixture CFA model was unsuccessful when classifying individuals into the correct latent class when the mean structure was invariant, except for when sample size and CFA-DIF were large. Entropy was not a good indicator of classification accuracy under any study condition in the presence of identical mean structures.

The true number of latent classes was used when estimating all models, which allowed for use of the likelihood-ratio statistic to conduct invariance testing. The likelihood-ratio test had adequate power when there was more heterogeneity across latent classes in the residual variances which allowed for the completely restricted model to be estimated successfully. In the test of one truly invariant factor loading, a model with all class-specific factor loadings was estimated and compared to a model with an individual constraint imposed on a noninvariant factor loading with the other factor loadings free to vary. Power in this test was low, especially when there were fewer truly noninvariant factor loadings. Combining results from all outcome measures, the invariant mean structure had the least negative consequences for successful estimation and invariance testing when total sample size was equal to 800, latent class proportions were equal, and CFA-DIF was 0.40.

Limitations of the research design

The design of the current study allowed for inferences about the performance of mixture CFA models in an extreme case under conditions that were thought to be important to applied researchers. Nonetheless, as with any study with simulated data, there are an infinite number of combinations of study conditions that could have been analyzed. The following is a discussion about some of the more substantial limitations of the current study.

Conditions in the current study were chosen in order to provide an analysis of the potential for using mixture CFA modeling to test factor loading heterogeneity as an alternative to multigroup CFA models. Given the current paucity of research on mixture CFA models for continuous data when the focus is on factor loading heterogeneity, the current study focused on an extreme case to offer preliminary analyses on the viability of testing for CFA-DIF across latent classes, with the expectation that the method would be used more frequently by applied researchers if it can be carried out similarly to multigroup CFA. The invariant mean structure only allowed for investigation of moderate to large sample sizes in the current study, with the smallest latent class comprised of 100 observations. Questions still remain about the performance of mixture CFA modeling for invariance testing of measurement model parameters when samples are small as well as the nature of the trade-off between mean separation and small sample sizes.

High rates of convergence to local maxima may have impacted the generalizability of inferences about parameter coverage and power in the current study. Specifically, using only replications that resulted in proper solutions may have

inflated parameter recovery percentages and power. In practice, nonconvergence may indicate that the model is misspecified or that the estimation algorithm is on the edge of the parameter space. Increasing the number of iterations or decreasing the change in stop criterion may result in convergence to a global solution in practice that is less accurate than the results of this study may suggest.

The current study was limited to one type of fit statistic for invariance testing, the χ^2 difference statistic. While popular in applications, there are conflicting results about its success in multigroup CFA, as described in Chapter 2. In addition, since the likelihood-ratio cannot be used to compare models with different numbers of latent classes, the results of this study cannot be generalized to exploratory modeling with mixture factor models.

Invariance tests were conducted under the assumption that the true number of latent classes was known. While this is a common assumption in simulation studies (e.g., Gagné, 2004; Lubke & Muthén, 2007) it ignores model selection problems common to mixture modeling. For example, Lubke and Neale (2006) showed that choosing the number of latent classes introduces the potential of equivalent models while Lubke and Muthén (2007) call for more research on choosing the number of classes in mixture CFA. Since true two -class models were estimated in the current study, there also was no possibility for the extraction of spurious latent classes, another potential problem that can occur when estimating latent class models.

A challenge when testing for CFA-DIF, whether groups are manifest or latent, is that problems can arise from setting restrictions on parameters for model identification. In CFA models, one loading for each factor is typically fixed to unity

to set the metric of the factors (Bollen, 1989). Since testing for nonuniform CFA-DIF involves the imposition and release of constraints on the factor loadings, the choice of setting the scale for model identification can potentially influence the results of CFA-DIF testing (e.g., Cheung & Rensvold, 1999, for manifest groups). The current study assumed a known correct referent variable, which is a typical assumption in simulation studies, while the choice of referent is not necessarily straightforward in empirical research. As such, results from the current study can be generalized only to situations in which the referent has been correctly identified. Misidentification of the referent is a type of model misspecification that can lead to incorrect parameter estimates or inferences from invariance tests.

Implications and recommendations

The results from the current study revealed that a total sample size of at least 800 with equal latent class proportions was adequate to derive accurate conclusions about parameter values and their cross-population equality when there were moderate to many noninvariant factor loadings. This suggests that any deviations of these characteristics towards fewer observations and less heterogeneity would lead to problems with convergence, accuracy of parameter estimates, and power to detect CFA-DIF when the mean structure is completely invariant. Problems with convergence and accuracy of estimates when there were only a few noninvariant factor loadings even with large sample sizes and large CFA-DIF suggests that, at least when the mean structure is invariant, mixture CFA cannot be used for exploratory analyses.

The lack of mean separation had a negative impact on inferences about individual posterior probabilities of latent class membership. The results suggest that the ability to cluster individuals based on a heterogeneous measurement model may depend more on the amount of separation in the mean structure than in variability caused by the covariance structure. As such, when impact does not exist and the substantive focus is on estimating and testing the measurement model parameters, the use of mixture CFA modeling to cluster individuals is tenuous.

When the number of classes is known or assumed to be known a priori, the likelihood-ratio test may be a viable option when it is expected that there are other sources of heterogeneity beyond the parameters being tested. When there is little or no heterogeneity in other parameters, then the likelihood-ratio test may not be a good choice, and researchers should instead consider estimating a model with fewer classes and using comparative fit indices such as AIC or sample size adjusted BIC that allow for model comparisons across different numbers of latent classes.

The low power rates when testing the invariance of one factor loading suggests that a backward approach to testing partial measurement invariance with mixture CFA may not be feasible when the mean structure is completely invariant. An alternative strategy for testing measurement invariance would be a forward approach in which a model with all factor loadings constrained equal across classes is compared to a model with individual or groups of factor loadings freely estimated across latent classes based on substantive theory. The choice of strategy may depend on the number of factor loadings hypothesized to vary across classes.

While mixture CFA has the potential to accurately model heterogeneous factor loadings and test for nonuniform CFA-DIF when the mean structure is invariant, many study conditions led to unsuccessful estimation in terms of convergence to the global solution, parameter recovery, and classification accuracy. In addition, success depended on conditions that are not able to be controlled through research design and data collection (e.g. characteristics of the factor loadings) as well as having moderate to large sample sizes, which may be hard to collect in practice. A potential way to improve convergence and accuracy when using mixture CFA models is to include observed covariates to explain latent class membership. For example, Lubke and Muthén (2007) showed improved convergence rates and percent correct assignment when covariates were included when modeling uniform CFA-DIF with mixture CFA and a noninvariant mean structure. While theory should dictate the inclusion of covariates, post hoc exploratory analyses can compare class-specific parameters across latent classes with respect to measured variables such as background characteristics (Lubke & Muthén, 2005). Both continuous and categorical covariates have been included in previous research on mixture CFA models (e.g. Lubke & Muthén, 2005, 2007). Covariates with theoretical underpinnings can be included in the model as predictors of latent class membership, as direct influences on the indicators, or as indirect influences on the indicators through the factor.

Together, the results of the current study suggest that the utility of mixture CFA modeling is questionable when substantive theory suggests that impact does not exist, unless latent class sizes are at least 400 or there is mixed pattern of large CFA-DIF in most of the factor loadings. It is expected that performance would improve

and requirements for sample size and magnitude of factor loading heterogeneity would decrease when testing for nonuniform CFA-DIF using mixture CFA models as the mean structure becomes more separated. Of these conditions, sample size is the only one that can be under the control of the researcher. Without further investigation of other strategies that may improve convergence and accuracy of parameter estimates, the mixture CFA model should only be used for measurement invariance testing when within class sample sizes are all at least 400.

Results from the simulation study offered evidence that mixture CFA models can be used for modeling and testing for heterogeneous factor loadings across latent classes and that the mean structure can be safely ignored when sample sizes are larger and the magnitude of heterogeneity is expected to be large in order to compensate for the potential lack of impact. That the mixture CFA model showed some success under an invariant mean structure points to the potential capabilities of mixture CFA modeling for invariance testing instead of multigroup CFA. More research would be needed to find strategies to improve the capabilities of the mixture CFA model for invariance testing in order for it to be used under a wider variety of situations found in practice, in particular because most study conditions that had a significant impact on successful estimation are not under a researcher's control. Nonetheless, the results provide some support for the consideration of latent classes, not just when important moderator variables are not included in the dataset or were not thought about during the data collection process, but to add theoretically important qualitative information in order to enrich inferences and improve parameter estimates.

Methodological extensions

The current study provided a preliminary evaluation of the viability of mixture CFA models for differential indicator functioning by studying the extreme case of a completely invariant mean structure. The findings showed a limited number of conditions, most of which are out of a researcher's control, that allowed for the success of the mixture CFA model to test for CFA-DIF without relying on separation in the mean structure. In order for mixture CFA modeling for invariance testing to have broad appeal, sample size requirements have to be much lower than 400 per class to accommodate a larger range of research designs since many applications in the social and behavioral sciences with continuous data have much smaller samples. As such, research on strategies to improve convergence and estimation when using mixture CFA to model CFA-DIF would help to make the method more useful to applied researchers.

Including observed covariates to predict latent class membership may help improve convergence and correct class assignment. Of particular interest would be the interaction of covariates, sample size, and amount of heterogeneity on the performance of the mixture CFA model for testing differential indicator functioning. It is expected that to the extent that observed covariates are good predictors of latent class membership, sample size requirements would decrease and the amount and size of heterogeneity needed would be smaller.

If convergence to the global maximum were to be improved for conditions with smaller sample sizes and less heterogeneity either through the addition of observed covariates or by using something other than maximum-likelihood estimation

with the EM algorithm that is available in Mplus 5.1 (Muthén & Muthén, 2007), then further analysis of accuracy and power would be valuable. In particular, evaluating the performance of other fit indices besides the likelihood-ratio statistic for invariance testing would be useful because model choice often involves different numbers of latent classes.

Multigroup CFA modeling is less complex than mixture CFA modeling in terms of estimation, model selection, and the inferences drawn from it. Multigroup CFA modeling is also currently available in more statistical software programs, making it a more desirable model to use than mixture CFA. If it can be established that mixture CFA modeling can be successful when testing for CFA-DIF under a wide variety of conditions, evaluating the extent to which the use of latent classes provides more accurate inferences than the use of manifest groups would help contribute to mainstream appeal of the model.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Ahmavaara, Y. (1954). The mathematical theory of factorial invariance under selection. *Psychometrika, 19,* 27-38.

Aitkin, M., & Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series D (Methodological), 47,* 67-75.

Allua, S., Stapleton, L., & Beretvas, S. (2008). Testing latent mean differences between observed and unobserved groups using multilevel factor mixture models. *Educational and Psychological Measurement, 68,* 357-378.

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika, 67,* 49-78.

Arminger, G., & Stein, P. (1997). Finite mixture of covariance structure models with regressors: loglikelihood function, distance estimation, fit indices, and a complex example. *Sociological Methods and Research, 26,* 148-182.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological Methods, 9,* 3-29.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bergan, J. R. (1983). Latent-class models in educational research. *Review of Research*

in Education, 10, 305-360.

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

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach.* Hoboken, NJ: John Wiley & Sons.

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195-212.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 464-504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95,* 1005-1018.

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

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence

response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Psychology, 31,* 187-212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9,* 233-255.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42,* 133-148.

Cowles, M. (2001). *Statistics in psychology: An historical perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.

Cudeck, R., & MacCallum, R. C. (Eds.). (2007). Factor analysis at 100: Historical developments and future directions. Mahwah, NJ: Erlbaum

Dayton, C. M. (1998). *Latent class scaling analysis.* Thousand Oaks, CA: Sage Publications.

Dempster, A. P., Laird, N.M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1-38.

Dolan, C. V., & van der Maas, H. L. J. (1998). Fitting multivariate normal finite

mixtures subject to structural equation modeling. *Psychometrika, 63*, 227-253.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95,* 134-135.

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

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68,* 940-958.

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

Finch, W. H., & French, B. F. (2008). Anomalous Type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68,* 743-759.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 378-402.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal, 15,* 96-113.

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25,* 9-28.

Gagné, P. E. (2004). General confirmatory factor mixture models: A tool for assessing factorial invariance across unspecified populations (Doctoral dissertation, University of Maryland, 2004). *Dissertation Abstracts International, 65,* 1389B.

Gagné, P. E. (2006). Mean and covariance structure mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 197-224). Greenwood, CT: Information Age Publishing, Inc.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Guttman, L. (1992). The irrelevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27,* 175-204.

Hancock, G. R. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 317-334). Thousand Oaks, CA: Sage Publications.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and manova in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 534-556.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 202-226.

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the

Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117-144.

Jedidi, K., Jagpal, H. S., & DeSarbo,W. S. (1997). STEMM: A general finite mixture structural equation model. *Journal of Classification, 14,* 23-50.

Jensen, A. R. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *The Behavioral and Brain Sciences, 8,* 193-263.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 57,* 409-426.

Keren, G.. & Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences: Statistical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement, 45,* 271-285.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton-Mifflin.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach.* West Sussex: John Wiley & Sons.

Li, H. –H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647-677.

Lindsay, B., Clogg, C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96-107.

Longford, N.T., & Muthén, B. (1992) Factor analysis for clustered observations. *Psychometrika, 57,* 581-597.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research, 36,* 299-324.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56,* 231-248.

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21-39.

Lubke, G. H., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 26-47.

Lubke, G. H., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41,* 499-532.

Lubke, G. H., & Neale, M. C. (2008). Distinguishing between latent classes and

continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research, 43,* 592-620.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84-99.

Magidson, J., &Vermunt, J.K. (2004) Latent class models. D. Kaplan (ed.), *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (Educational Testing Service Research Report No. RR-08-43). Princeton, NJ: Educational Testing Services

Maydeu-Olivares, A. & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11,* 344-362.

Mazor, K.M., Clauser, B.E., & Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443-451.

McClendon, M. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research, 20,* 60-103.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97-103.

McLachlan, G., & Krishnan, T. (2008). *The EM algorithm and extensions.* New York: John Wiley & Sons.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 611-635.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 11,* 60-72.

Meng, X. L., van Dyk, D. A. (1997). The EM algorithm -- an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological), 59,* 511-567.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 19,* 187–206.

Meredith,W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Millsap, R. E., & Kwok, O. –M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9,* 93-115.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–215.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41–71.

Moore, W. L. (1980). Levels of aggregation in conjoint analysis: An empirical comparison. *Journal of Marketing Research, 17,* 516-523.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Muthén, B.O. (1998-2004). Mplus Technical Appendices. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park, CA: Sage Publications.

Muthén, B.O. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1-26). Charlotte, NC: Information Age Publishing, Inc.

Muthén, B. O., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31,* 1050–1066.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407-419.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463–469.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide. Fifth Edition.* Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2007). Mplus version 5.1 [Computer software]. Los Angeles: Muthén & Muthén.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show non-uniforn DIF. *Applied Psychological Measurement, 20*, 257-274.

Neale, M.C., Boker, S.M., Xie, G., & Maes, H.H. (2002). Mx: Statistical Modeling [Computer software]. Richmond, VA: VCU, Department of Psychiatry.

Nylund, K., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 535-569.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A, 185*, 71-110.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517–529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multiple Behavioral Research, 13*, 403-418.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to

item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rubin, D. B., & Thayer, T. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47, 69-76.

Samuelson, K. (2005). Examining differential item functioning from a latent class perspective (Doctoral dissertation, University of Maryland, 2005). *Dissertation Abstracts International, 66,* 1734A.

Samuelsen, K. M. (2007). Examining differential item functioning from a latent mixture perspective. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 177-198). Charlotte, NC: Information Age Publishing.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222.

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

Sörbom, D. (1974). A general model for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229-239.

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

Steiger, J. H. (1989). EzPATH: Causal modeling. Evanston, IL: SYSTAT.

Thomson, G. H. (1939). *The factorial analysis of human ability*. London: University of London Press.

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

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317-342). Charlotte, NC: Information Age Publishing.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60,* 275-304.

Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139-158.

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*, 4-69.

Vermunt, J.K., & Magidson, J. (2000). *Latent GOLD's user's guide*. Boston: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2005). Structural equation models: Mixture models. In B. Everitt & D. Howell, (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1922-1927). Chichester, UK: John Wiley & Sons.

Webb, M-Y. L., Cohen, A., & Schwanenflugel, P. J. (2008). Latent class analysis of

differential item functioning on the Peabody Picture Vocabulary Test III. *Educational and Psychological Measurement 68,* 335-351.

Wedel, M. E., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced Methods of Marketing Research* (pp. 352-388). Cambridge, MA: Blackwell Publishing.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89,* 696-716.

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 587-614.

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

Yung, Y. F. (1997). Finite mixtures in confirmatory factor analysis models. *Psychometrika, 62,* 297-330.