MODEL IMPLIED INSTRUMENTAL VARIABLE ESTIMATION FOR MULTILEVEL CONFIRMATORY FACTOR ANALYSIS

Michael L. Giordano

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in Psychology and Neuroscience in the College of Arts and Sciences.

Chapel Hill
2018

Approved by:

Kenneth A. Bollen

Daniel J. Bauer

Kathleen M. Gates

ABSTRACT

Michael L. Giordano: Model Implied Instrumental Variable Estimation for Multilevel Confirmatory Factor Analysis
(Under the direction of Kenneth Bollen)

Multilevel Confirmatory Factor Analysis (MCFA) models are most commonly estimated with full information maximum likelihood (FIML). FIML is asymptotically efficient and asymptotically unbiased given correct model specification and no excessive multivariate kurtosis. When these assumptions are violated, we have no guarantee about the asymptotic properties of FIML. In single level SEMs, the Model Implied Instrument Variable (MIIV-2SLS) estimator has been shown to be an excellent alternative to maximum likelihood. Following prior work for single level SEMs, this paper develops two MIIV-2SLS estimators for MCFA models. I evaluate both estimators in comparison to FIML with a Monte Carlo simulation study varying number of clusters, cluster size, distribution of data, balance of clusters and correct versus incorrect model specifications. Results suggest that both MIIV estimators are good alternatives to FIML across a variety of conditions. Most importantly, they are more robust to model misspecification and offer local tests of fit with Sargan's test. The primary limitation found in this simulation study suggests that these estimators may underestimate standard errors given small number of clusters, unbalanced clusters, and skew/kurtosis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Multilevel confirmatory factor analysis (MCFA) has grown in popularity over the past several decades. Leveraging the strengths of multilevel modeling and confirmatory factor analysis, MCFAs allow for the simultaneous modeling of measurement error and hierarchically clustered data. In common applications, standard confirmatory factor analysis accounts for measurement error, but is not typically used to account for nested or clustered data. Standard multilevel models, on the other hand, are used for analyzing clustered data, but not typically used to simultaneously account for measurement error.[1] Many studies have shown the deep connections between multilevel models and latent variable models, often illustrating equivalence between the methods (Bauer, 2003; Curran, 2003; Kamata, Bauer, & Miyazaki, 2008; Mehta & Neale, 2005; Meredith & Tisak, 1990).

Given the overlap between latent variable models and MLM, modern research in psychometrics has sought to unify latent variable models with multilevel models (Asparouhov & Muthén, 2003; Rabe-Hesketh, Skrondal, & Pickles, 2004; Rabe-Hesketh, Skrondal, & Zheng, 2007). The developments in MCFAs have led to approaches for modeling clustered data with measurement error and heterogeneous factor structures at the within and between levels (Asparouhov & Muthén, 2003, 2007; Bentler & Liang, 2003; Goldstein, 1987; Goldstein & McDonald, 1988; Kim, Dedrick, Cao, & Ferron, 2016; Longford & Muthén, 1992; McDonald, 1994; McDonald & Goldstein, 1989; Muthén, 1989, 1990, 1994, Rabe-Hesketh et al., 2004,

[1] While it is not as common to use measurement models in most multilevel software packages, some software such as HLM and GLLAMM do allow for this (Rabe-Hesketh, Skrondal, & Pickles, 2004; S. Raudenbush, Bryk, Cheong, Congdon, & Du Toit, 2011).

2007, p. 207; Skrondal & Rabe-Hesketh, 2004). Of course, latent variable models are diverse. Modern multilevel latent variable models could include multilevel IRT, multilevel factor analysis, and multilevel structural equation modeling. For this study I will focus on multilevel factor analysis specifically, but it is my hope that future work could expand on this to include more general multilevel structural equation models.

First, it is important to establish why MCFA is needed. One of the assumptions made in CFA (or SEM more broadly), is that observations are independent and identically distributed. When data are clustered we violate this assumption. Given clustered measurement data, we might be tempted to ignore clustering, especially if we are not interested in all levels in the measurement model. However, this would be naïve. Julian (2001) showed that ignoring clustering in CFA results in biased loadings, unique variances, factor variances. In the more full SEM context, du Toit & Toit (2008) showed that ignoring clustering resulted in incorrect parameter estimates, standard errors and inappropriate fit statistics. In short, if data are clustered then researchers need to consider multilevel analyses (or other corrections).

The interpretation of MCFA models is important to consider, and best done through example. A nice illustration comes from Van Peet (1992), who analyzed six measures of intelligence of children nested within families. At the individual level the six intelligence measures correspond to numeric intelligence and perceptual intelligence (i.e., two correlated latent factors for intelligence). At the family level, intelligence corresponds to a single intelligence factor. In other words, at the within groups level (the individual level), we can imagine two factors of intelligence. At the group level, it makes less sense to imagine two factors because this level corresponds to average family intelligence. This last example also highlights one of the more unique features of MCFAs, which is the option to model different factor structures for within-groups and between-groups.

Given the relative complexity of these models, MCFA estimation is an area of ongoing methodological development. Before general maximum likelihood methods were available, early estimation approaches relied on decomposing variance-covariance matrices at different levels and fitting factor models with traditional SEM software, using the variance covariance matrices in place of raw data. These early methods have been called pseudo-maximum likelihood methods or ‘ad-hoc’ methods (Goldstein, 1987; McDonald & Goldstein, 1989; Muthén & Satorra, 1989). Today, full information maximum likelihood (FIML) estimation routines are available through Mplus, Stata’s GLLAMM routine, the ‘OpenMx’ package in R, and the ‘xxM’ package in R (Asparouhov & Muthén, 2003; Mehta, 2013; Muthén & Muthén, 2015; Neale et al., 2016; Rabe-Hesketh et al., 2004; StataCorp, 2015) [2].

Unsurprisingly, FIML estimation has become the de facto estimation method in most applications. Maximum likelihood offers the usual properties of being asymptotically efficient and asymptotically unbiased—excellent qualities to have. However, these properties require strong assumptions such as correct model specification and no excessive multivariate kurtosis. If these assumptions are violated, there is no guarantee about the asymptotic efficiency or asymptotic unbiasedness of the ML estimator. Additionally, convergence becomes an issue as models grow in complexity. In single level SEMs, convergence has been shown to be a barrier in many conditions such as small sample sizes, having only two indicators, and using mixture models (Anderson & Gerbing, 1984; Boomsma, 1982; Curran, Bollen, Paxton, Kirby, & Chen, 2002; Longford N. T. & Muthén, 1992; Yuan & Hayashi, 2005). Additionally, convergence can be problematic in multilevel contexts generally (Hox, 2010). It stands to reason to suspect that

[2] This list may not be all inclusive. These are simply a handful of the software routines I have come across.

convergence would also be a potential barrier in MCFA (Preacher, Zhang, & Zyphur, 2015; Preacher, Zyphur, & Zhang, 2010; Ryu & West, 2009).

Given the potential shortcomings of FIML for MCFA, I am proposing to look to the class of Two-Stage Least Squares (2SLS) instrumental variable estimators which have been shown to be effective estimators in single-level SEMs (Bollen, 2001; Bollen, Kirby, Curran, Paxton, & Chen, 2007). Bollen (1996b) proposed the Model Implied Instrumental Variable Two-Stage Least Squares (MIIV-2SLS) estimator for single level SEM’s. Under ideal conditions for ML, simulation studies found similar efficiency for MIIV estimators. It is more robust to model misspecification, it requires fewer distributional assumptions and it is computationally efficient. Also because it is non-iterative it is not subject to non-convergence issues. For these reasons, I believe that the MIIV-2SLS can be a useful estimator in MSEM’s.

In this study I will (1) briefly review the literature on MCFA/MSEM estimation, (2) review the literature on MIIV-2SLS and provide a simple example of its use for a single level SEM model, (3) develop two novel procedures for estimating MCFAs with MIIV-2SLS and (4) demonstrate the qualities of these procedures with a Monte Carlo simulation study varying several factors.

Chapter 1: MIIV-2SLS estimation for MCFA

Multilevel Latent Variable Models

The earliest work on multilevel latent variable models was in an unpublished dissertation by Schmidt (1969). Schmidt studied multivariate random effects and provided a computer program for maximum likelihood estimation—though this early work did not consider group level structures or variables, instead focusing on what I will call the ‘within groups’ model. Most MSEM applications specify separate within-groups and between-group structures and Schmidt’s dissertation did not deal with separate between and within components. Instead it dealt with a special case where the between and within components were equivalent. Aside from this very early work on multilevel models with latent variables, McDonald, Goldstein, Muthén, Rabe-Hesketh have been very active pushing forward the methodology for these models (Goldstein, 1987; Goldstein & McDonald, 1988; McDonald & Goldstein, 1989; Muthén, 1990; Rabe-Hesketh et al., 2007). See Hox (2010) for an excellent overview of both early and more modern approaches. I will review some of the notable techniques, focusing on the measurement model.

The particular multilevel factor analysis model that I will examine is the random intercepts MCFA. The expression for the random intercepts MCFA model is given by:

$$\begin{aligned} \boldsymbol{Y_{ij}} &= \boldsymbol{Y_{W_{ij}}} + \boldsymbol{Y_{B_j}} \\ \boldsymbol{Y_{ij}} &= \boldsymbol{\alpha} + (\boldsymbol{\Lambda_W L_{W_{ij}}} + \boldsymbol{\varepsilon_{W_{ij}}}) + (\boldsymbol{\Lambda_B L_{B_j}} + \boldsymbol{\varepsilon_{B_j}}) \end{aligned} \tag{1}$$

Where we make the following distributional assumptions

$$\begin{gathered}\boldsymbol{L}_{\boldsymbol{B}_j} \sim \boldsymbol{N}(\boldsymbol{0},\ \boldsymbol{\Phi}_{\boldsymbol{B}}) \\ \boldsymbol{L}_{\boldsymbol{W}_{ij}} \sim \boldsymbol{N}(\boldsymbol{0},\ \boldsymbol{\Phi}_{\boldsymbol{W}}) \\ \boldsymbol{\varepsilon}_{\boldsymbol{B}_j} \sim \boldsymbol{N}(\boldsymbol{0},\ \boldsymbol{\Theta}_{\boldsymbol{B}}) \\ \boldsymbol{\varepsilon}_{\boldsymbol{W}_{ij}} \sim \boldsymbol{N}(\boldsymbol{0},\ \boldsymbol{\Theta}_{\boldsymbol{W}})\end{gathered} \tag{2}$$

In this model α is a vector of measurement intercepts, $\boldsymbol{\Lambda}_{\boldsymbol{W}}$ is a vector of within groups factor loadings, $\boldsymbol{L}_{\boldsymbol{W}_{\boldsymbol{ij}}}$ is the vector of within groups latent variables, $\boldsymbol{\Lambda}_{\boldsymbol{B}}$ is a vector of between groups factor loadings, $\boldsymbol{L}_{\boldsymbol{B}_{\boldsymbol{j}}}$ is the vector of between groups latent variables, and $\boldsymbol{\varepsilon}_{\boldsymbol{W}_{\boldsymbol{ij}}}/\boldsymbol{\varepsilon}_{\boldsymbol{B}_{\boldsymbol{j}}}$ are vectors of within and between residuals. We can derive the model implied covariance matrix as

$$\begin{aligned} Cov(\boldsymbol{Y}_{\boldsymbol{ij}}) &= Cov\left(\boldsymbol{Y}_{\boldsymbol{b}_j} + \boldsymbol{Y}_{\boldsymbol{w}_{ij}}\right) \\ &= Cov\left(\boldsymbol{Y}_{\boldsymbol{b}_j}\right) + Cov\left(\boldsymbol{Y}_{\boldsymbol{w}_{ij}}\right) \\ &= \boldsymbol{\Sigma}_{\boldsymbol{B}} + \boldsymbol{\Sigma}_{\boldsymbol{W}} \\ &= \ \boldsymbol{\Lambda}_{\boldsymbol{B}}\boldsymbol{\Phi}_{\boldsymbol{B}}\boldsymbol{\Lambda}'_{\boldsymbol{B}} + \boldsymbol{\Theta}_{\boldsymbol{B}} + \boldsymbol{\Lambda}_{\boldsymbol{W}}\boldsymbol{\Phi}_{\boldsymbol{W}}\boldsymbol{\Lambda}'_{\boldsymbol{W}} + \boldsymbol{\Theta}_{\boldsymbol{W}} \end{aligned} \tag{3}$$

where $Cov\left(\boldsymbol{Y}_{\boldsymbol{b}_j} + \boldsymbol{Y}_{\boldsymbol{w}_{ij}}\right) = \ Cov\left(\boldsymbol{Y}_{\boldsymbol{b}_j}\right) + Cov\left(\boldsymbol{Y}_{\boldsymbol{w}_{ij}}\right)$ based on the assumption that the within and between models are orthogonal.

As I have already mentioned the most ubiquitous method for estimating MSEMs is FIML. Muthén (1990) derived the ML fitting function as

$$\mathrm{F}_{\mathrm{ML}} = \ \sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} (\boldsymbol{n_j} - \boldsymbol{1})\{\mathbf{log}|\boldsymbol{\Sigma}_{\boldsymbol{W}}(\boldsymbol{\theta})| + \boldsymbol{tr}(\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-\boldsymbol{1}}(\boldsymbol{\theta})\boldsymbol{S}_{\boldsymbol{yWj}})\} + \\ \textstyle\sum_{\mathbf{j=1}}^{\mathbf{J}}\{\mathbf{log}|\boldsymbol{\Sigma}_{\mathbf{gj}}(\boldsymbol{\theta})| + \mathbf{tr}(\boldsymbol{\Sigma}_{\mathbf{gj}}^{-\mathbf{1}}(\boldsymbol{\theta})\mathbf{S}_{\mathbf{gj}})\} \tag{4}$$

ML estimation has many excellent properties, including that the estimator is asymptotically unbiased and asymptotically efficient. However, in order for these qualities to hold we have to assume correct model specification and multivariate normality (alternatively, no excess multivariate kurtosis). If these qualities do not hold then we have no guarantee about the asymptotic properties of our estimator.

FIML estimation is also available via the generalized linear latent and mixed models (GLLAMM) approach developed by Skrondal & Rabe-Hesketh (2004). This modeling approach can be used to estimate multilevel latent variable models with full maximum likelihood. The FIML solution available in the Mplus uses a structural equation modeling approach, while the GLLAMM framework is more akin to traditional multilevel models setup. GLLAMM has some nice properties in that it can handle a larger number of levels, missing data, unbalanced designs and flexible factor structures (FIML can handle many of these as well). As with other ML approaches GLLAMM relies on correct model specification and distributional assumptions for significance tests. Additionally, MSEM models tend to take a long time to estimate in GLLAMM, limiting their practicality.

Before FIML was readily available there were other methods for estimating MSEMs including methods proposed by Goldstein (1987), McDonald & Goldstein (1989), and Muthén (1989, 1990), among others. These earlier methods have been called "ad hoc" methods or two-step approaches. FIML relies on simultaneously fitting and estimating within-groups and between-groups models. These ad-hoc approaches instead rely on (1) decomposing variances into separate within-groups and between-groups covariance matrices and (2) fitting these matrices either separately or together in a factor model with maximum likelihood. Though several approaches have been proposed, I will focus on the Goldstein method and the two step Muthén method (often referred to as MUML).

The idea of decomposing variances into different levels is nothing new (Cornfield & Tukey, 1956; Searle, Casella, & McCulloch, 1992). We can imagine decomposing an individual observed variable into parts as,

$$Y_{ij} = Y_{ijW} + Y_{ijB} \tag{5}$$

Where $Y_{ijW}$is varying at the individual level and $Y_{ijB}$is varying at the group level. In other words we can decompose variation in observed variables into their group means and group mean deviations. Another way of showing this is

$$Y_{ij} = \left(Y_{ij} - \bar{Y}_{.j}\right) + \bar{Y}_{.j} \tag{6}$$

where $\left(Y_{ij} - \bar{Y}_{.j}\right)$ is group mean centered and $\bar{Y}_{.j}$ represents group means. Importantly, Searle, Casella, & McCulloch (1992) showed these components are additive and orthogonal.

Similar to the single observed variable example above, early MSEM approaches relied on separating the population covariance matrices into their within-groups and between-groups components.

$$\boldsymbol{\Sigma_T} = \boldsymbol{\Sigma_W} + \boldsymbol{\Sigma_B}\ ^{3} \tag{7}$$

where $\boldsymbol{\Sigma_T}$ is the total covariance, $\boldsymbol{\Sigma_W}$ is the within-groups covariance matrix, and $\boldsymbol{\Sigma_B}$is the between-groups covariance matrix. This was shown in Equation 3, where we separated the covariances into their constituent parts based on the assumption of orthogonality. Goldstein's approach and Muthén's approach differ in how they decompose the covariance components.

Goldstein (1987, 1995) proposed using a three-level multilevel model to decompose the within-groups and between-groups covariance. The general two-step procedure is (1) estimate within-groups and between-groups covariance matrices via a univariate multilevel model and

[3] See, for example, McDonald & Goldstein (1989), p. 217-218 for the proof of this claim.

then (2) estimate MSEM with covariance matrices using ML and standard SEM software. Hox & Maas (2004) provide a nice description for fitting these models.

As an example of the Goldstein procedure, consider a simple MCFA model with three observed indicators. To obtain estimates of the within-groups and between-groups covariance matrices we specify a three-level univariate multilevel model, which I describe in more detail below. Thinking about this model hierarchically, level-1 could represent indicators, level-2 could represent individuals, and level-3 represents clusters of individuals. Alternatively, this could be a longitudinal study with indicators nested within individuals nested within time. The level-1 model is given by,

$$Y_{hij} = \pi_{1ij}D_{1ij} + \pi_{2ij}D_{2ij} + \pi_{3\mathrm{ij}}D_{3ij} \tag{8}$$

where *D* are dummy codes for indicators, *h* indexes indicators, *i* indexes individuals, and *j* indexes clusters. Also, note that we have removed the intercept from the model, in order to estimate the random component of each indicator instead of having one as a reference. Additionally, we are not estimating a level-1 error term and instead this is being fixed at zero. [4] Moving up to the second level, the individual level, we re-define all π_{pij} terms allowing them to randomly vary across individuals. Level-2 equations are then,

$$\begin{aligned} \pi_{1ij} &= \beta_{1\mathrm{j}} + u_{1ij} \\ \pi_{2ij} &= \beta_{2\mathrm{j}} + u_{2ij} \\ \pi_{3ij} &= \beta_{3\mathrm{j}} + u_{3ij} \end{aligned} \tag{9}$$

[4] The omission of the intercept and fixing level-1 error is a simple ‘trick’ for using a univariate multilevel model to estimate a multivariate outcome. The same strategy is used in a variety of models one example being David Kenny’s Actor Partner Interdependence Model (Hox, 2010; Kenny, Kashy, & Cook, 2006).

where β$_{\mathrm{pj}}$ is the average of item *p* varying over clusters j and u_{pij} captures the random variation in item *p* across individuals within clusters. Moving up again we redefine all β_{pj} terms at level-3 as,

$$\begin{aligned} \beta_{1j} &= \gamma_1 + u_{1j} \\ \beta_{2j} &= \gamma_2 + u_{2j} \\ \beta_{3j} &= \gamma_3 + u_{3j} \end{aligned} \tag{10}$$

where γ_p is the grand mean for item *p* and u_{pj} captures the random variation in item *p* across clusters. Applying simple substitution we combine all three levels obtaining the reduced form equation as,

$$\begin{aligned} Y_{hij} = {} & \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + u_{1ij} D_1 + u_{2ij} D_2 + u_{3ij} D_3 + u_{1j} D_1 + u_{2j} D_2 \\ & + u_{3j} D_3 \end{aligned} \tag{11}$$

Importantly, we assume that the level-2 random effects are distributed as

$$\begin{bmatrix} u_{1ij} \\ u_{2ij} \\ u_{3ij} \end{bmatrix} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) \tag{12}$$

where,

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{Var(u_{1ij})} & \boldsymbol{Cov(u_{2ij}, u_{1ij})} & \boldsymbol{Cov(u_{3ij}, u_{1ij})} \\ \boldsymbol{Cov(u_{2ij}, u_{1ij})} & \boldsymbol{Var(u_{2ij})} & \boldsymbol{Cov(u_{3ij}, u_{2ij})} \\ \boldsymbol{Cov(u_{3ij}, u_{1ij})} & \boldsymbol{Cov(u_{3ij}, u_{2ij})} & \boldsymbol{Var(u_{3ij})} \end{bmatrix} \tag{13}$$

Similarly, we assume that the level-3 random effects are distributed as,

$$\begin{bmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{B}}) \tag{14}$$

where,

$$\boldsymbol{\Sigma}_{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{Var(u_{1j})} & \boldsymbol{Cov(u_{2j}, u_{1j})} & \boldsymbol{Cov(u_{3j}, u_{1j})} \\ \boldsymbol{Cov(u_{2j}, u_{1j})} & \boldsymbol{Var(u_{2j})} & \boldsymbol{Cov(u_{3j}, u_{2j})} \\ \boldsymbol{Cov(u_{3j}, u_{1j})} & \boldsymbol{Cov(u_{3j}, u_{2j})} & \boldsymbol{Var(u_{3j})} \end{bmatrix} \tag{15}$$

In summary, as a part of this model we estimate the mean responses for each indicator (given by γ_p's) and the estimated variances and covariances of the random effects. Here we are mostly interested in the variances and covariances of random effects. These variances and covariances are direct estimates of within group covariances, $\boldsymbol{\Sigma_W}$, and the between group covariances, $\boldsymbol{\Sigma_B}$, which are used in the second step. With estimates of the within-groups and between-groups variance-covariance matrices, step 2 of the Goldstein approach is to fit within-groups and between-groups structural model using ML and $\boldsymbol{\Sigma_W}$ and $\boldsymbol{\Sigma_B}$ as data.

Here we have the additional choice of using ML or REML to estimate the multilevel models. In general, REML produces less biased estimates of the random components, especially at smaller sample sizes (Snijders & Bosker, 1999). ML has a tendency to negatively bias random components, though this bias shrinks as number of clusters approach infinity. ML also has a smaller MSE when considering variance components. Thus, there is a definite trade-off between bias and variance when considering estimation of the variance components.

This Goldstein approach has several advantages for decomposing covariances. The estimated covariances can be used with standard SEM software separately or in a two group model (Goldstein, 1987; Hox & Maas, 2004). Since the within and between models are estimated separately, we can obtain separate estimates and fit indices at each level. Additionally, given the way the covariances are computed we can accommodate missing data easily and we can adjust for dichotomous or categorical variables easily by fitting generalized multilevel models instead. Of course there are some disadvantages as well. One of the primary disadvantages to this approach is that the covariances are estimated values and not directly calculated. This means they are estimated with uncertainty, and our second step of using covariances in covariance analysis we treat the covariances as population values rather than sample values. Additionally, it may be difficult to scale up to more levels. Three-level multilevel models with several random effects is already computationally difficult to estimate. As more levels and more indicators are added this approach could become computationally intractable fairly quickly.[5]

Muthén, (1989, 1990) proposed another two-step solution to estimate MSEMs; Muthén's method is often referred to as the pseudo-balanced or MUML (Muthén's ML) approach (McDonald, 1994; Muthén, 1994). He showed an unbiased estimate of $\boldsymbol{\Sigma_W}$, can be computed from the pooled within group covariance matrix

$$\boldsymbol{S_{pw}} = \frac{\sum_j^G \sum_i^n (\boldsymbol{Y_{ij}} - \boldsymbol{\bar{Y}_j})(\boldsymbol{Y_{ij}} - \boldsymbol{\bar{Y}_j})'}{N - G} \tag{16}$$

and the between group covariance matrices can be estimated by

$$\boldsymbol{S_B} = \frac{\sum_j^G n\,(\boldsymbol{\bar{Y}} - \boldsymbol{\bar{Y}_j})(\boldsymbol{\bar{Y}} - \boldsymbol{\bar{Y}_j})'}{G - 1} \tag{17}$$

Muthén showed that $\boldsymbol{S_{pw}}$ is an asymptotic unbiased estimator of $\boldsymbol{\Sigma_W}$ and $\boldsymbol{S_B}$ is an asymptotic unbiased estimator of $\boldsymbol{\Sigma_B}$ *assuming groups have equal sizes*.

Once the variance has been decomposed, the two covariance matrices are used in multiple group model with ML as the estimator. Muthén (1990) showed how these two covariance matrices could be modeled with existing SEM software. This technique relies partly on having balanced group sizes. Accounting for unbalanced groups in this setup is burdensome, and Muthén even proposed ignoring unbalance and using the average group size when data are not balanced (1989, 1990). Several studies have confirmed that the within portion of the model is consistently estimated using the pooled within covariance matrix and this holds across many conditions including balanced and unbalanced group sizes (Hox & Maas, 2001; Hox, Maas, &

[5] With more levels and more random components, moving into a Bayesian framework is likely to make these models more tractable--although that adds a different layer of complexity with priors and posterior distributions (Gelman & Hill, 2006).

Brinkhuis, 2010; Yuan & Hayashi, 2005). For the between part of the model, residual variances tend to be underestimated and standard errors too small leading to alpha levels of close to 8%, instead of the expected 5%. In addition unbalanced groups do not lead to biased coefficients but they do lead to underestimated standard errors (Hox & Maas, 2001).

Each method—FIML, the Goldstein approach, and MUML—has advantages and disadvantages. While FIML is the gold standard at the moment I will pull from the Goldstein approach and the MUML approach to develop a MIIV-2SLS estimator for MCFAs. In order to do that I will first review MIIV-2SLS.

MIIV-2SLS

The model implied instrumental variable two-stage least squares (MIIV-2SLS) estimator was first proposed by (Bollen, 1996b) and has been further developed for many applications since (Bollen et al., 2007; Bollen, Kolenikov, & Bauldry, 2014; Bollen & Maydeu-Olivares, 2007). The general family of Two-Stage Least Squares (2SLS or sometimes TSLS) estimators are not new; they have been in use for decades, in many different forms. 2SLS estimators have mostly been used in simultaneous equations models without latent variables. The family of 2SLS estimators for estimating coefficients of a single equation in a simultaneous equations model, were independently introduced by Theil (1953b, 1953a, 1954, 1961), Basmann (1957), and Sargan (1958) (Anderson, 2005). Though less cited, Anderson & Rubin (1950) also played an important role by deriving the asymptotic distribution for the limited information maximum likelihood (LIML) estimator for simultaneous equation models, which was used in all of the aforementioned works (Anderson, 2005). For latent variable models, Hägglund (1982) applied a type of the 2SLS estimator for exploratory factor analysis. Later Jöreskog & Sörbom, (1987) incorporated a different 2SLS estimator into factor analysis estimation. These early applications of 2SLS estimators were applied to the measurement model only and assumed no correlated errors. One exception was Jöreskog & Sörbom (1993) who applied Hägglund's method to estimate the measurement model and generated the covariance matrix of latent variables. They then estimated the latent variable model by using the covariance matrices as input for a 2SLS procedure. This method still assumed no correlated errors and did not provide significance tests.

Bollen (1996b, 1996a) proposed the MIIV-2SLS estimator which is another instrumental variable estimator for latent variable models and part of the 2SLS family of estimators, though it differs from earlier ones. In contrast to earlier 2SLS methods for latent variable models, Bollen (1996b) showed how the MIIV-2SLS estimator could be applied to estimate full SEM's with latent variables in an equation by equation basis, while also deriving standard errors and significance tests. Bollen (2001) further developed MIIV-2SLS noting that there is a closed form to estimate the full model simultaneously instead of equation by equation. This form provides identical point estimates and standard errors. Further, Bollen (2001) provides general analytic condition for when MIIV-2SLS is robust to misspecification.

The MIIV-2SLS estimator differs from traditional instrumental variable techniques in several important ways. Early versions of a 2SLS estimator for SEM's with latent variables, did not allow for correlated errors of measurement; the MIIV-2SLS estimator does allow correlated errors. Jöreskog and Sörbom's versions of the 2SLS estimator required estimating the measurement model and then estimating the latent variable model; the MIIV-2SLS estimator does not require this. Finally, MIIV-2SLS offers asymptotic standard errors for all coefficients in the model which previous versions did not.

Since its first proposal, simulation and analytic work has illustrated its effectiveness and extended it in important ways. In a large Monte Carlo simulation study, Bollen et al., (2007) compared bias and efficiency of ML and MIIV-2SLS in conditions of correct model specification and incorrect model specification; they concluded that in correct model specification, results are similar and given misspecification MIIV-2SLS performed better. Cragg (1968) noted the same quality for 2SLS for simultaneous equations as compared to FIML. Bollen (1996b) developed an algorithm to automatically select model-implied instruments; Bollen & Bauer, (2004) later programmed this is SAS and Bauldry, (2014) programmed a similar procedure for general use in Stata, both of which made the process of using MIIV-2SLS more accessible. Bollen & Maydeu-Olivares, (2007) developed a polychoric instrumental variable estimator which can be used for models with categorical indicators. Bollen (1996b), proposed the use of Sargan's test with MIIV-2SLS and Kirby & Bollen (2009) further developed their use and evaluated their performance for systematically investigating model misspecification. Bollen et al. (2014) developed generalized method of moments MIIV estimator (MIIV-GMM). Finally, Nestler (2014a, 2014b) explored estimating growth models with MIIV-2SLS.

Though I have mentioned several of the beneficial qualities of MIIV-2SLS I think it is worth briefly revisiting each again. The MIIV-2SLS estimator is distribution free, that is, the asymptotic standard errors can be used in significance testing even when observed variables are non-normal or have excessive multivariate kurtosis (Bollen, 1996b). As comparison, Full-information maximum likelihood requires the assumptions of normality or no excessive kurtosis, though there are corrections for nonnormality (ex., Satorra & Bentler, 1994). MIIV-2SLS estimators are more robust to misspecification. Here I am referring to structural misspecification; and it does not mean we have full robustness, but that model misspecification is less likely to spread throughout the model than full information estimators. For example, FIML uses information from the full system to estimate all parameters. This means that if one part of the model is misspecified, this information might influence estimates throughout, and can spread bias. MIIV-2SLS might have bias in any given misspecified equation, but that bias is more isolated and is less likely to spread to other parts of the model. Finally, though this has not been discussed at length in the literature, it should be noted that MIIV-2SLS is computationally efficient. It does not require an iterative procedure and instead has a closed form solution. As models become more complex this property may prove more and more advantageous.

Although the MIIV-2SLS estimator has been around for over two decades now, its general use is not widespread. This is likely due to the fact that it has not been readily available in software, until very recently. Fisher, Bollen, Gates, & Rönkkö (2016) have programmed MIIV-2SLS estimation into the "MIIVsem" package in R. Given that its use is not familiar practice to all, I will provide a simple example of estimating a single level CFA model with MIIV-2SLS. Additionally, this will further facilitate the development of the MIIV-2SLS estimator for MSEMs.

Suppose we have a simple two dimensional confirmatory factor analysis with eight observed indicators and two latent variables. The general latent variable model is given by

$$\boldsymbol{Y_i} = \boldsymbol{\alpha} + \boldsymbol{\Lambda L_i} + \boldsymbol{\varepsilon_i} \tag{18}$$

And with the full matrices given by

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \end{bmatrix} + \begin{bmatrix} \Lambda_{11} & 0 \\ \Lambda_{21} & 0 \\ \Lambda_{31} & 0 \\ \Lambda_{41} & 0 \\ 0 & \Lambda_{52} \\ 0 & \Lambda_{62} \\ 0 & \Lambda_{72} \\ 0 & \Lambda_{82} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix} \tag{19}$$

L_1 is measured by $Y_1 - Y_4$ and L_2 is measured by $Y_5 - Y_8$. In order to identify the model and scale the latent variable we set $\alpha_1 = \alpha_5 = 0$ and $\Lambda_{11} = \Lambda_{52} = 1$. To apply the MIIV-2SLS estimator to this model we will go through two steps. (1) Is to transform the model equations from latent variable equations to observed variable equations. And (2) is to identify MIIV's and perform instrumental variable regression.

To re-express the model equations as observed variable equations, we algebraically manipulate the scaling indicators.

$$\begin{gathered} Y_1 = L_1 + \varepsilon_1 \\ L_1 = Y_1 - \varepsilon_1 \end{gathered} \tag{20}$$

And we perform a similar transformation with L_2 andY_5. With the latent variables expressed as observed variables, we simply substitute in the observed variable expression for every L_1 and L_2. For Y_2 we start with

$$Y_2 = \alpha_2 + \Lambda_{21} L_1 + \varepsilon_2 \tag{21}$$

And after substituting the observed variable expression of L_1, we have

$$Y_2 = \alpha_2 + \Lambda_{21}(Y_1 - \varepsilon_1) + \varepsilon_2 \tag{22}$$

Finally, to simplify the expression we distribute Λ_{21} and create a composite error term to get

$$Y_2 = \alpha_2 + \Lambda_{21} Y_1 + u_1 \tag{23}$$

which is an observed variable equation which suggests a simple regression model. However, we cannot use OLS regression because we would violate the typical assumption that the explanatory variables are uncorrelated with the composite error. Instead we can use instrumental variable regression to estimate the equation.

A valid instrument needs to be: correlated with the offending explanatory variable, uncorrelated with the equation disturbance term, and the rank of the covariance matrix of Z and X (where Z contains all instrumental variables and X contains all X variables) is equal to the total number of X variables (Bollen, 1996b, 2012). Given the structural nature of the equations in

CFA and SEM it is worth noting that instruments vary from one equation to the next. Identifying MIIV's is an important step in this procedure and as it has been mentioned there have been several developments for automating the task in a variety of statistical software environments (Bauldry, 2014; Bollen & Bauer, 2004; Fisher et al., 2016). Once we have identified our MIIV's we perform instrumental variable regression, equation by equation. For the Y_2 equation, we determine that $Y_3 - Y_8$ are possible instruments. We regress Y_1 onto instruments to form $\hat{Y}_1$. Finally, we regress Y_2 on $\hat{Y}_1$to obtain estimates of α_2 and Λ_{21}. The same steps can be performed with all equations in the model.

While the two-step process is a useful way to think of performing MIIV-2SLS, it can also be done in a one-step procedure using the following expression

$$\hat{\boldsymbol{A}} = (\hat{\boldsymbol{Z}}\,'\,\hat{\boldsymbol{Z}})^{-\mathbf{1}}\hat{\boldsymbol{Z}}\,'\,\boldsymbol{Y} \tag{24}$$

where

$$\hat{\boldsymbol{Z}} = \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-\mathbf{1}}\boldsymbol{V}'\boldsymbol{Z} \tag{25}$$

And $\boldsymbol{Z}$ is a block diagonal matrix of indicators, $\boldsymbol{V}$ is a block diagonal matrix of instrumental variables corresponding to various model equations. The derivations of this single equation form are developed and explained more fully in (Bollen, 2001).

MIIV-2SLS Estimator for MCFAs

In order to apply MIIV-2SLS to MSEM's I borrow from the pseudo-maximum likelihood MSEM methods which separated the within groups and between groups covariance matrices. The Goldstein approach and the MUML approach were both two-step methods which broadly involved (1) estimating within groups and between groups covariance matrices and (2) using maximum likelihood and within/between covariance matrices to estimate within and between parts of the factor model. I am proposing a simple extension of these methods where MIIV-

2SLS is applied in the second step in place of ML. Using the Goldstein approach and the MUML approach as two different starting points, I have two possible MIIV-2SLS estimators. I should note that here is that we will be using covariance matrices to fit MIIV-2SLS. Fox (1979) describes 2SLS estimation using only covariance matrices and those results can be adopted for this application; all of the general principles of MIIV-2SLS discussed prior still apply.

For the first possible estimator I will combine the Goldstein approach with MIIV-2SLS. From here I will refer to this as Goldstein-MIIV. First, estimate within-groups and between-groups variance covariance matrices with a three level multilevel model as we did in Equation (11). Second, apply MIIV-2SLS to the within-groups and between- groups covariance matrices separately to estimate the within and between portions of the MSEM. Based on prior research I might expect Goldstein to be better in cases of unbalance. On the other hand, using a parametric model in the first step requires additional distributional assumptions so Goldstein-MIIV might not perform as well given non-normality.

The second possible approach would be to combine the MUML estimator with MIIV-2SLS for another hybrid approach. I will refer to this estimator as MUML-MIIV. The first step would be to decompose the variance into the between and within levels as given in the MUML estimator in Equations (16) and (17). And again once the variance has been decomposed perform MIIV-2SLS on the covariance matrices separately. MUML-MIIV will be more computationally efficient and likely will perform less well given unbalanced clusters. Additionally it makes fewer distributional assumptions so we might expect it to be less affected given non-normality.

These two approaches may have several general advantages over FIML and the MUML and Goldstein approaches. Prior research has shown the general efficacy of the MUML and Goldstein estimators and we will leverage that work by using their strategies for decomposing variances. Further, I expect that MIIV-2SLS in this case will be more robust to model misspecification than FIML. By using MIIV-2SLS we will have access to Sargan's tests of over-identification which is has been shown to be useful for pinpointing model misspecification on an equation by equation basis in single level SEMs. That said, I would like to add a note of caution that Sargan's test was derived in the context of single level analyses. It has not been used in a case like mine where the between and within covariance matrices are estimated separately. It remains to be proven whether the assumptions of Sargan's test hold in these conditions. Finally, MUML-MIIV should be more computationally efficient than other approaches and require fewer distributional assumptions.

The proposed approaches are not without complication either. Standard errors of our model coefficients might not reflect the normal 0.05 alpha levels. As mentioned previously simulations with the MUML estimator found the nominal alpha levels for the between-groups model was 0.08 instead of the expected .05. While one of the proposed advantages are the fewer distributional assumptions, this may be less applicable to the Goldstein approach because this model starts with the 3-level multilevel which requires distributional assumptions in order to be estimated. This 3-level multilevel model may be burdensome to estimate as well, given the number of random coefficients.

Additionally, the procedures being tested here are multi-stage statistical procedures. Two stage least squares is obviously a two stage procedure; we are adding an additional stage prior to 2SLS, creating 3-stage estimation procedures. Complications arise in multiple stage estimation procedures as uncertainty is not carried forward between steps. For example with Goldstein-MIIV, variance/covariance parameters from the first stage are treated as if they were the usual single level sample covariance matrix when they are used with 2SLS. We know that these covariance parameters are estimates and we do not know whether they have similar or different sampling properties. In other words, no corrections are applied for the analysis being done on the decomposed covariance matrices in contrast to the usual single level covariance matrices upon which the test statistic is based. WE suspect that this has the potential to lead to problems with standard errors though it should not bias point estimates.

In sum, I have attempted to make the case that although ML has excellent properties it requires strong assumptions which are often unmet. Given this I developed two possible MIIV-2SLS style estimators, each with their own potential strengths and weaknesses. Next I evaluate the performance of both estimators with a Monte Carlo simulation study varying several factors that I describe in the next section. As a part of this simulation study, I had three primary hypotheses:

1. Given a correctly specified model, balanced clusters, and multivariate normality, Goldstein-MIIV and MUML-MIIV would perform similarly to FIML with a possible slight loss of efficiency.
2. Given structural misspecification, Goldstein-MIIV and MUML-MIIV would be more robust to bias than FIML.
3. Violations of multivariate normality will disrupt the efficacy of FIML and Goldstein-MIIV (MUML-MIIV being robust), and unbalanced groups will disrupt the efficacy of MUML-MIIV.

Chapter 2: Simulation Study

Simulation Design

Data were simulated based on the population model given in Equation 1. Figure 1 shows the model via path diagram. Population values are given as:

$$\boldsymbol{\Lambda_W} = \begin{bmatrix} 1.0 & 0 \\ 0.8 & 0.3 \\ 0.7 & 0 \\ 0.0 & 1.0 \\ 0.3 & 0.8 \\ 0 & 0.7 \end{bmatrix} \quad \boldsymbol{\Lambda_B} = \begin{bmatrix} 1.0 \\ 0.7 \\ 0.6 \\ 0.8 \\ 0.7 \\ 0.8 \end{bmatrix}$$

$$\boldsymbol{\Phi_W} = \begin{bmatrix} 2.0 & 0.3 \\ 0.3 & 2.0 \end{bmatrix} \quad \boldsymbol{\Phi_B} = [0.5]$$

$$\boldsymbol{\Theta_W} = \begin{bmatrix} 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0.3 & 0 & 0 & 0 \\ 0 & 0.3 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 \end{bmatrix} \boldsymbol{\Theta_B} = \begin{bmatrix} 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 \end{bmatrix}$$

The data were generated varying cluster size (CS: 30, 100), number of clusters (CN: 30, 100), balance of cluster sizes (CB: balanced vs unbalanced), and underlying data distribution. CS varied between 30 and 100, both of which would be considered large cluster sizes. CN varied between 30 and 100 as well. Again both of these are well within the range of good, however 30 clusters would be considered a smaller number of clusters, especially for a complex MCFA model (McNeish & Stapleton, 2016). Given the novelty of the approach the goal was to examine how these new estimators perform under ideal conditions. Even at the lowest possible total sample size (CS=30, CN=30), I have a fairly large N=900. I chose to examine more ideal CS and

CN, though in practice applications are likely to encounter smaller CS and CN. I revisit cluster and sample size in my discussion.

In addition to CS and CN, I varied the balance of clusters and the underlying data distribution. Both of these factors were chosen to differentiate possible weaknesses of MUML-MIIV and Goldstein-MIIV respectively. Previous single level MCFA studies suggested that unbalanced clusters led to underestimated SE's with MUML (Hox & Maas, 2001). In this study 'unbalanced' data were generated such that cluster sizes were equal to the average cluster size plus or minus 15. For example, when CS=100 the smaller cluster size was n=85 and the larger sample size was n=115. A more severe disparity in size happens when CS=30 because the smaller cluster size n=15 and the larger cluster size n=45.

Varying the data distribution was expected to have the most impact on the Goldstein-MIIV estimator and FIML estimation. Here we have the added complication that data at both levels can stray from multivariate normality. I used four different conditions: (i) multivariate normal at the within and between level (W-MVN; B-MVN), (ii) Within Level: Skew=2, Kurtosis=8; between level: multivariate normal (W-SK; B-MVN), (iii) within level: multivariate normal; between level: skew=2, kurtosis=8 (W-MVN; B-SK) and, (iv) skew=2, kurtosis=8, at both the within and between levels (W-SK; B-SK). The values of skew=2 and kurtosis=8 were chosen to be within the range of skew and kurtosis found commonly in applications (Micceri, 1989). In addition, previous simulation studies have found these levels of skew and kurtosis to be problematic in both single level SEMs as well as multilevel SEMs (Bauer & Curran, 2003; Curran & West, 1996; Ryu, 2011). Multivariate non-normal data were generated following the procedures outlined by Fleishman (1978) and Vale & Maurelli (1983).

In addition to between cell conditions I examined two within cell conditions: model specification and estimator. I fit each dataset with four different model specifications, namely the true model and 3 separate misspecified models, each missing a single model parameter. Dotted or Dashed lines in Figure 1 represent each of the 3 misspecified paths (while the true model contains all solid, dotted and dashed lines). The first misspecification was missing the cross loading $L_{W1} by\ Y5$. The second misspecification was missing $L_{W2} by\ Y2$, and the third misspecification was missing the correlated errors between Y2 and Y3.

Conditions were not fully crossed, and I will briefly describe the exact combinations of factors examined[6]. The first set of results pertain to ideal conditions and model misspecification. I generated data from the 4 cells corresponding to all combinations of CN/CS holding clusters balanced and assuming multivariate normality (i.e., isolating the effects of sample size and model specification). I generated 600 data sets for each cell, and then fit each data set with all combinations of model specification and model estimators (2400 unique datasets and 28,000 unique models). Next I examined the effects of skew and kurtosis by generating all data for the 16 cells corresponding to all combinations of CN/CS/Skew/Kurtosis holding clusters balanced. Again I generated 600 data sets for each cell and this time fit each with all three estimators and only the true model specification (9600 unique datasets and 28,800 unique models). Finally, I examined the effects of unbalanced data by generating data from the 4 cells corresponding to all combinations of CN/CS, with unbalanced clusters assuming multivariate normality. I generated 600 datasets for each cell and fit each with all three estimators and only the true model specification (2400 unique datasets and 7200 unique models).

[6] In reality, I did run this simulation fully crossed. That's 384 experimental conditions, 19200 independent datasets and 230,400 models fit. However, this was far too much to summarize in any reasonable way. Here I'm presenting what I see as some of the most relevant results, generally isolating factors and considering it with respect to sample size.

Data were generated such that intraclass correlations were kept around 0.1 and 0.15. Hox et al. (2010) found no effect of this in a similar MSEM simulation and Preacher et al., (2010) have suggested that ICC's lower than .05 are likely to cause convergence problems. All data were generated in R. Multivariate normal data was generated using the 'mvrnorm' function from the 'MASS' package and non-normal data were generated using the 'mvrnonnorm' function from the 'semTools' package (Jorgensen, 2016; Venables & Ripley, 2002). FIML models were fit in Mplus with the 'MLR' estimator (robust maximum likelihood, which is the default in Mplus for any multilevel procedure). MUML-MIIV and Goldstein-MIIV estimators were fit in R with self-programmed covariance decomposition functions in conjunction with the 'MIIVsem' package (Fisher, Bollen, Gates, & Rönkkö, 2016). Goldstein-MIIV decompositions were estimated with REML.

There is one final important note about how instrumental variables were selected for the MIIV procedures. Table 1 shows each indicator and all possible MIIVs across true and misspecified conditions. For within group's factor loading I used all MIIVs for each equation (varying over true and misspecified conditions). For any given equation there were between one and four instruments. For the between level factor loadings, two instruments were randomly selected (different for every iteration) from the list of all possible MIIV's. Previous studies have shown that using too many instruments with small sample sizes can lead to downward bias. Thus to mitigate this problem, I used a random procedure to select a different subset of MIIV's for each iteration.

To summarize and compare results I examined a few outcomes. Since MIIV-2SLS has been shown to be more robust to misspecification I used percent bias as one of my primary outcome measures.

$$\% \, bias = \left(\frac{\hat{\theta} - \theta}{\theta}\right) X \, 100 \qquad (26)$$

I also looked at the Root Mean Squared Error as another measure of discrepancy between the true and computed parameters.

$$RMSE = \sqrt{E((\hat{\theta} - \theta)^2)} \qquad (27)$$

I examined the variability of the parameters by looking at the empirical standard deviation of simulation estimates as compared to the mean standard error reported for each estimate. Finally, I summarized the performance of Sargan’s test by examining proportion rejection rates across all of the conditions discussed.

Results

Given that two of the estimators involved iterative estimation procedures (FIML and Goldstein-MIIV), as a first step I checked the integrity of the models by sweeping through all to find non-convergent solutions. Surprisingly, non-convergence was less of an issue than was expected. Table 2 reports the number of models that converged across a variety of conditions. Only a handful of FIML models failed to converge and all Goldstein-MIIV models converged. Given this result I decided to summarize all converged solutions (all solutions for MUML-MIIV). For the models that did not converge with FIML, sensitivity analysis suggested that their inclusion for MUML-MIIV and Goldstein-MIIV results did not unduly influence my findings. Thus, I included them in all reported analyses.

True Model, No Skew/Kurtosis, Balanced Clusters

The first subset of results I examined was the ‘best case’ scenario. Here I compare the Goldstein-MIIV and MUML-MIIV estimators to FIML in the conditions with no violations of any structural or distributional assumptions. Figure 2 illustrates the percent relative bias by CN and

CS for factor loadings. I found evidence that MIIV estimators are performing well across all sample sizes examined. I did not find any mean relative bias across different sample sizes. In general, I note that the spread of relative bias scores increases as sample size decreases, this is normal and to be expected. Additionally, the spread of parameter estimates are slightly larger for MIIV-estimates, as compared to FIML, suggesting a *very slight* loss of efficiency; this effect is most pronounced at CN=30 and CS=30 and disappears as the cluster size increases. For between level factor loadings I found similar results and less of a difference in terms of efficiency. The difference in variability between within groups factor loadings and between groups factor loadings is also immediately apparent. This is due to the fact that the effective sample size at the between group level is much smaller than at the within level. Across different CN and CS I found negligible mean bias for FIML as well as MIIV estimators. Interestingly, at the smallest sample size in this simulation FIML starts to be positively biased while MIIV estimates have no mean bias.

Figure 3 displays relative bias for the variance/covariance parameters. Similar to factor loadings we find these parameters to have no overall mean bias and a similar variability in estimates across estimators. The variance of the between latent factor, ‘LB’, does have a small amount of negative median bias in the CN = 30, CS = 100 condition; despite the median being negatively biased, the mean is still roughly zero. The other general trend we note is that the variability in the between groups latent factor estimates is considerably larger than within, which is due to the smaller effective sample size.

Table 3 reports the RMSE for all loadings across the four CN/CS conditions here. I generally found RMSE’s between FIML and MIIV estimators to be roughly equal. FIML had a slight tendency to have smaller RMSE’s, though these differences were tiny and in most cases non-existent. This is most pronounced at the smallest N and is least pronounced at the largest N and for between groups factor loadings.

Table 4 reports the empirical standard deviations with the mean standard error reported from each model. For within factor loadings, FIML and MUML-MIIV standard deviations vs mean standard errors are very close to one another across conditions. This is what I would expect and hope for, because it suggests the reported standard errors reflect the actual variability of estimates. The Goldstein-MIIV estimates had a slight tendency to underreport standard errors for these within factor loadings. In general, Goldstein-MIIV and MUML-MIIV have similar empirical standard deviations. Goldstein-MIIV then reports a mean standard deviation too small whereas MUML-MIIV reports a mean standard deviation more similar to the actual spread of estimates. At the between level, both MIIV estimators had a tendency to under report standard errors. At the largest sample sizes, this was not much of a problem. However, this became more pronounced as CN/CS decreased and was especially noticeable at CN=30/CS=30. Although, at this sample size even FIML had a tendency to underreport standard errors, but the discrepancy was not as large as it was for MIIV estimators. For the time being it is unclear why this happens. In the future a bootstrap SE or cluster robust SE might do well to help alleviate this disconnect.

Performance under structural misspecification

I fit the exact same datasets from above with three different misspecified models, still holding clusters balanced and all data as multivariate normal. This allowed me to assess the effects of model misspecification alone.

Figure 4 displays percent relative bias for loadings given a model which omits a single cross loading, L1 measured by Y5. Immediately, we can see the effects of misspecification for all three estimators on the factor loadings L2 measured by Y5. All three estimators exhibit mean relative bias at or greater than 9% for this particular loading. FIML has the most bias, then MUML and then Goldstein. Beyond this single loading directly affected by the misspecification, several other paths have mean bias with FIML; we can see the spread of bias which has been noted as a potential shortcoming of FIML. In contrast to this pattern, for Goldstein-MIIV and MUML-MIIV none of the unaffected paths have mean relative bias. In other words, MIIV's appear to be more robust to this structural misspecification. Interestingly, the mean percent bias is stable across sample sizes. Finally, FIML had the most mean relative bias at the between level as well (still less than 5%) suggesting a possible cross level contamination of misspecification. MIIV between level loadings are unaffected, suggesting MIIV's are not as susceptible to such cross level contamination of misspecification.

Table 5 reports the RMSE for this misspecified #1 condition. In addition to being most affected by in terms of mean bias, FIML also has the greatest increase in RMSE—this is mostly true for within factor loadings. Under ideal conditions FIML had general lower RMSE's; however with this misspecification, within loadings also have an increase RMSE. Except for the single affect paths for MIIV's, RMSE stayed almost identical to the true model case. For between factor loadings, FIML generally still had smaller RMSE, though this difference was still small.

Finally Table 6 reports the empirical standard deviations vs mean standard errors of the estimates. Similar to my findings in the true model scenario, FIML and MUML-MIIV estimates of mean standard error match the empirical standard deviation well for within groups factor loadings. Goldstein-MIIV still reports smaller than normal mean standard errors. At the between level, I found all estimators underreporting SE's as the smaller sample sizes, but MIIV estimators do worse than FIML.

I considered two additional structurally misspecified models. In the second misspecification I omitted a different factor loading. In the third misspecification I omitted a single correlated error. The findings here reflect much of the same pattern for misspecification #1. For this reason I summarize these misspecified models together and more briefly.

Figure 5 and Figure 6 show relative bias for misspecification #2 and #3 respectively. In misspecification #2 I found a pattern of bias spread similar to misspecification #1 though slightly more severe. The omitted path here occurs with an item that also has a correlated residual, meaning more paths are effect to this omission. As with before we see that bias spreads to more than a single path for FIML but is retained for just a single path with Goldstein-MIIV and MUML-MIIV. Misspecification #3 has the most unique pattern. The two loadings directly affected by the missing correlated residual are biased for all three estimators. Interestingly, the cross loaded path is more biased for MUML-MIIV and Goldstein-MIIV. This was a slightly unexpected result. As with previous models FIML has the most bias for between level factor loadings suggesting a possible cross level spread while this does not happen as much for MIIV estimates.

Table 7 and Table 9 display the RMSE’s for misspecification #2 and #3 respectively. As a general pattern FIML has larger RMSE’s than both the Goldstein-MIIV and the MUML-MIIV. Both MIIV estimators perform similarly across items. Table 8 and Table 10 display the empirical standard deviations and mean standard errors for misspecification #2 and #3 respectively. As with the other conditions examined so far, for within factor loadings I found that the empirical standard deviations match the mean standard errors for FIML and MUML-MIIV while the Goldstein-MIIV has mean standard errors that are too small. Finally I found mean SE’s are underreported for MIIV estimators especially at smaller sample sizes.

Taken together these misspecified models generally confirm my hypothesis that MIIV-style estimators would be more robust to structural misspecification. These results are in line with previous simulation studies examining single level SEMs. Additionally, these results suggest that the particular misspecification is an important factor in the performance of all of these estimators. The robustness of the MIIV estimators relies on selecting unaffected instruments. When the model misspecification affects the instrument selection for many equations the MIIV estimators are going to be less robust as we saw in misspecification #3. This could perhaps be mitigated by selecting fewer MIIV's.

Performance under distributional misspecification

For this section the four skew/kurtosis conditions were crossed with CS/CN, holding clusters balanced and only using the true model specification. I ran these conditions with across all sample sizes, however for simplicity tables and figures fix CN=100 and CS=30. Analysis of other cluster sizes revealed that general trends were consistent across cluster size and number of clusters. As a general trend results in this section are similar to pattern seen in the true model condition. Generally bias and RMSE are barely affected, and some of the most pronounced effects can be seen in estimates of variability and uncertainty.

Figure 7 displays the relative bias for conditions with skew and kurtosis. In terms of relative bias all three estimators appear to be fairly robust to skew and kurtosis at the levels I specified in this simulation. Perhaps the largest effect can be seen for the between factor loadings in conditions of between group skew. Here, mean relative bias tends to hover between one and three percent for most between factor loadings. However, this is still a very limited amount of mean percent bias. Additionally, it appears as though FIML has slightly higher mean percent bias at the between level, but again these differences are small.

Table 11 reports RMSE for the skew/kurtosis conditions. Here I did not find as much differentiation between the estimators. For a handful of factor loadings I found a slight advantage for FIML (ex., L1 measured by Y2). However for the majority of parameters I found similar RMSE's across.

Table 12 reports empirical standard deviations and mean standard errors for this condition. Effects of skew and kurtosis were most pronounced for these metrics. In general skew/kurtosis creates more variability in estimates, as expected. However, while there is more variability, mean standard errors do not always reflect this. All three estimators have a tendency to under-report standard errors given these distributional misspecifications. These effects are most pronounced for MIIV estimators and this is especially prominent for the between groups factor loadings given between level skew and kurtosis.

Based on prior research I expected skew of 2 with kurtosis of 8 to have more of an effect than I saw here. Though this result does not confirm the hypothesis, I find it positive to know that all three estimators are generally robust at levels of skew and kurtosis considered high. Often studies of skew and kurtosis in SEM deal more with the effects on global model fit parameters, which I did not examine here (Curran & West, 1996; Ryu, 2011). In single level SEM, Satorra (1990), showed that parameters remained consistent under distributional misspecification (though not structural), and showed that if errors are generated independently from explanatory variables in equation then estimates of precision were also robust. More work could be done to examine more extreme levels of skew and kurtosis.

Performance given unbalanced data

Figure 8 contains the relative bias plots for the unbalanced data. I predicted that unbalanced clusters would create the most problems for MUML-MIIV; instead I found that unbalance at the levels specified here did not have as much of an effect. I will refer to average cluster size (Avg-CS), implying that the clusters sizes are actually bifurcated around that average. In terms of mean percent relative bias there was no effect for the within groups factor loadings across as Avg-CS and CN. This was expected given that cluster sizes typically have an effect at the between level. At the between level I found low levels of mean relative bias. At the larger sample sizes FIML and MUML-MIIV had slightly more mean relative bias than the Goldstein-MIIV. This trend generalized somewhat to other sample sizes, but was not universal (i.e., for some loadings Goldstein-MIIV had the most mean relative bias). In general I expected Goldstein to be more robust to cluster unbalance, and I found some evidence of this at CN=100, but less evidence at CN=30. I suspect that as CN increases, so does Goldstein's robustness to cluster unbalance, though future work will have to examine this more closely.

Table 13 contains the RMSE's for unbalanced cells. I continued to find the general trend that FIML has lower RMSE's at the within level, although in most cases this difference was slight. Between level factor loadings had almost no differentiation between RMSE at the highest Avg-CS/CN; at Avg-CS=30 and CN=30 I found the opposite effect that Goldstein and MUML generally have smaller RMSE.

Table 14 contains the empirical standard deviations vs the mean standard errors of estimates for unbalanced cells. In terms of standard errors I found more similar patterns as before. At the within level and across all Avg-CS and CN, FIML and MUML-MIIV have mean standard errors similar to the empirical standard deviation of estimates, while Goldstein-MIIV has a tendency to under report mean standard errors. At the between level I found performance of all estimators to be relatively even. At Avg-CS=30 and CN=30 I found that all estimators under-report standard errors. This trend is actually most pronounced for FIML which has the largest empirical standard deviation of between groups factor loading estimates.

Examining Sargan's test performance

Sargan's test for misspecification is a major advantage of using MIIV estimation. While FIML offers global fit statistics, if researchers have over-identified equations with MIIV's then they have the ability to test for local misspecification one equation at a time. Given that MIIV estimation has never been done for multilevel CFA. I wanted to examine the performance of the test with these estimation procedures. Obviously, Sargan's test was not developed with the procedure of decomposing data into within groups and between groups covariance matrices. That is, I was unsure if the proposed estimators would meet the usual assumptions of Sargan's test and still provide a useful tool in this context. To examine the functioning of Sargan's test I summarize the percent rejection rates for each indicator across various conditions.

Figure 9 presents Sargan's rejection rates for the true model specification. Give that this model has no misspecification I would expect to reject the null hypothesis at the alpha level of 5%. Starting with the largest sample size two distinct findings jump out. 1. For within groups' factor loadings, the Goldstein-MIIV estimator has a much higher base rejection rate than the expected alpha level (rejection rate was between 20% and 30%). 2. MUML-MIIV Sargan's has almost exactly 5% rejection for within loadings and both Goldstein-MIIV and MUML-MIIV have roughly 0.06-0.08 rejection rates at the between level. At the between level this is obviously a bit elevated, though not egregious. Previous studies examining Sargan's test with single level MIIV estimation found similar alpha levels at small samples sizes (Kirby & Bollen, 2009). Across the other sample sizes we see the same pattern described. When the number of clusters is small, alpha levels were elevated for both Goldstein-MIIV and MUML-MIIV. For within groups factor loadings, MUML-MIIV alpha rates stay consistently around 0.05 across all sample sizes, and the Goldstein-MIIV alpha rates remain elevated. The failure of the Goldstein MIIV Sargan's test is a bit perplexing given the relative success of the MUML-MIIV. This is likely related to the smaller reported standard errors—correcting the standard errors in the Goldstein procedure will likely help alleviate the shortcomings of Sargan's here. This problem persists throughout the rest of the comparisons.

The true advantages of using Sargan's test are more apparent in Figure 10, Figure 11 and Figure 12, which illustrate alpha levels across the various misspecifications. The general trends found in the true model specification remain, however for each of the misspecification Sargan's test flags one or more paths as misspecified. In misspecification #1, both Goldstein-MIIV and MUML-MIIV flag the L2 measured by Y5 path 100% of the time. Recall that the misspecified path in this model is the cross loading L1 measured by Y5. Looking back at Figure 2, the mean relative bias plot for misspecification #1, we see that L2 measured by Y5 is the only path which is biased for MIIV estimators. This pattern is the similar across misspecification #2 and #3. Sargan's test flags paths which also happen to be the most biased paths in the model.

Finally, Figure 13 and Figure 14 display the performance of Sargan's test given skew/kurtosis and cluster unbalance respectively. Skew and kurtosis raises the rejection rates for the level of factor loadings with skew/kurtosis. As with Goldstein, this may be related to problems with standard errors in the skew/kurtosis conditions. Unbalanced clusters has a less pronounced effect, but it does lead to slightly higher rejection rates for the between groups factors mostly. Interestingly this seems to be more related to the size of the clusters than to the number of clusters.

Chapter 3: Conclusions

Full information maximum likelihood has become the de facto estimation routine for multilevel CFA, and this is not without reason. We know that given the correct model and no excessive multivariate kurtosis, FIML will be asymptotically efficient and asymptotically unbiased. FIML handles missing data effortlessly. FIML can incorporate random slopes and more than two levels. FIML is flexible and in many circumstances it makes perfect sense that it is used widely. However, I have tried to make the argument that despite its usefulness and flexibility it also makes strong assumptions. Without meeting these assumptions, researchers have no guarantee about the asymptotic efficiency and asymptotic unbiasedness of FIML estimates. Therefore, in circumstances when it is possible to use limited information estimators it may make sense to do so.

Limited information estimators have been shown to be excellent estimators often requiring fewer and less rigid assumptions (Bollen 1996a). In this paper I developed and evaluated two such estimators: The Goldstein-Model Implied Instrumental Variable Estimator and the MUML-Model Implied Instrumental Variable estimator. Following analytic developments of the MIIV estimator for single level SEM's, I developed two novel procedures for using MIIV's to estimate multilevel CFA models. Finally, I evaluated both new estimators in a direct comparison to FIML with a Monte Carlo simulation study varying a variety of factors.

It should be emphasized that the MIIV-estimators evaluated here are not meant to be replacements for FIML in all multilevel CFA models. Indeed, in their current form they are limited in several ways compared to FIML. While FIML can incorporate random slopes, the

MIIV approaches as implemented here cannot. FIML can incorporate more than two levels and MIIV approaches here cannot. FIML can easily handle missing data while MUML-MIIV would need multiple imputation. This is to say that MUML-MIIV and Goldstein-MIIV may be useful for many multilevel CFA models but cannot compete with the general flexibility of FIML and thus FIML is certainly still necessary in many multilevel CFA models, particularly more complex models.

That said, our results suggest that both MIIV estimators performed well in the conditions I considered, with the MUML-MIIV being a clear favorite between the two. Starting with the best case scenario (i.e., True model, no distributional misspecification, and balanced clusters), Goldstein-MIIV and MUML-MIIV performed comparably to FIML. This best-case scenario is unlikely to resemble modeling in practice as it is highly unlikely we ever have the true model. However, this was an opportunity to compare ideal properties. The MIIV estimators had a slight loss in efficiency as compared with FIML, however the difference was minute, found mostly in this true model condition. The real advantages of using the MUML-MIIV or Goldstein-MIIV estimators was apparent in the misspecified model conditions. In these cases, which are arguably much more likely to resemble CFA with real data, I found that MIIV estimators were much more robust to bias due to misspecification. When we do not know the 'True' model, it is all but guaranteed that our models will include some misspecification. These types of scenarios are the norm, and they happen to be when the MIIV estimators outperformed FIML.

My results were largely focused on factor loadings as the primary outcome, although I also considered variance covariance parameters in the True Model condition. These results indicated that we could also estimate variance covariance parameters with MIIV's similarly to FIML. I should note that there are important considerations here. Computing variance parameters with MIIV procedure requires an additional stage in the estimation process, where factor loadings are treated as fixed and covariance parameters are estimate. This is an additional stage in an estimation routine that already has 3 stages. Though our results suggest variance covariance parameters were estimated with no mean bias and similar efficiency to FIML, in some cases the multi-stage process might cause more issue. Additionally, standard errors are not as easy to obtain on variance components with MIIVs. We did not compute them here, though it is possible to estimate them with bootstrapping. FIML, on the other hand, provides these for free.

Despite the positive performance of both MIIV estimators in terms of mean relative bias and RMSE, they did have a problematic tendency to underestimate standard errors. For MUML-MIIV this tendency was most noticeable at small sample sizes for between groups factor loadings and given non-normality and cluster unbalance. In addition to those circumstances, Goldstein-MIIV had a slight tendency to underestimate standard errors for within loadings even at large sample sizes. It should be noted that even FIML had a tendency to underestimate standard errors for between groups factor loadings at CS=30, albeit the discrepancy was less. Though corrections are likely possible for the MIIV estimators, it is also possible that thirty clusters is an inadvisably low number of clusters for fitting complex multilevel factor models. Obviously, underestimating standard errors is problematic and something which needs to be considered heavily. Future studies will need to investigate this further and perhaps consider corrections such as bootstrapping which has its own complications with multilevel data (van der Leeden, Meijer, & Busing, 2008).

Of course, the particular type of model misspecification is a major factor in how much bias is introduced into the model. In this simulation, a single omitted cross loading created several biased paths for FIML and a single biased path for MIIVs. A single omitted cross loading with an indicator with a cross loading created more bias for FIML and more bias for MIIVs (though still less than FIML). Finally, an omitted correlated residual created the most mean bias for all three estimators, and in one particular path MIIV’s had more mean bias than FIML. While MIIV’s in general had less mean bias due to misspecification, MIIVs did not always have the least mean bias, and it may vary model to model.

Additionally, there are a variety of types of misspecification and no study is able to simulate all the possible discrepancies models and reality. This study and many MIIV-2SLS studies have been more focused on model misspecification as it arises from omitting true relationships and included null relationships. It might be that this is the type of misspecification MIIV-2SLS is most robust to. However, discrepancies between models and reality can come in many forms (Cudeck & Henly, 1991; Linhart & Zucchini, 1986; MacCallum, 2003; MacCallum & Tucker, 1991; Meehl, 1990). MacCallum (2001) discussed the importance of considering model imperfection and offers an excellent overview of many others who have considered this. Macallum and Tucker (1991) call the lack of correspondence between reality and our simplified models *model error*. Cudeck and Henly (1991) call the same phenomena *approximation discrepancy*. MacCullum and Tucker (1991) discussed five sources of model error concluding that the true phenomena producing the population covariance matrix may never be captured by the linear factor analysis. The true population values come from relationships and factors far too complex to ever be captured completely in our models—this type of model error might not be well represented by the misspecifications we have offered here. These more complex types of model error may appear as small error across all parameters instead of simply one missing misspecified path. MacCallum and Tucker provide a framework for including diverse sources of error in single level factor analysis. Future work should consider a more diverse set of misspecifications and model errors to more fully understand how MIIV-2SLS performs given a variety of model errors.

There may be more ways to control the spread of bias with MIIV's that were not employed in this particular study. Bias with MIIV's is a direct result of the instruments used in each equation. When bad instruments are used, this creates bias for MIIVs. In this study, to estimate each within equation, I used all possible MIIVs. This increases the chance of using a bad instrument, and increasing the chance that bias may spread. One possible option would be to limit the subset of MIIVs to be a smaller subset, decreasing the chances that bias spreads (though not a guaranteed result). Although, this has the negative consequence that Sargan's test is less useful in identifying model misspecification. Future research should examine particular strategies for deciding how many MIIVs to include.

Given model misspecification, I cannot overlook the importance of Sargan's test as a useful tool when using MIIV estimation. In the single level case, Sargan's Test has been shown to offer a more specific strategy for testing model fit as opposed to typical global fit statistics. At the outset, I was not sure if the conditions of splitting covariance matrices at two levels, would meet the typical assumptions of Sargan's test. In this simulation I showed that there was a direct correspondence to Sargan's test flagging misspecification and biased paths. Sargan's test with MUML-MIIV also had roughly appropriate alpha levels, while the Sargan's with Goldstein-MIIV had elevated alpha levels. In general, Sargan's offers researchers an excellent way to test specific paths in their model. With FIML, we have very little information about which paths are likely to be biased, or where our model is misspecified. This is even more pronounced in the multilevel case where model fit describes both levels simultaneously. With FIML a poor model fit statistic could be due to poor fit at the within level, poor fit at the between level, or both! With

MUML-MIIV in particular I showed that each path at the between and within levels can be tested for model misspecification.

The failure of Sargan’s test with the Goldstein-MIIV estimation routine is somewhat perplexing. I suspect that this is related to the underestimation of standard errors; further I suspect that both of these problems are related to the fact that we are not accounting for sampling variability from the first stage of the analysis. By not taking into account the sampling distribution of the between and within covariance matrices, we have no guarantee that the Sargan test would work. I speculate that this is the root of both of these problems. With Goldstein we are underestimating the amount of uncertainty in our estimates which directly effects our standard errors and indirectly causes over-inflated rejection rates with Sargan’s Test. This is in line with previous findings from McDonald and Goldstein (1989), who note that a primary shortcoming in their procedure is the failure to account for sampling variability from the first stage (Hox, 2010).

Though I have largely talked about the Goldstein-MIIV and MUML-MIIV in tandem so far, the two were differentiated by two important factors: mean standard errors and baseline alpha for Sargan’s test. In both of these cases the MUML-MIIV performed better. The Goldstein-MIIV consistently under reported standard errors while MUML-MIIV did not. Further, Sargan’s test with the Goldstein MIIV was successful in flagging true model misspecification, however the baseline alpha for indicators in correctly specified models was far too high (around 0.3). It is very possible that these two shortcomings are connected. The exact mechanism for this is yet to be determined, but future work might be able to correct this with bootstrapping or other possible procedures. The Goldstein-MIIV was proposed as an estimator which I expected to perform better given unbalanced cluster sizes. I do feel as though there was minor evidence of Goldstein offering some robustness to unbalanced clusters as CN increases, further work needs to probe this more fully.

In addition to performance in this simulation, it should be emphasized that MUML-MIIV is far more computationally efficient (than both Goldstein-MIIV and FIML). Goldstein-MIIV involves fitting complex multilevel models requiring additional parametric assumptions. MUML-MIIV is non-parametric and relies on simple matrix computations. Taken together with performance it is obvious that between MUML-MIIV and Goldstein-MIIV, the MUML-MIIV is the better choice given the current development of each. Perhaps future work can modify the Goldstein-MIIV offer better performance given imbalanced clusters.

An additional advantage of the MIIV estimators proposed in this paper is that they do not require fitting both the between and within models simultaneously. In practice, if my hypothesis only involved the within groups factor structure, I could fit only the within groups factor model (this goes for between as well). This cannot be done FIML as the likelihood function uses information at both levels simultaneously. In fact, several authors have suggested a correct within groups factor structure is a necessary but not sufficient condition to having a correctly specified between groups model with FIML (Muthén & Satorra, 1995; Preacher, Zyphur, & Zhang, 2010). With MIIVs we could only fit one or the other.

Similarly, by separating the between and within covariance matrices first and fitting them with MIIVs second, misspecification should not spread from one level to the next. Theoretically, one level of the model could be catastrophically incorrect and inferences at the opposing level would be unaffected. We only looked at misspecification at the within level and have evidence that bias did not spread to the between level. Future, research should also consider misspecification at the between level. With FIML there was slight evidence of bias spreading across levels. This is likely due to the above mentioned fact that the likelihood uses information from both levels simultaneously. In direct contrast to that MIIV estimation clearly separates the between and within levels offering a robustness to misspecification spreading across levels. This is a major difference between these general approaches and it needs to be emphasized. With FIML, a researcher needs to specify both the within and between level models and both of these models need to be specified correctly. Without specifying both correctly, neither level has any guarantee about the quality of the estimates. Given how MIIV's treat both levels, we can specify a single level or both, and misspecification at one level is less likely to affect model parameters at the opposing level.

Though convergence wasn't a huge issue in this specific simulation prior research has found the convergence can be a problem with FIML in these complex models. An additional benefit of MUML-MIIV estimation is that it does not require any sort of iterative routine to converge. In this way it might be useful to researchers who cannot fit models with FIML. Goldstein-MIIV does require an iterative routine, but the multilevel model required to decompose the covariance matrix is much simpler than a full MCFA model. In this way, it is possible that models are still more likely to converge with Goldstein-MIIV.

This study's primary purpose was to introduce a possible MIIV estimation procedure and evaluate it across a number of conditions. Future work should consider a larger variety of issues. I believe that several extensions could be done with relative ease. The first would be to evaluate these MIIV procedures' ability to handle missing data. Goldstein-MIIV should be able to handle missingness through maximum likelihood in the first stage of the analysis while MUML-MIIV would require something more complicated such as multiple imputation. I briefly discussed the possibility of using a bootstrap procedure to correct standard errors. Finally, future work could make it possible to use cross level equality constraints. In this paper we estimated the between and within models completely separately, however it would be simple to add equality constraints by estimating the within and between models together with a block diagonal covariance matrices.

As with all studies, the current study is not without limitation. The first issue is the generalizability to other latent variable models. I believe the majority my conclusions should generalize to other random intercept confirmatory factor analysis models. However, I only tested one primary structure leaving the possibility that my conclusions may not generalize to all factor analysis models. Future studies should examine other MCFA structures (e.g., more indicators per latent variable and more latent variables). In a similar vein, the current approaches were developed specifically to deal with random intercept MCFA models and not to deal with random factor loadings. It is unlikely that these specific approaches would be able to handle random slopes, and thus the proposed estimators will be most useful when models do not have the complexity of random slopes. Finally, it is possible these approaches generalize to some more general multilevel structural equation models, but that remains to be tested.

The overall effects of cluster size, skew/kurtosis, and unbalanced clusters were minimal in my results. On one hand, this is a generally positive result for this type of research in practice. On the other hand, CS/CN, skew/kurtosis, and cluster unbalance all can be more extreme than the specific conditions tested in this study. I caution the use of these estimators in more extreme conditions. Future work will most certainly be needed to test these MIIV estimators in smaller cluster sizes, with more skew, with more kurtosis, and with more imbalance. These estimators are certainly not invincible and finding conditions where they fail could be very informative. In particular, I suspect that certain combinations might be the especially problematic (unbalanced clusters with non-normality and misspecified models). For example, I suspect that unbalanced clusters with a small average cluster size would have more of an effect than either of those in the current study. More generally the interaction between various misspecifications could be especially important to examine in future work.

Finally, the current study only dealt with two level models. Though not very common in practice, it is possible in some frameworks (e.g., FIML and GLLAMM) to specify more than two level models. The ability of MIIV approaches is currently limited to only two level models. MUML-MIIV cannot be scaled up to multiple levels. Goldstein-MIIV can theoretically include more levels, though this remains to be tested. In a similar way, this study only dealt with random-intercept type models and did not consider random slopes (factor loadings). Varying factor loadings across clusters is less common and methods that allow for this are on the forefront of psychometrics at this time (Hox, 2010). The current MIIV approaches require the constraint that factor loadings are equal across levels.

In sum, this paper laid the foundation for using MIIV-2SLS to estimate MCFA models. My Monte Carlo simulation showed that the proposed techniques are a good alternative to the usual FIML estimation procedure, especially with large samples, balanced clusters, and multivariate normality. Future work will examine a greater variety of models and conditions to more fully understand the use of MIIV-2SLS to estimate multilevel confirmatory factor analysis models.

APPENDIX A: TABLES

Table 1. Reviewing estimation equations and instruments for MIIV-2SLS procedures. Bolded indicators are false instruments based on the True model, but used in the corresponding condition because they are implied by the specified model.

		All possible instruments			
Indicator	Regressed On (scaling indicator)	True Model	Misspecified #1	Misspecified #2	Misspecified #3
Within Factor Loadings					
Y2	Y1	Y5, Y6	Y5, Y6	**Y4**, Y5, Y6	**Y3**, Y5, Y6
Y2	Y4	Y5, Y6	Y5, Y6	-	**Y3**, Y5, Y6
Y3	Y1	Y4, Y5, Y6	Y4, Y5, Y6	Y4, Y5, Y6	**Y3**, Y4, Y5, Y6
Y5	Y1	Y2, Y3, Y6	-	Y2, Y3, Y6	Y2, Y3, Y6
Y5	Y4	Y2, Y3, Y6	**Y1**, Y2, Y3, Y6	Y2, Y3, Y6	Y2, Y3, Y6
Y6	Y4	Y1, Y2, Y3, Y5	Y1, Y2, Y3, Y5	Y1, Y2, Y3, Y5	Y1, Y2, Y3, Y5
Between Factor Loadings					
Y2	Y1	Y3, Y4, Y5, Y6			
Y3	Y1	Y3, Y4, Y5, Y6			
Y4	Y1	Y3, Y4, Y5, Y6			
Y5	Y1	Y3, Y4, Y5, Y6			
Y6	Y1	Y3, Y4, Y5, Y6			

Table 2. Number of converged FIML models by various conditions and sample sizes, for true model specifications.

	CN = 100; CS = 100	CN = 100; CS = 30	CN = 30; CS = 100	CN = 30; CS = 30
W-MVN; B-MVN, balanced [a]	600	600	600	599
W-SK; B-MVN, balanced	600	600	600	600
W-MVN; B-SK, balanced	600	600	600	599
W-SK; B-SK, balanced	600	600	600	600
W-MVN; B-MVN, unbalanced	600	600	597	592

[a] The data in this row were used for all of the misspecified model comparisons. Convergence patterns for misspecified models were identical to those reported here. There was a single dataset which caused problems for all model specifications.

Note: All Goldstein-MIIV converged and MUML-MIIV is non-iterative, so convergence is never a problem.

Table 3. RMSE across CN/CS condition, for the true model specification. Clusters are balanced and all data are multivariate normal.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.02	0.03	0.03	0.04	0.06	0.06	0.04	0.06	0.06	0.07	0.11	0.11
L1 by Y3	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.06	0.07	0.07
L1 by Y5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.04	0.04
L2 by Y2	0.01	0.01	0.01	0.02	0.03	0.03	0.02	0.03	0.03	0.03	0.05	0.05
L2 by Y5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.04
L2 by Y6	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
LB by Y2	0.09	0.09	0.10	0.10	0.11	0.11	0.19	0.19	0.20	0.22	0.23	0.32
LB by Y3	0.08	0.09	0.09	0.09	0.10	0.10	0.17	0.18	0.18	0.20	0.21	0.21
LB by Y4	0.11	0.11	0.11	0.11	0.12	0.13	0.20	0.20	0.21	0.26	0.26	0.27
LB by Y5	0.09	0.10	0.10	0.10	0.11	0.11	0.19	0.19	0.20	0.23	0.28	0.23
LB by Y6	0.10	0.11	0.11	0.11	0.12	0.12	0.19	0.20	0.22	0.24	0.25	0.24

Table 4. Empirical SD of estimates vs (Mean SE) across each CN/CS for the true model specification. All data are multivariate normal and clusters are balanced.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.02 (0.02)	0.03 (0.02)	0.03 (0.03)	0.03 (0.04)	0.06 (0.04)	0.06 (0.06)	0.04 (0.04)	0.06 (0.04)	0.06 (0.05)	0.07 (0.06)	0.11 (0.07)	0.11 (0.10)
L1 by Y3	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)	0.06 (0.06)	0.07 (0.04)	0.06 (0.06)
L1 by Y5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.04 (0.03)	0.04 (0.04)
L2 by Y2	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)	0.03 (0.02)	0.03 (0.03)	0.02 (0.01)	0.03 (0.02)	0.03 (0.03)	0.03 (0.03)	0.05 (0.04)	0.05 (0.05)
L2 by Y5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.04 (0.04)	0.04 (0.03)	0.04 (0.04)
L2 by Y6	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)
LB by Y2	0.09 (0.09)	0.09 (0.09)	0.10 (0.09)	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.19 (0.17)	0.19 (0.17)	0.19 (0.17)	0.21 (0.20)	0.23 (0.17)	0.31 (0.18)
LB by Y3	0.08 (0.08)	0.09 (0.09)	0.09 (0.09)	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.17 (0.16)	0.18 (0.16)	0.18 (0.16)	0.20 (0.18)	0.21 (0.16)	0.21 (0.16)
LB by Y4	0.11 (0.10)	0.11 (0.10)	0.11 (0.10)	0.11 (0.12)	0.12 (0.10)	0.13 (0.10)	0.20 (0.18)	0.20 (0.18)	0.21 (0.18)	0.26 (0.24)	0.26 (0.18)	0.27 (0.18)
LB by Y5	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.19 (0.17)	0.19 (0.17)	0.20 (0.17)	0.22 (0.21)	0.28 (0.19)	0.23 (0.18)
LB by Y6	0.10 (0.10)	0.11 (0.10)	0.11 (0.10)	0.11 (0.11)	0.12 (0.10)	0.12 (0.10)	0.19 (0.19)	0.20 (0.18)	0.22 (0.18)	0.24 (0.23)	0.25 (0.19)	0.24 (0.19)

Table 5. RMSE across CN/CS condition, for misspecification #1 (missing L1 by Y5 factor loading). Clusters are balanced and all data are multivariate normal.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.06	0.03	0.03	0.08	0.06	0.06	0.08	0.06	0.06	0.13	0.11	0.11
L1 by Y3	0.03	0.02	0.02	0.06	0.03	0.03	0.06	0.03	0.03	0.11	0.07	0.07
L1 by Y5												
L2 by Y2	0.02	0.01	0.01	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.05	0.05
L2 by Y5	0.12	0.08	0.09	0.12	0.08	0.09	0.12	0.08	0.09	0.13	0.08	0.09
L2 by Y6	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
LB by Y2	0.09	0.10	0.10	0.10	0.11	0.11	0.19	0.20	0.19	0.22	0.23	0.22
LB by Y3	0.08	0.09	0.09	0.09	0.10	0.10	0.17	0.18	0.18	0.20	0.21	0.24
LB by Y4	0.11	0.11	0.11	0.11	0.13	0.13	0.20	0.20	0.20	0.27	0.30	0.27
LB by Y5	0.09	0.10	0.10	0.10	0.11	0.11	0.19	0.20	0.20	0.23	0.22	0.28
LB by Y6	0.10	0.11	0.11	0.11	0.12	0.12	0.19	0.20	0.21	0.25	0.26	0.25

Table 6. Empirical SD of estimates vs (Mean SE) across each CN/CS for misspecification #1 (Missing L1 by Y5 factor loading). All data are multivariate normal and clusters are balanced.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.04 (0.04)	0.03 (0.02)	0.03 (0.03)	0.06 (0.07)	0.06 (0.04)	0.06 (0.06)	0.06 (0.06)	0.06 (0.04)	0.06 (0.05)	0.13 (0.12)	0.11 (0.07)	0.11 (0.10)
L1 by Y3	0.03 (0.03)	0.02 (0.01)	0.02 (0.02)	0.06 (0.06)	0.03 (0.02)	0.03 (0.03)	0.06 (0.06)	0.03 (0.02)	0.03 (0.03)	0.11 (0.11)	0.07 (0.04)	0.06 (0.06)
L1 by Y5												
L2 by Y2	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.03 (0.02)	0.03 (0.03)	0.02 (0.02)	0.03 (0.02)	0.03 (0.03)	0.03 (0.03)	0.05 (0.04)	0.05 (0.05)
L2 by Y5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.03 (0.03)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)	0.02 (0.02)	0.02 (0.02)	0.05 (0.05)	0.04 (0.03)	0.04 (0.04)
L2 by Y6	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)
LB by Y2	0.09 (0.09)	0.09 (0.09)	0.09 (0.09)	0.10 (0.10)	0.10 (0.09)	0.11 (0.09)	0.19 (0.17)	0.20 (0.17)	0.19 (0.17)	0.22 (0.20)	0.23 (0.18)	0.22 (0.17)
LB by Y3	0.08 (0.08)	0.09 (0.09)	0.09 (0.09)	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.17 (0.16)	0.17 (0.16)	0.18 (0.16)	0.20 (0.19)	0.21 (0.16)	0.24 (0.17)
LB by Y4	0.11 (0.10)	0.11 (0.10)	0.11 (0.10)	0.11 (0.12)	0.13 (0.10)	0.13 (0.10)	0.20 (0.18)	0.20 (0.18)	0.20 (0.18)	0.26 (0.24)	0.30 (0.20)	0.27 (0.18)
LB by Y5	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.19 (0.17)	0.20 (0.17)	0.20 (0.17)	0.23 (0.21)	0.22 (0.18)	0.28 (0.19)
LB by Y6	0.10 (0.10)	0.11 (0.10)	0.11 (0.10)	0.11 (0.11)	0.12 (0.10)	0.12 (0.10)	0.19 (0.19)	0.20 (0.18)	0.21 (0.18)	0.24 (0.23)	0.26 (0.19)	0.25 (0.19)

Table 7. RMSE across CN/CS condition, for misspecification #2 (missing L2 by Y2 factor loading). Clusters are balanced and all data are multivariate normal.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.42	0.29	0.35	0.43	0.29	0.36	0.43	0.29	0.35	0.45	0.30	0.36
L1 by Y3	0.22	0.02	0.02	0.23	0.03	0.03	0.23	0.03	0.03	0.25	0.07	0.07
L1 by Y5	0.03	0.01	0.01	0.03	0.02	0.02	0.03	0.02	0.02	0.05	0.04	0.04
L2 by Y2												
L2 by Y5	0.05	0.01	0.01	0.05	0.02	0.02	0.05	0.02	0.02	0.06	0.04	0.04
L2 by Y6	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
LB by Y2	0.09	0.10	0.10	0.11	0.11	0.11	0.19	0.20	0.19	0.23	0.28	0.28
LB by Y3	0.08	0.09	0.09	0.09	0.10	0.10	0.17	0.18	0.18	0.20	0.21	0.23
LB by Y4	0.11	0.11	0.11	0.12	0.12	0.12	0.20	0.21	0.21	0.28	0.26	0.30
LB by Y5	0.09	0.10	0.10	0.10	0.11	0.11	0.19	0.20	0.20	0.23	0.23	0.23
LB by Y6	0.10	0.11	0.11	0.11	0.12	0.13	0.19	0.21	0.21	0.26	0.24	0.25

Table 8. Empirical SD of estimates vs (Mean SE) across each CN/CS for misspecification #2 (missing L2 by Y2 factor loading). All data are multivariate normal and clusters are balanced.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.04 (0.04)	0.03 (0.02)	0.03 (0.02)	0.07 (0.07)	0.05 (0.03)	0.05 (0.05)	0.07 (0.06)	0.05 (0.03)	0.05 (0.04)	0.13 (0.13)	0.09 (0.06)	0.09 (0.08)
L1 by Y3	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.04 (0.04)	0.03 (0.02)	0.03 (0.03)	0.04 (0.04)	0.03 (0.02)	0.03 (0.03)	0.08 (0.08)	0.07 (0.04)	0.06 (0.06)
L1 by Y5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.04 (0.04)	0.04 (0.03)	0.04 (0.04)
L2 by Y2												
L2 by Y5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.04 (0.04)	0.04 (0.03)	0.04 (0.04)
L2 by Y6	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)
LB by Y2	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.19 (0.17)	0.20 (0.17)	0.19 (0.17)	0.23 (0.21)	0.28 (0.18)	0.28 (0.18)
LB by Y3	0.08 (0.08)	0.09 (0.09)	0.09 (0.09)	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.17 (0.16)	0.18 (0.16)	0.18 (0.16)	0.20 (0.18)	0.21 (0.16)	0.23 (0.17)
LB by Y4	0.11 (0.10)	0.11 (0.10)	0.11 (0.10)	0.11 (0.12)	0.12 (0.10)	0.12 (0.10)	0.20 (0.19)	0.21 (0.18)	0.21 (0.18)	0.27 (0.25)	0.26 (0.18)	0.30 (0.20)
LB by Y5	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.19 (0.17)	0.20 (0.17)	0.20 (0.17)	0.23 (0.21)	0.23 (0.18)	0.23 (0.17)
LB by Y6	0.10 (0.10)	0.11 (0.10)	0.11 (0.10)	0.11 (0.12)	0.12 (0.10)	0.13 (0.10)	0.19 (0.19)	0.21 (0.18)	0.21 (0.19)	0.25 (0.24	0.24 (0.19)	0.25 (0.19)

)

Table 9. RMSE across CN/CS condition, for misspecification #3 (missing Y2 and Y3 correlated residual). Clusters are balanced and all data are multivariate normal.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.22	0.22
L1 by Y3	0.18	0.17	0.17	0.18	0.17	0.17	0.18	0.17	0.17	0.19	0.17	0.17
L1 by Y5	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.04
L2 by Y2	0.02	0.08	0.08	0.03	0.08	0.08	0.03	0.08	0.08	0.03	0.08	0.08
L2 by Y5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.04
L2 by Y6	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
LB by Y2	0.09	0.10	0.10	0.10	0.11	0.11	0.19	0.19	0.20	0.21	0.27	0.31
LB by Y3	0.08	0.09	0.09	0.09	0.10	0.10	0.17	0.18	0.18	0.20	0.22	0.22
LB by Y4	0.11	0.12	0.12	0.11	0.13	0.12	0.20	0.20	0.20	0.26	0.27	0.26
LB by Y5	0.09	0.10	0.10	0.10	0.11	0.11	0.19	0.20	0.20	0.22	0.23	0.28
LB by Y6	0.10	0.11	0.11	0.11	0.12	0.13	0.19	0.20	0.20	0.24	0.26	0.25

Table 10. Empirical SD of estimates vs (Mean SE) across each CN/CS for misspecification #3 (missing Y2 and Y3 correlated residual). All data are multivariate normal and clusters are balanced.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.04 (0.04)	0.05 (0.03)	0.05 (0.04)
L1 by Y3	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.04 (0.04)	0.04 (0.03)	0.04 (0.04)
L1 by Y5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.04 (0.03)	0.04 (0.04)
L2 by Y2	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.01 (0.01)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)	0.04 (0.03)	0.04 (0.04)
L2 by Y5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.04 (0.04)	0.04 (0.03)	0.04 (0.04)
L2 by Y6	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)
LB by Y2	0.09 (0.09)	0.10 (0.09)	0.09 (0.09)	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.19 (0.17)	0.19 (0.17)	0.20 (0.17)	0.21 (0.20)	0.27 (0.18)	0.31 (0.18)
LB by Y3	0.08 (0.08)	0.09 (0.09)	0.09 (0.09)	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.17 (0.16)	0.18 (0.16)	0.18 (0.16)	0.20 (0.18)	0.22 (0.16)	0.22 (0.16)
LB by Y4	0.10 (0.10)	0.12 (0.10)	0.12 (0.10)	0.11 (0.12)	0.13 (0.10)	0.12 (0.10)	0.20 (0.18)	0.20 (0.18)	0.20 (0.18)	0.26 (0.24)	0.27 (0.19)	0.26 (0.19)
LB by Y5	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.19 (0.17)	0.20 (0.17)	0.19 (0.17)	0.22 (0.21)	0.23 (0.18)	0.28 (0.19)
LB by Y6	0.10 (0.10)	0.11 (0.10)	0.11 (0.10)	0.11 (0.11)	0.12 (0.10)	0.13 (0.10)	0.19 (0.19)	0.20 (0.18)	0.20 (0.18)	0.24 (0.23)	0.26 (0.19)	0.25 (0.19)

Table 11. RMSE across various combinations of Skew/Kurtosis. Within groups multivariate normal (W-MVN) vs. within skew=2 kurtosis=8 (W-SK) crossed with between groups multivariate normal (B-MVN) vs. between skew=2 kurtosis=8 (B-SK). Other fixed factors are balanced clusters, true model specification, CN=100; CS=30. Trends are representative of findings at other CN/CS.

	W-MVN; B-MVN[a]			**W-SK**; B-MVN			W-MVN; **B-SK**			**W-SK**; **B-SK**		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.04	0.06	0.06	0.05	0.07	0.07	0.04	0.05	0.05	0.05	0.07	0.07
L1 by Y3	0.03	0.03	0.03	0.04	0.04	0.04	0.03	0.03	0.03	0.04	0.04	0.04
L1 by Y5	0.02	0.02	0.02	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.03
L2 by Y2	0.02	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.03
L2 by Y5	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03
L2 by Y6	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03
LB by Y2	0.10	0.11	0.11	0.10	0.11	0.10	0.15	0.15	0.15	0.16	0.17	0.17
LB by Y3	0.09	0.10	0.10	0.10	0.11	0.10	0.15	0.15	0.16	0.15	0.15	0.15
LB by Y4	0.11	0.12	0.13	0.12	0.13	0.13	0.18	0.19	0.18	0.18	0.18	0.18
LB by Y5	0.10	0.11	0.11	0.11	0.11	0.12	0.16	0.17	0.17	0.16	0.17	0.17
LB by Y6	0.11	0.12	0.12	0.12	0.12	0.13	0.18	0.18	0.18	0.17	0.18	0.18

[a] The W-MVN; B-MVN condition presented here is the same as the RMSE presented in the True model CN=100; CS=30 condition.

Table 12. Empirical SD of estimates vs (Mean SE) across skew/kurtosis conditions. CN=100, CS=30. All clusters are balanced and fit with fit with true model specification.

	W-MVN; B-MVN			**W-SK**; B-MVN			W-MVN; **B-SK**			**W-SK**; **B-SK**		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.03 (0.04)	0.06 (0.04)	0.06 (0.06)	0.05 (0.05)	0.07 (0.04)	0.07 (0.05)	0.03 (0.04)	0.05 (0.04)	0.05 (0.05)	0.05 (0.05)	0.07 (0.04)	0.07 (0.06)
L1 by Y3	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)	0.04 (0.04)	0.04 (0.02)	0.04 (0.03)	0.03 (0.03)	0.03 (0.02)	0.03 (0.03)	0.04 (0.04)	0.04 (0.02)	0.04 (0.03)
L1 by Y5	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.03 (0.01)	0.03 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.02 (0.02)	0.03 (0.01)	0.03 (0.02)
L2 by Y2	0.02 (0.01)	0.03 (0.02)	0.03 (0.03)	0.02 (0.02)	0.03 (0.02)	0.03 (0.03)	0.02 (0.01)	0.03 (0.02)	0.03 (0.03)	0.02 (0.02)	0.03 (0.02)	0.03 (0.03)
L2 by Y5	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)	0.03 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)	0.03 (0.02)	0.03 (0.02)
L2 by Y6	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.03 (0.01)	0.03 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.02)	0.03 (0.03)	0.03 (0.01)	0.03 (0.02)
LB by Y2	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.10 (0.10)	0.11 (0.09)	0.10 (0.09)	0.15 (0.14)	0.15 (0.09)	0.15 (0.09)	0.16 (0.14)	0.17 (0.09)	0.16 (0.09)
LB by Y3	0.09 (0.09)	0.10 (0.09)	0.10 (0.09)	0.10 (0.09)	0.11 (0.09)	0.10 (0.09)	0.15 (0.13)	0.15 (0.09)	0.16 (0.09)	0.15 (0.13)	0.15 (0.09)	0.15 (0.09)
LB by Y4	0.11 (0.12)	0.12 (0.10)	0.13 (0.10)	0.12 (0.12)	0.13 (0.10)	0.13 (0.10)	0.18 (0.16)	0.19 (0.10)	0.18 (0.10)	0.17 (0.16)	0.18 (0.10)	0.18 (0.10)
LB by Y5	0.10 (0.10)	0.11 (0.09)	0.11 (0.09)	0.11 (0.10)	0.11 (0.09)	0.11 (0.09)	0.16 (0.15)	0.17 (0.09)	0.17 (0.09)	0.16 (0.14)	0.17 (0.09)	0.17 (0.09)
LB by Y6	0.11 (0.11)	0.12 (0.10)	0.12 (0.10)	0.12 (0.11)	0.12 (0.10)	0.13 (0.10)	0.18 (0.16)	0.18 (0.10)	0.18 (0.10)	0.17 (0.15)	0.18 (0.10)	0.18 (0.10)

Table 13. RMSE across CN/CS condition with unbalanced clusters. All data are multivariate normal and fit with true model specification.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.02	0.03	0.03	0.04	0.06	0.06	0.04	0.06	0.06	0.07	0.10	0.10
L1 by Y3	0.02	0.02	0.02	0.03	0.04	0.04	0.03	0.04	0.04	0.06	0.06	0.06
L1 by Y5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.04	0.04
L2 by Y2	0.01	0.01	0.01	0.01	0.03	0.03	0.02	0.03	0.03	0.03	0.05	0.05
L2 by Y5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.04
L2 by Y6	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
LB by Y2	0.09	0.10	0.10	0.10	0.11	0.12	0.18	0.19	0.19	0.55	0.22	0.23
LB by Y3	0.08	0.09	0.09	0.09	0.10	0.11	0.18	0.19	0.18	0.65	0.21	0.24
LB by Y4	0.10	0.11	0.11	0.12	0.13	0.14	0.20	0.21	0.21	0.63	0.30	0.29
LB by Y5	0.10	0.10	0.10	0.10	0.11	0.11	0.18	0.18	0.19	0.46	0.24	0.25
LB by Y6	0.10	0.11	0.11	0.11	0.13	0.13	0.19	0.21	0.21	0.40	0.37	0.26

Table 14. Empirical SD of estimates vs (Mean SE) across each CN/CS for unbalanced clusters. All data are multivariate normal and fit with the true model specification.

	CN = 100; CS = 100			CN = 100; CS = 30			CN = 30; CS = 100			CN = 30; CS = 30		
	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV	FIML	Gold-MIIV	MUML-MIIV
L1 by Y2	0.02	0.03	0.03	0.04	0.05	0.06	0.04	0.06	0.06	0.07	0.10	0.10
	(0.02)	(0.02)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.07)	(0.10)
L1 by Y3	0.02	0.02	0.02	0.03	0.04	0.04	0.03	0.04	0.03	0.06	0.06	0.06
	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.06)	(0.04)	(0.06)
L1 by Y5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.03)	(0.04)
L2 by Y2	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.03	0.03	0.05	0.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.01)	(0.02)	(0.03)	(0.03)	(0.04)	(0.05)
L2 by Y5	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.04
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.03)	(0.04)
L2 by Y6	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)
LB by Y2	0.09	0.10	0.10	0.10	0.11	0.12	0.18	0.19	0.19	0.55	0.22	0.23
	(0.09)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.17)	(0.17)	(0.17)	(0.33)	(0.17)	(0.17)
LB by Y3	0.08	0.09	0.09	0.09	0.10	0.11	0.18	0.19	0.18	0.64	0.21	0.24
	(0.08)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.16)	(0.16)	(0.16)	(0.35)	(0.16)	(0.16)
LB by Y4	0.10	0.11	0.11	0.12	0.13	0.14	0.20	0.20	0.21	0.63	0.30	0.29
	(0.10)	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.19)	(0.18)	(0.18)	(0.40)	(0.18)	(0.20)
LB by Y5	0.09	0.10	0.10	0.10	0.11	0.11	0.18	0.18	0.18	0.46	0.24	0.25
	(0.09)	(0.09)	(0.09)	(0.11)	(0.09)	(0.09)	(0.18)	(0.17)	(0.17)	(0.31)	(0.17)	(0.17)
LB by Y6	0.10	0.11	0.11	0.11	0.13	0.13	0.19	0.21	0.21	0.40	0.37	0.25
	(0.10)	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.18)	(0.18)	(0.18)	(0.32)	(0.20)	(0.18)

APPENDIX B: FIGURES

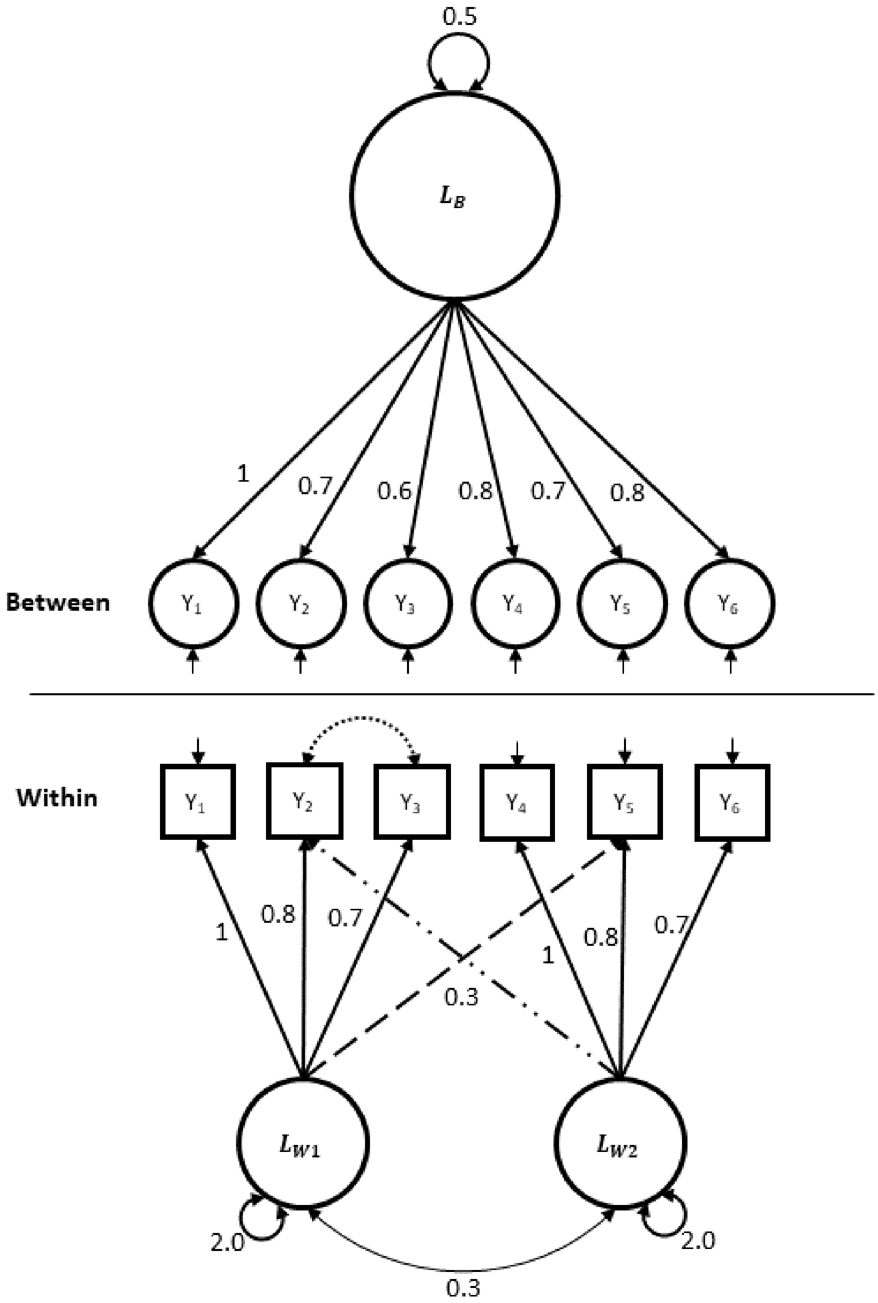

Figure 1. Population model data generating model. True model contains all solid dashed and dotted lines. Misspecification #1 omits the path between L_{W1} and Y5. Misspecification #2 omits the path between L_{W2} and Y2. Misspecification #3 omits the correlated residual between Y2 and Y3.

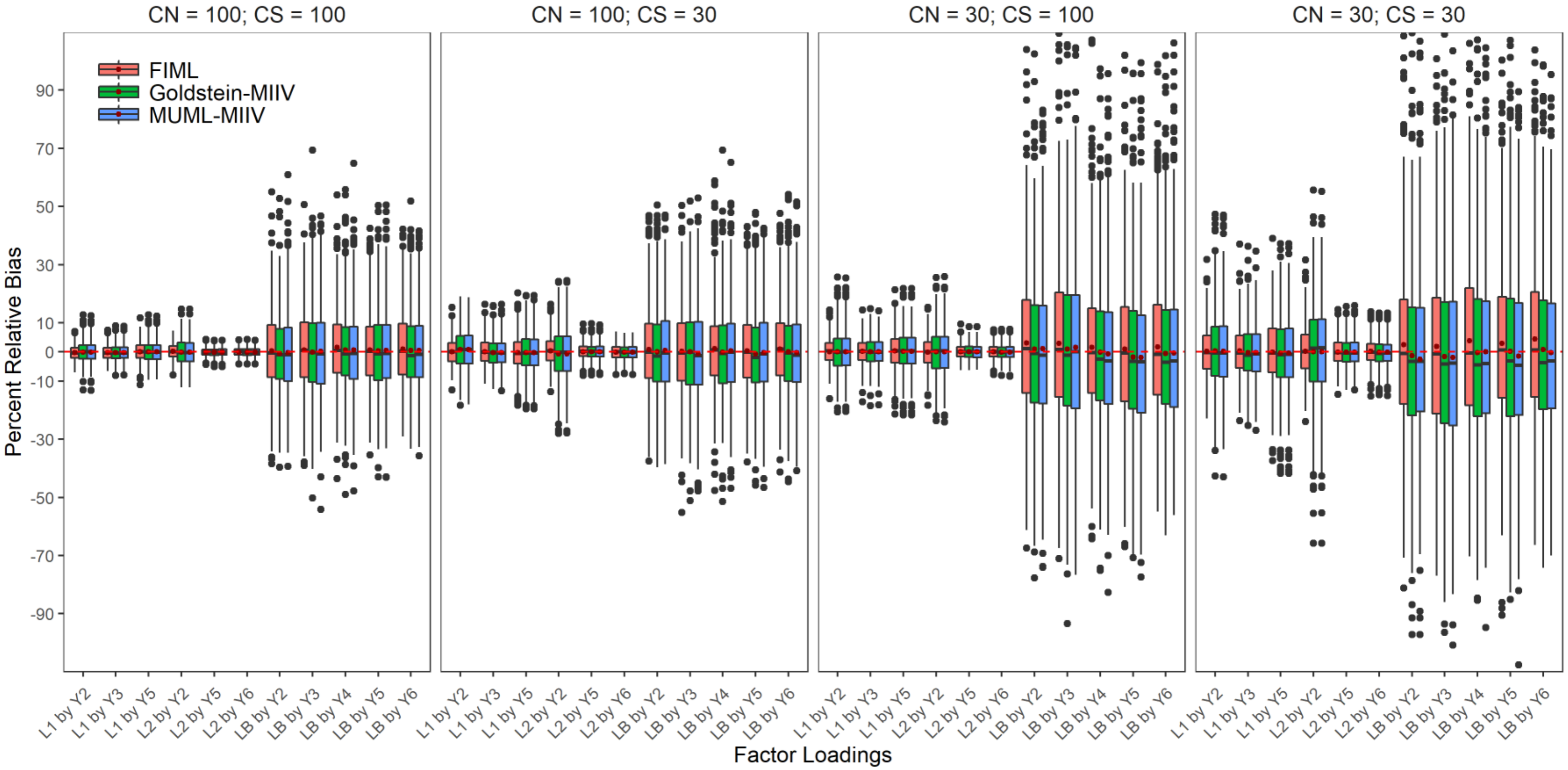

Figure 2. Relative bias boxplots for each factor loading in the True Model condition. Black line represents median relative bias. Black dot represents mean relative bias. Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

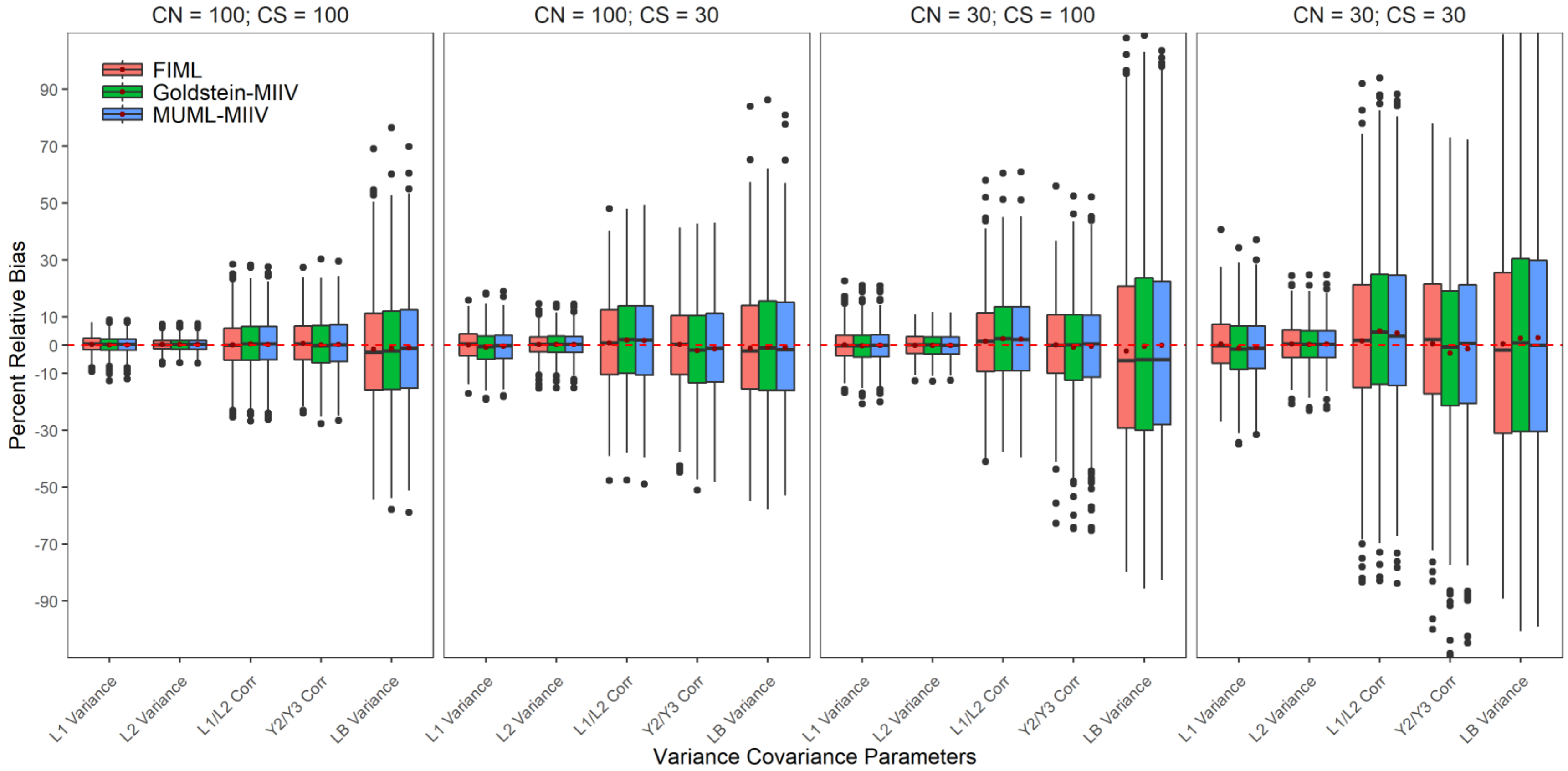

Figure 3. Relative bias boxplots for variance/covariance parameters in the True Model condition. Black line represents median relative bias. Black dot represents mean relative bias. Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

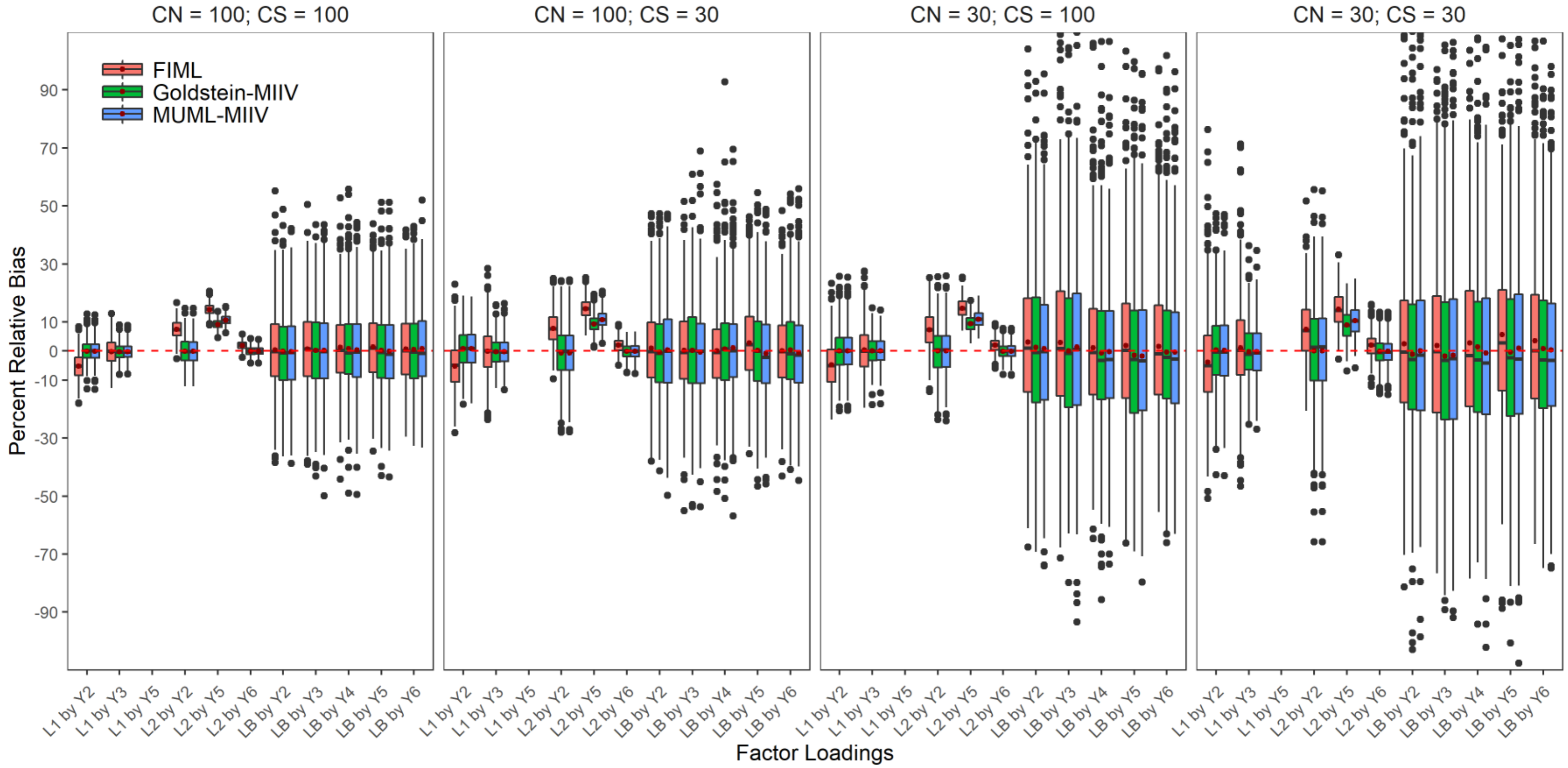

Figure 4. Relative bias boxplots for each factor loading in the Misspecified # 1 condition (missing the L1 by Y5 factor loading). Black line represents median relative bias. Black dot represents mean relative bias. Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

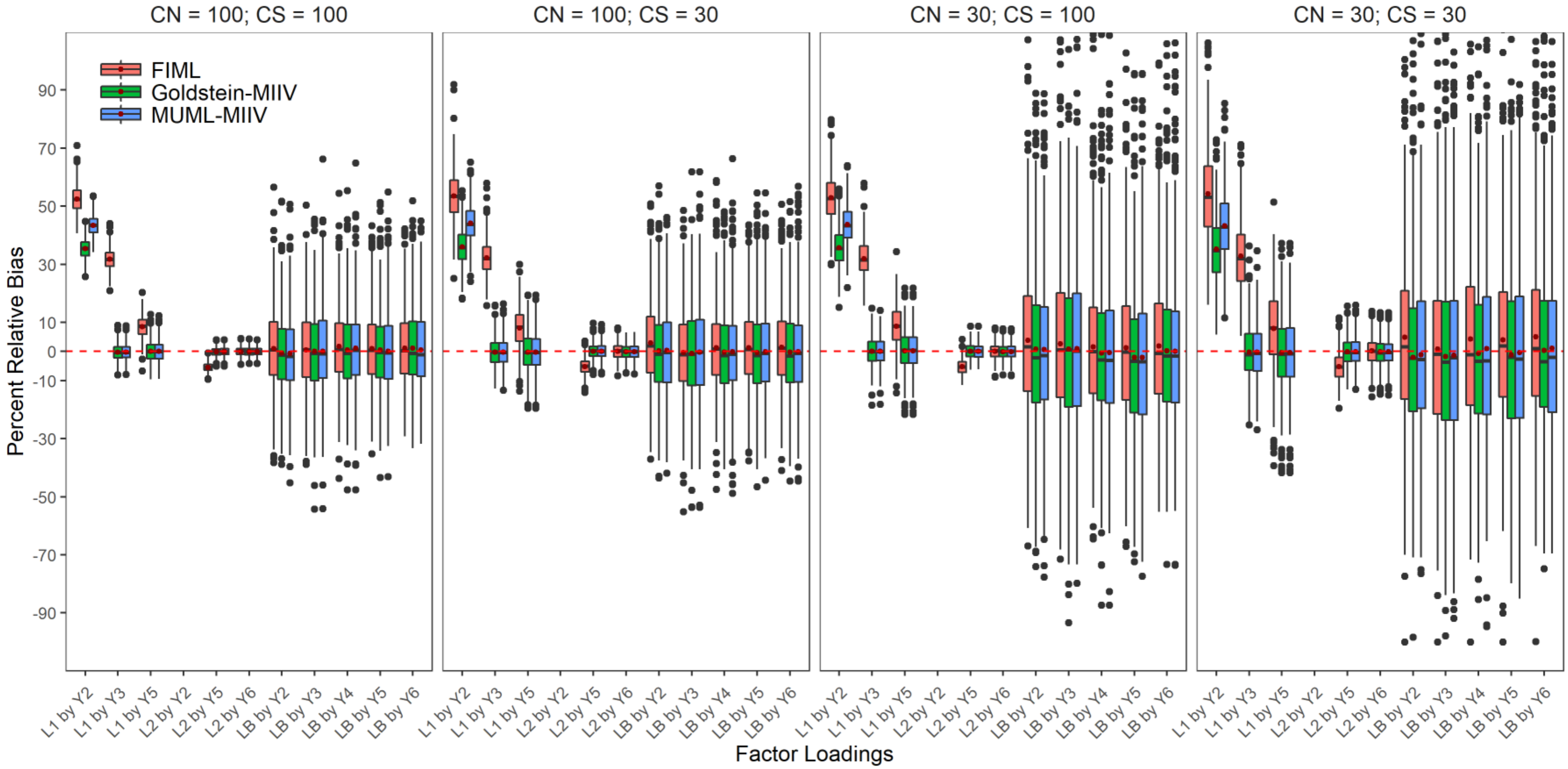

Figure 5. Relative bias boxplots for each factor loading in the Misspecified # 2 condition (missing the L2 by Y2 factor loading). Black line represents median relative bias. Black dot represents mean relative bias. Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

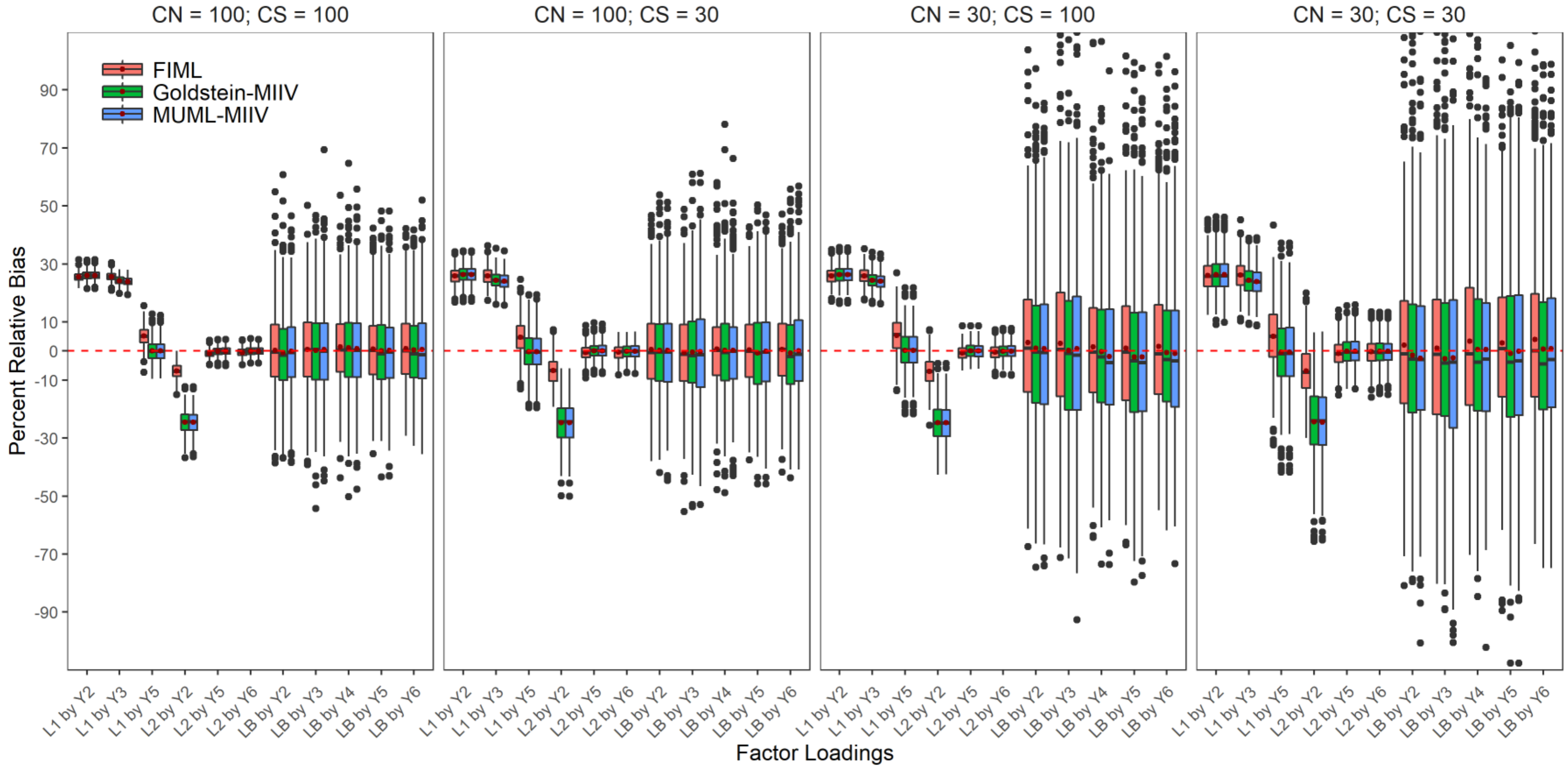

Figure 6. Relative bias boxplots for each factor loading in the Misspecified # 3 condition (missing the correlated residual between Y2 and Y3). Black line represents median relative bias. Black dot represents mean relative bias. Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

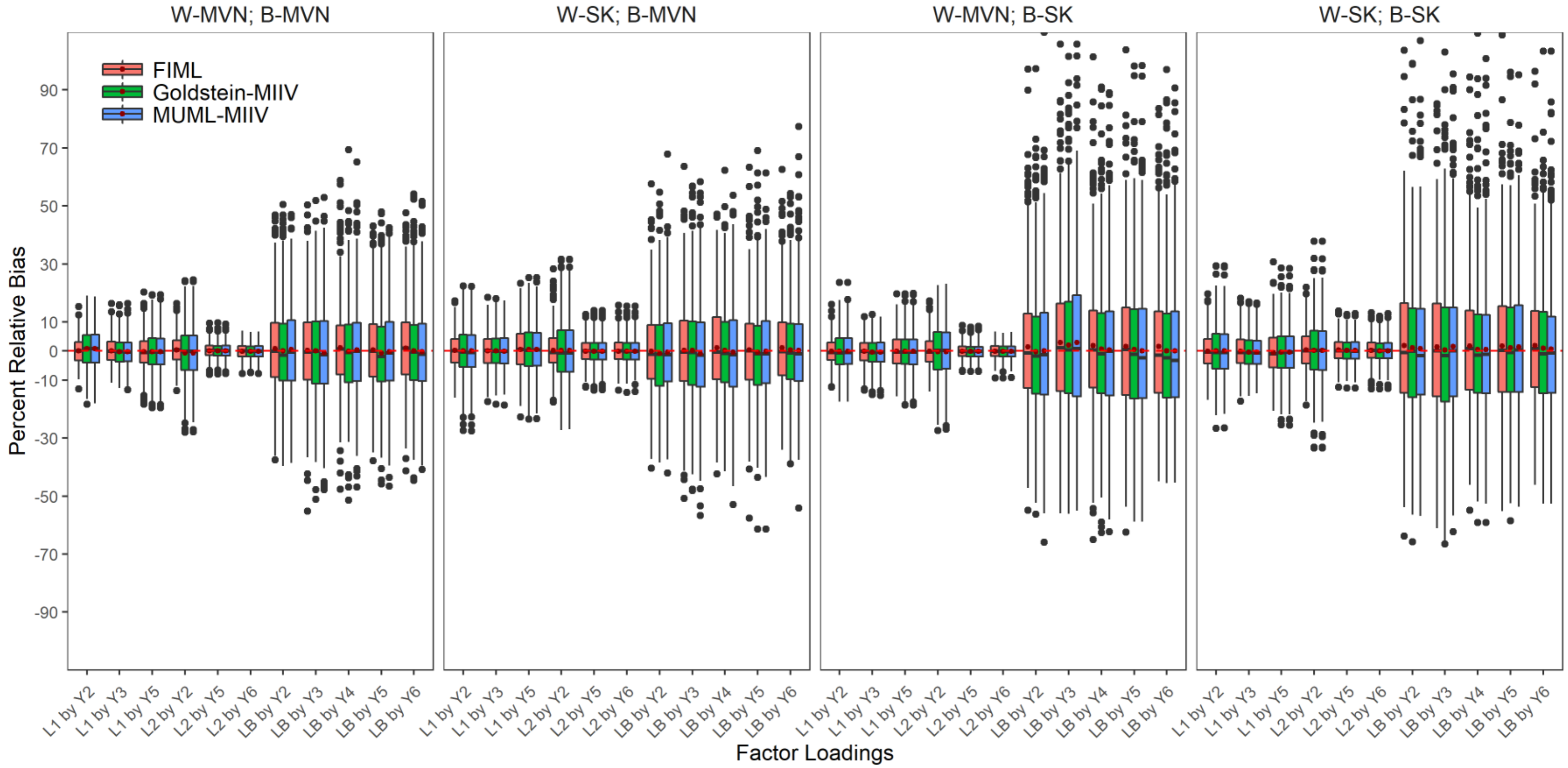

Figure 7. Relative bias boxplots for each factor loading across the Skew/Kurtosis conditions. CN=100 and CS=30, while trends here generalized to other CN/CS conditions. Black line represents median relative bias. Black dot represents mean relative bias. Colors separate different estimators and each cell is a different combination skew/kurtosis condition. All clusters are balanced and data are fit with the true model specification.

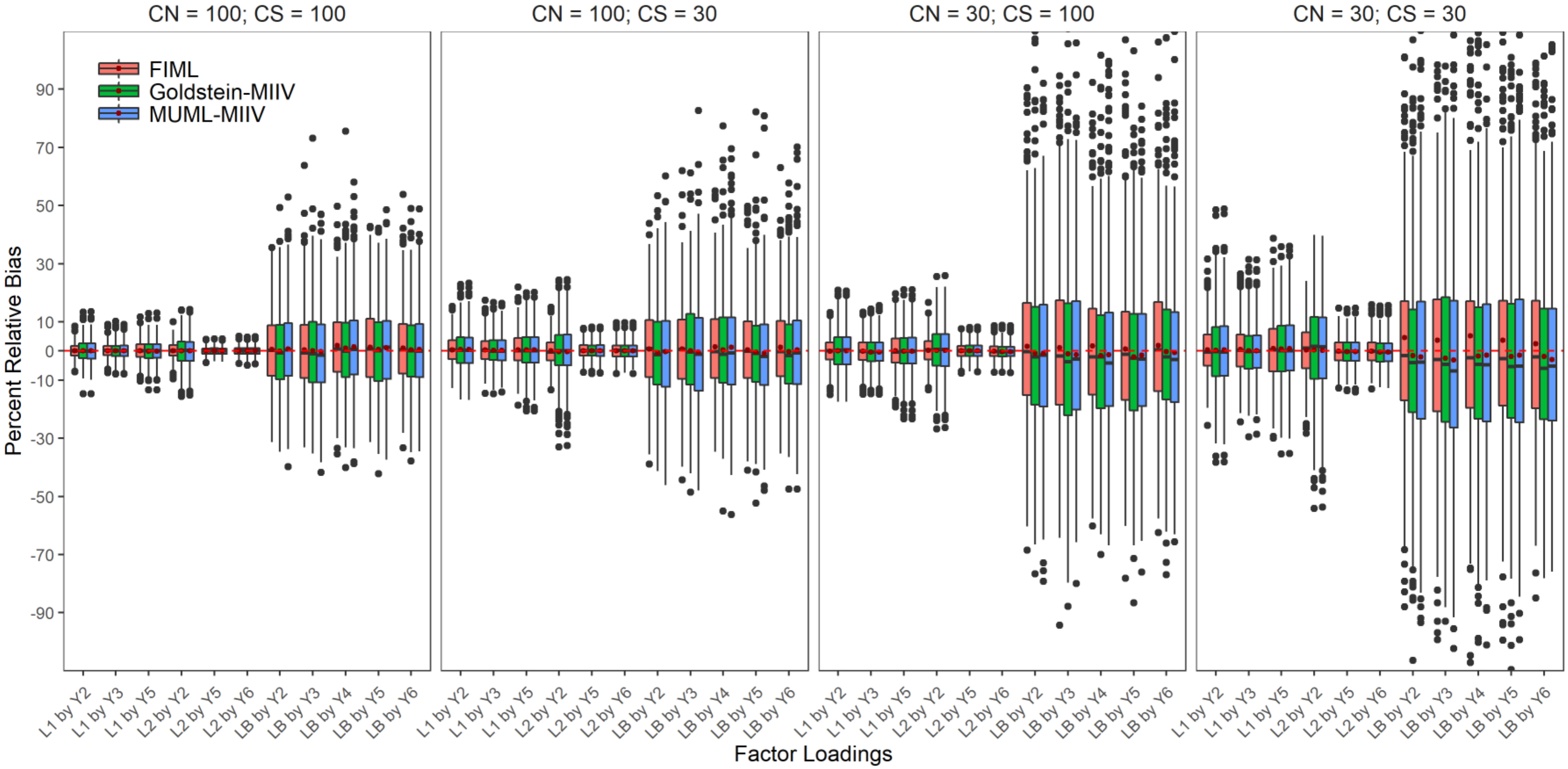

Figure 8. Relative bias boxplots for each factor loading in the unbalanced clusters condition. Black line represents median relative bias. Black dot represents mean relative bias. Colors separate different estimators and each cell is a different combination of CN and CS. All data are multivariate normal and fit with the true model specification.

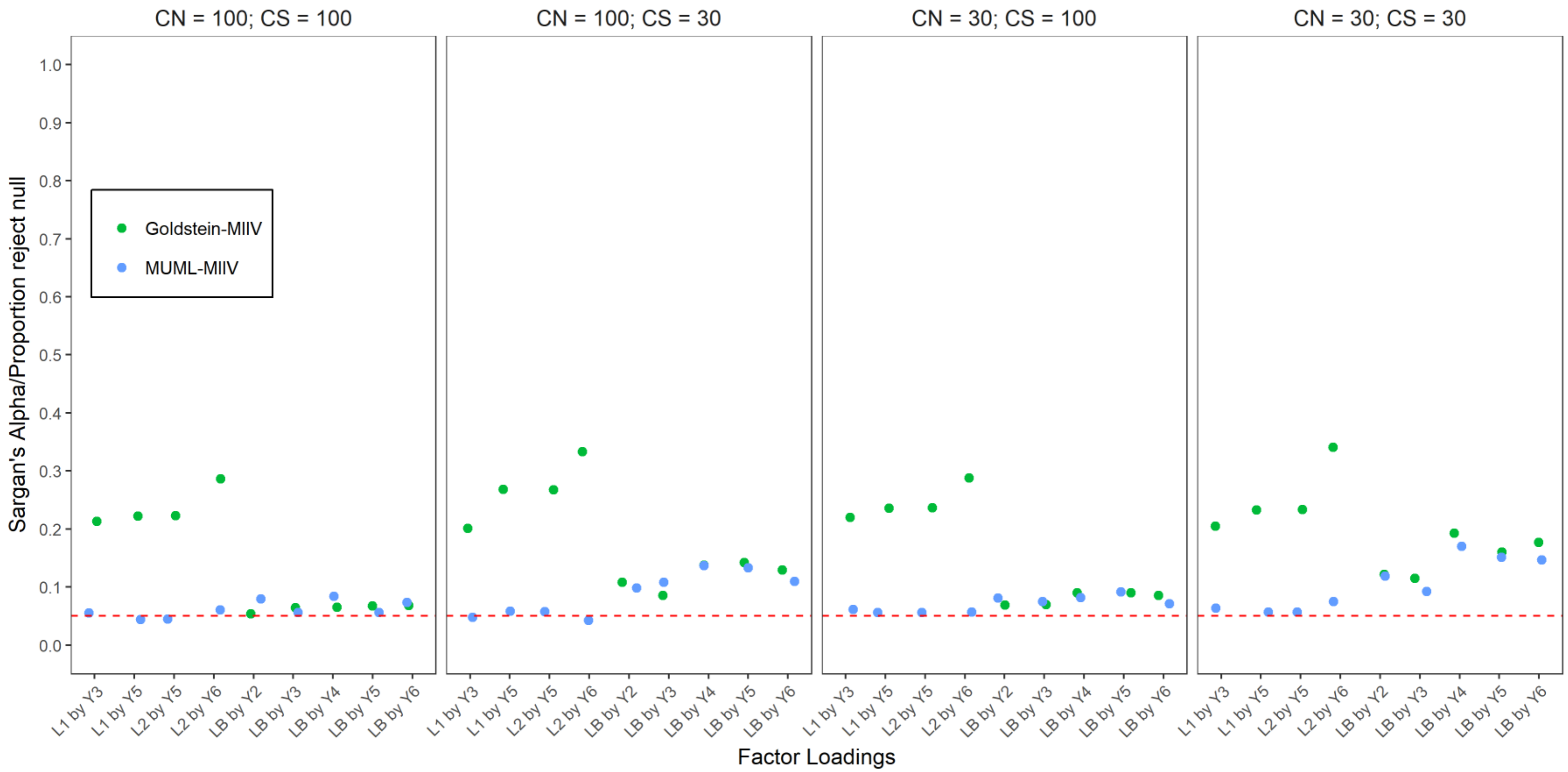

Figure 9. Sargan's test rejection rates for each factor loading in the True Model condition. Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

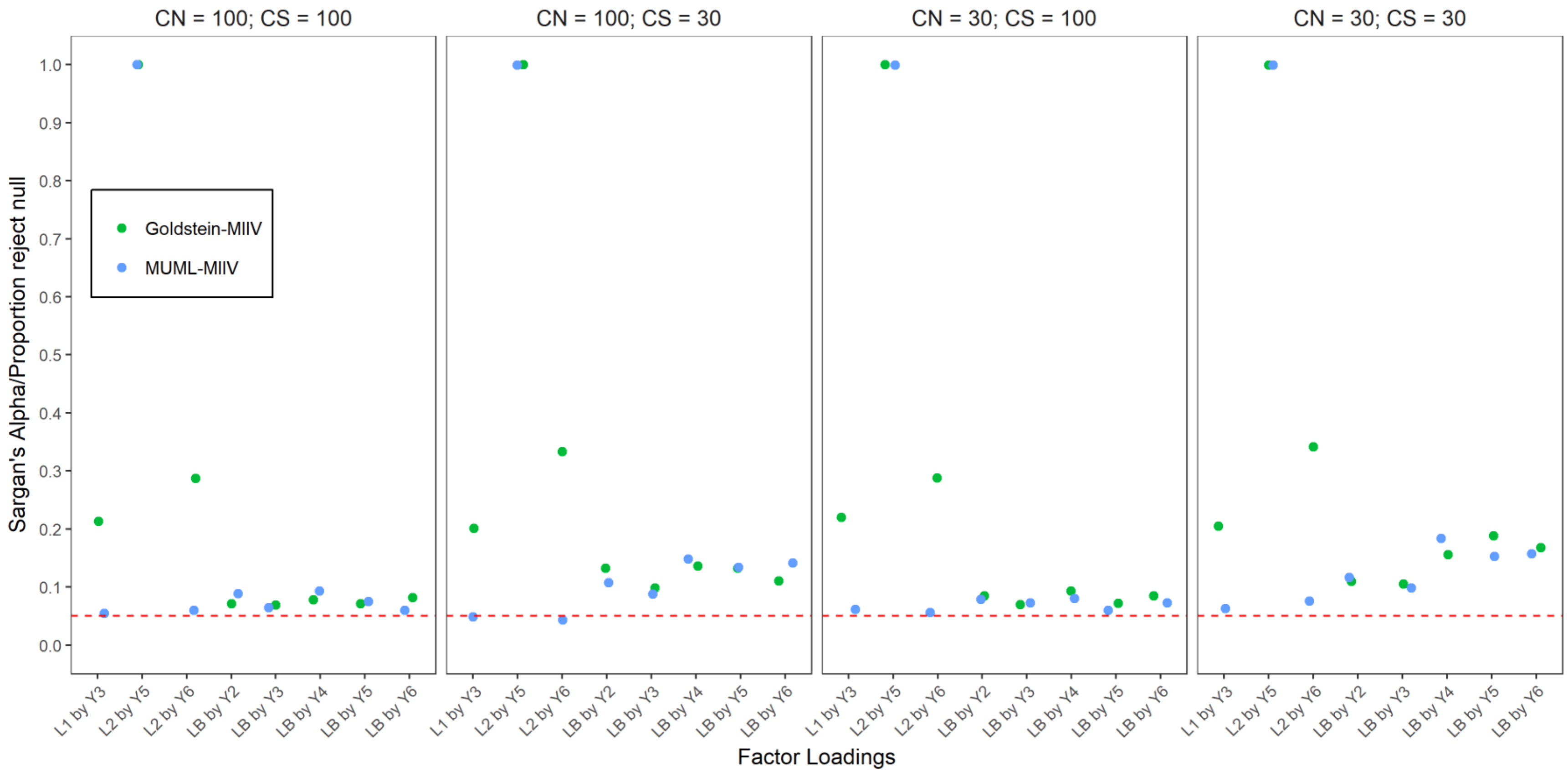

Figure 10. Sargan's Test Rejection Rates for each factor loading in the Misspecified # 1 condition (missing the L1 by Y5 factor loading). Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

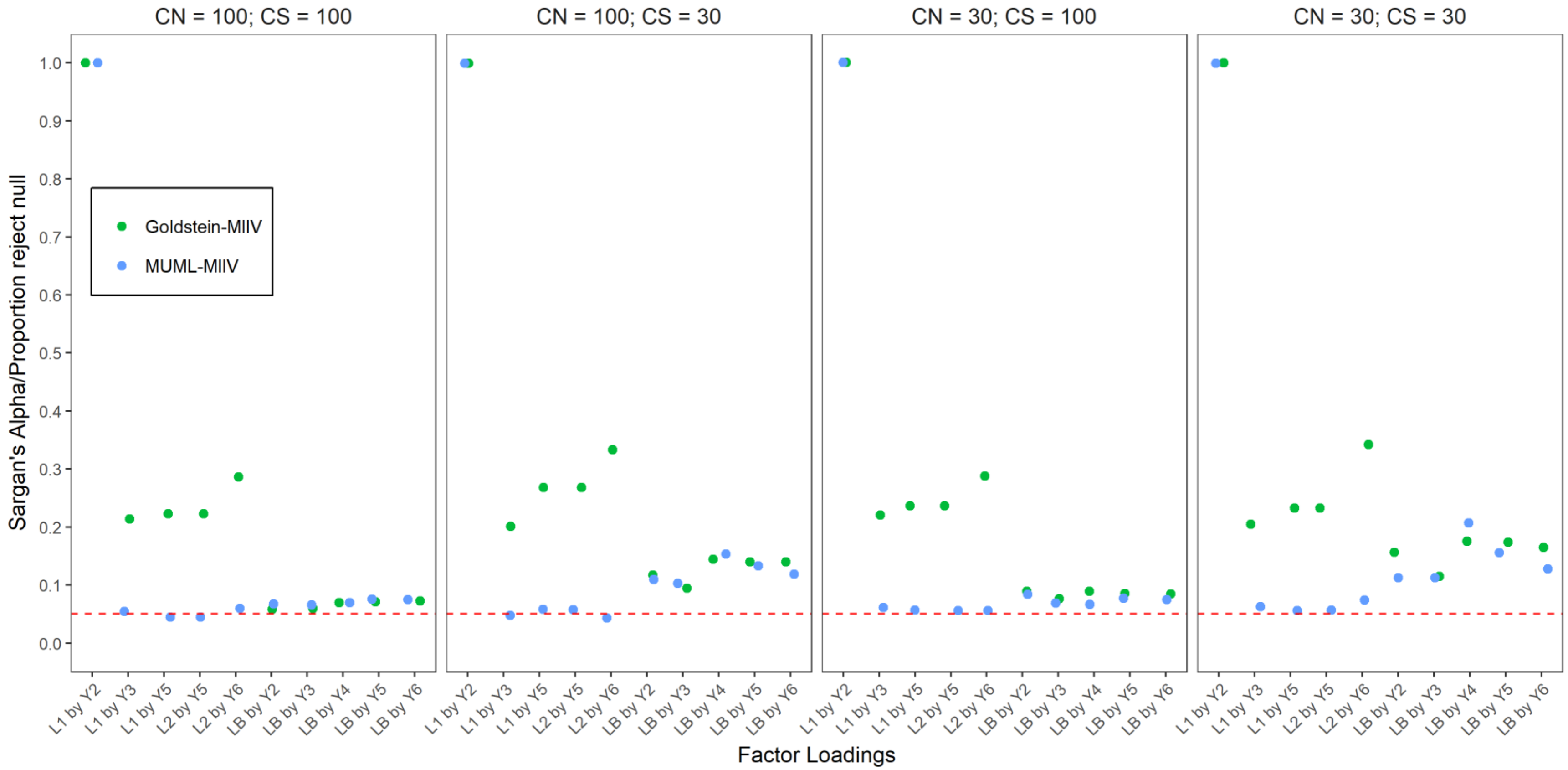

Figure 11. Sargan's Test Rejection Rates for each factor loading in the Misspecified # 2 condition (missing the L2 by Y2 factor loading). Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

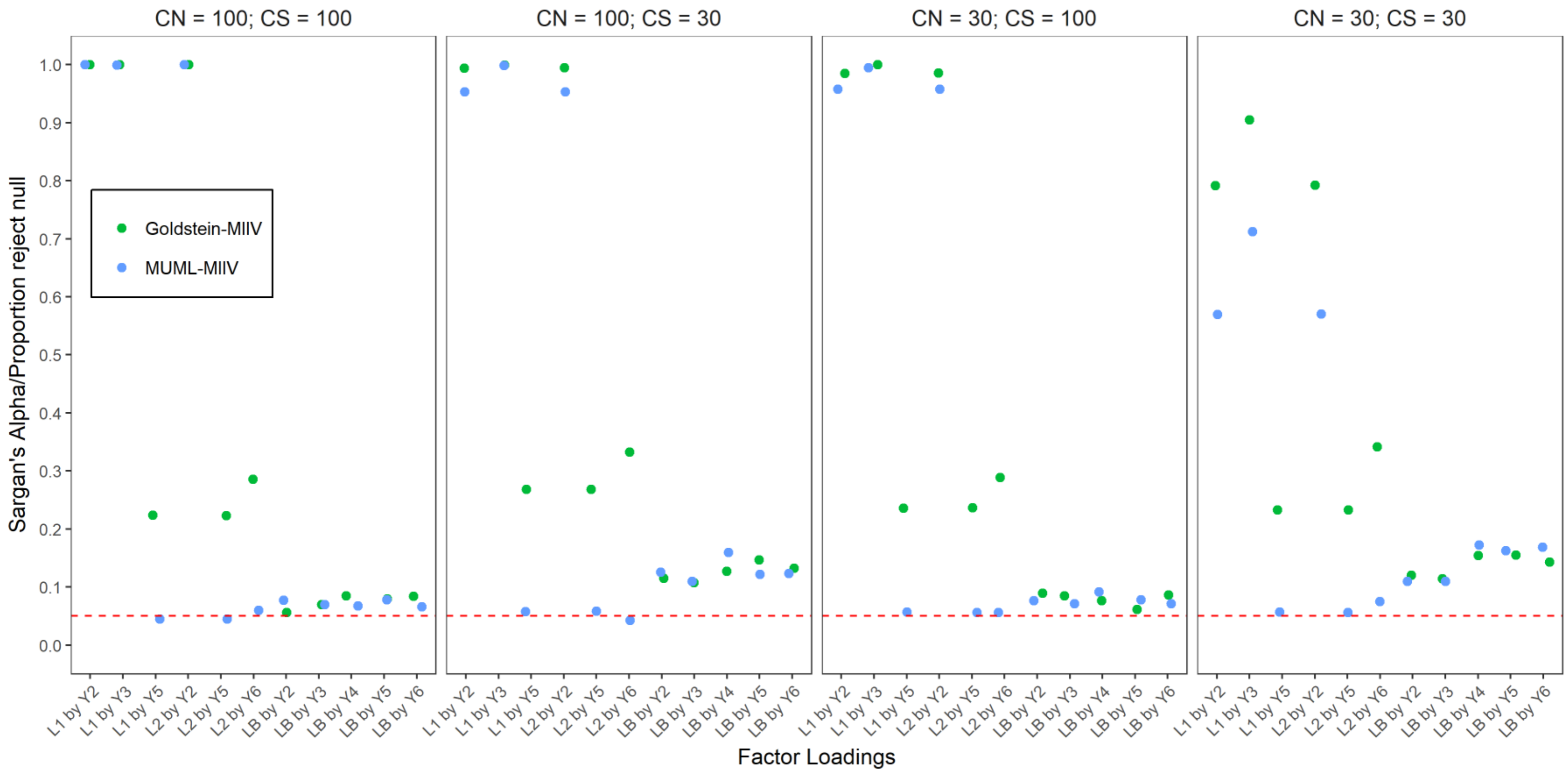

Figure 12. Sargan's Test Rejection Rates for each factor loading in the Misspecified # 3 condition (missing the correlated residual between Y2 and Y3). Colors separate different estimators and each cell is a different combination of CN and CS. All clusters are balanced and multivariate normal.

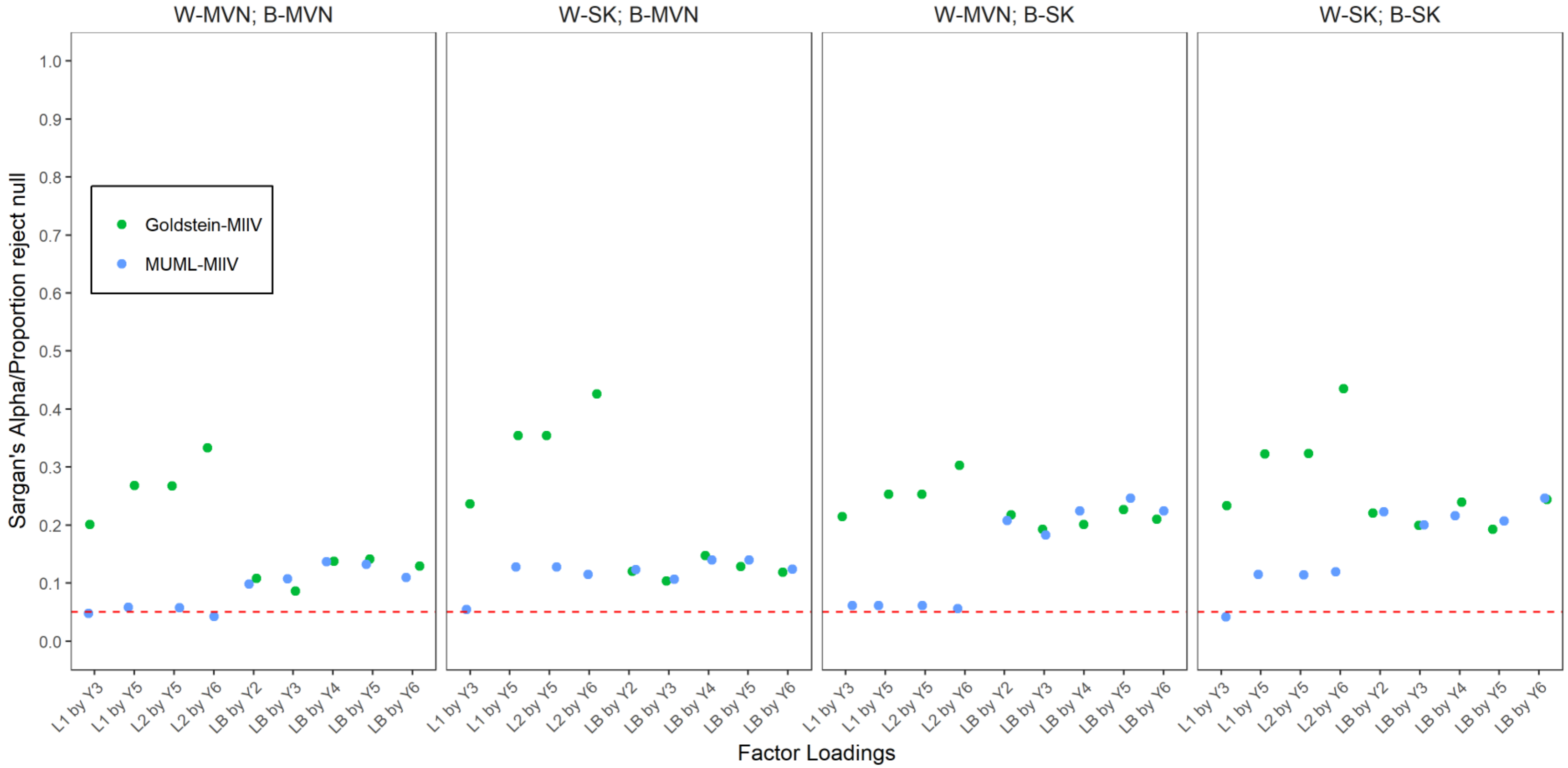

Figure 13. Sargan’s Test Rejection Rates for each factor loading across the Skew/Kurtosis conditions. CN=100 and CS=30, while trends here generalized to other CN/CS conditions. Colors separate different estimators and each cell is a different combination skew/kurtosis condition. All clusters are balanced and data are fit with the true model specification.

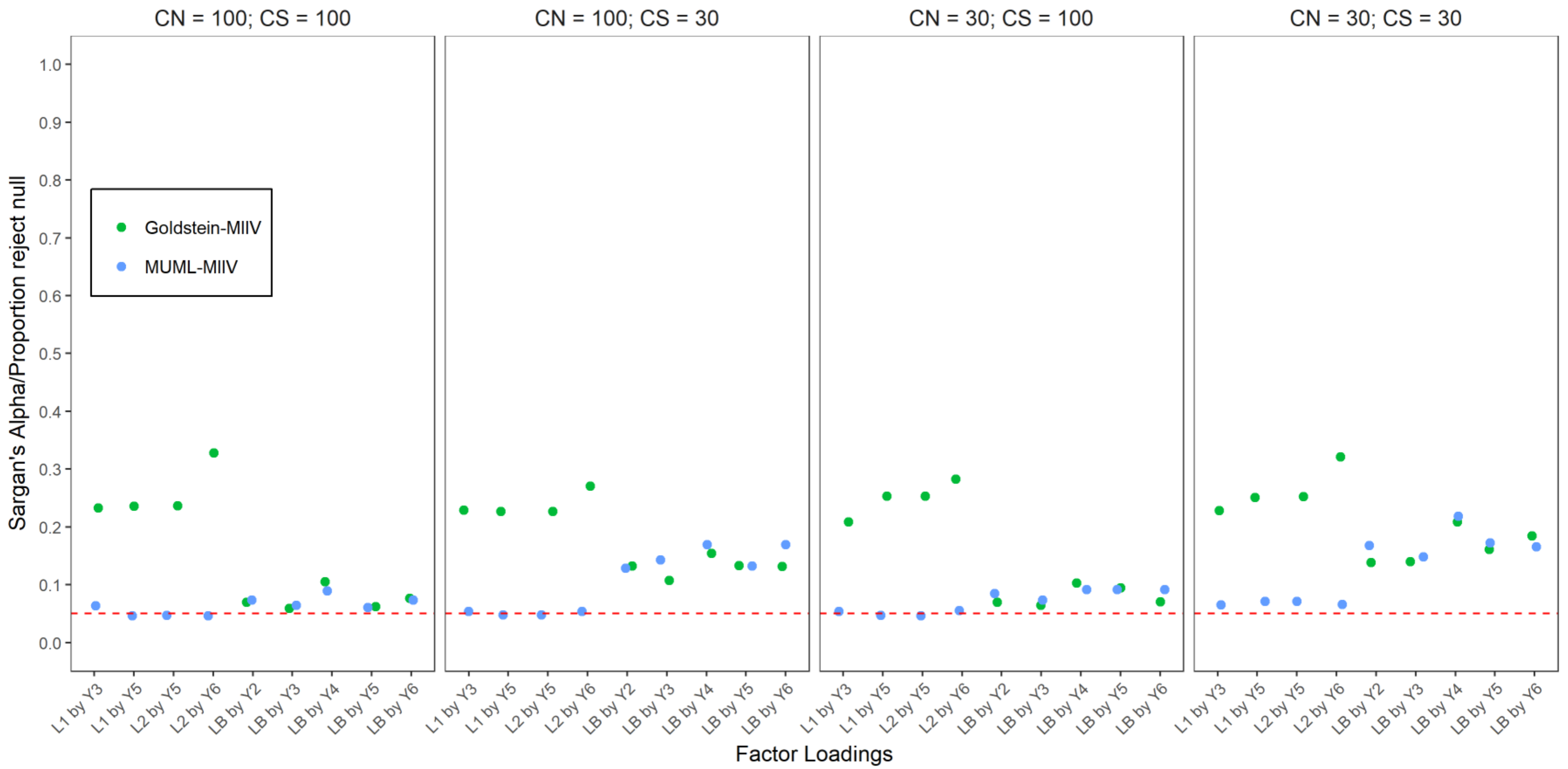

Figure 14. Sargan's Test Rejection Rates for each factor loading in the unbalanced clusters condition. Colors separate different estimators and each cell is a different combination of CN and CS. All data are multivariate normal and fit with the true model specification.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49(2), 155–173.

Anderson, T. W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. Journal of Econometrics, 127(1), 1–16. https://doi.org/10.1016/j.jeconom.2004.09.012

Anderson, T. W., & Rubin, H. (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. The Annals of Mathematical Statistics, 570–582.

Asparouhov, T., & Muthén, B. O. (2003). Full-information maximum-likelihood estimation of general two-level latent variable models with missing data: A technical report. MPLUS working paper.

Asparouhov, T., & Muthén, B. O. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology (pp. 2531–2535).

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica: Journal of the Econometric Society, 77–83.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. Journal of Educational and Behavioral Statistics, 28(2), 135–167.

Bauer, D. J., & Curran, P. J. (2003). Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes. Psychological Methods, 8(3), 338–363. https://doi.org/10.1037/1082-989X.8.3.338

Bauldry, S. (2014). miivfind: A command for identifying model-implied instrumental variables for structural equation models in Stata. Stata Journal, 14(1), 60–75.

Bentler, P. M., & Liang, J. (2003). Two-level mean and covariance structures: Maximum likelihood via an EM algorithm. In S. P. Reise & N. Duan (Eds.), Multilevel Modeling: Methodological Advances, Issues, and Applications. (pp. 53–70). Mahwah, N.J: Lawrence Erlbaum Associates.

Bollen, K. A. (1989). Structural Equations with Latent Variables (1 edition). New York: Wiley-Interscience.

Bollen, K. A. (1996a). A Limited Information Estimator for LiISREL Models with and Without Heteroscedasticity. In G. A. Marcoulides & R. E. Schumacker (Eds.), Advanced Structural Equation Modeling: Issues and Techniques (pp. 227–241). Psychology Press.

Bollen, K. A. (1996b). An alternative two stage least squares (2SLS) estimator for latent variable equations. Psychometrika, 61(1), 109–121. https://doi.org/10.1007/BF02296961

Bollen, K. A. (2001). Two-stage least squares and latent variable models: simultaneous estimation and robustness to misspecifications. In R. Cudeck, K. G. Jöreskog, & D. Sörbom (Eds.), Structural Equation Modeling: Present and Future : a Festschrift in Honor of Karl Jöreskog. Scientific Software International.

Bollen, K. A. (2012). Instrumental Variables in Sociology and the Social Sciences. Annual Review of Sociology, 38(1), 37–72. https://doi.org/10.1146/annurev-soc-081309-150141

Bollen, K. A., & Bauer, D. J. (2004). Automating the Selection of Model-Implied Instrumental Variables. Sociological Methods & Research, 32(4), 425–452. https://doi.org/10.1177/0049124103260341

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent Variable Models Under Misspecification: Two-Stage Least Squares (2SLS) and Maximum Likelihood (ML) Estimators. Sociological Methods &amp; Research, 36(1), 48–86. https://doi.org/10.1177/0049124107301947

Bollen, K. A., Kolenikov, S., & Bauldry, S. (2014). Model-Implied Instrumental Variable—Generalized Method of Moments (MIIV-GMM) Estimators for Latent Variable Models. Psychometrika, 79(1), 20–50.

Bollen, K. A., & Maydeu-Olivares, A. (2007). A polychoric instrumental variable (PIV) estimator for structural equation models with categorical variables. Psychometrika, 72(3), 309.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. Systems under Indirect Observation: Causality, Structure, Prediction, 1, 149–173.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. The Annals of Mathematical Statistics, 907–949.

Cragg, J. G. (1968). Some effects of incorrect specification on the small-sample properties of several simultaneous-equation estimators. International Economic Review, 9(1), 63–86.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude* treatment interaction: Reanalysis of a study by GL Anderson.

Cudeck, R., & Henly, S. J. (1991). Model Selection in Covariance Structures Analysis and the “Problem” of Sample Size: A Clarification. Psychological Bulletin, 109(3), 512–519.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? Multivariate Behavioral Research, 38(4), 529–569.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. Multivariate Behavioral Research, 37(1), 1–36.

Curran, P. J., & West, S. G. (1996). The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis. Psychological Methods, 1 (1).

du Toit, S. H. C., & Toit, M. du. (2008). Multilevel Structural Equation Modeling. In J. de Leeuw & E. Meijer (Eds.), Handbook of Multilevel Analysis (pp. 435–478). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-73186-5_12

Fisher, Z., Bollen, K., Gates, K., & Rönkkö, M. (2016). MIIVsem: Model Implied Instrumental Variable (MIIV) Estimation of Structural Equation Models.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. Psychometrika, 43(4), 521–532.

Fox, J. (1979). Simultaneous Equation Models and Two-Stage Least Squares. Sociological Methodology, 10, 130. https://doi.org/10.2307/270769

Gelman, A., & Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Leiden: Cambridge University Press.

Goldstein, H. (1987). Multilevel models in educational and social research. London: Griffin.

Goldstein, H. (1995). Multilevel statistical models (2nd ed.). London: E. Arnold.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. Psychometrika, 53(4), 455–467.

Hägglund, G. (1982). Factor analysis by instrumental variables methods. Psychometrika, 47(2), 209–222.

Hox, J. J. (2010). Multilevel analysis: techniques and applications (2. ed). New York: Routledge, Taylor & Francis.

Hox, J. J., & Maas, C. (2004). Multilevel structural equation models: The limited information approach and the multivariate multilevel approach. In Recent developments on structural equation models (pp. 135–149). Springer.

Hox, J. J., & Maas, C. J. M. (2001). The Accuracy of Multilevel Structural Equation Modeling With Pseudobalanced Groups and Small Samples. Structural Equation Modeling: A Multidisciplinary Journal, 8(2), 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. Statistica Neerlandica, 64(2), 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x

Julian, M. (2001). The Consequences of Ignoring Multilevel Data Structures in Nonhierarchical Covariance Modeling. Structural Equation Modeling: A Multidisciplinary Journal, 8(3), 325–352. https://doi.org/10.1207/S15328007SEM0803_1

Jöreskog, K. G., & Sörbom, D. (1987). SIMPLIS: Estimating linear structural relationships the easy way using two-stage least squares. Chicago: International Educational Services.

Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Scientific Software International.

Jorgensen, T. D. (2016). semTools: Useful tools for structural equation modeling.

Kamata, A., & Bauer, D. J. (2008). A Note on the Relation Between Factor Analytic and Item Response Theory Models. Structural Equation Modeling: A Multidisciplinary Journal, 15(1), 136–153. https://doi.org/10.1080/10705510701758406

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel Measurement Modeling. In A. A. O'Connell & D. B. McCoach (Eds.), Multilevel Modeling of Educational Data (pp. 345–386). Charlotte, NC, US: Information Age Publishing Inc.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). Dyadic data analysis. New York: Guilford Press.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel Factor Analysis: Reporting Guidelines and a Review of Reporting Practices. Multivariate Behavioral Research, 881–898. https://doi.org/10.1080/00273171.2016.1228042

Kirby, J. B., & Bollen, K. A. (2009). Using Instrumental Variable (IV) Tests to Evaluate Model Specification in Latent Variable Structural Equation Models. Sociological Methodology, 39(1), 327–355.

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. Psychometrika, 57(4), 581–597.

Linhart, H., & Zucchini, W. (1986). Model Selection. New York: Wiley.

MacCallum, R. C. (2003). 2001 Presidential Address: Working with Imperfect Models. Multivariate Behavioral Research, 38(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

MacCallum, R. C., & Tucker, L. R. (1991). Representing Sources of Error in the Common-Factor Model: Implications for Theory and Practice. Psychological Bulletin, 109(3), 502–511.

McDonald, R. P. (1994). The bilevel reticular action model for path analysis with latent variables. Sociological Methods & Research, 22(3), 399–413.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. British Journal of Mathematical and Statistical Psychology, 42(2), 215–232.

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. Educational Psychology Review, 28(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. Psychological Inquiry, 1(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Mehta, P. D. (2013). xxM [Computer software]. Retrieved from http://xxm.times.uh.edu/.

Mehta, P. D., & Neale, M. C. (2005). People Are Variables Too: Multilevel Structural Equations Modeling. Psychological Methods, 10(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Meredith, W., & Tisak, J. (1990). Latent curve analysis. Psychometrika, 55(1), 107–122.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. Psychometrika, 54(4), 557–585.

Muthén, B. O. (1990). Means and covariance structure analysis of hierarchical data. Los Angeles: UCLA Statistics series, #62.

Muthén, B. O. (1994). Multilevel covariance structure analysis. Sociological Methods & Research, 22(3), 376–398.

Muthén, B. O., & Muthén, L. K. (2015). Mplus User's Guide (Seventh Edition). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.), Multilevel Analysis of Educational Data (pp. 87–99). San Diego, CA, US: Academic Press.

Muthén, B. O., & Satorra, A. (1995). Complex Sample Data in Structural Equation Modeling. Sociological Methodology, 25, 267. https://doi.org/10.2307/271070

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. Psychometrika, 81(2), 535–549. https://doi.org/10.1007/s11336-014-9435-8

Nestler, S. (2014a). How the 2SLS/IV estimator can handle equality constraints in structural equation models: A system-of-equations approach. British Journal of Mathematical and Statistical Psychology, 67(2), 353–369. https://doi.org/10.1111/bmsp.12023

Nestler, S. (2014b). Using Instrumental Variables to Estimate the Parameters in Unconditional and Conditional Second-Order Latent Growth Models. Structural Equation Modeling: A Multidisciplinary Journal, 0(0), 1–13. https://doi.org/10.1080/10705511.2014.934948

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2015). Multilevel Structural Equation Models for Assessing Moderation Within and Across Levels of Analysis. Psychological Methods. https://doi.org/10.1037/met0000052

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. Psychological Methods, 15(3), 209–233. https://doi.org/10.1037/a0020141

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

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. Elsevier.

Ryu, E. (2011). Effects of skewness and kurtosis on normal-theory based maximum likelihood test statistic in multilevel structural equation modeling. Behavior Research Methods, 43(4), 1066–1074. https://doi.org/10.3758/s13428-011-0115-7

Ryu, E., & West, S. G. (2009). Level-Specific Evaluation of Model Fit in Multilevel Structural Equation Modeling. Structural Equation Modeling: A Multidisciplinary Journal, 16(4), 583–601. https://doi.org/10.1080/10705510903203466

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. Econometrica: Journal of the Econometric Society, 393–415.

Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments. Quality and Quantity, 24(4), 367–386. https://doi.org/10.1007/BF00152011

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In Latent variables analysis: Applications for developmental research. (pp. 399–419). Thousand Oaks, CA, US: Sage Publications, Inc.

Schmidt, W. H. (1969). Covariance structure analysis of the multivariate random effects model. University of Chicago, Department of Education.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). Variance components (Vol. 391). John Wiley & Sons.

Skrondal, A., & Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. CRC Press.

Snijders, T., & Bosker, R. (1999). Multilevel analysis: An introduction to basic and applied multilevel analysis. London, UK: SAGE Publications Ltd.

StataCorp. (2015). STATA 14 [Computer software]. College Station, TX: StataCorp LP.

Theil, H. (1953a). Estimation and simultaneous correlation in complete equation systems. Central Planning Bureau. The Hague, mimeo.

Theil, H. (1953b). Repeated least squares applied to complete equation systems. The Hague: Central Planning Bureau.

Theil, H. (1954). Estimation of parameters of econometric models. Bulletin of International Statistics Institute, 34, 122–8.

Theil, H. (1961). Economic forecast and policy, vol. XV of Contributions to Economic Analysis. Ed: North-Holland Pub. Co., Amsterdam.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. Psychometrika, 48(3), 465–471.

Van Peet, A. A. (1992). De potentieeltheorie van intelligentie (The potentiality theory of intelligence). PhD dissertation, University of Amsterdam.

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In Handbook of multilevel analysis. (pp. 401–433). New York, NY, US: Springer Science + Business Media. https://doi.org/10.1007/978-0-387-73186-5_11

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (Fourth). New York: Springer.

Yuan, K.-H., & Hayashi, K. (2005). On muthén's maximum likelihood for two-level covariance structure models. Psychometrika, 70(1), 147–167. https://doi.org/10.1007/s11336-003-1070-8