Original Research Article

Educational and Psychological Measurement
1–25

DOI: 10.1177/00131644221080451
journals.sagepub.com/home/epm
SAGE

Croon's Bias-Corrected Estimation for Multilevel Structural Equation Models with Non-Normal Indicators and Model Misspecifications

Kyle Cox[1] and Benjamin Kelcey[2]

Abstract

Multilevel structural equation models (MSEMs) are well suited for educational research because they accommodate complex systems involving latent variables in multilevel settings. Estimation using Croon's bias-corrected factor score (BCFS) path estimation has recently been extended to MSEMs and demonstrated promise with limited sample sizes. This makes it well suited for planned educational research which often involves sample sizes constrained by logistical and financial factors. However, the performance of BCFS estimation with MSEMs has yet to be thoroughly explored under common but difficult conditions including in the presence of non-normal indicators and model misspecifications. We conducted two simulation studies to evaluate the accuracy and efficiency of the estimator under these conditions. Results suggest that BCFS estimation of MSEMs is often more dependable, more efficient, and less biased than other estimation approaches when sample sizes are limited or model misspecifications are present but is more susceptible to indicator non-normality. These results support, supplement, and elucidate previous literature describing the effective performance of BCFS estimation encouraging its utilization as an alternative or supplemental estimator for MSEMs.

Keywords

multilevel structural equation model, Croon's estimation, model misspecification, non-normality

[1]University of North Carolina at Charlotte, USA
[2]University of Cincinnati, OH, USA

Corresponding Author:
Kyle Cox, Educational Research, Measurement, and Evaluation, University of North Carolina at Charlotte, 266 Cato College of Education, Charlotte, NC 28223, USA.
Email: kyle.cox@uncc.edu.

Multilevel structural equation models (MSEMs) are well suited for research in education because they accommodate hierarchical structures (e.g., teachers nested within schools, students nested within classrooms, and principals nested within districts), complex theories relating individuals and groups and variables that are latent (i.e., not directly observable). With large sample sizes, MSEMs have proven effective in evaluating theses complex multilevel systems while accounting for the measurement error associated with latent variables (e.g., Cheung & Lau, 2017; Hox et al., 2010; Li & Beretvas, 2013). However, educational research often has financial and logistical constraints that limit feasible sample sizes. For example, large experimental multilevel studies can be expensive while budgets are limited (e.g., Kelcey & Phelps 2013a, 2013b), and recruiting large sample sizes is difficult (e.g., Autio & Deussen, 2017).

These constraints are somewhat intractable, but several methodological solutions have been proposed in literature. MSEM are typically estimated using maximum likelihood (ML) estimation, but it requires large sample sizes at each level to provide accurate and dependable parameter estimates (e.g., Hox et al., 2010; Li & Beretvas, 2013; Meuleman & Billiet, 2009; van de Schoot & Miocević, 2020). Multilevel path analysis using factor scores (FSs) in place of measurement models (i.e., uncorrected FS approach) provides an alternative method to examine complex multilevel systems connecting latent variables. While this approach typically requires smaller sample sizes, it ultimately provides biased results because it disregards the measurement error associated with the latent variables (e.g., Devlieger et al., 2016; Devlieger & Rosseel, 2017). Croon's bias-corrected factor score (BCFS) path estimation tracks and corrects for the bias introduced in a typical FS path analysis using key measurement model properties (Croon, 2002).

BCFS has shown promise in select settings and conditions. It provided nearly unbiased coefficient estimates for a variety of MSEMs with various cluster and individual per cluster sample sizes (Devlieger & Rosseel, 2019; Kelcey et al., 2021) and outperformed ML in terms of bias, efficiency, convergence rate, and robustness to model misspecification. The performance of BCFS estimation was influenced by sample size, indicator weights, and type of model misspecification. While these results are encouraging, it is important to establish the relative and absolute performance of BCFS estimation of MSEMs under a more complete range of conditions that are common in planned educational research. We expand on work by Kelcey et al. (2021) and Devlieger and Rosseel (2019) by considering BCFS with a fully crossed combination of factors that influence the estimation of MSEMs including non-normal indicators, the number of indicators per factor, model misspecifications, and limited sample sizes. The influence of these factors on BCFS estimation is considered in an MSEM similar to the model employed by Kelcey et al. (2021) but with additional indicators per factor. This multilevel mediation model reflects a complex theory of action examined in an experimental study likely to have a limited sample size (e.g., Schochet, 2011; Spybrook et al., 2016).

We also consider the performance of BCFS estimation with various model misspecifications. Misspecified measurement and structural models are common in social science research, and these misspecifications can bias estimates of MSEM parameters when using ML (e.g., Bollen et al., 2007; Ropovik, 2015). Previous research has found BCFS estimation to be more robust to structural misspecifications and perform similarly or better than ML under misspecified measurement models (Devlieger & Rosseel, 2017, 2019; Hayes & Usami, 2020a; Kelcey et al., 2021). However, it is unclear if these results hold with additional complexity in the measurement model and increasingly severe model misspecifications (e.g., two misspecified indicators in the measurement model).

Therefore, the purpose of this study is to extend understanding of BCFS estimation in MSEMs with limited sample sizes when also facing indicator non-normality or a misspecified model. Specifically, we examine the performance of BCFS estimation with MSEMs when sample size is limited, the number, normality, and factor weights of measurement model indicators vary, and the structural and measurement models include misspecifications. We accomplish this through two simulation studies focusing on the estimation of MSEM path coefficients using Croon's BCFS approach, ML estimation, and an uncorrected FS approach. The first simulation focuses on small sample size conditions and various measurement model indicator complications, while the second simulation includes three types of structural model misspecifications and two types of measurement model misspecifications. Criteria for evaluating estimator performance include convergence failure rate, bias, and efficiency. Preceding the two simulations is a description of BCFS estimation, and the MSEM used in the simulation studies. Following each simulation study are supplemental studies to investigate specific areas of interest and guide future research. To conclude, we summarize results then discuss their implications, limitations, and future research possibilities.

BCFS Path Estimation

Croon's BCFS path estimation has been developed for single-level mediation, single-level moderation, sequential mediation, multilevel mediation, and other MSEMs (e.g., Cox & Kelcey, 2021; Devlieger & Rosseel, 2017, 2019; Kelcey, 2019; Kelcey et al., 2021) and is applicable in a variety of MSEMs including those with more complex multidimensional factor structures, multiple outcome models, models with multiple endogenous variables (e.g., treatments), and models incorporating several of these features (e.g., Devlieger et al. 2019; Hayes & Usami, 2020b). The BCFS process begins with the estimation of factor models for each latent variable and subsequent calculation of FSs. A variance–covariance matrix is constructed using these FSs and values from any observed variables under consideration. Parameter estimates based on this variance–covariance matrix are biased because latent variable unreliability has been disregarded. The correction of BCFS uses results from the measurement models to adjust the variance–covariance matrix to

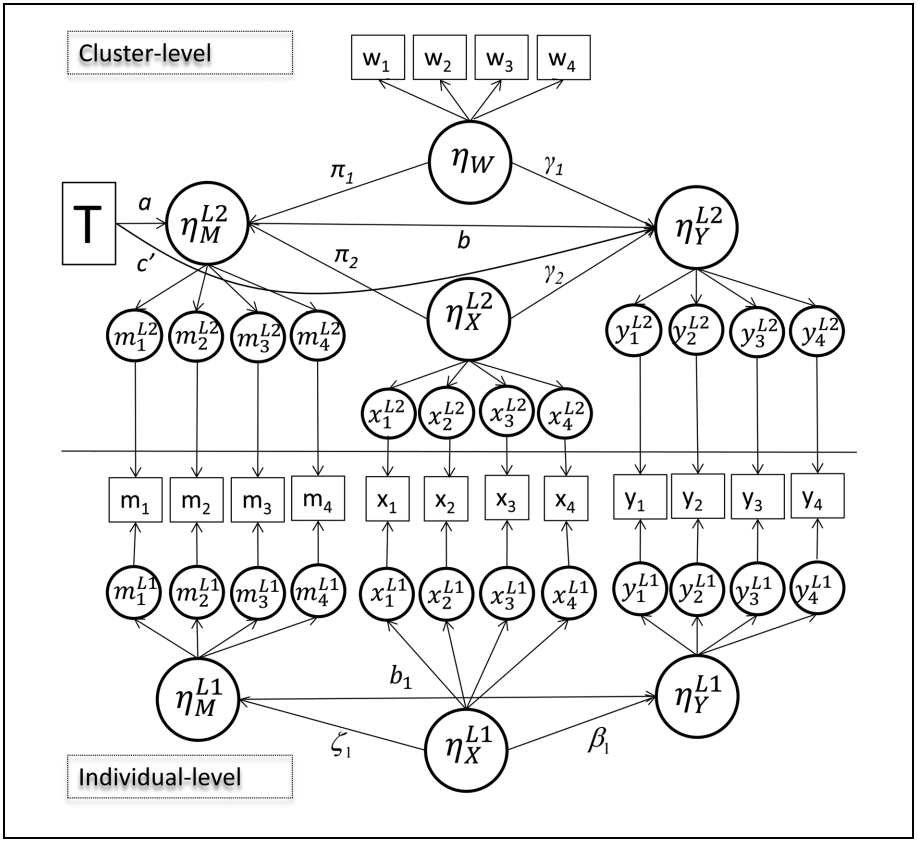

Figure 1. Conceptual Representation of a MSEM With a Manifest Variable (T) and Latent Variables Y, X, W, and M Measured by Four Indicators.
Note. MSEM = multilevel structural equation model.

account for this unreliability. Finally, the corrected variance–covariance matrix is used to estimate the structural model. If inferential testing is of interest, bootstrap methods are applicable and appropriate for determining standard errors with BCFS results (see Kelcey et al., 2021 for details).

In our examination of BCFS, we employ a modified version of the multilevel mediation model seen in Kelcey et al. (2021) with more indicators per latent factor. The modified multilevel mediation model with four indicators per latent factor is illustrated in Figure 1 (adapted from part of Figure 1 by Kelcey et al., 2021). The model includes a treatment (T) that is assigned at the cluster-level and influences an outcome (Y), through an individual-level mediator (M) and includes covariates at the cluster level (W) and individual level (X). BCFS with this model begins with the single-level and multilevel factor models representing the latent variables. The

individual-level latent variables (i.e., η_M, η_Y, and η_X) require a multilevel measurement model to properly account for variance among and within clusters while the cluster-level latent covariate (η_W) only varies between clusters enabling the use of a single-level measurement model. The factor models are estimated using ML with FSs obtained using the regression predictor method.

BCFS utilizes a Croon (2002) method of moments correction to adjust the FS covariance matrix to account for attenuation brought about by additional uncertainty in the FSs. Measurement model error indicates the magnitude of the bias and correction necessary for each latent variable (Croon, 2002). These corrections are the crucial component of BCFS and are required for each variable pairing that includes a latent variable. For example, when estimating the mediation model (see Figure 1), there is a cluster-level covariance between two latent variables measured at the individual-level ($\text{cov}(\eta_Y^{L2}, \eta_M^{L2})$). The measurement step provides the necessary values to determine the covariance between the cluster-level latent variable FSs ($\text{cov}[\tilde{Y}^{L2}, \tilde{M}^{L2}]$). The corrected covariance between the latent variables ($\text{cov}(\eta_Y^{L2}, \eta_M^{L2})$) is a function of ($\text{cov}(\tilde{Y}^{L2}, \tilde{M}^{L2})$ and their FSs and factor loading values (Devlieger et al., 2016; Kelcey et al., 2021). The corrections for other latent variable pairings operate in a similar fashion. Following these corrections, a bias-corrected path analysis is conducted using the corrected covariance matrixes (see Devlieger & Rosseel, 2019 and Kelcey et al., 2021 for a more detailed explanation of BCFS for MSEMs).

Measurement and Structural Models

Our investigation of BCFS utilizes the MSEM depicted in Figure 1 and a similar model with 10 indicators for each latent variable. The single-level common factor model used for the cluster-level latent covariate (η_W) was

$$\mathbf{w}_j = \boldsymbol{\mu}_W + \boldsymbol{\Lambda}_W \boldsymbol{\eta}_{W_j} + \boldsymbol{\varepsilon}_j^W \tag{1}$$

with j indexing clusters, w_j representing observed indicators, Λ_W for factor loadings, μ_W capturing indicator intercepts, with error terms $\boldsymbol{\varepsilon}_j^W$. We fixed the variance of η_{W_j} to one to set the scale, used ML estimation to fit the model, and the regression predictor method to obtain FSs.

A multilevel factor model was utilized for the individual-level latent variables with each decomposed into a cluster- and individual-level components (see configural constructs in Stapleton et al., 2016). For example, M is decomposed into cluster- and individual-level components η_M^{L2} and η_M^{L1} with

$$m_{ij} = \boldsymbol{\mu}_{M_j} + \boldsymbol{\Lambda}_M^{L2} \eta_{M_j}^{L2} + \boldsymbol{\Lambda}_M^{L1} \eta_{M_{ij}}^{L1} + \boldsymbol{\varepsilon}_j^{M^{L2}} + \boldsymbol{\varepsilon}_{ij}^{M^{L1}} \tag{2}$$

where m_{ij} represents the latent variable indicators of the mediator for individual i in cluster j, with cluster- and individual-level factor loadings $\boldsymbol{\Lambda}_M^{L2}$ and $\boldsymbol{\Lambda}_M^{L1}$, intercepts $\boldsymbol{\mu}_{M_j}$, and cluster- and individual-level error terms $\boldsymbol{\varepsilon}_j^{M^{L2}}$ and $\boldsymbol{\varepsilon}_{ij}^{M^{L1}}$. The variance of the cluster- and individual-level factors is again set to one to establish the scale. Parallel

model formulations were utilized for the other latent variables requiring multilevel factor models (η_X^{L2} and η_X^{L1} for X and η_Y^{L2}, and η_Y^{L1} for Y).

We connect the latent and manifest variables under consideration (i.e., T, W, X, M, and Y) with multilevel structural models (e.g., Preacher et al., 2010) such that the path model for M is

$$\begin{aligned} \eta_{M_{ij}}^{L1} &= \zeta_0 + \zeta_1 \eta_{X_{ij}}^{L1} + \varepsilon_{ij}^{M} \quad \varepsilon_{ij}^{M} \sim N(0, \sigma_{M|}^2) \\ \eta_{M_j}^{L2} &= \pi_0 + aT_j + \pi_1 \eta_{W_j} + \pi_2 \eta_{X_j}^{L2} + u_j^{M} \quad u_j^{M} \sim N(0, \tau_{M|}^2) \end{aligned} \tag{3}$$

At the individual-level ($\eta_{M_{ij}}^{L1}$), ζ_0 represents the intercept with an individual-level error term of ε_{ij}^{M}, the only path coefficient, ζ_1, captures the relationship between the mediator and the individual-level component of the individual-level covariate ($\eta_{X_{ij}}^{L1}$). At the cluster-level, π_0 represents the intercept with a random effect of u_j^{M}, the treatment indicator for each cluster is T_j, the path coefficient, a, captures the relationship between the treatment and mediator (M), the remaining path coefficients at the cluster-level, π_1 and π_2, capture the relationship between the mediator and latent cluster-level covariate (η_{W_j}) and the cluster-level component of the individual-level covariate ($\eta_{X_j}^{L2}$).

The corresponding multilevel structural model for the outcome, Y, is similar with

$$\begin{aligned} \eta_{Y_{ij}}^{L1} &= \beta_0 + b_1 \eta_{M_{ij}}^{L1} + \beta_1 \eta_{X_{ij}}^{L1} + \varepsilon_{ij}^{Y} \quad \varepsilon_{ij}^{Y} \sim N(0, \sigma_{Y|}^2) \\ \eta_{Y_j}^{L2} &= \gamma_{00} + b\eta_{M_j}^{L2} + c' T_j + \gamma_1 \eta_{W_j} + \gamma_2 \eta_{X_j}^{L2} + u_j^{Y} \quad u_j^{Y} \sim N(0, \tau_{Y|}^2) \end{aligned} \tag{4}$$

β_0 represents the intercept at the individual-level ($\eta_{Y_{ij}}^{L1}$), the b_1 coefficient captures the relationship between the individual-level component of M ($\eta_{M_{ij}}^{L1}$) and the outcome while the β_1 coefficient captures the relationship between the individual-level component of the individual-level latent covariate ($\eta_{X_{ij}}^{L1}$) and the outcome, and the error term at the individual-level error is ε_{ij}^{Y}. At the cluster-level, γ_{00} represents the intercept with a cluster-level random effect of u_j^{Y}, b is paired with the cluster-level component of M ($\eta_{M_j}^{L2}$) capturing its relationship to the outcome, the direct effect of the treatment on the outcome is represented by c', coefficients paired with the cluster-level latent covariate (η_{W_j}) and the cluster-level component of the individual-level covariate ($\eta_{X_j}^{L2}$) are γ_1 and γ_2. Note that in our multilevel structural models for both the mediator and outcome, we do not specify random slopes. This is a necessary restriction because BCFS cannot currently accommodate MSEMs with random slopes.

Simulation Study I: Limited Sample Size and Measurement Model Factors

The first simulation study seeks to better understand BCFS performance when sample size is limited and the number, normality, and factor weights of measurement

model indicators vary. We estimate the MSEM with BCFS, ML, and an uncorrected FS approach and capture convergence failure rate, bias, and efficiency of coefficient estimates as performance criteria. ML is a prevalent estimation approach often serving as the default for MSEM (e.g., *Mplus* and the *lavaan* package in *R*). As a full information method, it estimates the measurement and structural components of the MSEM simultaneously. While ML is both efficient and consistent, its simultaneous approach to highly parameterized MSEMs has been found to produce biased estimates and high convergence failure rates when sample sizes are limited (Rosseel, 2020). An alternative to ML is the FS approach which estimates the measurement models and then the structural components similar to BCFS but does not include a correction step to address measurement error related to any latent variables.

Data Generation

Data for the first simulation were generated based on the depiction in Figure 1, the structural models in Equations 3 and 4, and the measurement models in Equations 1 and 2. We began by generating the manifest treatment indicator T_j coded as ± 0.5 and the exogenous latent covariates η_W and η_X such that ~$N(0,1)$. The intraclass correlation coefficients of M and Y were set to be equal and included two conditions $\rho = 0.2$ and 0.4. The remaining endogenous latent variables were generated using Equations 3 and 4 with $a = 0.5$, $b = 0.4$, $b_1 = 0.1$, $c' = 0.1$, $\pi_1 = 0.2$, $\pi_2 = 0$, $\gamma_1 = 0.3$, $\gamma_2 = 0$, $\zeta_1 = 0.2$, and $\beta_1 = 0.2$. We based our limited sample sizes on previous simulation literature (e.g., Devlieger & Rosseel, 2019; Kelcey et al., 2021) with cluster sample sizes of $n_2 = 100,\ 80,\ 50,\ 30,\ 20,$ and 10 and individual per cluster sample sizes of $n_1 = 80,\ 40,\ 20,\ 10,$ and 5.

We varied three measurement model factors with these limited sample sizes: the number of indicators per latent variable, indicator weights, and distribution of indicator residuals. We included four indicators per latent trait reflecting typical values from past simulation literature (e.g., Devlieger et al., 2016; Devlieger & Rosseel, 2019; Kelcey, 2019; Kelcey et al., 2021) and also 10 indicators per latent trait to examine estimator performance under a novel large number of indicators condition. The measurement model had three conditions for indicator weights because they have been shown to influence estimator performance in MSEMs (e.g., Devlieger & Rosseel, 2019; Kelcey et al., 2021). The first condition was a mix of 1.0, 0.666, and 1.5, while the second and third conditions had uniform indicator weights of 0.666 and 1.5, respectively. We imposed cross-level invariance on the factor loadings (e.g., $\mathbf{\Lambda}_M^{L2}$ and $\mathbf{\Lambda}_M^{L1}$) by setting indicator weights at the cluster- and individual-level equal with variability of cluster-level components set at 0.2 and individual-level components at 0.8 such that ICC(1) = 0.2 (see Shrout & Fleiss, 1979; Stapleton et al., 2016).

We also considered non-normal distributions for the indicator residuals by generating residuals with distributions having skewness of 2 and kurtosis of 7 (Fleishman, 1978). Considering non-normal indicator distributions while varying other

measurement model factors (e.g., indicators per factor and factor weights) properly reflects the conditions in which BCFS is likely to be applied (e.g., Blanca et al., 2013; Micceri, 1989). Literature has indicated ML is somewhat robust to indicator non-normality, but it is unclear if these results hold under the conditions considered here and extend to BCFS with MSEMs (e.g., Lei & Lomax, 2005).

To summarize, we conducted a simulation study to investigate BCFS, ML, and FS estimation of MSEMs when sample sizes are limited and several influential factors vary. Performance criteria included convergence rate, bias, and efficiency. We considered novel conditions with non-normal indicator residuals and 10 indicators per latent trait while also developing a more comprehensive understanding of BCFS performance in MSEMs by employing a fully crossed design with more than 720 conditions. In total, 1000 data sets were generated for each condition using *R* (R Core Team, 2019) with estimation conducted in the *lavaan* (Rosseel, 2012) and *lme4* (Bates et al., 2015) packages.

Results

Select results are presented by performance criteria with sections below detailing convergence rate, bias, and efficiency. The general results were not substantially influenced by the intraclass correlation coefficient of *M* and *Y*, so we focus on the $\rho = 0.2$ condition. Given the variety of conditions, we rely on several figures to efficiently communicate these results. Unless specifically identified, the default conditions for Simulation study 1 figures were four indicators per latent trait, mixed indicator weights (i.e., a mix of 1.0, 0.666, and 1.5), indicator residuals with normal distributions, correctly specified MSEMs, intraclass correlation coefficients of *M* and *Y* of,$\rho = 0.2$, and 40 individuals per cluster ($n_1 = 40$). Complete results and simulation code are available upon request.

Convergence Rate. We tracked the convergence rate or more specifically the failure of an estimation method to provide a solution across all conditions. Convergence failure rates increased for each estimator as sample sizes decreased. This relationship is illustrated in Figure 2 by the black markers identifying the convergence failure rate of each estimation approach by cluster sample size with different individual per cluster sample sizes (Figure 2A–C). Any condition missing a black marker indicates the estimation approach failed to converge in more than 50% of the replications. Convergence failure rate was influenced by the sample of individuals per cluster (n_1) but driven more by the sample of clusters (n_2) such that all approaches had unacceptable convergence failure rates (e.g., $> 25\%$) at the lowest cluster sample size considered ($n_2 = 10$).

While the relationship between sample size and convergence failure rate was similar for each estimator, the convergence failure rate did vary substantially. BCFS had the smallest convergence failure rate in almost every condition followed closely by the FS approach with ML often incurring the largest convergence failure rate. It

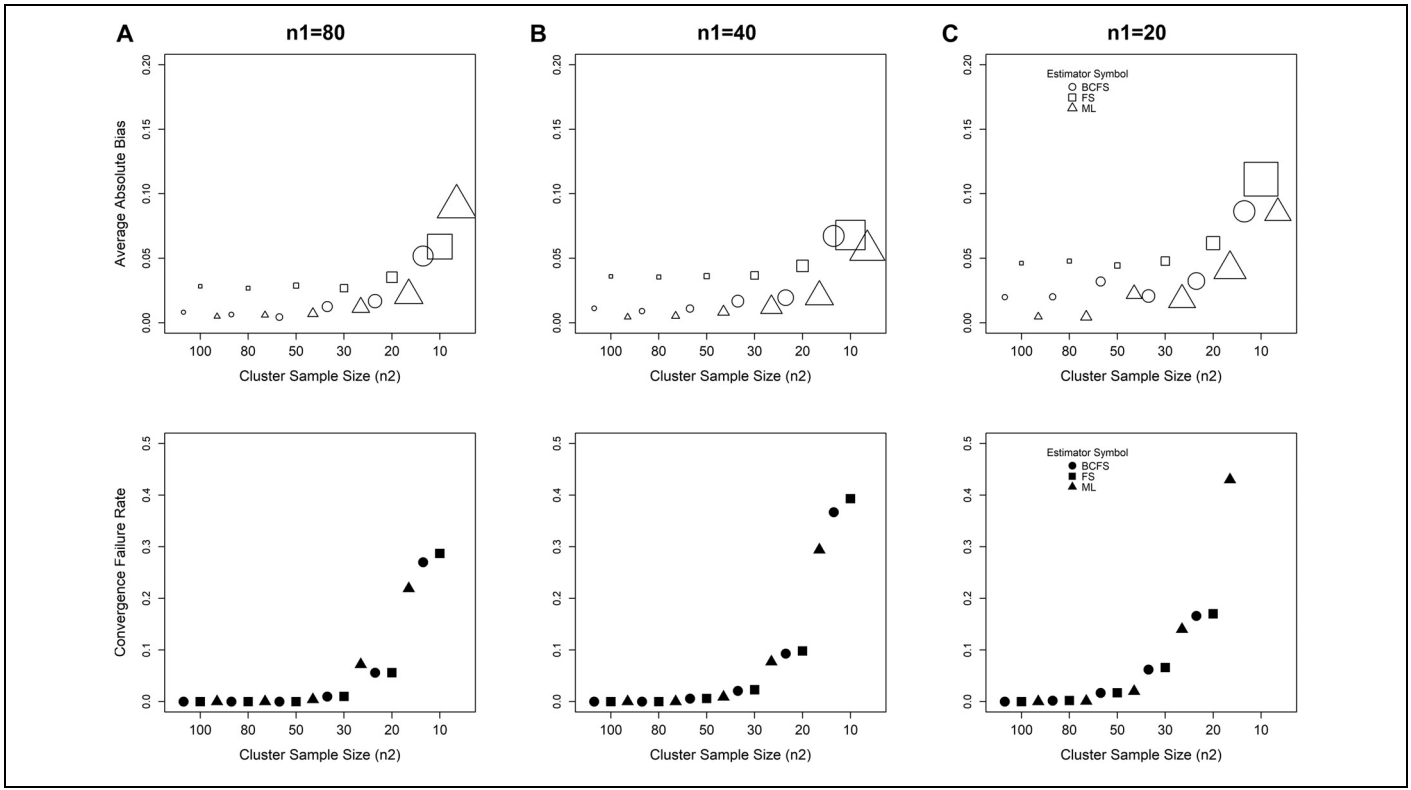

Figure 2. Bias and Convergence Failure Rate for Each Estimator by Cluster Sample Size With (A) 80, (B) 40, and (C) 20 Individuals Per Cluster, Four Indicators Per Latent Trait and Mixed Indicator Weights.

Note. Results in this figure were based on normally distributed indicator residuals. The size of each point marking bias reflects the average *SD* of the coefficient estimates for that estimator under the model and conditions indicated. A larger *SD* results in larger points on the plot with smaller *SD*s producing smaller points.

should be noted that the convergence failure rate for ML far exceeded BCFS and FS approaches especially when $n_2 \leq 30$. For example, with a sample size of $n_2 = 20$ when $n_1 = 80$, ML failed to converge over 20% of the time while BCFS failed to converge just over 5% of the time (see the black markers in Figure 2A).

Convergence failure rate was also influenced by indicator weights (see Figure 3) and number of indicators per latent variable (see Figure 4) but demonstrated no qualitative differences when the intraclass correlation coefficient of the outcome and mediator varied. Larger indicator weights decreased convergence failure rates especially with ML. For example, the black markers in Figure 3B indicate convergence failure rates diverge from zero with as many as 50 clusters when using ML. Conversely, in Figure 3C, the convergence failure rates remained near 0 for all three estimators with cluster sample sizes around 20 and individual per cluster sample sizes of 40.

Increasing the number of indicators for each latent variable also helped reduce convergence failure rates (see Figure 4). Reading Figure 4 from left to right, we see reduced convergence failure rates when using 10 indicators. This benefit is especially pronounced for ML. However, this relationship can be influenced by non-normality of the indicators. Reading Figure 4 from top to bottom, non-normal indicator residuals had little effect on convergence rate in the four indicator conditions (compare Figure 4A–C) but generally led to higher convergence failure rates in the 10-indicator condition (compare Figure 4B–D). Unlike the previous factors under consideration, non-normality in indicator residuals had a greater detrimental effect on BCFS and FS approaches. It is important to note that convergence failure rates for the BCFS and FS approach generally increased only a few percentage points.

Bias. We tracked bias in the estimates of all structural coefficients as the difference between the estimated value and the true coefficient value in the generated data set. Using these bias values, we calculated the average absolute bias for each estimation approach across a 1000 data sets. Focusing on those conditions with reasonable convergence rates, we found bias increased as sample size decreased for all three estimators. Within this result we found, reductions in cluster sample size were more detrimental to estimator accuracy than reductions in individual per cluster sample size.

In Figure 2, white markers indicate the average absolute bias of the estimation approach for each condition and are sized based on the standard deviation of the estimates with larger markers indicating a greater *SD* or less precision (subsequently detailed). Bias for each estimator actually held relatively steady when cluster sample size was greater than 50 with BCFS and ML performing similarly and the FS approach typically incurring the most bias. However, with extremely small individual per cluster sample sizes ($n_1 = 5$ or 10), each estimator saw steep increases in bias as cluster sample sizes fell below $n_2 = 80$. With cluster sample sizes less than 50 and individual per cluster sample sizes greater than 10, several factors influenced estimator bias including indicator weight, and number of indicators per latent variable.

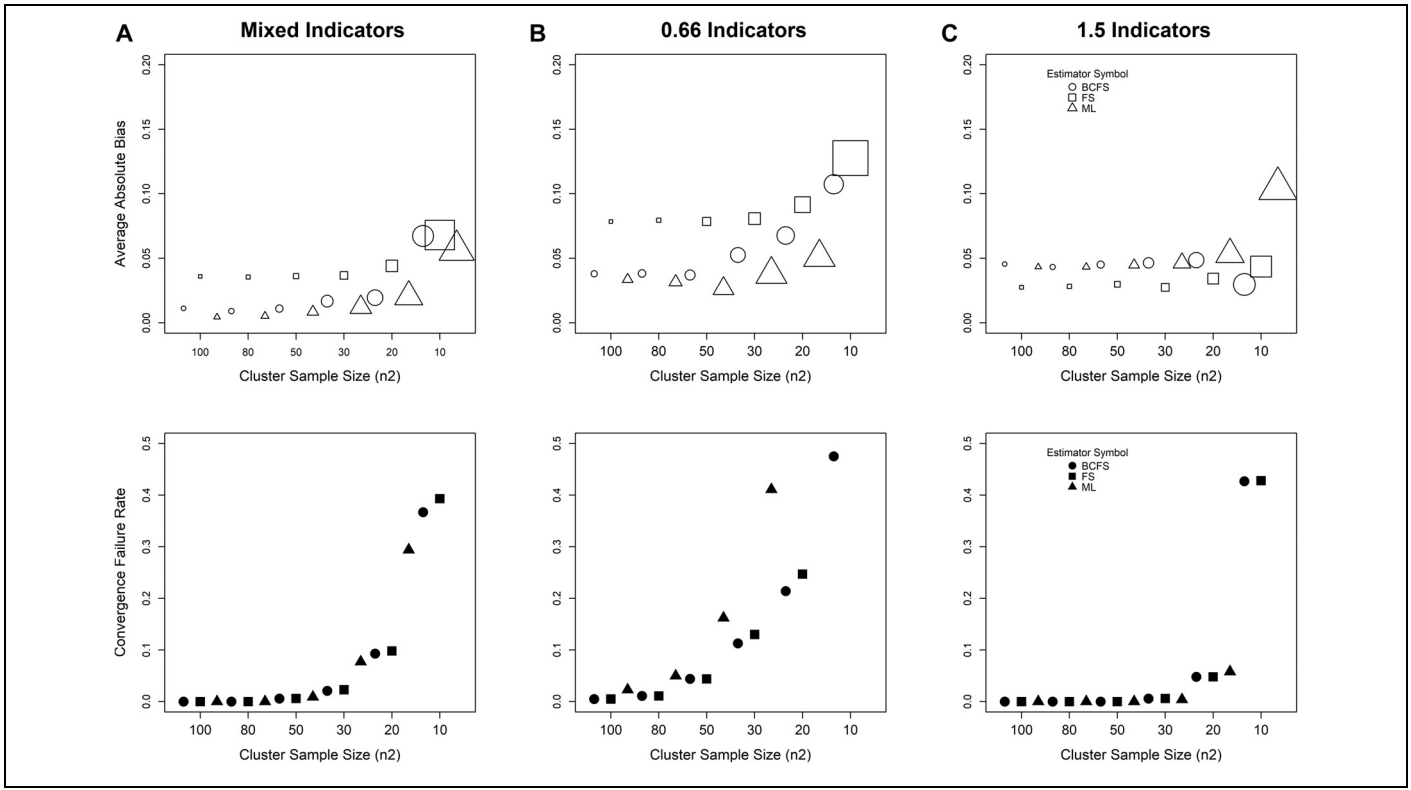

Figure 3. Bias and Convergence Failure Rate for Each Estimator by Cluster Sample Size With (A) Mixed Indicator Weights; (B) Indicator Weights of 0.66; and (C) Indicator Weights of 1.5, Four Indicators Per Latent Trait and 40 Individuals Per Cluster.

Note. The size of each point marking bias reflects the average *SD* of the coefficient estimates for that estimator under the model and conditions indicated. A larger *SD* results in larger points on the plot with smaller *SD*s producing smaller points.

The relationship between estimator bias and indicator weights was somewhat nebulous. Under the small indicator weight condition, each estimator suffered increased bias (see Figure 3). However, estimators generally incurred more bias in the large indicator weight condition when compared to the mixed indicator condition. The exception being the FS approach which was the most susceptible to influence by indicator weights. For example, in the smaller indicator weight condition, the FS approach suffered the greatest bias, but under the large indicator weight condition, the FS approach suffered the least bias. The relationship between estimator bias and the number of indicators and the distribution of their residuals was more apparent (see Figure 4). Increasing from four indicators per latent variable to 10 reduced bias—albeit minimally—while non-normal indicator residuals caused a slight increase in bias for each estimator.

Interpretation of results involving bias requires consideration of convergence rates and error variance of the estimates. Results suggest BCFS and ML often achieve similar levels of bias with the FS approach generally incurring more bias. The strong performance of ML in terms of bias is often overshadowed by high convergence failure rates in the same conditions. For example, in Figure 2B when $n_2 = 20$ and 10, ML seems to outperform BCFS in terms of bias but under these conditions ML rarely converged (e.g., convergence failure rate of 97% and 99%, respectively). Results involving estimator efficiency further call into question the performance of ML in terms of bias.

Efficiency. We tracked the standard deviation (*SD*) of path coefficient estimates across the 1000 data sets to understand the efficiency of each estimator. All of the estimators were less efficient with smaller sample sizes (e.g., increased *SD*). ML performed relatively well in terms of bias, but it consistently had the largest *SD* of estimates. Conversely, the FS approach incurred the most bias but consistently had the smallest *SD* results. The BCFS approach tended to balance these considerations. It had relatively small amounts of bias and outperformed ML in terms of *SD* of estimates, but it was typically less efficient than the FS approach.

The results involving *SD* of path coefficient estimates are illustrated in Figures 2 to 4 by the size of the symbols plotted for bias. For example, in Figure 4C when the sample cluster size is 10, we see that the triangle marking bias for the estimates using ML is at the highest point on the figure and is larger than the circle and square marking bias for BCFS and FS, respectively. This indicates that under these conditions, ML suffered the largest amount of bias and was the least efficient. As for other factors, smaller indicator weights resulted in small decreases to efficiency across estimators while increasing the number of indicators per latent variable increased efficiency of each estimator. Interestingly, non-normality in indicator residuals had no discernible influence on coefficient estimator efficiency. Reading Figure 4 from top to bottom, we see the size of each marker maintain the same size when moving from a normal distribution of indicator residuals to the non-normal condition.

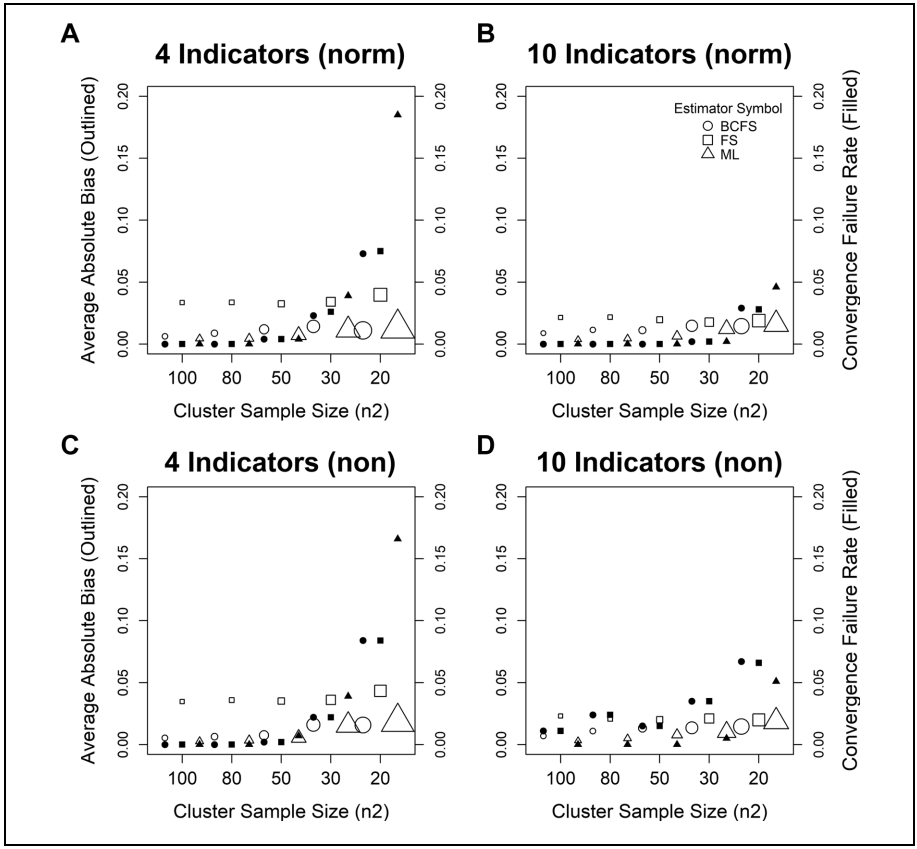

Figure 4. Bias and Convergence Failure Rate for Each Estimator by Cluster Sample Size With Differing Numbers of Indicators and Normal (Norm) and Non-Normal (Non) Residuals, Mixed Indicator Weights, and 40 Individuals Per Cluster.

Note. The size of each point marking bias reflects the average *SD* of the coefficient estimates for that estimator under the model and conditions indicated. A larger *SD* results in larger points on the plot with smaller *SD*s producing to smaller points.

Simulation Study I: Supplemental Investigations

The conditions considered in any simulation study are necessarily limited. To expand the scope of this work and better guide future research, we considered three supplemental conditions representing initial investigations into areas of interest. To supplement the first simulation study, we considered multilevel measurement invariance (see Jak et al., 2013; Stapleton et al., 2016) through a scenario in which the assumption of cross-level invariance on the factor loadings is not met. We set different indicator weights for latent variable components on the within and between levels.

Specifically, we set four within-level indicator weights as a mix of 1.0, 0.666, and 1.5, while the four between-level indicator weights were a uniform 0.666. Other conditions and the MSEM from the first simulation study were retained. We examined this cross-level invariance condition with $n_2 = 100, 80, 50, 30, 20,$ and 10 and an individual per cluster sample size of $n_1 = 80$. We also examined the cross-level invariance scenario with non-normal distributions for the indicator residuals.

Estimating MSEMs with multilevel measurement noninvariance was found to be more difficult (see the Supplement: noninvariance condition in Table 1 for select results). All of the estimation methods incurred more bias, higher convergence failure rates, and larger *SD*s of estimates when compared to similar simulation results with cross-level invariance (Simulation I: normal condition in Table 1). However, these increases were all relatively minor. For example, convergence failure rates began to increase in the $n_2 = 50$ condition as opposed to $n_2 = 30$. Relative estimator performance remained similar with BCFS still outperforming or paralleling FS and ML approaches with multilevel measurement noninvariance. The combination of non-normal data with multilevel measurement noninvariance lead to a slight deterioration in estimator performance but still did not change relative estimator performance.

Our second supplemental simulation condition related to Simulation study I focused on non-normal data. Specifically, we considered indicator residuals sampled from a bimodal distribution to form a different type of non-normal data. We considered four and 10 indicators per latent variable with a mix of indicator weights, and $n_2 = 100, 80, 50, 30, 20,$ and 10 and an individual per cluster sample size of $n_1 = 80$. The MSEM and other parameter values were retained from the first simulation study. Estimator performance was very similar with both types of non-normal data (i.e., bimodal verse non-normal with skewness of 2 and kurtosis of 7) with select results presented in Table 1.

Simulation Study II: Limited Sample Sizes and Misspecified Models

The second simulation study again considered bias, efficiency, and convergence failure rates for BCFS, ML, and FS estimation but under a misspecified MSEM. BCFS has been considered with misspecified MSEMs with encouraging initial results (Devlieger & Rosseel, 2019; Hayes & Usami, 2020a; Kelcey et al., 2021). This simulation develops a more comprehensive understanding of BCFS performance in the presence of model misspecifications by considering two measurement model misspecifications and three structural model misspecifications (see Figure 5). Simulation study conditions are again guided by previous literature (e.g., Devlieger & Rosseel, 2019; Kelcey et al., 2021) and adapted based on the first simulation study. For example, sample sizes of $n_2 = 10$, and often $n_2 = 20$, were not feasible for any of the estimators with correctly specified models so they were excluded from the second simulation study. We also found that as n_1 exceeded 20, there was relatively little change in estimator performance so to match previous literature and avoid excessive

Table 1. Select Results From Simulation Study I and Supplemental Simulation Conditions.

Simulation condition	n_2	Convergence failure rate			Average absolute bias			Average *SD* of estimates		
		BCFS	FS	ML	BCFS	FS	ML	BCFS	FS	ML
Simulation I:	100	0.00	0.00	0.00	0.01	0.03	0.00	0.09	0.07	0.10
Normal	80	0.00	0.00	0.00	0.01	0.03	0.01	0.11	0.08	0.11
	50	0.00	0.00	0.00	0.00	0.03	0.01	0.15	0.11	0.16
	30	0.01	0.01	0.07	0.01	0.03	0.01	0.22	0.16	0.28
	20	0.06	0.06	0.22	0.02	0.04	0.02	0.29	0.22	0.45
	10	0.27	0.29	0.96	0.05	0.06	0.09	0.43	0.53	0.62
Supplement:	100	0.00	0.00	0.00	0.01	0.08	0.05	0.13	0.07	0.17
Noninvariance	80	0.01	0.01	0.02	0.05	0.08	0.05	0.16	0.09	0.24
	50	0.04	0.04	0.06	0.05	0.08	0.05	0.21	0.17	0.33
	30	0.12	0.13	0.22	0.06	0.09	0.05	0.28	0.22	0.52
	20	0.17	0.23	0.43	0.06	0.08	0.05	0.34	0.32	0.66
	10	0.39	0.53	1.00	0.10	0.10	0.17	0.37	0.63	0.37
Simulation I:	100	0.00	0.00	0.00	0.01	0.03	0.00	0.09	0.07	0.10
Non-normal	80	0.00	0.00	0.00	0.01	0.03	0.00	0.11	0.08	0.11
	50	0.00	0.00	0.00	0.01	0.03	0.01	0.15	0.11	0.17
	30	0.01	0.01	0.06	0.01	0.03	0.01	0.22	0.15	0.28
	20	0.05	0.06	0.22	0.02	0.04	0.03	0.29	0.22	0.43
	10	0.24	0.27	0.98	0.05	0.06	0.12	0.44	0.56	0.54
Supplement:	100	0.00	0.00	0.00	0.01	0.03	0.00	0.09	0.07	0.10
Bimodal	80	0.00	0.00	0.00	0.00	0.03	0.00	0.11	0.08	0.11
	50	0.00	0.00	0.00	0.01	0.03	0.01	0.14	0.11	0.17
	30	0.02	0.02	0.03	0.01	0.03	0.01	0.21	0.15	0.34
	20	0.05	0.05	0.14	0.01	0.03	0.04	0.27	0.20	0.50
	10	0.21	0.27	0.96	0.06	0.05	0.05	0.42	0.53	0.69

Note. Conditions for these results include $n_1 = 80$, four indicators per latent trait with a mix of indicator weights, and intraclass correlation coefficients of *M* and *Y* set at 0.2. All other simulation parameters match those described for Simulation Study I. BCFS = bias-corrected factor score; FS = factor score; ML = maximum likelihood.

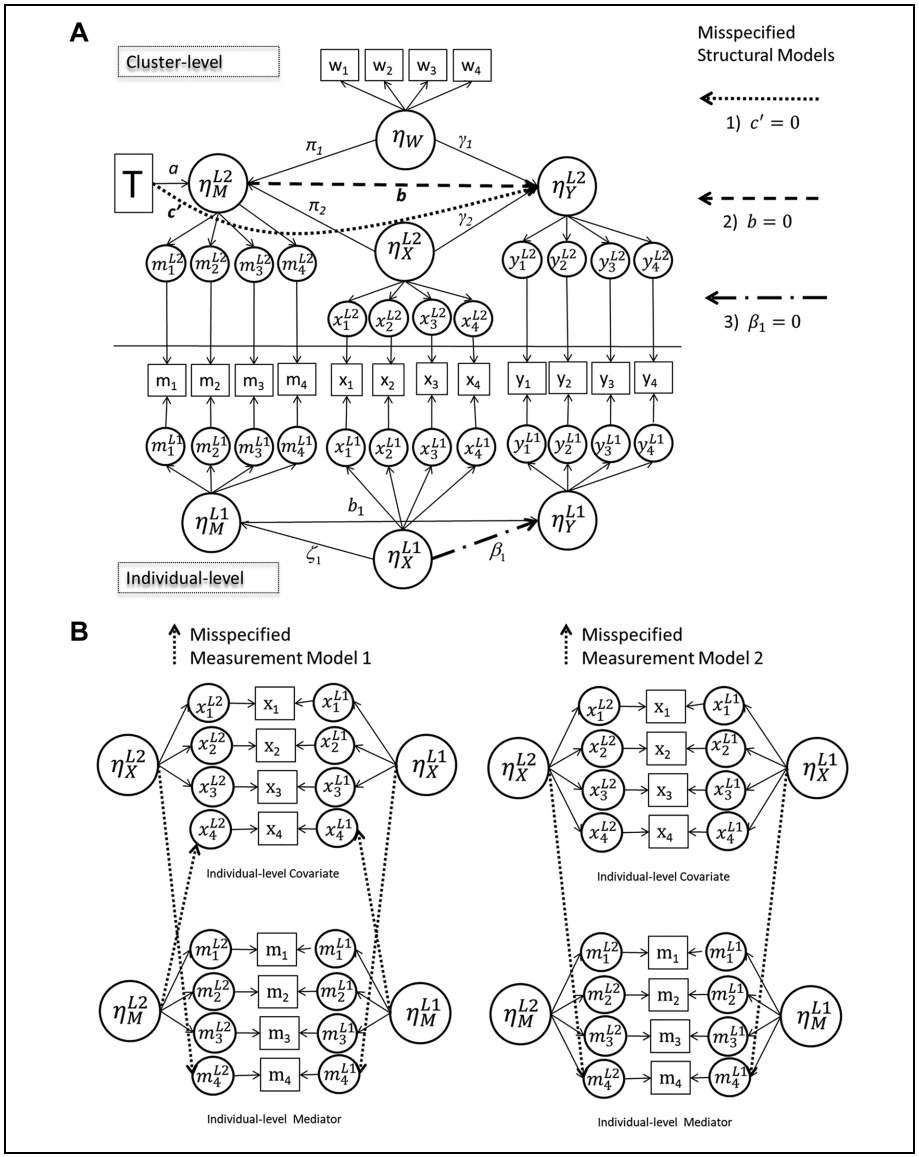

Figure 5. Conceptual Representation of (A) Structural and (B) Measurement Model Misspecifications in a MSEM With a Manifest Variable (*T*) and Latent Variables *Y*, *X*, *W*, and *M* Measured by Four Indicators.

Note. MSEM = multilevel structural equation model.

simulation conditions, we set two small n_1 conditions. The structural and measurement models from the first simulation are employed with model coefficient values of $a = 0.5$, $b = 0.4$, $b_1 = 0.1$ $c' = 0.1$, $\pi_1 = 0.2$, $\pi_2 = 0$, $\gamma_1 = 0.3$, $\gamma_2 = 0$, $\zeta_1 = 0.2$ and

$\beta_1 = 0.2$ and the intraclass correlation coefficients of the mediator and outcome set at 0.2. We include cluster sample sizes of $n_2 = 200, 100, 50,$ and 30 and the individual per cluster sample sizes of $n_1 = 20$ and 5. As for measurement conditions, four indicators were assigned to each latent factor with the three weight conditions noted in the first simulation.

The first measurement misspecification involved swapped indicators in which x_4 was set to measure η_M while indicator m_4 was set to measure η_X (see Figure 5B). The second measurement misspecification involved a missing cross-loading. Data for this analysis were generated with m_4 loading on both η_M and η_X while analytic model matched the original measurement model (see Figure 1 and Equation 2). The misspecified structural models all involve a missing path but at various levels and with varying coefficient magnitudes. Specifically, $c' = 0$, $b = 0$, and $\beta_1 = 0$ in the first, second, and third misspecified structural models, respectively (see Figure 5A). We generated 1000 data sets for each of the 24 conditions described and analysed these data using six MSEM specifications: the true MSEM (i.e., correctly specified model), two models with measurement misspecifications, and three models with structural misspecifications.

Results

We again start by considering the convergence failure rate of each estimation approach. These results along with bias and efficiency results are illustrated in Figure 6 with $n_2 = 50$ and $n_1 = 20$ across different indicator weight conditions. Similar to the previous simulation, the convergence failure rate increased as both cluster and individual per cluster sample size decreased. This was true for each estimator and under each model specification. Overall, we found BCFS generally had the lowest convergence failure rate followed closely by the FS approach (see black markers in Figure 6). Conversely, the convergence failure rate for ML was $> 25\%$ in several conditions. Differences in estimator performance dissipated and eventually became indistinguishable as sample size increased.

The influence of model misspecification type on convergence failure rate varied by estimation approach. Overall, the measurement misspecifications undermined estimator convergence more than structural misspecifications (see Mea1 and Mea2 in Figure 6). The limited information approaches (e.g., BCFS and FS) were fairly robust to all three types of structural misspecifications as their convergence failure rates under the true model and misspecified structural models were nearly equal. Measurement misspecifications substantially increased convergence failure rates across all estimators but the BCFS approach was the most robust to these misspecifications. There were specific conditions in which the structural misspecifications actually reduced the convergence failure rate of ML. For example, the convergence failure rate for ML was less under the structural misspecifications with $c' = 0$ and $b = 0$ (i.e., str1 and str2) when indicator weights were 0.66 compared to its

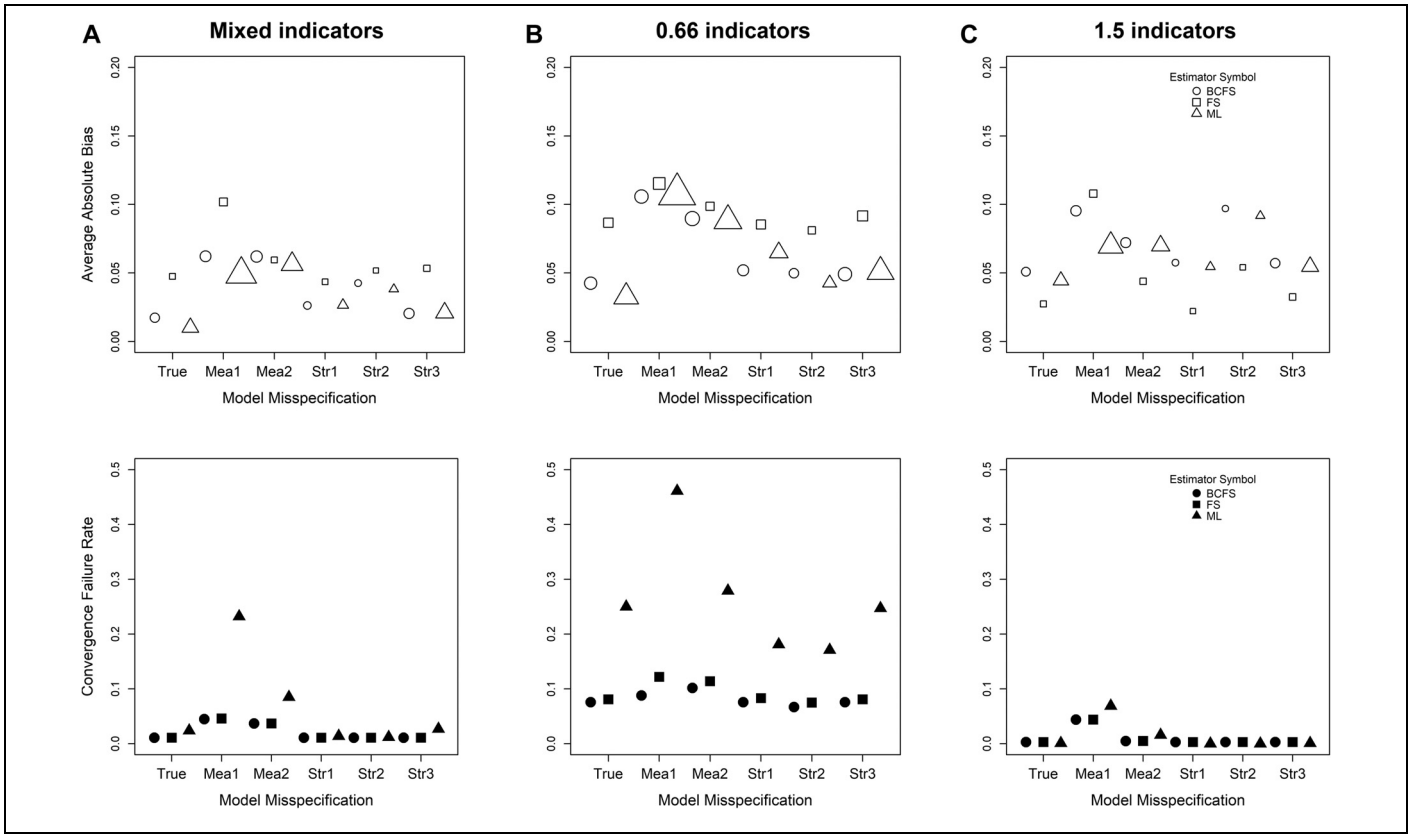

Figure 6. Bias and Convergence Failure Rate for Each Estimator by Model Misspecification With (A) Mixed Indicator Weights; (B) Indicator Weights of 0.66; and (C) Indicator Weights of 1.5.

Note. Results in this figure were based on a sample size of 50 clusters with 20 individual per cluster. The size of each point marking bias reflects the average *SD* of the coefficient estimates for that estimator under the model and conditions indicated. A larger *SD* results in larger points on the plot with smaller *SD*s producing to smaller points.

convergence failure rate under the true model. This is likely due to the misspecified model being less complex.

The relationships between model misspecification, estimation approach, and convergence failure rate were further complicated by changes in indicator weighting (see Figure 6A–C). As noted in the first simulation, convergence rates for ML benefited from increased indicator weights. This relationship held true for BCFS and FS approaches but was moderated by the individual-level sample size (n_1). For example, when $n_1 = 5$ both the BCFS and FS approach suffered higher convergence failure rates under the strongest indicator weights condition.

Bias. We found some similarities between estimator performance in terms of bias and convergence failure rate. Once again, larger individual sample sizes improved the performance of all three estimators, measurement misspecifications were more detrimental than structural misspecifications, and bias was greater with the small indicator weights. In Figure 6, we again use white markers sized by the *SD* of the estimates to mark bias for each estimation approach but track bias across the different model misspecifications. Relative performance of the estimators in terms of bias was dependent on indicator weights (see Figure 6A–C). The FS approach generally incurred the most bias with ML incurring the least bias followed closely by BCFS. However, under the largest indicator weight condition, the FS approach often had the least amount of bias.

As with the first simulation, results involving bias and ML are somewhat misleading. First, the advantages of ML in terms of bias tended to be reduced and reversed with misspecified models and smaller sample sizes. Under these conditions BCFS estimation incurred the same or less bias than ML. Second, the small amount of bias achieved by ML was accompanied by inefficiencies. The inflated triangles in Figure 6 marking the bias in ML estimates illustrate that under many of the conditions in which ML had the least amount of bias it was also the most inefficient. In addition, across several conditions in which ML demonstrated little bias or performed well relative to the other estimators, it had the highest convergence failure rate. The BCFS approach consistently incurred similar amounts of bias compared to ML but avoided inflated *SDs* and high convergence failure rates.

Efficiency. We again tracked the *SD* of path coefficient estimates for each method across the 1000 data sets to evaluate estimator efficiency. Results involving the *SD* of the estimates were unambiguous. We found improved estimator performance with larger sample sizes, more detrimental effects when the measurement model was misspecified, and generally more efficient estimation with larger indicator weights.

Results by estimator approach followed a consistent pattern across conditions and model misspecifications with *SD* of FS estimates being the smallest followed closely by BCFS estimates. The largest *SD* results consistently came from ML. These results reflect the trade-off between accuracy and precision in the limited information approaches (e.g., Kelcey et al., 2021). The BCFS approach consistently provides

less-biased estimates but proper consideration of the measurement error associated with the latent variables in the model led to decreased efficiency.

Simulation Study II: Supplemental Investigation

Our third and final supplemental simulation condition extended Simulation study II by combining model misspecifications and non-normal data considerations. We separated these areas of focus in our two primary simulation studies to better understand the specific detrimental effects of both. However, a supplementary investigation of this combination was conducted to investigate any compounding detrimental effects. In the combination model misspecification-non-normal data supplemental simulation, we paralleled Simulation study II and its misspecified model conditions but utilized non-normal indicator residuals using the approach from simulation study I (i.e., non-normal indicator residuals with distributions having skewness of 2 and kurtosis of 7). We included cluster sample sizes of $n_2 = 200, 100, 50,$ and 30 and the individual per cluster sample sizes of $n_1 = 20$. The measurement models were limited to four indicators per latent variable with mixed indicator weights.

Results were similar between the combination supplemental condition and parallel misspecified model conditions in Simulation study II without non-normal data (see Figure 7). Overall, BCFS and FS had minor increases in convergence failure rates but almost no change in bias when estimating misspecified models with non-normal data. While ML also avoided major increases to bias, the model misspecification-non-normal data condition did lead to decreases in efficiency (i.e., increases in the *SD* of ML estimates). The primary takeaway from this supplemental simulation is that the detrimental effect of model misspecification far exceeds the detrimental effects of the non-normal data considered here. Put differently, estimators performed relatively well with non-normal data and struggled against specific model misspecification conditions (e.g., the first measurement model misspecification and model misspecifications in the 0.666 indicator weight condition), so the combination of the two conditions mostly reflects estimator difficulty with model misspecifications.

Discussion

Multilevel structural equation modeling serves as an appropriate and effective approach to delineate and test complex theories involving multiple latent variables in multilevel settings. These conditions are common in educational research but the scope of even well-designed and well-resourced studies is often limited. The highly parameterized nature of MSEMs require large sample sizes at each level to be estimated using ML methods. Under the limited sample sizes common in educational research (e.g., less than 100 clusters and 20 individuals per cluster), ML estimation can suffer convergence issues and produce biased parameter estimates. The more recently developed Croon's BCFS approach has shown promise in estimating

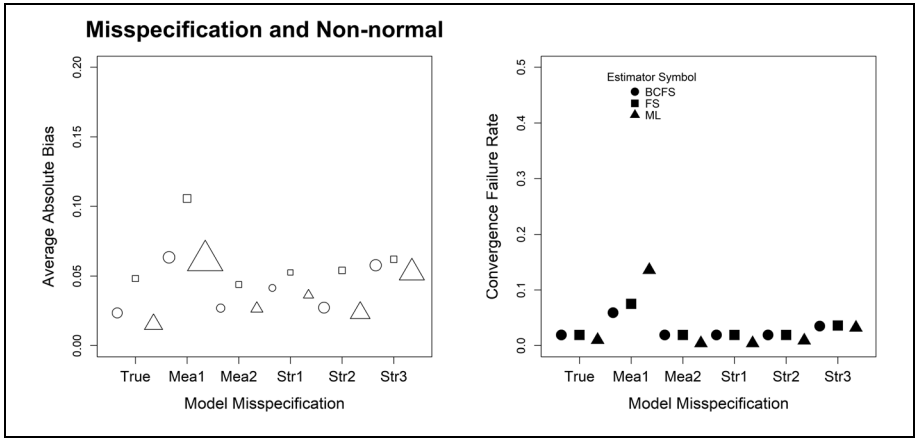

Figure 7. Bias and Convergence Failure Rate for Each Estimator by Model Misspecification With Four Indicators Per Latent Variable that Have Non-Normal Indicator Residuals and Mixed Indicator Weights.

Note. The cluster sample size was 50 with 20 individuals per cluster. The size of each point reflects the average *SD* of the coefficient estimates for that estimator under the model and conditions indicated. A larger *SD* results in larger points on the plot with smaller *SD*s leading to smaller points.

MSEMs with the smaller sample sizes common in planned educational research (e.g., Schochet, 2011).

This study sought to better map out the performance of BCFS for MSEMs in terms of bias, efficiency, and convergence failure rate under a combination of difficult conditions. Specifically, we examined the performance of BCFS with MSEMs when limited sample sizes were combined with non-normal indicators and model misspecifications. To complete this investigation, we employed two simulation studies. In the first simulation, we focused on the estimation of MSEMs with limited sample sizes while varying measurement model conditions, and in the second simulation study, we focused on estimation of MSEMs with different measurement and structural model misspecifications. We evaluated BCFS, in comparison with the (uncorrected) FS approach, and ML to gauge relative and absolute performance. BCFS provided results more often (i.e., high convergence rate) than the FS or ML estimators while balancing bias and efficiency.

Through improved dependability, increased efficiency, and reductions in bias, BCFS increases the feasibly of planned studies that utilize an MSEM. Utilization of MSEMs in educational research brings about better alignment between theory and research. For example, a cluster-randomized study on the effect of a school-wide teacher professional development program on student achievement is best operationalized using an MSEM. This model allows considerations for the hierarchical structure of the educational setting and the many latent variables involved in theories of teaching and learning. A sample of 100 schools may not be feasible, but employing

BCFS estimation reduces the scale of schools needed to ensure consistent and accurate results. In other words, an MSEM more appropriately reflects the structure of the educational setting, latent nature of many variables in educational research, and the complexity of teaching and learning theory while utilizing BCFS makes the study more feasible.

The fully crossed structure of our simulation studies produced results supporting and supplementing previous literature investigating BCFS with MSEMs. BCFS performed relatively well with smaller sample sizes but no estimation approach worked well with cluster sample sizes less than 10, and ideal conditions were needed for any estimation approach to work with cluster sample sizes around 20. BCFS was more feasible with cluster sample sizes around 30 when large individual per cluster sample sizes and larger indicator weights were present. Relatedly, BCFS, like the other estimation approaches, incurred substantial bias and had high convergence failure rates when individual per cluster sample sizes were very limited. We caution against BCFS with $n_1 \leq 10$. BCFS has increased potential as cluster sample sizes approach 50. Here, ML estimation was still problematic, but BCFS performed consistently well both relative to other approaches and in an absolute sense.

We considered four and 10 indicators per latent variable, various indicator weights, and a non-normal distribution of indicator residuals and found increasing the number of indicators per latent trait up to 10 can be beneficial to BCFS, but a large number of indicators was not necessary. Relatedly, we found BCFS to be more susceptible to non-normal indicator residuals as the number of indicators per latent variable increased. The detrimental effects were relatively small but the result is worth noting because the influence of non-normal indicator residuals on BCFS has not been previously considered, and it was one of the few conditions in which BCFS was more susceptible to detrimental conditions.

The other novel contribution of this work involved BCFS estimation with limited samples sizes and model misspecifications. Broadly, we found BCFS was more robust than the other estimation approaches to several measurement and structural misspecifications. This is a noteworthy result as previous research has indicated BCFS performs relatively well only under structural model misspecifications (Devlieger & Rosseel, 2019; Kelcey et al., 2021).

Of course, BCFS is not a universal solution for problems estimating MSEMs. Under the most difficult conditions considered here, BCFS estimation still failed to converge at a high rate, incurred bias, and was inefficient. There was also a notable trade-off between accuracy and efficiency in BCFS. BCFS estimates incurred less bias across most of the simulation conditions but were often less precise than those of the FS approach. BCFS is also currently limited to multilevel models without random slopes. We see the extension of BCFS to structural models that include random slopes as an important area of future research with considerations of estimator robustness an important follow-up investigation. Our supplemental investigations of BCFS with multilevel measurement invariance, varying types of non-normal indicator residuals, and combinations of difficult conditions (e.g., model misspecification and

non-normal data) also deserve more comprehensive examinations. Despite these shortcomings, the results here support, supplement, and elucidate the effective performance of BCFS further encouraging its utilization with MSEMs.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Foundation (grant nos. 1552535 and 1760884). The opinions expressed herein are those of the authors and not the funding agency.

ORCID iD

Kyle Cox https://orcid.org/0000-0002-7173-4701

References

Autio, E., & Deussen, T. (2017). Recruiting rural schools for education research: Challenges and strategies. In G. Nugent, G. Kunz, S. Sheridan, T. Glover & L. Knoche (Eds.), *Rural education research in the United States* (pp. 77–93). Springer.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78–84.

Bollen, K., Kirby, J., Curran, P., Paxton, P., & Chen, F. (2007). Latent variable models under misspecification: Two-Stage Least Squares (2SLS) and Maximum Likelihood (ML) Estimators. *Sociological Methods & Research*, *36*, 48–86.

Cheung, G., & Lau, R. (2017). Accuracy of parameter estimates and confidence intervals in moderated mediation models: A comparison of regression and latent moderated structural equations. *Organizational Research Methods*, *20*, 746–769.

Cox, K., & Kelcey, B. (2021). Croon's bias corrected estimation of latent interactions. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 863–874.

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Lawrence Erlbaum.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*, 741–770.

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology*, *13*, 31–38.

Devlieger, I., & Rosseel, Y. (2019). Multilevel factor score regression. *Multivariate Behavioral Research*, *55*(4), 600–624.

Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement*, *79*, 1017–1037.

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

Hayes, T., & Usami, S. (2020a). Factor score regression in connected measurement models containing cross-loadings. *Structural Equation Modeling*, *27*, 942–951.

Hayes, T., & Usami, S. (2020b). Factor score regression in the presence of correlated unique factors. *Educational and Psychological Measurement*, *80*, 5–40.

Hox, J., Maas, C., & Brinkhuis, M. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modelling*, *20*, 265–282.

Kelcey, B. (2019). A robust alternative estimator for small to moderate sample SEM: Bias-corrected factor score path analysis. *Addictive Behaviors*, *94*, 83–98.

Kelcey, B., Cox, K., & Dong, N. (2021). A bias-corrected limited information estimator for small to moderate multilevel structural equation models. *Organizational Research Methods*, *24*(1), 55–77.

Kelcey, B., & Phelps, G. (2013a). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis,* 35(3), 370–390.

Kelcey, B., & Phelps, G. (2013b). Strategies for improving power in school randomized studies of professional development. *Evaluation Review*, 37, 520–554.

Lei, M., & Lomax, R. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*, 1–27.

Li, X., & Beretvas, S. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling*, *20*, 241–264.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*(1), 45–58.

Micceri, T. (1989.). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233.

R Development Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, Article 1715.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A how to guide for applied researchers and practitioners* (pp. 226–238). Routlegde.

Schochet, P. Z. (2011). Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? *Journal of Educational and Behavioral Statistics*, *36*, 441–471.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*, 255–267.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*, 481–520.

van de Schoot, R., & Miocević, M. (2020). *Small sample size solutions*. Taylor & Francis.