Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Small Sample Estimation of Structural Equation Models: A Comparison of Alternative Estimation Approaches

Graham G. Rifenbark [*†] and Yves Rosseel [‡]

†Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison, Madison, 53703, WI, USA

‡Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, East Flanders, Belgium

*Corresponding author. Email: rifenbark@wisc.edu

Abstract

Typically, structural equation model (SEM) parameters are estimated using maximum likelihood (ML) which possess desirable attributes related to point estimation and inferences; however, ML is a large sample estimation method. In practice, researchers are often confronted with small sample sizes. When ML is used in these settings a host of issues can surface, including model nonconvergence, improper solutions, biased point estimates, and poor inference quality. Ultimately, latent parameters (e.g., regressions) are of most interest and the accuracy of these estimates is paramount. Alternative modeling approaches such as factor score regression (FSR) and the structural-after-measurement framework (SAM, Rosseel & Loh, 2024) have been proposed and can be useful in the context of small sample size. In lieu of traditional (unbounded) ML, bounded ML estimation is a viable alternative for SEM, FSR, and SAM; whereas, non-iterative estimation is feasible for two-stage approaches (e.g., SAM). Therefore, we executed a Monte Carlo simulation to compare the performance of SEM, FSR, and SAM in which we systematically varied sample size, construct reliability, number of measurement blocks, and estimation bound type. For the SAM approach, we evaluated two different estimation procedures at stage-1: a non-iterative estimation method (multiple group method) and ML; whereas, we used ML for both stages of estimation for FSR. Outcomes of interest included convergence rates and estimation accuracy of structural parameters. We report the results from this Monte Carlo simulation and discuss implications for research with limited sample sizes.

Keywords: structural equation models, two-stage estimation, structural-after-measurement, small sample size, parameter recovery

1. Introduction

Structural Equation Models (SEMs) are commonly employed in the educational and social sciences to investigate unobservable phenomenon (i.e., latent constructs) affording the opportunity to elucidate the relationship between measured constructs.

An SEM contains both a measurement and a structural model, where the measurement model links the p observed (or manifest) variables to the m latent variables as a linear function. The structural model specifies the relations among the m latent variables, allowing them to affect each other. Using the *all-y* LISREL parameterization used by `lavaan` (Rosseel, 2012) these models can be written as:

$$\boldsymbol{y} = \boldsymbol{\nu} + \boldsymbol{\Lambda\eta} + \boldsymbol{\varepsilon}, \tag{1}$$

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{B\eta} + \boldsymbol{\zeta}. \tag{2}$$

Equation 1 describes the measurement model where, $\boldsymbol{\nu}$ is a $m \times 1$ vector of indicator intercepts; $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of factor loadings (linear regression slopes) relating indicators to latent variables; and $\boldsymbol{\varepsilon}$ is a $p \times 1$ random vector of indicator residuals with a mean of zero and a variance-covariance ($\boldsymbol{\epsilon} \sim [0, \boldsymbol{\Theta}]$). Equation 2 describes the structural model where, $\boldsymbol{\alpha}$ is a $m \times 1$ vector of latent-variable intercepts; $\mathbf{B}$ is a $m \times m$ matrix of linear regressions in which diagonal elements must be zero; and $\boldsymbol{\zeta}$ is a $m \times 1$ random vector of latent-variable residuals (or disturbances) with a mean of zero and a variance-covariance ($\boldsymbol{\zeta} \sim [0, \boldsymbol{\Psi}]$).

In order for unique parameter estimates to result, it is necessary for the model to be identified which can be accomplished by fixing one factor loading per factor (i.e., the marker variable) to a value (typically 1.0) as well as its corresponding intercept (typically 0.0). Using the SEM parameters, model-implied (or conditional) variance-covariance matrix and mean vector can be constructed as follows:

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}) &= \boldsymbol{\mu} = \boldsymbol{\nu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha}, \\ \mathrm{Var}(\boldsymbol{y}) &= \boldsymbol{\Sigma} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}\left[(\mathbf{I} - \mathbf{B})^{-1}\right]'\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \end{aligned} \tag{3}$$

where $\mathbf{I}$ is a $m \times m$ identity matrix and further, it is assumed that $(\boldsymbol{I} - \boldsymbol{B})$ is invertible. Ultimately, SEM attempts to reproduce the observed variance-covariance matrix ($\boldsymbol{S}$) and mean vector ($\bar{\boldsymbol{y}}$), therefore, model estimation attempts to minimize the discrepancy between the model-implied moments: $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$.

1.1 Modeling Approaches

SEMs are routinely estimated using *maximum likelihood* (ML) and its discrepancy fit function is:

$$F_{\mathrm{ML}} = \log|\hat{\boldsymbol{\Sigma}}| - \log|\boldsymbol{S}| + \mathrm{trace}(\boldsymbol{S}\hat{\boldsymbol{\Sigma}}^{-1}) - p + (\bar{\boldsymbol{y}} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}). \tag{4}$$

Effectively, ML estimation needs an iterative optimization procedure and terminates once a predefined set of criteria has been met resulting in a converged solution. A test statistic of exact model fit can be constructed: $T_{ML} = F_{ML} \times N$ and assuming multivariate normality, T_{ML} is χ^2 distributed with P - q degrees of freedom, where P is defined as $\frac{p(p+3)}{2}$ (i.e., the unique pieces of information in the observed covariance matrix and mean vector).

With a large enough sample size and observed data that is multivariate normal, ML estimation has many benefits such as unbiased point estimates, trustworthy standard errors which leads to sensible confidence intervals, and finally, a test of exact fit with nominal Type-I error rates. However, sufficient sample size will depend on attributes of the model and the data available. As the observed variance-covariance matrix increases in size (e.g., more indicators-per-factor) there are more elements for the SEM to reproduce - therefore, SEMs with a larger measurement model will require a larger sample size. With respect to observed data, holding model size constant, SEMs with reliable measures will require a smaller sample size compared to SEMs with low reliability. Ultimately, when estimating SEMs with ML in the context of insufficient sample sizes will impact the accuracy of the parameter estimates as well as the inferences from the model. Specifically, parameter estimates will be biased as a result of being under- or over-estimated (Kosmidis, 2014), standard errors will exhibit unwarranted precision (i.e., too small) leading to confidence intervals that are too narrow; and finally, a test statistic that is untrustworthy (Rosseel, 2020). Next, we describe alternative modeling approaches and estimation procedures that can be used in the context of limited sample sizes.

While there are a variety of approaches researchers could use when faced with insufficient sample sizes, we focus on approaches that require multiple stages of estimation; which is counter to SEM as all measurement and structural parameters are estimated simultaneously. Effectively, an initial model is fitted and the resulting information is used to inform the estimation of a subsequent model. By fitting

sequential models, this decreases the number of parameters to be estimated which in-turn increases the likelihood the model will converge. The initial model estimated for multi-step approaches is often an unrestricted confirmatory factor analysis (CFA) in which all bi-directional relationships are estimated and latent variances are exogenous. Therefore, Equation 3 becomes:

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}) &= \boldsymbol{\mu} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha}, \\ \mathrm{Var}(\boldsymbol{y}) &= \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \end{aligned} \tag{5}$$

A complication inherent in multi-step approaches is the need to account for the uncertainty in the estimates from the initial model when fitting the subsequent model; if not, bias will be introduced. Therefore, care should be taken to account for this uncertainty in the subsequent modeling stage. A feature of multi-stage approaches is the possibility of avoiding *interpretational confounding*. Interpretational confounding occurs when the definition of a latent variable changes in a non-negligible way when additional variables are entered into the model (Burt, 1976; Levy, 2017). For example, if an observed variable is entered into an SEM as an outcome, the SEM is not able to distinguish it from its indicators; therefore, its inclusion could systematically change the meaning of the latent variable itself. Burt (1976) advocated for fitting an individual CFA for each latent variable in the SEM. In effect, this deconstructs the original measurement model into as many measurement blocks as there are latent variables, which are fitted independently. Afterwards, the SEM is specified with all measurement model parameters fixed to the estimates obtained from the one-factor CFAs and ultimately, the structural parameters are estimated using ML. By proceeding in this manner, the estimation problem is simplified while also preserving the meaning of the latent variables.

1.1.1 Factor Score Regression

Instead of fixing measurement model parameters to previously obtained values, factor score regression (FSR) is accomplished by extracting factor scores from fitted CFAs, resulting in a set of factor scores for each latent variable. These factor scores contain an estimate for each observation and are combined into a case-level data set. Afterwards, a subsequent model (e.g., regression or path analysis) is fitted using ML treating the factor scores as 'observed' data. Two common methods of extracting factor scores are: Bartlett (1937) and Regression factor scores (Thurstone, 1934). When Bartlett scores are used as independent variables; or when Regression scores are used as dependent variables, bias is introduced into the model (Devlieger et al., 2016). Therefore, Skrondal and Laake (2001) proposed the *bias avoiding* approach in which Bartlett scores are used for dependent variables and Regression scores are used for independent variables thus mitigating the issue of bias; however, this approach is tedious for mediation models due to variable(s) serving both as a dependent and independent variables (Devlieger et al., 2016). Therefore, for the purposes of this paper (see Figure 1) we used Bartlett factor scores for all latent variables to carry out FSR. While this approach (uncorrected FSR) is known to produce biased regression coefficients, the variability around these point estimates is small and may be conducive to small samples.

1.1.2 Structural After Measurement

Rosseel and Loh (2024) proposed the structural-after-measurement (SAM) framework which provides great flexibility for researchers to implement multi-stage approaches to SEM and is implemented in `lavaan::sam()` (Rosseel, 2012). Specifically, SAM makes it possible to use different estimators for the measurement and structural parameters as they are fitted separately.

In the initial stage, measurement model parameters $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\nu}$ (if the mean structure is included) are estimated separately for each measurement block and ultimately are pooled together. In the second stage, there are two options for estimating the structural parameters: global and local. Global

SAM is similar to Burt (1976), as all measurement parameters are fixed to their estimate from the first stage, and only the structural parameters are free in the second stage. However, in this paper, we only consider local SAM, which relies on summary statistics: the variance-covariance matrix of the latent variables [$Var(\boldsymbol{\eta})$] and the mean vector of the latent variables [$E(\boldsymbol{\eta})$]. These summary statistics are computed based on the estimated measurement model parameters and the observed sample statistics ($\boldsymbol{S}$) and ($\bar{\boldsymbol{y}}$), see Rosseel and Loh (2024) for details. These summary statistics are used as input for the second stage, where we estimate the remaining structural parameters. Several estimators can be used in the second stage, but in this paper we always use ML. When using many measurement blocks, misfit in one part of the measurement model will not spread to other parts of the model; but at the same time, estimation is not able to incorporate information from other parts of the model. By default, the number of measurement blocks is equal to the number of latent constructs.

1.2 Estimation Procedures

Similar to alternative approaches to modeling, there too are options for estimation that can increase the likelihood of model convergence. As previously mentioned, ML estimation is an iterative procedure which given challenging contexts (e.g., complex models and/or limited sample sizes) may fail to converge on a solution. Further, ML estimation can also be characterized as an unregularized estimation procedure meaning it has no knowledge of what may or may not be admissible solutions. Next, we detail options available to researchers which show promise in such contexts: non-iterative estimation and bounded estimation.

1.2.1 Non-iterative Estimation

By definition, non-iterative estimation procedures are not prone to issues regarding non-convergence because they change the problem at hand into a series of smaller tasks which have a closed form (Dhaene & Rosseel, 2024). Because of this, such estimators can be beneficial both with limited samples and with very large samples as its implementation is not computationally burdensome (Dhaene & Rosseel, 2024). In fact, Dhaene and Rosseel (2023) show that regardless of the number of indicators-per-factor in an SEM, non-iterative estimators saw negligible differences in elapsed time compared to ML estimation; whereas computation time for SEM increased significantly. A downside of non-iterative estimators is the unavailability of standard errors as no analytic solution exists to date. Instead, resampling methods (e.g., bootstrap) can be used to inform inferences. Through simulation, Dhaene and Rosseel (2023) report promising findings for the *Multiple Group Method* (MGM) proposed by Guttman (1952) relative to other non-iterative estimation procedures. For more details on the MGM estimation procedure interested readers should consult Dhaene and Rosseel (2023).

1.2.2 Bounded Estimation

Naturally, SEM parameters have theoretical bounds, for example variance parameters (i.e., diagonal elements of $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$) cannot be negative. On the other hand, off-diagonal elements of $\boldsymbol{\Psi}$ (i.e., latent correlations), when standardized, can not be greater than |1.0|. It is possible for a model to converge on a solution that contain these Heywood cases and/or result in warnings about non-positive definite matrices. De Jonckere and Rosseel (2022) proposed the use of bounded estimation in which constraints are placed on free parameters in the model (e.g., the residual of an endogenous latent variable). Therefore, the purpose of imposing bounds is to increase the likelihood of model convergence without introducing bias in the point estimates of parameters. If model convergence can be achieved by implementing lower and upper boundaries for model parameters that are freely estimated, this then provides an opportunity to enhance standard errors by employing resampling methods (De Jonckere & Rosseel, 2022). Using the observed data, it is possible to manually determine "'standard"' upper and lower bounds (De Jonckere & Rosseel, 2022). The method of identification plays a role in determining the lower and upper bounds and we focus on the case in which the

marker variable approach is used which is the default in `lavaan`. In order to construct the bounds, it is necessary to make an assumption about the *minimum reliability of the marker variable* and have access to the sample variance-covariance matrix. We detail how to arrive at the boundary values per De Jonckere and Rosseel (2022):

Standard By default, the minimum reliability of the marker variable is set to 0.1, therefore, bounds would be set as follows:

- Lower:
  - Residuals: 0.0
  - Endogenous latent variance: 0.005 (default)
  - Exogneous latent variance: $S^2_{marker} - 0.90 \times S^2_{marker}$
  - Factor Loading: $-1 \times (\sqrt{\frac{var}{\Psi}})$
- Upper:
  - Residuals: $(1 - 0.1) \times S^2_p$ [marker]; S^2_p [non-marker]
  - Endogenous latent variance: S^2_{marker}
  - Exogenous latent variance: S^2_{marker}
  - Factor Loading: $\sqrt{\frac{var}{\Psi}}$

It is important to note that due to the use of the marker variable method of identification in the above example, both exogenous latent variance and endogenous latent variance (latent residual) are free parameters and thus, are not a function of other parameters.

Wide Using standard bounds as a basis, it is straightforward to determine bounds as boundaries are relaxed. Specifically, upper bounds are increased for factor loadings (10%), manifest residuals (20%), and latent variances (30%); while lower limits are increased for manifest residuals (5%) and factor loadings (10%). Therefore, by loosening boundaries this makes it possible for variances that are slightly negative to result and as such, wide bounds should result in higher convergence rates than standard bounds.

Positive Variance When using positive variance bounds, no upper boundaries are placed on any parameter; while lower bounds are only placed on manifest residuals and latent variances/disturbances to ensure they are not negative. However, by omitting an upper limit it is possible for parameter estimates to exceed what is feasible given the characteristics of the data (i.e., sample variances).

To date, there exists gaps in the literature related to the performance of these alternative modeling approaches and estimation methods in the context of small sample sizes. First, the impact of bounded estimation has only been examined in the context of an SEM with a small number of free parameters; for example, De Jonckere and Rosseel (2022) used an SEM with at most 23 free parameters (model M0) and only considered standard and wide bounds. Second, the impact of using different measurement block specifications has yet to be studied. Third, no studies have systematically investigated the interaction of these factors, while also varying levels of reliability and sample size. Therefore, the purpose of this study was to fill these gaps by investigating the performance of alternative modeling and estimation strategies in the context of small sample sizes relative to SEM. Of specific interest was convergence rates with and without bounds for SEM and two-stage approaches, estimation accuracy of the structural parameters, and to test if the use of bounded estimation impacts parameter recovery.

2. Method

To accomplish the study aims, we executed a Monte Carlo simulation in which we systematically varied design factors to examine their impact on study outcomes - described in Section 2.3. The specific modeling approaches (or estimation methods) and estimation strategies used in the study were:

- SEM with ML
  - Bounds: None, Standard, Wide, and Positive Variance
  - Measurement Blocks: 1
- FSR with ML (stage-1 and stage-2)
  - Bounds: None, Standard, Wide, and Positive Variance
  - Measurement Blocks: 1, 2, and 5
- Local SAM with ML (stage-2)
  - Stage-1 Estimator: ML (LSAM-ML) and MGM (LSAM-MGM)
  - Bounds: None, Standard, Wide, and Positive Variance
  - Measurement Blocks: 1, 2, and 5

The SEM used in the simulation was motivated by previous research (Lance et al., 2016; McNeish & Hancock, 2018; Rifenbark, 2022). The SEM contains five latent variables: 3 exogenous (X1, X2, X3) and 2 endogenous (Y1 and Y2). The structural relations among latent variables in the population can be found in Figure 1.

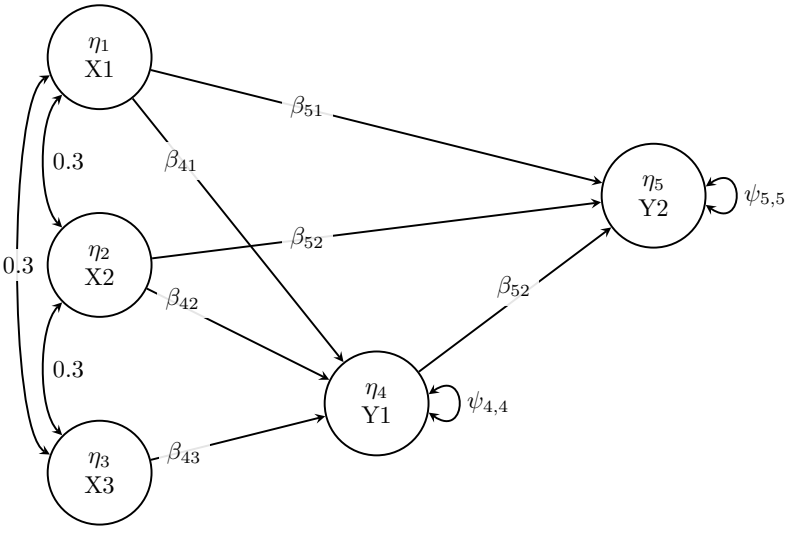

Figure 1. Structural Path Diagram in the Population.

With respect to the measurement model of the SEM, we held the number of indicators-per-factor fixed at 6, therefore, there were a total of 30 manifest (or observed) variables. To identify the model, the marker variable method was used in which the first factor loading per factor was fixed at 1.0, allowing for all structural parameters to be freely estimated. In total, there were 14 free parameters in the structural component and 55 free parameters in the measurement component.

2.1 Manipulated Factors

2.1.1 Sample Size

A total of four sample size conditions were used in the study: 75, 100, 150, and 1,000. When estimating the full SEM (69 parameters), globally there would be as few as 1.09 (N = 75) or as many as 14.5 (N = 1,000) observations per parameter; whereas structurally (14 parameters), there would be as few as 5.36 (N = 75) or as many as 71.43 (N = 1,000) observations per parameter.

2.1.2 Reliability

Measure reliability has a large impact on the estimation of SEMs, specifically with respect to convergence rates (Gagne & Hancock, 2006) and bias (Devlieger & Rosseel, 2017; Dhaene & Rosseel, 2022); therefore, it was necessary to investigate varying levels of reliability. Specifically, we varied levels of *construct reliability* (CR) to be high (CR = 0.90), moderate (CR = 0.70), or low (CR = 0.60). Construct reliability is determined as a function of the number of indicators-per-factor and the strength of factor loadings (Gagne & Hancock, 2006). To accomplish these levels of CR, it was necessary to vary population values for the diagonal elements of $\boldsymbol{\Theta}$ given that all factor loadings were 1.0 in the population. We expand on the data generation process below in Section 2.2.

2.1.3 Measurement Blocks

Due to the modeling flexibility two-stage estimation approaches afford, we varied the number of measurement blocks used in the first stage of estimation to be either 1, 2, or 5. Given a single measurement block, all latent variables are estimated together using a correlated factor model; when using two measurement blocks, exogenous (X1, X2, X3) and endogenous variables (Y1 and Y2) were grouped together and were estimated in separate correlated factor models; finally, when five measurement blocks were used, five separate factor models were estimated. Traditional SEM by definition employs a single measurement block due to all model parameters being estimated simultaneously and therefore, could not be varied.

2.1.4 Estimation Bounds

We investigated a total of four types of estimation bounds: unbounded, standard, wide, and positive variance bounds.

2.2 Data Generation

First, population values were specified for $\boldsymbol{B}$ and Ψ and were selected based on previous research studies (Lance et al., 2016; Rifenbark & Jorgensen, 2023).

Table 1. Population values for $\boldsymbol{B}$ and $\boldsymbol{\Psi}$

B	X1	X2	X3	Y1	Y2	**Ψ**	X1	X2	X3	Y1	Y2
X1	0.0	0.0	0.0	0.0	0.0	X1	1.0	0.300	0.30	0.0	0.0
X2	0.0	0.0	0.0	0.0	0.0	X2	0.3	1.000	0.30	0.0	0.0
X3	0.0	0.0	0.0	0.0	0.0	X3	0.3	0.300	1.00	0.0	0.0
Y1	0.40	0.60	0.40	0.0	0.0	Y1	0.0	0.0	0.0	0.320	0.0
Y2	0.60	0.40	0.0	0.40	0.0	Y2	0.0	0.0	0.0	0.0	0.462

Using the population matrices found in Table 1, we took the following steps:

1. Computed the population covariance matrix: $\boldsymbol{\Phi} = (\boldsymbol{I} - \mathbf{B})^{-1} \times \boldsymbol{\Psi} \times \left[(\mathbf{I} - \mathbf{B})^{-1}\right]'$
2. Using $\boldsymbol{\Phi}$, computed $\boldsymbol{\Lambda\Phi\Lambda}$ with block diagonal elements of $\boldsymbol{\Lambda}$ set to 1.0.
3. Solved for diagonal elements of $\boldsymbol{\Theta}$ via $(\frac{1}{Reliability} - 1) \times diag(\boldsymbol{\Lambda\Phi\Lambda})$, where $diag(\boldsymbol{\Lambda\Phi\Lambda})$ corresponds to true score variance - see Table 2 for more details.
4. Solved for the population variance-covariance matrix: $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta}$

Finally, Σ was used to generate data assuming multivariate normality using the `rockchalk` package (Johnson, 2022) in R. For each simulation condition a total of 1,000 data sets were generated. Further details can be found on our dedicated OSF page.

Table 2. Population values for $Diag(\boldsymbol{\Theta})$ across levels of construct reliabilities (CR)

CR	Reliability per Indicator	X1:X3	Y1	Y2
.60	.200	4.000	5.536	7.809
.70	.281	2.560	3.543	4.998
.90	.600	0.667	0.923	1.301

2.3 Outcomes

2.3.1 Convergence

Across all simulation conditions we tracked the number of replications in which estimation converged on a solution without warnings or errors. This was done for each unique combination of modeling approach and estimation strategy. Convergence results are reported as rates of success to ease interpretation.

2.3.2 Parameter Recovery

To assess the performance of the modeling and estimation approaches, it was necessary to examine the variability of parameter estimates across replications. Total error of estimation can be partitioned into two components: systematic error (or bias) and random error (or variance) with the standard deviation across replications interpreted as the standard error of estimation. *Bias* is computed as the difference between the mean estimate of a parameter ($\bar{\hat{\theta}}$) over replications and its population value (θ). Therefore, estimators can systematically under- or over-estimate parameters and exhibit negative or positive bias, respectively. Because bias is influenced by the scale (or magnitude) of the parameter we operationalize bias as a percentage, allowing us to compare bias across all parameters regardless of their population values – see Equation 6 for its computation. *Standard deviation* is computed by summing the squared difference between the mean estimate and estimate ($\hat{\theta}$) across all (R) replications and taking its square root – see Equation 7; whereas, *root mean squared error* is determined in a similar fashion, however, the differences are between the population value and estimate over replications – see Equation 8.

$$Percent\ Bias = \frac{\bar{\hat{\theta}} - \theta}{\theta}, \tag{6}$$

$$SD = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\theta} - \bar{\hat{\theta}})^2}{R}}, \tag{7}$$

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\theta} - \theta)^2}{R}}, \tag{8}$$

3. Results

3.1 Convergence Rates

Clear patterns emerged related to the proportion of replications that converged on a solution without warnings or errors. First, convergence problems were not observed for any estimation method when sample size was large (N = 1,000) or when construct reliability was high (CR = .90). Second, local SAM with MGM estimation at stage-1 resulted in a 100% convergence rate across all simulation conditions. Finally, convergence rates for FSR and local SAM with ML estimation at stage-1 were identical across all simulation conditions. Next we report findings by whether or not bounded estimation was used. Additional results can be found on our dedicated OSF page.

3.1.1 Unbounded Estimation

Table 3. SEM Convergence Rates: Sample Size (N) X Construct Reliability (CR)

	CR = .60			CR = .70		
	N = 75	**N = 100**	**N = 150**	**N = 75**	**N = 100**	**N = 150**
Unbounded SEM	43.6%	60.4%	76.0%	73.9 %	87.2%	95.3%

Note. SEM with unbounded estimation achieved 100% convergence rates in all other simulation conditions. SEM with bounded estimation resulted in convergence rates of 99% or greater across all simulation conditions.

A stark difference in convergence rates was observed between SEM and the two-stage estimation methods. The lowest convergence rate for two-stage estimation methods observed was 97.1% compared to 43.6% for SEM which occurred when construct reliability was low (CR = .60) and sample size was small (N = 75). Even with moderate construct reliability (CR = .70) and a larger sample size (N = 150), the observed convergence rate for unbounded SEM still did not surpass the lowest rate observed for two-stage estimation methods. Table 3 contains convergence rates for unbounded SEM in settings in which rates were lower than 99%. Interestingly, a trend was observed for two-stage estimation methods with respect to measurement blocks. Specifically, as the number of measurement blocks increased, convergence rates decreased slightly - 98.7% (single measurement block) to 97.1% (5 measurement blocks). Interested readers can consult the supplemental materials for this information.

3.1.2 Bounded Estimation

SEM convergence rates dramatically improved when using bounded estimation with a minimum of 99.2% replications converging. Use of wide and standard bounds each resulted in two conditions in which convergence rates were below 100%; whereas, positive variance bounds resulted in four such conditions. In terms of two-stage estimation methods, the difference between bounded and unbounded estimation was negligible. Standard and wide bounds resulted in 100% convergence rates across all simulation conditions. Interestingly, when five measurement blocks were specified positive variance bounds produced a convergence rate that was slightly lower than when no bounds were used: 96.2% versus 97.1%.

3.2 Estimation Accuracy

3.2.1 Impact of Bounded Estimation

Using all converged solutions we investigated estimation accuracy and provide plots in a similar fashion to Dhaene and Rosseel (2022). First, we investigated the degree to which boundary type, on average, had an impact on percent bias - see Figure 2. In Figure 2 , the row factor corresponds to estimation method (FSR, LSAM-MGM, LSAM-ML and SEM) and the column factor corresponds to construct reliability (.60, .70, and .90). In each plot, line color corresponds to boundary type: none (red), standard (cyan), wide (purple), and positive variance (green); line type according to the number of measurement blocks (1, 2, or 5) with sample size and percent bias on the x-axis and y-axis, respectively. It is important to note that the scale of the y-axis was allowed to vary by estimation method, therefore, caution should be exercised when consulting Figure 2. Finally, dotted lines are drawn which correspond to -10%, 0%, and 10% bias to provide a reference.

Upon visual inspection of Figure 2 we find that FSR was the most biased estimation method, however, the use of bounded estimation or type of bound used had no impact on bias. Local SAM with MGM estimation resulted in non-negligible bias when construct reliability was low (CR = .60) and sample size was small ($N \leq 150$); however, was its bias was not impacted by the use of bounded

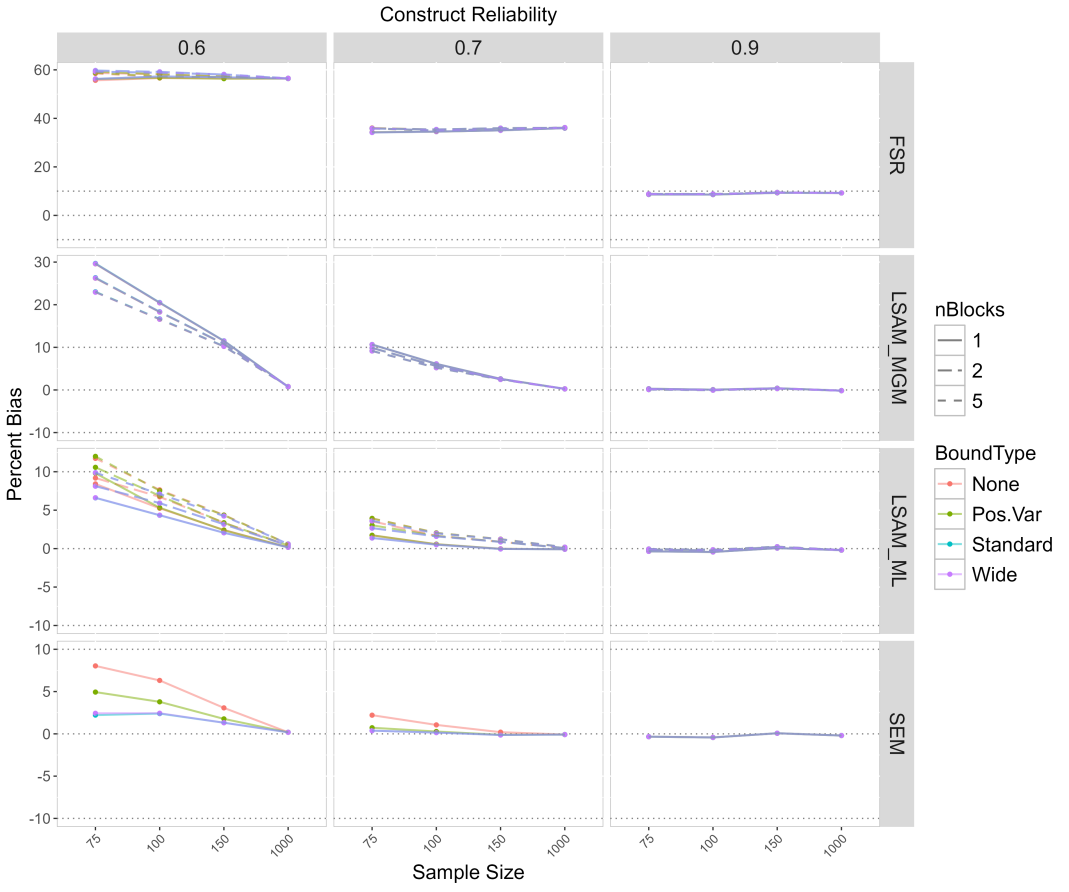

Figure 2. Bias across structural parameters: X-axis = method-by-nBlocks; Y-axis = percent bias

estimation or bound type. In terms of LSAM-ML, use of bounded estimation and bound type had the largest impact on bias when construct reliability was low (CR = .60); however, as sample size and construct reliability increases this impact dissipates. In terms of SEM, the use of bounded estimation and bound type had a similar impact on bias as was observed for LSAM-ML. Given that bias differed by no more than ∼ 6% and 5% for LSAM-ML and SEM, respectively; we concluded that the use of bounded estimation and bound type had a minimal impact on bias, on average, across methods.

Collapsing across all boundary types, we investigated the impact the remaining design factors had on percent bias (Figure 3), standard deviation (Figure 4), and root mean squared error (Figure 5). In a similar fashion to Figure 2, each figure contains 12 plots, however, the row factor now corresponds to parameter type and line color corresponds to estimation method: FSR (red), LSAM-MGM (blue), LSAM-ML (green) and SEM (black) with the performance measure on the y-axis which was allowed to vary based on parameter type. In Figure 3, dotted lines have been included corresponding to -10%, 0%, and 10% bias. We now report findings by parameter type.

3.2.2 Latent Covariances

A total of three latent covariances were estimated, each with a population value of .30. Across most design factors percent bias for all estimation methods was observed to be was within ± 10%. However, when sample size was small (N = 75) and construct reliability was low (CR = 0.60) FSR,

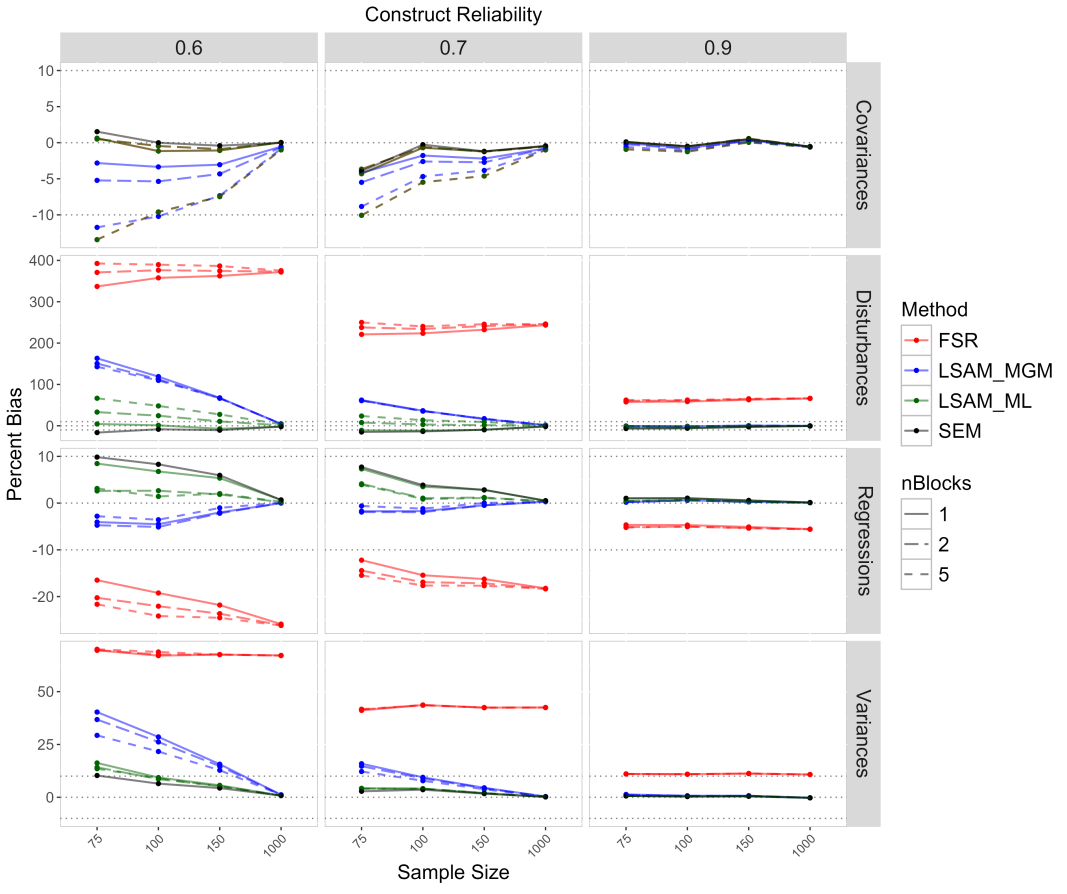

Figure 3. Percent bias: Construct Reliability X Parameter X Sample Size X Method

LSAM-ML, and LSAM-MGM underestimated latent covariances by approximately 13% when using five measurement blocks.

With respect to standard deviation, estimation methods were found to perform similar to one another, regardless of the number of measurement blocks. All SD estimates were $\sim$ 0.26 or smaller. As sample size and construct reliability increased, standard deviation decreased. In terms of RMSE, the pattern observed for SD remained true.

3.2.3 Latent Variances

Focusing on exogenous variances, estimation methods were found to result in varying amounts of bias. First, SEM resulted in the least amount of bias with estimates being between $\sim$ 10% and 0% across all conditions. Second, LSAM-ML was found to have slightly more bias than SEM when construct reliability was low (CR = .60) and at smaller sample sizes (N < 1,000). Local SAM with MGM estimation at stage-1 resulted in bias estimates ranging between $\sim$ 40% and 0% across all conditions, with more bias being observed in the context of fewer measurement blocks and smaller sample sizes. Finally, FSR was found to be the most biased with estimates ranging from $\sim$ 60% (CR = .60), $\sim$ 40% (CR = .70), and $\sim$ 12% (CR = .90) - regardless of sample size or the number of measurement blocks.

Standard deviation was found to differ between estimation methods when construct reliability

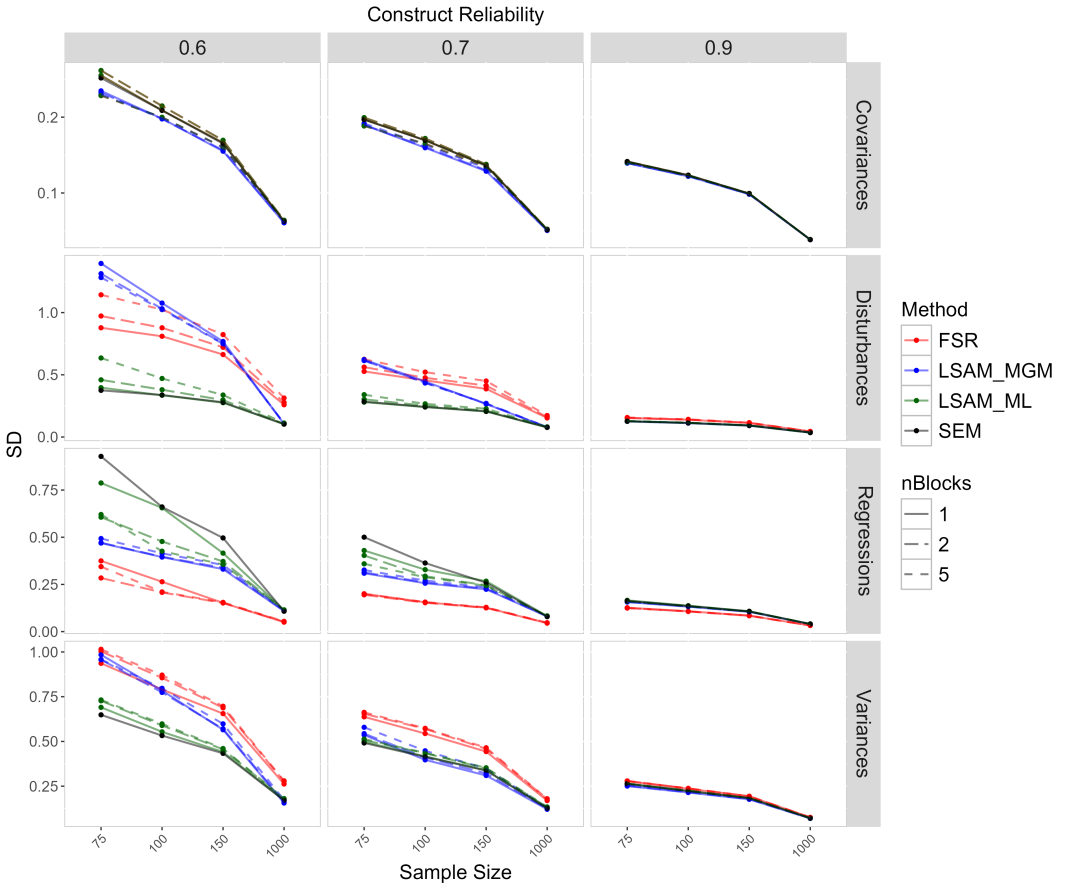

Figure 4. SD: Construct Reliability X Parameter X Sample Size X Method

was below .90. When construct reliability was .60, LSAM-MGM and FSR had similar standard deviations which were higher than the other methods; however, when construct reliability was .70, LSAM-MGM possessed standard deviations that were at or below estimates for SEM and LSAM-ML. Across the board, standard deviation for all estimation methods decreased as sample size and construct reliability increased.

In terms of RMSE, FSR possessed the largest estimates across all simulation conditions. When construct reliability was .60, RMSE for FSR ranged from approximately 1.25 to 0.75; LSAM-MGM ranged from 1.1 to 0.25; whereas, SEM and LSAM-ML were more similar to one another with estimates ranging between 0.75 and 0.25.

3.2.4 Latent Disturbances

Switching to endogenous variances, we observed clear differences in bias between estimation methods. Specifically, FSR was found to overestimate latent disturbances, on average, by approximately 380% (CR = .60), 250% (CR = .70), and 50% (CR = .90). Local SAM with MGM estimation was found to overestimate latent disturbances by between approximately 160% and 100% when construct reliability was .60 and sample sizes were 150 or lower; however, when construct reliability was .70, percent bias dropped significantly. Of the two-stage methods, LSAM with ML performed the best. Interestingly, two-stage methods with ML estimation at stage-1 tended to be more biased as the

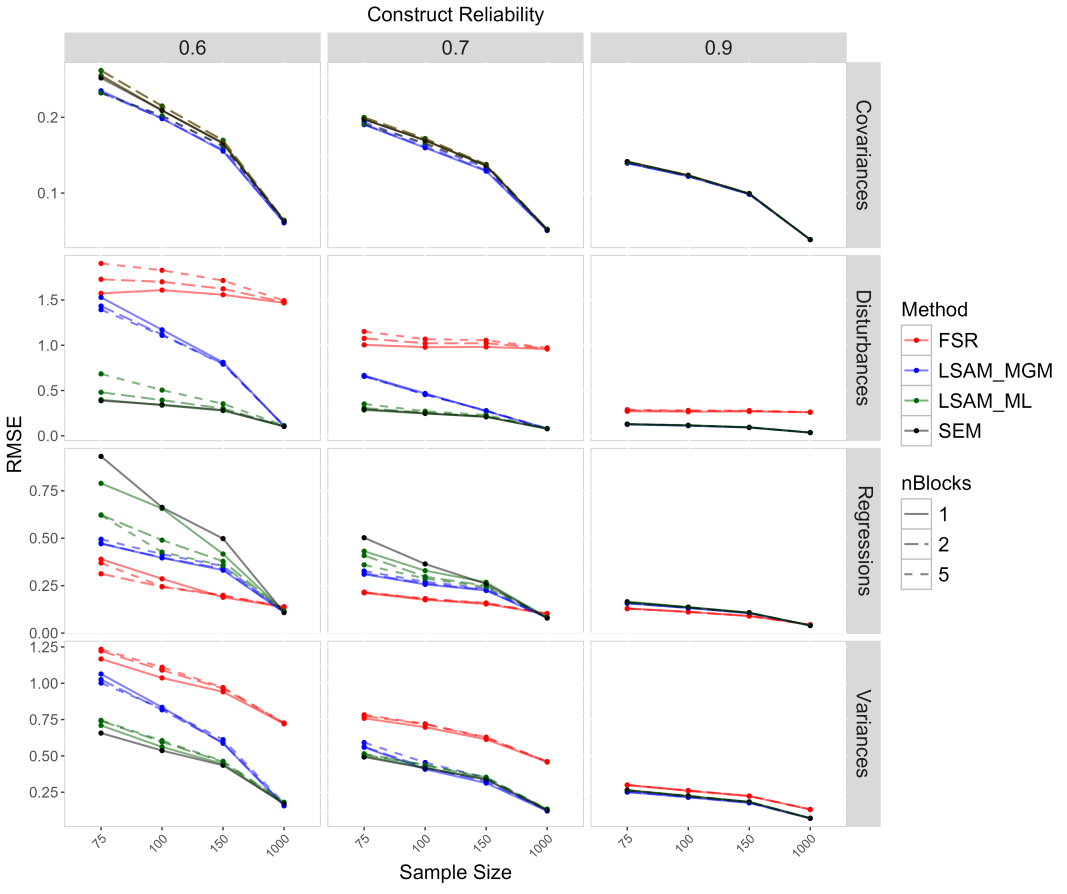

Figure 5. RMSE: Construct Reliability X Parameter X Sample Size X Method

number of blocks increased, however the opposite was true for LSAM-MGM.

Standard deviation estimates for SEM and LSAM-ML were found to be the lowest across all methods when construct reliability was .70 or lower and sample sizes were 150 or less. Interestingly, LSAM-ML with one measurement block possessed nearly identical estimates to SEM, whereas, more measurement blocks resulted in larger SD estimates. When construct reliability was .60, LSAM-MGM was found to have larger SD estimates than FSR, on average, when sample size was 75 or 100. Like with bias, the impact of the number of measurement blocks differed between FSR and LSAM-MGM, with the latter experiencing larger SD estimates with fewer measurement blocks. With higher construct reliability, LSAM with MGM estimation tended to have lower SD estimates than FSR.

In terms of RMSE, SEM and LSAM-ML possessed smaller RMSE estimates than LSAM-MGM and FSR; while LSAM-MGM nearly always had lower RMSE than FSR. The impact the number of measurement blocks had on RMSE differed based on the estimation method at stage-1, with ML estimation resulting in lower RMSE with fewer blocks, whereas MGM estimation resulted in higher RMSE with fewer measurement blocks. Finally, sample size had no impact on RMSE for FSR, whereas, larger sample sizes resulted in lower RMSE for the remaining estimation methods.

3.2.5 Latent Regressions

Bias was negligible for SEM, LSAM-ML, and LSAM-MGM with percent bias ranging between ±10%. LSAM-ML and SEM tended to overestimate, while LSAM-MGM tended to underestimate when sample sizes were 150 or smaller. On the other hand, FSR underestimated latent regression estimates, on average, by -20% (CR = .60), -15% (CR = .70), and -5% (CR = .90) with more bias resulting with larger sample sizes.

In terms of standard deviation, larger sample sizes and greater construct reliability resulted in smaller standard deviations. Factor score regression possessed the smallest estimates, followed by LSAM with MGM, LSAM with ML, and finally SEM. Interestingly, LSAM-ML appeared to be more impacted by the number of measurement blocks than FSR and LSAM-MGM. With respect to RMSE, a similar pattern was observed as was for SD. Where, FSR had lower RMSE compared to SEM, LSAM-ML, and LSAM-MGM.

4. Discussion

The goal of this study was to investigate the performance of two-stage estimation approaches to SEM in the context of small (or limited) sample sizes. Viable estimation approaches (iterative versus non-iterative; unbounded versus bounded) were used in the Monte Carlo simulation to better understand how these alternatives could be leveraged for the sake of model convergence and point estimation of structural parameters. Additionally, we investigated the impact the number of measurement blocks had on the performance of the two-stage estimation approaches.

In the context of unbounded estimation, two-stage approaches converged at a much higher rate than SEM with the former not dropping below 97%. Although SEM converged only $\sim$ 40% of the time in the most challenging setting (CR = .60; N = 75) when using unbounded ML, convergence rates increased dramatically when using bounds and did not drop below 99%. Use of the non-iterative (multiple group method) estimator at stage-1 for LSAM resulted in a 100% convergence rate across all simulation settings and bound types, highlighting the advantage of using non-iterative estimation methods. With respect to bound types, we found standard and wide bounds to result in 100% convergence rates for SEM, FSR, and LSAM-ML across all simulation approaches; while use of positive variance bounds possessed convergence rates of 96% or higher.

Importantly, we found that the use of bounded estimation did not, on average, have an impact on the recovery of structural parameters when controlling for estimation method. Collapsing across bounded estimation conditions, all estimation methods outperformed FSR with respect to bias and this difference was most obvious for latent disturbances, latent variances, and latent regressions. In terms of Local SAM, we observed that ML estimation at stage-1 appeared to be less biased and had smaller RMSE than when MGM estimation was used at stage-1 for all parameter types, with latent regressions being the exception. In terms of the impact the number of measurement blocks had on bias, we observed that Local SAM, regardless of stage-1 estimator, underestimated latent covariances to a greater degree when there was five measurement blocks (i.e., greater than -10%) compared to when there were fewer measurement blocks. If the goal is to recover latent covariances, latent disturbances, and latent variances SEM and LSAM-ML with a single measurement block appear to be the least biased with the smallest RMSE and should be considered. If the goal is to recover latent regressions, LSAM with MGM estimation at stage-1, regardless of the number of measurement blocks, should be considered as it appears to provide better parameter recovery with lower RMSE and SD than LSAM-ML, SEM, and FSR.

4.1 Limitations

As with all studies, there are certain limitations which should be acknowledged. First, we only considered one type of model which was not small in scope, nor was it large, therefore, results may not generalize to all SEMs. Second, there were a total of 14 free structural parameters leading to

minimum of 5.35 observations per free parameter, perhaps investigations with sample sizes smaller than 75 might be worth while. Third, we generated complete data that was multivariate normal, therefore, we are unable to comment on convergence and/or structural accuracy given non-normal observed data nor the impact missing data might have on results. Fourth, factor loadings exhibited tau equivalence and therefore, did not vary in magnitude. Fifth, we exclusively looked at point estimates of structural parameters and did not investigate the performance of standard errors when using ML or bootstrapping when using the multiple group method. Finally, the fitted models in our simulation matched the data generation model preventing us from examining the degree to which a misspecified structural model impacts the measurement model.

In conclusion, the use of Local SAM shows promise in the context of limited sample sizes and outperforms FSR in terms of bias. Additionally, the use of bounded estimation is a tool that is accessible to researchers in the `lavaan` (Rosseel, 2012) and was found to have no impact on structural parameter recovery while nearly guaranteeing model convergence.

References

Bartlett, M. S. (1937). The statistical conception of mental factors. *British journal of Psychology*, *28*(1), 97.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological methods & research*, *5*(1), 3–52.

De Jonckere, J., & Rosseel, Y. (2022). Using bounded estimation to avoid nonconvergence in small sample structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 412–427.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770.

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis. *Methodology*.

Dhaene, S., & Rosseel, Y. (2022). Resampling based bias correction for small sample sem. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(5), 755–771.

Dhaene, S., & Rosseel, Y. (2023). An evaluation of non-iterative estimators in the structural after measurement (sam) approach to structural equation modeling (sem). *Structural Equation Modeling: A Multidisciplinary Journal*, 1–15.

Dhaene, S., & Rosseel, Y. (2024). An evaluation of non-iterative estimators in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(1), 1–13.

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65–83.

Guttman, L. (1952). Multiple group methods for common-factor analysis: Their basis, computation, and interpretation. *Psychometrika*, *17*(2), 209–222.

Johnson, P. E. (2022). Package 'rockchalk'.

Kosmidis, I. (2014). Bias in parametric estimation: Reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(3), 185–196.

Lance, C. E., Beck, S. S., Fan, Y., & Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological methods*, *21*(3), 388.

Levy, R. (2017). Distinguishing outcomes from indicators via bayesian modeling. *Psychological methods*, *22*(4), 632.

McNeish, D., & Hancock, G. R. (2018). The effect of measurement quality on targeted structural model fit indices: A comment on lance, beck, fan, and carter (2016).

Rifenbark, G. G. (2022). Impact of construct reliability on proposed measures of structural fit when detecting group differences: A monte carlo examination. *The Annual Meeting of the Psychometric Society*, 313–328.

Rifenbark, G. G., & Jorgensen, T. D. (2023). The plausibility and feasibility of remedies for evaluating structural fit. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology: The 87th annual meeting of the psychometric society, bologna, italy, 2022* (pp. 147–159). https://doi.org/10.1007/978-3-031-27781-8_14

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 *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 226–238). Routledge.

Rosseel, Y., & Loh, W. W. (2024). A structural after measurement approach to structural equation modeling. *Psychological Methods*.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*, 563–575.

Thurstone, L. L. (1934). The vectors of mind. *Psychological review*, *41*(1), 1.