Handling dependent samples in meta-analytic structural equation models: A Wishart based approach

James Ohisei Uanhoro
Research, Measurement & Statistics, Department of Educational Psychology
University of North Texas

Abstract

We present an approach to meta-analytic structural equation models that relies on hierarchical modeling of sample covariance matrices under the assumption that the matrices are Wishart. The approach handles the commonplace fixed- and random- effects meta-analytic SEMs, and solves the problem of dependent covariance matrices where more than one covariance matrix is obtained from a single study or study author. The ability of the approach to adequately recover parameters is examined via a simulation study. The approach is implemented in the bayesianmasem R package and a demonstration shows applications of the model.

Structural equation modeling (SEM) is a popular multivariate data analysis technique for modeling covariance structures. Meta-analytic SEM (MASEM, Cheung & Chan, 2005; Viswesvaran & Ones, 1995) combines ideas from meta-analysis (Hedges & Olkin, 1985) and SEM to identify the covariance or correlation structure underlying observed covariance or correlation matrices. When collating sample covariance or correlation matrices, several matrices may be obtained from the same paper, or the same (set of) authors. This creates dependencies between observed matrices and has the potential to distort MASEM estimation and inference. In this paper, we present a Wishart-based MASEM to handle the case of dependent covariance matrices. We focus on latent variable models, thus excluding meta-analytic path models, though the model may easily be extended to meta-analytic path models. Before discussing the approach, we briefly review some established MASEM methods.

The most well-known MASEM method is two-stage SEM (TSSEM, Cheung & Chan, 2005, 2009). In TSSEM, the meta-analysis is usually done on sample correlation matrices. In the first step, the practitioner computes the average correlation matrix from the constituent correlation matrices. In the second step, the practitioner estimates the hypothesized SEM on the average correlation matrix using weighted-least squares estimation which accounts for uncertainty about the computed average correlation matrix. The most important decision in TSSEM is made in the first step where the modeller either assumes a fixed-effects or random-effects model. In the fixed-effects model, the same correlation matrix is believed to underlie the observed correlation matrices with differences between the observed matrices attributed to sampling error often assumed to be of known form (e.g. equations 3 – 5 in Olkin & Finn, 1995). In the random-effects model, each correlation matrix is assumed to arise from a distinct population correlation matrix, such that

The Version of Record of this manuscript has been published and is available in Behavior Research Methods (2024) https://doi.org/10.3758/s13428-024-02340-4

differences in the observed correlation matrices are due to both different populations and sampling error. The different populations are typically operationalized as multivariate normal deviations from the average true correlations. Practically, the recommendation between a fixed- or random-effects model is based on the size of the gap between the observed correlation matrices after accounting for sampling error. This gap is operationalized using standard SEM goodness of fit indices such as the root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR). When these fit indices are judged acceptable, a fixed-effects model is assumed to be a reasonable approximation to the patterns in the data.[1]

An alternative MASEM method is one-stage maximum-likelihood method (ML-MASEM, Oort & Jak, 2016), which fits a multi-group SEM to the observed correlation matrices with the parameter estimates constrained to be identical across studies. Hence, ML-MASEM is a fixed-effects approach. ML-MASEM has been recently extended to include random-effects models and study-level moderators (OSMASEM, Jak & Cheung, 2020), such that heterogeneity of parameters across studies can be accounted for and explained. ML-MASEM is asymptotically equivalent to TSSEM (Yuan & Kano, 2018), while OSMASEM improves over TSSEM. Both TSSEM and OSMASEM are implemented in the metaSEM R package (Cheung, 2015).

A conceptually different approach to MASEM is parameter-based MASEM (Cheung & Cheung, 2016). Conceptually, a multi-group SEM is fit to the correlation matrices with the parameter estimates freed to be different across studies. Alternatively, the parameter estimates may be directly extracted from multiple study reports. The resulting parameter estimates are then meta-analyzed. An interesting contribution to this tradition is to hierarchically model the parameter estimates across studies using Bayesian estimation within a one-stage model (Ke, Zhang, & Tong, 2019). Each parameter estimate may then be summarized via an estimated mean and standard deviation. This approach is conceptually appealing as it extends the benefits of hierarchical modeling to MASEM parameter estimation.

In the next section of the paper, we lay out the proposed Wishart methods for MASEM. Then we provide simulation results that show the method allows for valid inference and parameter recovery. Afterward, we present a data analysis example that shows some of the properties of the proposed MASEM methods. We then conclude with discussion of the approach and some thoughts for further developing the approach. Finally, code for simulation studies and data analysis examples is available at `https://osf.io/yd5q4/`.

Wishart-based MASEM

Our work builds on a Wishart-based Bayesian approach to random-effects MASEM (Uanhoro, 2023). Here, we develop the Wishart-based MASEM as a complete system for MASEMs including fixed-effects, random-effects and dependent samples MASEM. For Wishart-based MASEM, one meta-analyzes the covariance matrices directly when they are available, as opposed to the correlation matrices as typical in TSSEM.

We begin with the observation that the $p \times p$ sample covariance matrix, $\mathbf{S}$, for $n \times p$ multivariate normal data is a Wishart matrix:

$$n^*\mathbf{S} \sim \mathcal{W}_p(\mathbf{\Sigma}, n^*), \tag{1}$$

[1] Although this is the convention with TSSEM, we believe that random-effects models should be preferred by default as it is unlikely that different studies sample the exact same population. Moreover, the classification of a fit index as 'small' is not straightforward (e.g. McNeish, An, & Hancock, 2018; Savalei, 2012; Ximénez, Maydeu-Olivares, Shi, & Revuelta, 2022).

where $n^* = n - 1$, $\boldsymbol{\Sigma}$ (scale matrix) is the population covariance matrix underlying the data. As $n^* \to \infty$, $\mathbf{S} \to \boldsymbol{\Sigma}$, and the effect of sampling error is negligible.

Fixed-effects model

The sum of k Wishart matrices that share a common scale matrix is itself Wishart with the same scale matrix (Gupta & Nagar, 1999, Theorem 3.3.8):

$$\sum_{i=1}^{k} \left[n_i^* \mathbf{S}_i\right] \sim \mathcal{W}_p \left(\boldsymbol{\Sigma},\ \sum_{i=1}^{k} n_i^*\right) \text{ if } (n_i^* \mathbf{S}_i) \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n_i^*) \text{ for } i \in \{1, \ldots, k\}, \tag{2}$$

where $\boldsymbol{\Sigma}$ in equation 2 is the pooled covariance matrix under a fixed-effects model. $\boldsymbol{\Sigma}$ may be further assumed to be a structured covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, such that one directly estimates the SEM parameters, $\boldsymbol{\theta}$. Hence, this would be a one-stage fixed-effects MASEM. For meta-analytic confirmatory factor analysis (CFA) (the most common latent-variable MASEM[2]), the one-stage fixed-effects MASEM for k covariance matrices is:

$$\sum_{i=1}^{k} \left[n_i^* \mathbf{S}_i\right] \sim \mathcal{W}_p \left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Delta},\ \sum_{i=1}^{k} n_i^*\right), \tag{3}$$

where $\boldsymbol{\Lambda}$ is the loading matrix with certain elements set to 0 based on model identification and substantive considerations, $\boldsymbol{\Phi}$ is the inter-factor correlation matrix and $\boldsymbol{\Delta}$ is the residual covariance matrix. The uncertainty in the observed covariance matrices is assumed a function of study sample sizes (and the Wishart distribution).

Random-effects model

Given equation 1, the population covariance matrix, $\boldsymbol{\Sigma}$, may be assumed to be inverse-Wishart (Wu & Browne, 2015):

$$\boldsymbol{\Sigma} \sim \mathcal{W}_p^{-1}(\boldsymbol{\Omega} \times m, m), \tag{4}$$

where $\boldsymbol{\Omega}$ is the true covariance matrix underlying the population covariance matrix and $m > p - 1$ is the degrees of freedom and functions as a precision parameter – as $m \to \infty$, $\boldsymbol{\Sigma} \to \boldsymbol{\Omega}$. Wu and Browne (2015) assumed $\boldsymbol{\Omega}$ to be a structured covariance matrix, $\boldsymbol{\Omega}(\boldsymbol{\theta})$, such that differences between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}(\boldsymbol{\theta})$ are due to what Wu and Browne (2015) term *adventitious error* – error that arises because the exact study population differs from the hypothetical population for which the psychometric theory holds.

The models in equations 1 and 4 form a hierarchical model for $\mathbf{S}$ – the primary interest is in estimating $\boldsymbol{\Omega}$ not $\boldsymbol{\Sigma}$. Integrating $\boldsymbol{\Sigma}$ out, the resulting marginal distribution for $\mathbf{S}$ is a generalized matrix variate beta type II (GMB-II) distribution (Granström & Orguner, 2011; Wu & Browne, 2015):

$$\mathbf{S} \sim \mathrm{GB}_p^{\mathrm{II}} \left(\frac{n^*}{2}, \frac{m}{2}, \frac{m}{n^*}\boldsymbol{\Omega}, \mathbf{0}_{p \times p}\right), \tag{5}$$

with log-likelihood:

[2] Although path and regression models are the most common MASEMs, these models do not involve latent variables.

$$\begin{aligned}\ln \mathcal{L} = f(p, m+n^*) - f(p,m) - f(p,n^*) + \frac{1}{2}\Bigg(\\ (n^*-p-1)\ln|\mathbf{S}| + m\ln|\mathbf{\Omega}| - (n^*+m)\ln\left|\frac{m\mathbf{\Omega}+n^*\mathbf{S}}{m+n_i^*}\right|\Bigg),\end{aligned} \tag{6}$$

where $f(p,x) = \ln\Gamma_p(x/2) - \frac{1}{2}\left[xp\ln(x/2) - xp\right]$, and Γ_p is the multivariate gamma function (Gupta & Nagar, 1999, definition 1.4.2).

The model in equation 5 is a hierarchical model for $\mathbf{S}$ and may be extended to a one-stage random-effects MASEM for k covariance matrices (Uanhoro, 2023):

$$\mathbf{S}_i \sim \text{GB}_p^{\text{II}}\left(\frac{n_i^*}{2}, \frac{m}{2}, \frac{m}{n_i^*}\mathbf{\Omega}(\boldsymbol{\theta}), \mathbf{0}_{p\times p}\right) \text{ for } i \in \{1,\ldots,k\} \tag{7}$$

where $\mathbf{\Omega}(\boldsymbol{\theta})$ is the pooled structured covariance matrix, e.g. $\mathbf{\Omega}(\boldsymbol{\theta}) = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Delta}$.

The uncertainty in this model is a function of the sample sizes of the constituent studies and the m parameter. m is a precision parameter that captures the average gap of the different population covariances underlying each study to the shared covariance structure, $\mathbf{\Omega}(\boldsymbol{\theta})$. A simple transformation of m eases its interpretation, $\varepsilon = (m-p+1)^{-1/2}$, where ε is the RMSEA (Wu & Browne, 2015).[3] When ε is low, a fixed-effects model may provide a reasonable approximation to the patterns underlying the observed covariance matrices. Additionally, ε may be assumed different for different studies, and can be modeled to create a meta-regression (Uanhoro, 2023). However, our intent here is to broadly lay out Wishart-based MASEMs, so we assume a constant ε across studies.

Dependent-samples model

Assuming k covariance matrices were obtained from c clusters of covariance matrices, we propose the following hierarchical model for dependent covariance matrices:

$$\begin{aligned}\mathbf{S}_{ij} \sim \text{GB}_p^{\text{II}}\left(\frac{n_i^*}{2}, \frac{m_1}{2}, \frac{m_1}{n_i^*}\mathbf{\Psi}_{j[i]}, \mathbf{0}_{p\times p}\right) \text{ for } i \in \{1,\ldots,k\} \\ m_2\mathbf{\Psi}_j \sim \mathcal{W}\left(\mathbf{\Omega}(\boldsymbol{\theta}), m_2\right) \text{ for } j \in \{1,\ldots,c\}\end{aligned} \tag{8}$$

where $\mathbf{\Psi}_j$ is an unstructured covariance matrix that varies by cluster j. $\mathbf{\Psi}_j$ is assumed Wishart with a scale parameter that is the true structured covariance matrix, $\mathbf{\Omega}(\boldsymbol{\theta})$. Hence, this remains a hierarchical model; the within-cluster variation is controlled by m_1 (precisely: $v_1 = (m_1 - p + 1)^{-1}$) and the between-cluster variation is controlled by m_2 ($v_2 = (m_2 - p + 1)^{-1}$); such that $v_1 + v_2 = v$, where v is the variation between population covariances and the pooled structured covariance matrix in the standard random-effects model. Additionally, the overall (ε), within-cluster (ε_1) and between-cluster (ε_2) RMSEAs can be computed, $\varepsilon_{(1/2)} = 1/\sqrt{m_{(1/2)} - p + 1}$. Finally, the proportion of variance that is between-cluster may also be of interest, $\rho = v_2(v_1+v_2)^{-1}$. ρ is akin to the intraclass correlation coefficient in multilevel modeling, and the higher the value of ρ, the less adequate a random-effects model when covariance matrices are clustered.

[3] ε is often close in value to the RMSEA obtained from a multi-group SEM where parameters are constrained equal across groups.

Notes on the Wishart methods

The Wishart approach makes clear the data generation process for MASEM and sheds light on the nature of the pooled structured covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ in the fixed-effects case and $\boldsymbol{\Omega}(\boldsymbol{\theta})$ in the random-effects and dependent-samples cases. The approach assumes the pooled structured covariance matrix underlies the observed covariance matrices. In the fixed-effects case, only sampling error is responsible for differences between the observed covariance matrices. In the random-effects case, adventitious error or study-specific context e.g. non-random sampling of cases generates an intermediate population covariance matrix (true for the specific population sampled) between $\boldsymbol{\Omega}(\boldsymbol{\theta})$ and the observed covariance matrix. In the dependent-samples case, clustering creates another level of variation.

We apply Bayesian estimation to these models. Notably, the fixed-effects and random-effects models can easily be estimated using maximum likelihood. But the dependent-samples model requires the estimation of $\boldsymbol{\Psi}_j$ – a covariance matrix that varies by cluster, a problem that is more readily amenable to Bayesian estimation.

Comparison to extant MASEM methods

We briefly compare our Wishart approach to TSSEM since it is the most popular MASEM method, and to the one-stage method of Ke et al. (2019) as it is an alternative Bayesian approach.

TSSEM. The major difference in the construction of our Wishart approach and TSSEM is how deviations due to different populations (or random-effects) are parameterized. In TSSEM, the random-effect deviations of a group's correlation vector from the true correlation vector are a zero-mean multivariate normal vector with an unstructured covariance matrix shared across studies. Hence, there are as many random-effect variance parameters as there are unique elements in a correlation matrix. In our Wishart approach, the random-effect dispersion of covariance elements between studies from the true structure is controlled by a single parameter, m, equation 7. Hence the random-effect dispersion is of a more restrictive form, as determined by the inverse-Wishart distribution.[4] In practice, m may be permitted to vary across studies in a structured form (e.g. Uanhoro, 2023), or in an unstructured form where each group has a unique value of m. However, the dispersion of covariance elements is still determined by the inverse-Wishart assumption. This restrictiveness compares less favourably to TSSEM, however, it may provide some regularization benefits. For example, when the number of items is large, it may be quite difficult to correctly estimate the unstructured random-effect covariance matrix in TSSEM (Cheung, 2015) necessitating the need for simpler covariance structures e.g. a diagonal covariance matrix. Finally, the two-stage nature of TSSEM allows the modeller to utilize the pooled correlation matrix for analysis in their preferred SEM software – this is a very practical benefit of TSSEM.

Bayesian one-stage method of Ke et al. (2019). Ke et al. (2019)'s method assumes that the hypothesized structure holds within each group though the structural parameters are different across studies, e.g. as in weak or configural invariance. And each set of structural parameters (e.g. all loadings) is a multivariate normal vector; the true parameter vector (the effect sizes of interest) and the unstructured covariance matrix underlying the structural parameters in each study are then estimated. Conceptually, the hypothesized configuration between observed and latent variables is realized in each study, though parameters may vary across studies. Our approach alternatively assumes that it is the structured covariance matrix that is perturbed in each study not model parameters, such that the hypothesized structure may not hold in certain studies. We

[4] As m gets larger, the random-effects deviations as implied by the inverse-Wishart distribution is increasingly approximately a zero-mean multivariate normal vector (Wu & Browne, 2015) but with a more constrained covariance matrix than the unstructured covariance matrix estimated by TSSEM.

consider our assumption that the hypothesized structure may not hold in certain studies to be more realistic (Wu & Browne, 2015), making the Wishart approach less restrictive than the method of Ke et al. (2019).

Missing data handling

The input data for MASEM applications are correlation or covariance matrices. Missing data are most often a problem in meta-analytic path models where the different studies may collect different variables such that the correlation matrix is not fully available for all constituent studies. And the biggest challenge with extending the Wishart approach to path models is identifying credible procedures for handling the commonplace missing data challenges in path models. However, we focus on meta-analytic latent variable models in this paper – of which CFAs are the most common – and missing data are less likely to be a problem. However, we lay out a missing data strategy that should be adequate when the missing data mechanism is either missing completely at random or missing at random.

There are two general missing data scenarios in MASEM: (i) a variable may not be present in one or more of the constituent studies; or (ii) the covariance matrix is partially reported in one or more constituent studies (Jak & Cheung, 2018). Case ii usually occurs when only the covariance between predictors and outcomes are reported, as opposed to the full covariance matrix of all variables – this practice is common in regression analysis. For any given study, we only analyze variables that are at least partially reported in the study. In case i above, only the complete covariance matrix for each study is used to estimate the pooled structured covariance matrix. In case ii, missing covariances can be imputed within the Bayesian model based on the posterior predictive distribution.

Extant MASEMs result in poor performance when the missingness in covariance matrices are missing not at random (Furlow & Beretvas, 2005), and we do not expect the Wishart methods here to be any different. One example of missing not at random occurs when meta-analyzing long and short form versions of instruments together. This is because short form versions of instruments may include items chosen precisely for their stronger correlations (Widaman, Little, Preacher, & Sawalani, 2011). A related example occurs when researchers decide to focus on specific variables in a study because their observed correlations are large relative to other correlations which go unreported.

Implementation details

The methods here are implemented in the bayesianmasem R package.[5] The package relies on Markov Chain Monte Carlo estimation as implemented in Stan (Carpenter et al., 2017).

Simulation study

We conducted a simulation study to evaluate the adequacy of inference using the proposed models when applied to clustered covariance matrices. Uanhoro (2023) used simulation-based calibration (SBC, Talts, Betancourt, Simpson, Vehtari, & Gelman, 2018) to evaluate the original presentation of the Wishart approach, but we opted for traditional Monte Carlo simulation. This allowed us assess parameter recovery under purposely varied parameter estimates related to potential real world scenarios. However, given that Bayesian inference is better validated using SBC, we describe the results of an SBC study in appendix B – the results suggest proper calibration for the studied problem.

[5] bayesianmasem does not handle the problem of missing data.

Data generation and design conditions

The data generation process was based on the dependent-samples model in equation 8:

$$\begin{aligned}
&\mathbf{S}_{ij} \sim \text{GB}_p^{\text{II}}\left(\frac{n_i^*}{2}, \frac{m_1}{2}, \frac{m_1}{n_i^*}\boldsymbol{\Psi}_{j[i]}, \mathbf{0}_{p\times p}\right) \text{ for } i \in \{1,\ldots,k\} \\
&\quad m_2\boldsymbol{\Psi}_j \sim \mathcal{W}\left(\boldsymbol{\Omega}(\boldsymbol{\theta}), m_2\right) \text{ for } j \in \{1,\ldots,c\} \\
&\quad \boldsymbol{\Omega}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Delta},\ \boldsymbol{\Lambda}' = \left[\begin{smallmatrix} 0.7 & 0.8 & 0.6 & 0.9 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 & 0.7 & 0.75 & 0.85 & 0.6 \end{smallmatrix}\right], \\
&\quad \boldsymbol{\Phi} = \left[\begin{smallmatrix} 1 & \\ .3 & 1 \end{smallmatrix}\right],\ \boldsymbol{\Delta} = \text{diag-matrix}\left(\text{diag}\left(\mathbf{I}_{p\times p} - \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'\right)\right)
\end{aligned} \tag{9}$$

We varied three features within the simulation study:

1. number of clusters ranging from very small to moderate, $c \in \{5, 15, 25\}$;
2. size of the RMSEA ranging from small to non-ignorable, $\varepsilon \in \{0.05, 0.08, 0.15\}$; and
3. proportion of dispersion between clusters ranging from small to large, $\rho \in \{5\%, 35\%, 75\%\}$

This resulted in 27 ($3 \times 3 \times 3$) design conditions, with each condition repeated 1000 times. The model parameters m_1 and m_2 were computed according to the following equations: $m_1 = ((1-\rho) * \varepsilon^2)^{-1} + p - 1,\ m_2 = (\rho * \varepsilon^2)^{-1} + p - 1$. While exploring common MASEM datasets, we found that it is common for most clusters to contain only one covariance matrix. So we set the average number of covariance matrices per cluster to 1.4 on average. The exact cluster sizes for c clusters were:

$$[\underbrace{1,\ldots,1}_{a \text{ times}}, x_1, \ldots, x_b],\ \text{where } b = c - a,\ [x_1,\ldots,x_b] \sim 2 + \text{Poisson}(\alpha) \tag{10}$$

When $c = 5,\ a = 3,\ \alpha = 0$ with 50% probability and $a = 4,\ \alpha = 1$ with 50% probability. When $c = 15,\ a = 10,\ \alpha = 0.2$, and when $c = 25,\ a = 17,\ \alpha = 0.25$. As an example, when $c = 25$ clusters, 17 clusters contained only 1 covariance matrix, and the 8 remaining clusters had an average cluster size of 2.25, resulting in $17(1)+8(2.25) = 35$ sample covariance matrices; $35/25 = 1.4$ covariance matrices per cluster. Finally, the average sample size was 300, and no covariance matrix had a sample size less than 100 cases. Precisely, the sample size, n_{ij}, for covariance matrix i in cluster j was generated according to the following equation:

$$n_{ij} \sim 100 + \text{Poisson}(q_{j[i]}),\ q_j \sim \text{Negative-binomial}(\mu = 200, \phi = 20)$$

where ϕ is the dispersion parameter. The sample size generation is a hierarchical model such that studies in the same cluster had more similar sample sizes.

Analytical methods

For each generated dataset, we analyzed the data with both the random-effects (equation 7) and dependent-samples (equation 8) models using the following priors:

$$\begin{aligned}
&\boldsymbol{\lambda} \sim \mathcal{N}(0, \sigma_\lambda),\ \frac{\phi_{2,1}+1}{2} \sim \text{Beta}(2,2),\ \sqrt{\text{diag}(\boldsymbol{\Delta})} \sim t^+(3,0,2.5), \\
&\sigma_\lambda \sim t^+(3,0,1),\ \ln\left([m, m_1, m_2] - p + 1\right) \sim t^+(3,0,2.5)
\end{aligned} \tag{11}$$

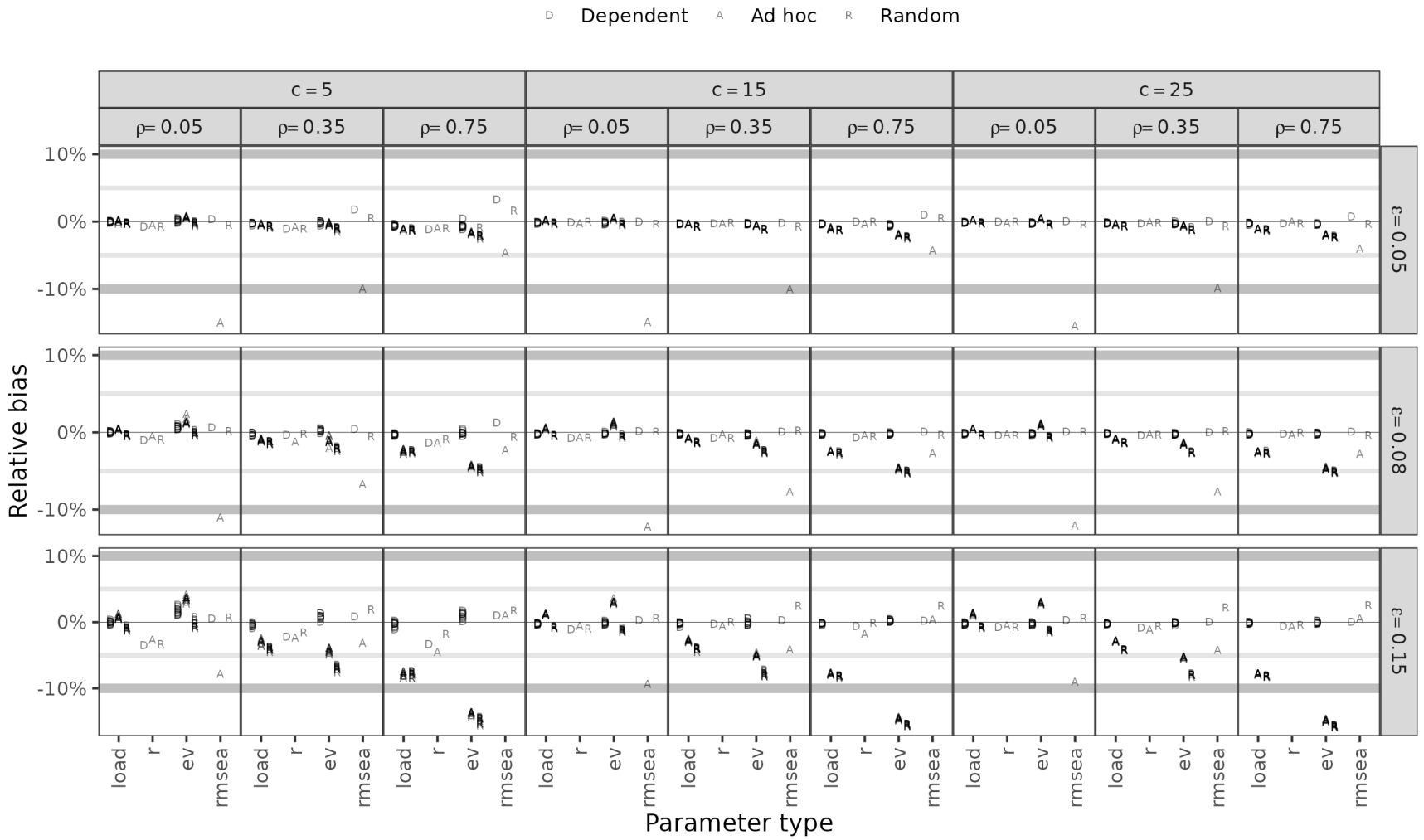
where non-zero loadings ($\boldsymbol{\lambda}$) have a normal prior with scale hyperparameter, σ_λ. Both σ_λ and residual standard deviations ($\sqrt{\text{diag}(\boldsymbol{\Delta})}$) have half-t priors (Gelman, 2006) creating a weakly regularizing effect (Lemoine, 2019) given that true the total variable variances are 1. The prior for the interfactor correlation ($\phi_{2,1}$) is boundary-avoiding. Loadings and the interfactor correlation are sign-corrected post-sampling (e.g. Conti, Frühwirth-Schnatter, Heckman, & Piatek, 2014; Merkle, Fitzsimmons, Uanhoro, & Goodrich, 2021) to correct for rotational indeterminacy of latent variables (Peeters, 2012).

We also included an *ad hoc* approach where we reduced all the covariance matrices in a cluster to a single covariance matrix by assuming a fixed-effects model for all covariance matrices in the cluster. Precisely, given h covariances in a cluster, the reduced covariance matrix was $\left[\sum_{i=1}^{h}\left(n_i^*\mathbf{S}_i\right)\right] \Big/ \left[\sum_{i=1}^{h} n_i^*\right]$, and the sample size for this covariance matrix was $1 + \sum_{i=1}^{h} n_i^*$. This eliminates the dependence problem in the analyzed covariance matrices, reducing the number of covariance matrices to the number of clusters. And the covariance matrices were analyzed using a random-effects model.

We ran all models in parallel across 3 chains. For the random-effects and ad hoc models, we dropped the first 500 iterations per chain and retained the final 500 iterations resulting in 1,500 posterior samples per parameter. For the dependent-samples models, we dropped the first 750 iterations per chain and retained the final 750 iterations resulting in 2,250 posterior samples per parameter. We retained more samples for the dependent-samples model because the model is more complex than the random-effects model.

Figure 1

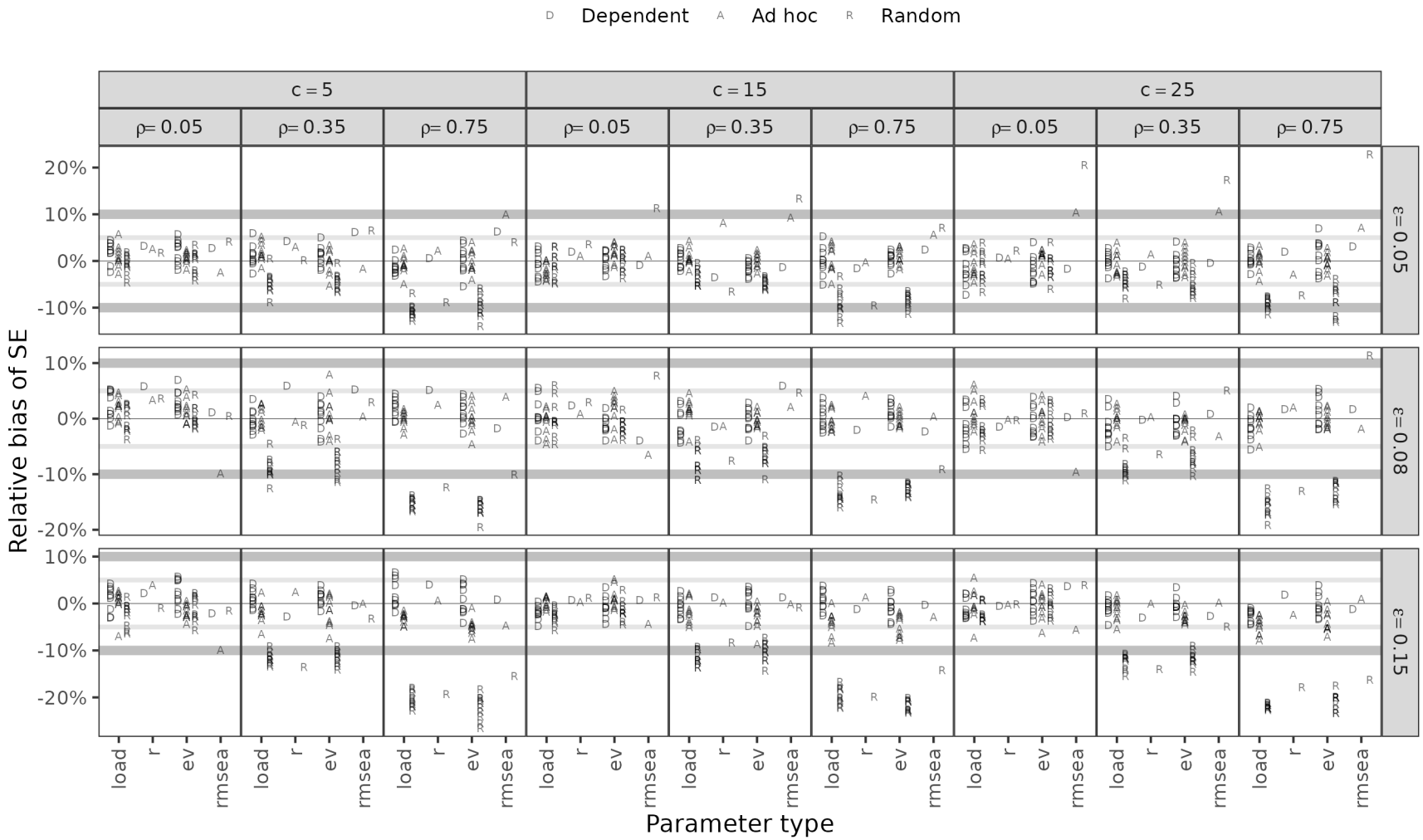
Relative bias of mean of posterior distribution



Note. load. = 10 loading estimates, r = inter-factor correlation, ev. = 10 error variance estimates.

Figure 2

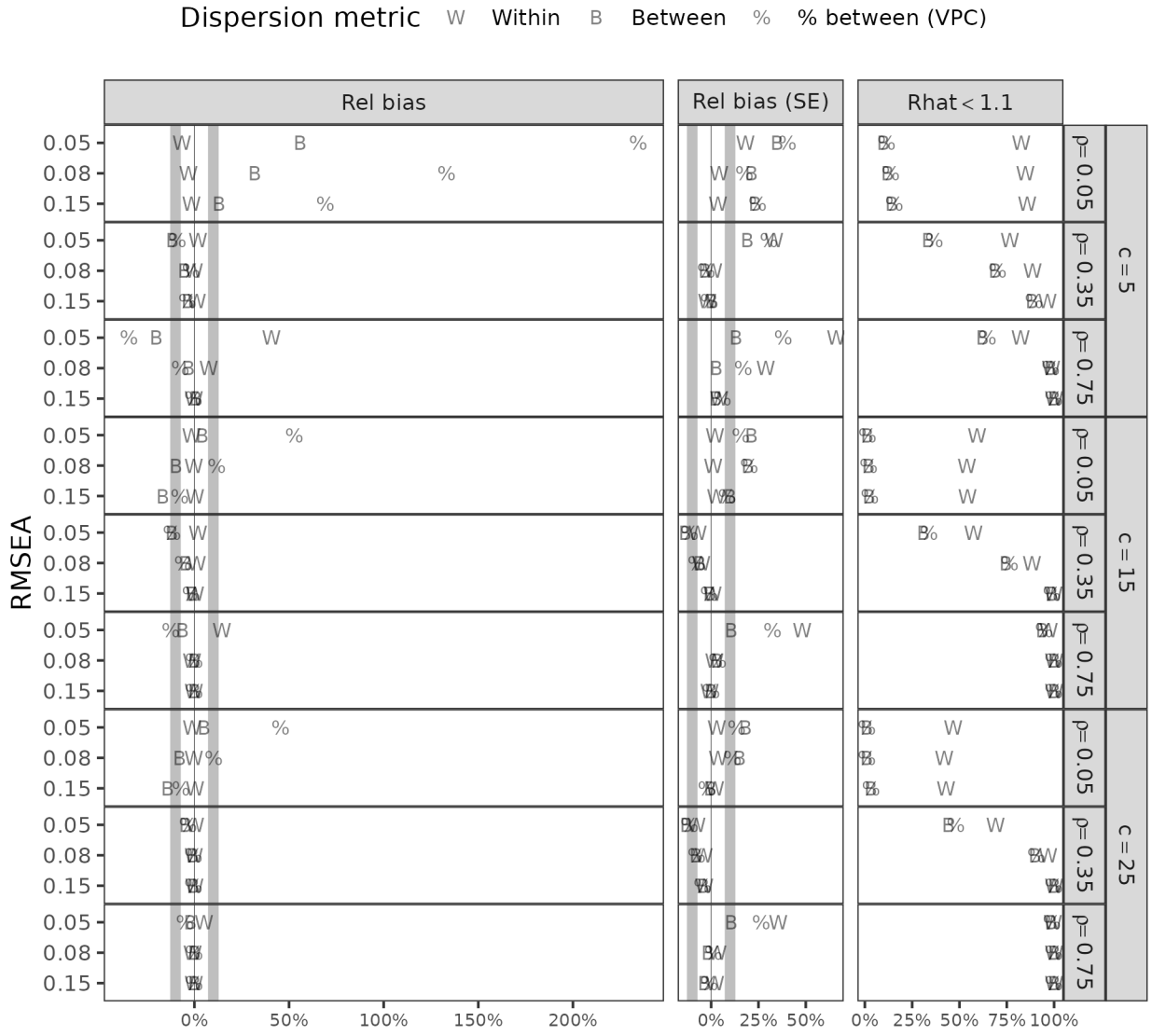
Relative bias of standard deviation of posterior distribution



Note. load. = 10 loading estimates, r = inter-factor correlation, ev. = 10 error variance estimates.

Figure 3

Recovery of dispersion parameters for the dependent-samples model



Note. Within = ε_1, Between = ε_2, % between = ρ

Evaluation metrics

We were interested in the recovery of the structural parameters: loadings, interfactor correlation and residual standard deviations for all models. We were also interested in the recovery of dispersion parameters: ε for all models, and ε_1, ε_2 and ρ for the dependent-samples model.

For each parameter, we had three assessment metrics: bias of the mean of the posterior distribution, bias of the standard deviation of the posterior distribution, and the empirical coverage rate (ECR) of the 90% credible interval. We transformed bias to relative bias deeming relative bias within ±5% as ideal and ±10% as acceptable. For coverage, we set (87.5%, 92.5%) and (85%, 95%) as ideal and acceptable limits for the 90% ECR respectively.

Simulation results

For the random-effects and ad hoc models, $\widehat{R}$ was less than 1.05 more than 99% of the time for all parameters – parameter convergence problems were ignorable. For the dependent-samples model, $\widehat{R}$ was less than 1.05 about 95% of the time for all parameters with some exceptions, suggesting minor problems with parameter convergence. The exceptions were the dispersion parameters, ε_1, ε_2 and ρ. Hence, in addition to reviewing results related to bias and inference of structural parameters, we also review bias, inference and convergence of these dispersion parameters in the dependent-samples model.

Relative bias of mean of posterior distribution of structural parameters and ε. As shown in Figure 1, structural parameter posterior means were largely unbiased ($< |10\%|$) for all models with some exceptions. There was an increasing downward bias in the loadings and error variances from the random-effects and ad hoc models at higher levels of ρ.[6] This bias was especially magnified when ε is large (0.15). Hence, the bias is a function of both ρ and ε.[7] The estimate of ε was never biased for both the random-samples and dependent-samples models. However, the estimate of ε for the ad hoc model was downwardly biased with increased bias at lower levels of ρ and ε. This bias is to be expected as the process of reducing all covariance matrices in a single cluster eliminates variability. And the more variability within clusters, the greater variability is eliminated by the ad hoc approach.

Relative bias of standard deviation of posterior distribution of structural parameters and ε. As shown in Figure 2, structural parameter posterior standard deviations were largely unbiased ($< |10\%|$) for the ad hoc and dependent-samples models. However, for the random-effects model, there was an increasing downward bias in the posterior standard deviations at higher levels of ρ. As with the bias in posterior means, the bias in posterior standard deviations was magnified when ε is large (0.15). Finally, the posterior standard deviation of ε was upwardly biased (about 20%) for the random-effects model when $\varepsilon = 0.05$ and $c = 25$.

Recovery of dispersion parameters for the dependent-samples model. The relative bias of the posterior mean, posterior standard deviation and the proportion of times $\widehat{R}$ was less than 1.10 for ε_1, ε_2 and ρ are reported in Figure 3. Adequate convergence of these parameters was the exception not the rule. Precisely, $\widehat{R} < 1.10$ was only commonplace when $c \in \{15, 25\}$ and $\rho = 75\%$. Generally, convergence for ε_1 and bias about the posterior mean and standard deviation for ε_1 were often better than for the other two parameters. This matches the observation that it is easier to accurately estimate residual or within variances in multilevel models than it is to estimate cluster-level variances and ICCs. Interestingly, estimation problems for these specific parameters did not distort the results for the structural parameters in the dependent-samples model (see Figures 1 and 2). Additionally, bias and inference for dispersion parameters was worst for the smallest number of clusters ($c = 5$) – this parallels the finding in multilevel models that estimation is more difficult with a small number of clusters. And convergence for dispersion parameters was worst when $\varepsilon = 0.05$. To demonstrate these patterns in an example, we report the parameter traceplots for a sample run of the dependent-samples model when $c = 5$, $\varepsilon = 0.05$, $\rho = 5\%$ – a simulation condition with poor recovery of dispersion parameters. The traceplots are clearly inadequate for ε_2 but appear adequate for other reported parameters, see Figure 4.

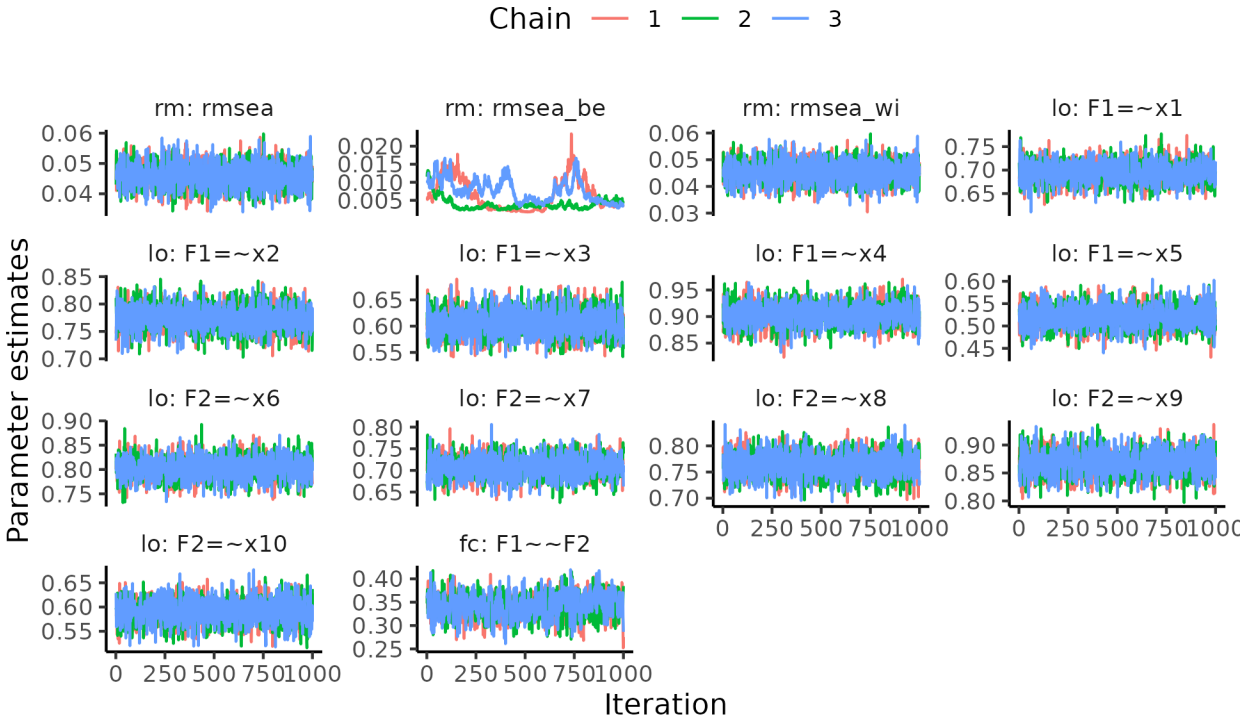
Summarizing the simulation findings. The dependent-samples model may be used for MASEM under the assumption that the covariance matrices are clustered. Assuming the approach

[6] The difference in relative bias for loadings and error variances occurs because variances are on the squared loading scale. Alternatively stated, the relative bias of loadings was the same as the relative bias of error standard deviations.

[7] Following from the mean of an inverse-Wishart distribution, the relative bias of model-implied covariance matrix elements in the random-effects model is: $\left[{}^{(m_1}/_{(m_1-p-1))} \Big/ {}^{(m}/_{(m-p-1))} \right] - 1$, $m_{(1)} = \varepsilon_{(1)}^{-2} + p - 1$.

Figure 4

Parameter traceplots for a sample run of the dependent-samples model with simulated data under unfavourable conditions for estimating dispersion parameters.



Note. Note poor convergence of ε_2 (rm:rmsea_be) for a randomly generated dataset under conditions: $c = 5,\ \varepsilon = 0.05,\ \rho = 5\%$.

we laid out is the data generating process, parameter estimates should be unbiased and inference about parameters should be adequate. The same cannot be said for the simpler random-effects model applied to such data. Random-effects model parameter estimates and inference are increasingly poor when there is higher variation due to clustering effects. However, the results also show the adequacy of the random-effects model applied to non-clustered data; this follows from the adequacy of the random-effects model when much of the variance was within clusters ($\rho = 5\%$). Like the random-effects model, the ad hoc approach produces biased estimates of the true parameters in the data generating process. However, the parameter posterior standard deviations in the ad hoc approach adequately convey uncertainty about the biased estimates. Parameter estimate bias in the ad hoc approach is determined by the true value of ε and ρ (which are not estimated by the ad hoc approach), so there is no way to bias correct estimates in practice. However, given the adequate inference about these biased estimates, the ad hoc approach may serve as a reasonable medium between the inadequate random-effects and the dependent-samples model. Finally, there are many conditions when the dependent-samples model struggles to adequately estimate all its dispersion parameters, precisely: ε_1, ε_2, and ρ – there is no problem with estimating ε.

Data demonstration

The data are 14 inter-factor correlation matrices of a five-factor model in Digman (1997), previously analyzed by Cheung (2014). The first three indicators: agreeableness (A), conscientious-

Table 1

Model comparison results sorted by LOOIC

Model	LOOIC	ΔLOOIC	SE(ΔLOOIC)
Dependent (two factors)	−162.4	–	–
Dependent (one factor)	−144.7	17.7	7.4
Random-effects (two factors)	−113.2	49.2	24.7
Random-effects (one factor)	−47.6	114.8	31.3
Fixed-effects (two factors)	993.9	1156.3	363.7
Fixed-effects (one factor)	1662.3	1824.7	496.8
Ad hoc (two factors)	−92.6	–	–
Ad hoc (one factor)	−50.4	42.2	5.5

As with other commonplace information criteria, smaller values of LOOIC suggest better predictive performance of a model. ΔLOOIC are LOOIC differences from the best fitting model, dependent-samples with two-factors.

ness (C) and emotional stability (ES) are hypothesized to load onto an *Alpha* factor; the last two indicators: extroversion (E) and intellect (I) load onto a *Beta* factor. The 14 correlation matrices were clustered within 7 authors with the following cluster sizes: 1 (4 authors); 2 (1 author); and 4 (2 authors). We fit the hypothesized model, and also fit an incorrect unidimensional model for demonstration purposes. The same priors were retained from the simulation study.

We fit each of four Wishart methods (fixed-effects, random-effects, ad hoc, dependent-samples) to the hypothesized two-factor and incorrect unidimensional model – eight models in all. For each model, there were 1000 warmup iterations then 1000 iterations retained for inference across 4 chains. Sampler-specific diagnostics (Betancourt, 2017) were adequate for all estimated models. Across all models and parameters, the maximum $\widehat{R}$ (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2020) was 1.007 and minimum effective sample size was 794 suggesting parameter convergence for all parameters across models.

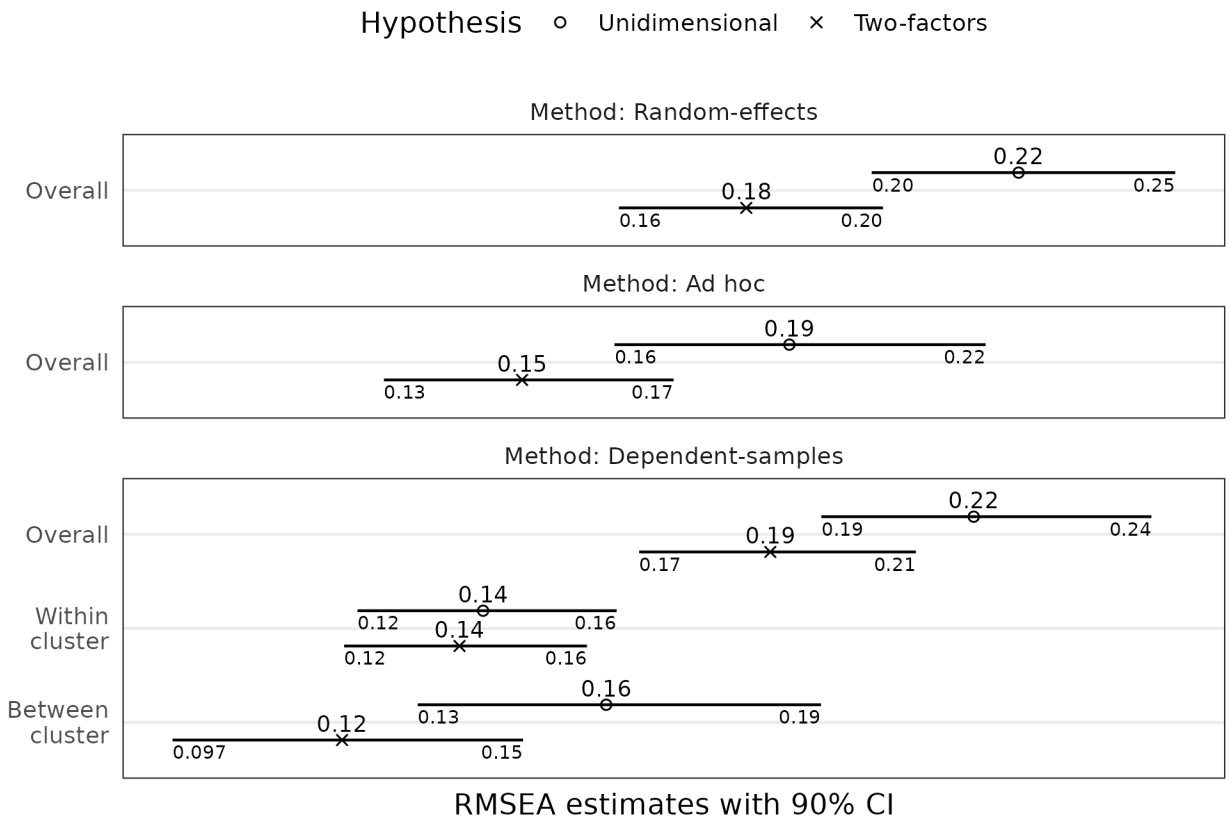
We compared the estimated models using approximate leave-one-out information criterion (LOOIC, Vehtari, Gelman, & Gabry, 2017) using the *loo* package (Vehtari, Gabry, et al., 2020) in R. We had to compare the ad hoc models separately since the number of correlation matrices for the ad hoc approaches (7) is different from the number of correlation matrices for other approaches (14). Comparing the fixed-effects, random-effects and dependent-samples models, the worst models were the fixed-effects models, while the dependent-samples models were the best performing models, see Table 1. Within each of the four types of model, the two-factor model hypothesis always had the lower LOOIC value, suggesting that the two-factor hypothesis was a better fit to the data than the one-factor hypothesis.

The RMSEA can also be used for comparing several random-effects models or comparing several dependent-samples models, see Figure 5. For the random-effects models, the two-factor model had the lower RMSEA, suggesting that the average distance between the pooled structured covariance matrix and the population covariance matrices underlying individual observed covariance matrices was lower for the two-factor model compared to the unidimensional model. We can conclude similarly based on the ad hoc models.

There are three sets of RMSEA values for the dependent-samples models. The random-effects RMSEAs should match the dependent-samples overall RMSEAs – both sets of metrics capture the same information. The within-cluster RMSEAs (ε_1) should be identical across different

Figure 5

RMSEA estimates from Digman (1997) example

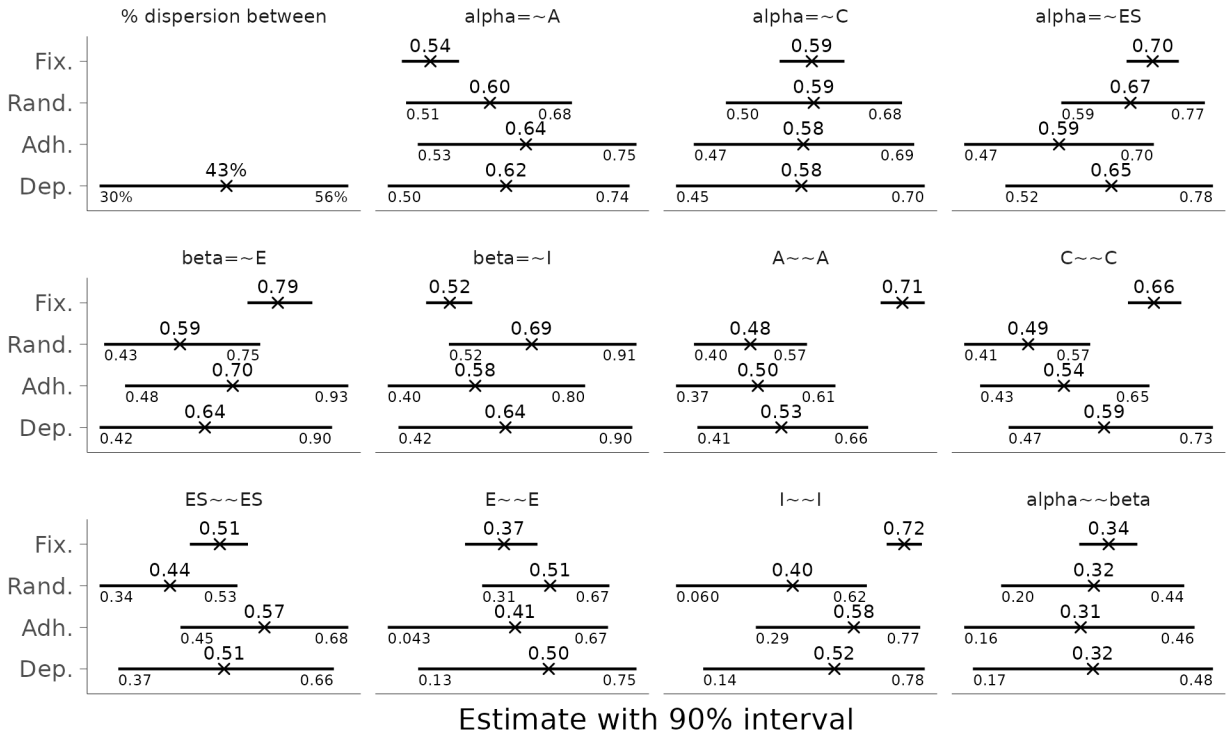


hypotheses about the same data as this is the gap between the cluster-level population covariance matrices and the population covariance matrices underlying each observed covariance matrix. Finally, the between-cluster RMSEA (ε_2) shows how the pooled structured covariance matrix differs from the cluster-level population covariance matrices, and this will differ by hypothesized model. Given that clustering of covariance matrices will create some shared variation, ε_2 is the best measure of how well the pooled structured covariance matrix differs from the individual population covariance matrices, as this metric reflects the RMSEA assuming there was no clustering. Based on this metric, the pooled two-factor model is less distant than the pooled unidimensional model from the individual population covariance matrices underlying observed covariance matrices given the lower value of ε_2 for the two-factor model. Hence, of the two hypotheses, the two-factor model better reflects the patterns responsible for generating the observed covariance matrices. Additionally, a considerable proportion of the variation is between clusters for the two-factor model, 43%, 95% CI [30%, 56%].

Finally, we report parameter estimates for all four two-factor models in Figure 6. Both the dependent-samples and ad hoc models had the widest credible intervals, while the fixed-effects model has the narrowest credible intervals. This pattern reflects the amount of uncertainty in the data generation process that each model captures. Ignoring dependent-samples wrongly reduces the uncertainty about estimated structural parameters.

Figure 6

Structural parameter estimates from Digman (1997) example



Note. Fix. = Fixed-effects, Rand. = Random-effects, Adh. = Ad hoc, Dep. = Dependent-samples. F =∼ X: factor loading, F reflected in X; F1 ∼∼ F2: correlated factors; X ∼∼ X: error variance of X.

Discussion

In this paper, we have presented an approach to MASEM that uses hierarchical Wishart models to capture the two most common MASEMs (fixed- and random- effects) and solves an extant problem in the MASEM literature (dependent covariance matrices). The approach has been demonstrated with a dataset, the demonstration showed how information-criteria based model comparison and the RMSEA produced by the models may be used for model comparison. Additionally, the capacity of the approach to yield adequate inference for dependent-samples has been tested via a simulation study.

The approach here builds off previous research on hierarchical modeling of covariance matrices applied to meta-analytic latent variable models (Uanhoro, 2023). Both methods are one-stage Bayesian meta-analytic SEM approaches for latent variable models. Our unique contribution is the ability to handle the case of dependent-samples. However, the work of Uanhoro (2023) includes two features that are worth extending to the case of dependent-samples. First, it is possible to model the dispersion parameter (ε) as a function of study-characteristics to create a meta-regression model. It would be worth extending this to the dependent-samples case. In the dependent-samples case, one could create two meta-regression models: one for the within-cluster dispersion using sample-level characteristics as predictors; and another for between-cluster dispersion using cluster level characteristics as predictors. Second, it is possible to model misspecification at the level of the pooled covariance structure concurrently with the hypothesized structure. Misspecification modeling allows for (i) capturing the degree of misspecification in the hypothesized structure; (ii) estimating

structural parameters whose uncertainty reflects the degree of model misspecification. This extension is worth studying in the dependent-samples case, because it is reasonable to assume that there is misspecification present at the level of the true covariance structure due to the influence of minor factors (e.g. MacCallum & Tucker, 1991). Additionally, it allows for exploration of model fit – a missing aspect in our current approach that we hope to address in future work.

One open question is how to identify clusters in MASEM applications. We have assumed that clustering occurs at the level of study authors. But this assumption may be too simple for some applications. For example, consider the case where there is clustering by both study authors and the country where the study was conducted. The approach we have presented can only account for one grouping factor at a time. Should the modeller then cluster at the level of country or author in this case? In a standard multilevel context, the simplest answer would be to account for both grouping factors. In the approach we have presented, the modeller would be forced to choose. In theory, one could create another hierarchy in our approach but this may lead to problems with accurately estimating dispersion parameters. For this reason, we intend to explore additional MASEM approaches for dependent samples. One promising approach is the parameter-based MASEM approach of Ke et al. (2019) which hierarchically models structural parameters. We expect that such an approach may more easily accommodate truly complex data structures; we intend to test this expectation in the future.

On correlation matrices as data for MASEM

Finally, we note that the data for meta-analytic CFAs are often correlation rather than covariance matrices. In this paper, we laid out a data-generation process (DGP) for sample covariance matrices that represents a plausible mechanism for the generation of such matrices. The Wishart approach then follows from this process. In practice, one can compute the covariance matrix from the correlation matrix and item standard deviations. Moreover, item standard deviations are often reported alongside item correlation matrices, or both sets of statistics can be retrieved from study authors. Thus the need for the sample covariance matrix as input should not overly limit the applicability of the recommended approach.

The challenge with proposing a coherent dependent-samples MASEM for correlation matrices partially lies with the difficulty of identifying a credible hierarchical MASEM DGP that results in correlation matrices, without inadequately reducing the data. The most commonly invoked DGP in MASEM is that the observed correlations ($\mathbf{r}_i$) for study i underlying p variables are multivariate normal: $\mathbf{r}_i \sim \mathcal{N}_{p^*}(\boldsymbol{\rho}, \boldsymbol{\Gamma}_i)$, where $p^* = p \times (p-1)/2$ and $\boldsymbol{\Gamma}_i = \boldsymbol{\Delta} + \boldsymbol{\Phi}_i$. $\boldsymbol{\Phi}_i$ is sampling variation assumed to be of known form, while $\boldsymbol{\Delta}$ is variation due to random-effects and is assumed to be a zero-matrix in the fixed-effects model. For dependent-samples MASEM, the extended DGP for observed correlations belonging to study i in cluster j would be: $\mathbf{r}_{ij} \sim \mathcal{N}_{p^*}(\boldsymbol{\rho} + \boldsymbol{\kappa}_{j[i]}, \boldsymbol{\Gamma}_i)$ and $\boldsymbol{\kappa}_j \sim \mathcal{N}_{p^*}(\mathbf{0}_{p^*}, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ captures variation due to clustering.

One challenge with DGPs of this form is that the multivariate normal distribution is unbounded while correlations are bounded, thus non-correlations are possible under this process. A second challenge is that even when all generated estimates are correlations, there is no guarantee that the resulting correlation matrix will be positive-definite or valid. Moreover, these problems are likely exacerbated in the dependent samples case. While a model assuming such a process may be pragmatically applied to analysis of dependent correlation matrices (e.g. Wilson, Polanin, & Lipsey, 2016), it is not a credible DGP for such data. This discrepancy between process and model is often not a challenge for frequentist methods, which can rely on robust variance estimation given approximately unbiased parameter estimates. However, the coherence of Bayesian inference partially rests on the belief that the likelihood encodes a plausible DGP. Hence, it would be incoherent

to propose a Bayesian model atop a likelihood known to be implausible.

Building off Archakov and Hansen (2021), Archakov, Hansen, and Luo (2022) laid out a DGP for observed correlation matrices centered about a target correlation matrix. This approach reduces to the Fisher r to z transformation for 2-by-2 correlation matrices. In MASEM, the target correlation matrix would be assumed to be a structured correlation matrix. In the future, we intend to explore a MASEM approach based on this process that would be adequate for meta-analysis of correlation matrices. In the meantime, we conducted a simulation study where we generated covariance matrices according to the Wishart DGP, transformed them into correlation matrices and evaluated the performance of the proposed Wishart approach. The results suggest low bias for structural parameters and overly conservative inference about loading parameters. These findings are elaborated in Appendix C. In summary, the Wishart approach is better applied to covariance matrices when available. When only correlation matrices are available, it may be better to use the Wishart approach if the correlation matrices are dependent than to employ an approach that ignores dependencies between observed matrices.

Declarations

Open Practices Statement. All code for simulation studies and data analysis are available at `https://osf.io/yd5q4/`.

References

Archakov, I., & Hansen, P. R. (2021). A new parametrization of correlation matrices. *Econometrica*, *89*(4), 1699–1715. doi: 10.3982/ECTA16910

Archakov, I., Hansen, P. R., & Luo, Y. (2022, October). *A new method for generating random correlation matrices* (Tech. Rep. No. arXiv:2210.08147). arXiv. Retrieved 2023-08-23, from `http://arxiv.org/abs/2210.08147` (arXiv:2210.08147 [econ, stat] type: article)

Betancourt, M. (2017, January). *A conceptual introduction to Hamiltonian Monte Carlo.* (arXiv: 1701.02434)

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi: 10.18637/jss.v076.i01

Cheung, M. W.-L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods*, *46*(1). doi: 10.3758/s13428-013-0361-y

Cheung, M. W.-L. (2015, January). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.01521

Cheung, M. W.-L., & Chan, W. (2005, March). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*(1), 40–64. doi: 10.1037/1082-989X.10.1.40

Cheung, M. W.-L., & Chan, W. (2009, January). A two-stage approach to synthesizing covariance matrices in meta-analytic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(1), 28–53. doi: 10.1080/10705510802561295

Cheung, M. W.-L., & Cheung, S. F. (2016). Random-effects models for meta-analytic structural equation modeling: review, issues, and illustrations. *Research Synthesis Methods*, *7*(2), 140–155. doi: 10.1002/jrsm.1166

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, *183*(1), 31–57. doi: 10.1016/j.jeconom.2014.06.008

Cook, S. R., Gelman, A., & Rubin, D. B. (2006, September). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, *15*(3), 675–692. doi: 10.1198/106186006X136976

Digman, J. M. (1997, December). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*(6), 1246–1256. doi: 10.1037/0022-3514.73.6.1246

Furlow, C. F., & Beretvas, S. N. (2005, June). Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. *Psychological Methods*, *10*(2), 227–254. doi: 10.1037/1082-989X.10.2.227

Gelman, A. (2006, September). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. doi: 10.1214/06-BA117A

Granström, K., & Orguner, U. (2011). *Properties and approximations of some matrix variate probability density functions* (LiTH-ISY-R-3042). Linköping, Sweden: Division of Automatic Control, Linköping University. Retrieved from `http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-88735`

Gupta, A. K., & Nagar, D. K. (1999). *Matrix variate distributions*. CRC Press.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Jak, S., & Cheung, M. W.-L. (2018, January). Accounting for missing correlation coefficients in fixed-effects masem. *Multivariate Behavioral Research*, *53*(1), 1–14. doi: 10.1080/00273171.2017.1375886

Jak, S., & Cheung, M. W. L. (2020). Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological Methods*, *25*(4), 430–455. doi: 10.1037/met0000245

Ke, Z., Zhang, Q., & Tong, X. (2019, May). Bayesian meta-analytic SEM: A one-stage approach to modeling between-studies heterogeneity in structural parameters. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(3), 348–370. doi: 10.1080/10705511.2018.1530059

Kim, S., Moon, H., Modrák, M., & Säilynoja, T. (2023). SBC: Simulation based calibration for rstan/cmdstanr models [Computer software manual]. (https://hyunjimoon.github.io/SBC/, https://github.com/hyunjimoon/SBC/)

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*(7), 912–928. doi: 10.1111/oik.05985

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502–511. doi: 10.1037/0033-2909.109.3.502

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, *100*(1), 43–52. doi: 10.1080/00223891.2017.1281286

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021, November). Efficient Bayesian structural equation modeling in Stan. *Journal of Statistical Software*, *100*(6), 1–22. doi: 10.18637/jss.v100.i06

Olkin, I., & Finn, J. D. (1995, July). Correlations redux. *Psychological Bulletin*, *118*(1), 155–164. doi: 10.1037/0033-2909.118.1.155

Oort, F. J., & Jak, S. (2016). Maximum likelihood estimation in meta-analytic structural equation modeling. *Research Synthesis Methods*, *7*(2), 156–167. doi: 10.1002/jrsm.1203

Peeters, C. F. W. (2012). Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika*, *77*(2), 288–292. doi: 10.1007/s11336-012-9259-3

Säilynoja, T., Bürkner, P.-C., & Vehtari, A. (2022, April). Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. *Statistics and Computing*, *32*(2), 32. doi: 10.1007/s11222-022-10090-6

Savalei, V. (2012, December). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement*, *72*(6), 910–932. doi: 10.1177/0013164412452564

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018, April). *Validating Bayesian inference algorithms with simulation-based calibration.* arXiv. (arXiv: 1804.06788)

Uanhoro, J. O. (2023, July). Hierarchical covariance estimation approach to meta-analytic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(4), 532–546. doi: 10.1080/10705511.2022.2142128

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.* Retrieved from `https://mc-stan.org/loo/`

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*, 1–28. doi:

10.1214/20-BA1221

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, *48*(4), 865–885. doi: 10.1111/j.1744-6570.1995.tb01784.x

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In *Secondary data analysis: An introduction for psychologists* (pp. 39–61). Washington, DC, US: American Psychological Association. doi: 10.1037/12350-003

Wilson, S. J., Polanin, J. R., & Lipsey, M. W. (2016, June). Fitting meta-analytic structural equation models with complex datasets. *Research Synthesis Methods*, *7*(2), 121–139. doi: 10.1002/jrsm.1199

Wu, H., & Browne, M. W. (2015, September). Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika*, *80*(3), 571–600. doi: 10.1007/s11336-015-9451-3

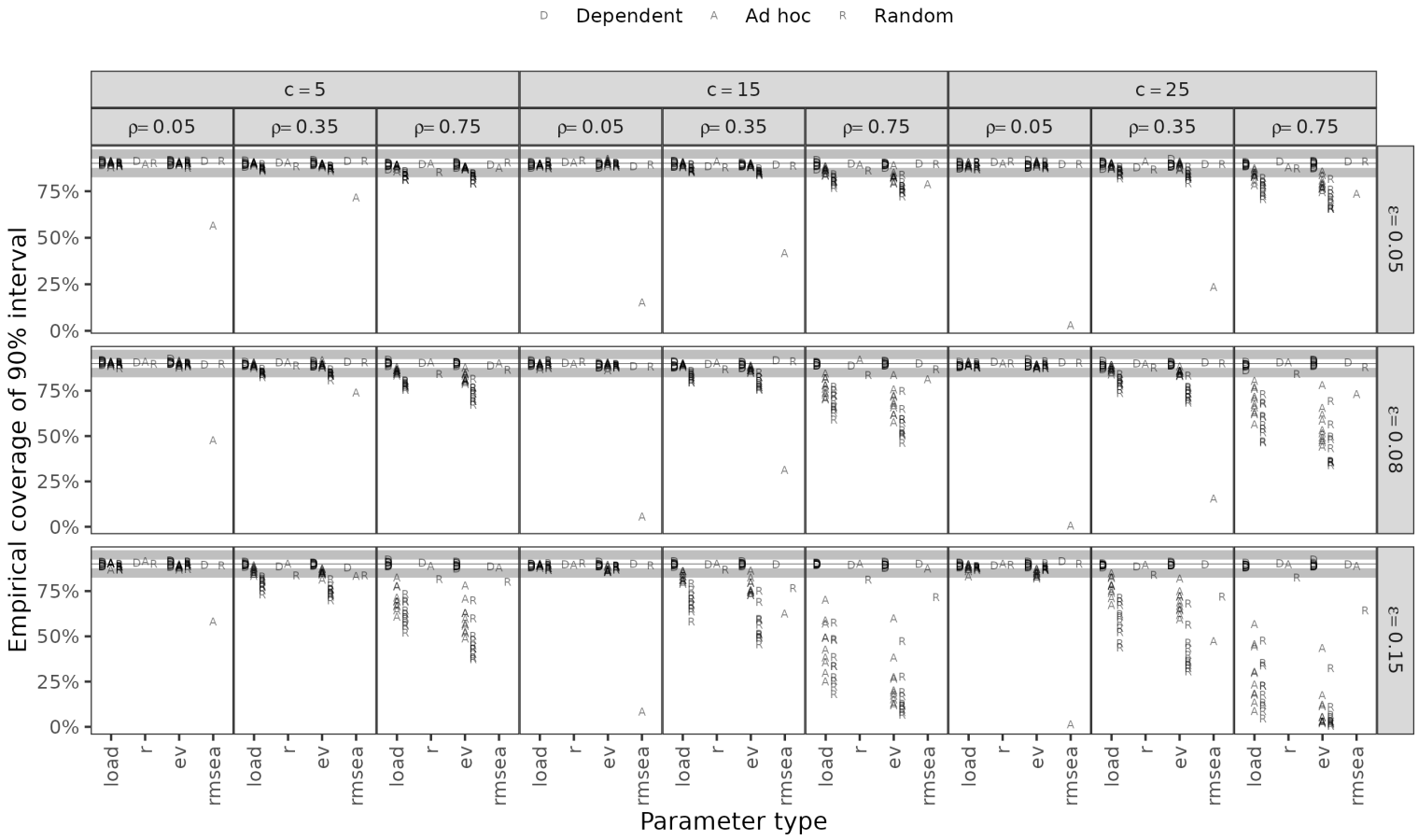
Ximénez, C., Maydeu-Olivares, A., Shi, D., & Revuelta, J. (2022, May). Assessing cutoff values of SEM fit indices: Advantages of the unbiased SRMR index and its cutoff criterion based on communality. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 368–380. doi: 10.1080/10705511.2021.1992596

Yuan, K.-H., & Kano, Y. (2018, December). Meta-analytical SEM: Equivalence between maximum likelihood and generalized least squares. *Journal of Educational and Behavioral Statistics*, *43*(6), 693–720. doi: 10.3102/1076998618787799

Appendix A
Additional simulation results

Figure A1

Empirical coverage rate of the 90% credible interval in the simulation study



Note. load. = 10 loading estimates, r = inter-factor correlation, ev. = 10 error variance estimates. The ECR for structural parameters is sometimes poor for ad hoc approach only because the estimates are systematically biased.

Appendix B
Simulation-based calibration – Digman (1997) application

The data generation process (DGP) for the SBC study was based on the Digman (1997) example. The exact DGP was:

$$\begin{aligned}
\mathbf{S}_{ij} &\sim \mathrm{GB}_p^{\mathrm{II}}\left(\frac{n_i^*}{2}, \frac{m_1}{2}, \frac{m_1}{n_i^*}\boldsymbol{\Psi}_{j[i]}, \mathbf{0}_{p\times p}\right) \text{ for } i \in \{1,\ldots,14\} \\
m_2\boldsymbol{\Psi}_j &\sim \mathcal{W}\left(\boldsymbol{\Omega}(\boldsymbol{\theta}), m_2\right) \text{ for } j \in \{1,\ldots,7\} \\
\boldsymbol{\Omega}(\boldsymbol{\theta}) &= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Delta},\ \boldsymbol{\Phi} = \begin{bmatrix} 1 & \\ \phi_{2,1} & 1 \end{bmatrix}
\end{aligned} \tag{B1}$$

Priors were chosen such that the generated data would produce valid covariance matrices (e.g. Merkle et al., 2021; Uanhoro, 2023). Loadings and residual standard deviations had median values of 0.8 and 0.6 respectively. And m_1 and m_2 priors were chosen such that the median value of ρ would be about 0.25, $\exp(-6)/(\exp(-5)+\exp(-6))$.

$$\begin{aligned}
\boldsymbol{\lambda} \sim &\mathcal{N}^+(0.8, \sigma_\lambda),\ \sigma_\lambda \sim \mathcal{N}^+(0, 0.5),\ \sqrt{\mathrm{diag}(\boldsymbol{\Delta})} \sim \mathcal{N}^+(0.6, 0.25), \\
&\ln([\begin{smallmatrix} m_1 \\ m_2 \end{smallmatrix}] - p + 1) \sim \mathcal{N}^+([\begin{smallmatrix} 5 \\ 6 \end{smallmatrix}], 0.5),\ \frac{\phi_{2,1}+1}{2} \sim \mathrm{Beta}(5,5)
\end{aligned} \tag{B2}$$

The distribution of parameters based on equation B2 is shown in Figure B1.

For the SBC study, each model was estimated using a single chain. We requested 5000 iterations, 1000 iterations were discarded for warmup, while the remaining 4000 iterations were thinned at every second iteration to reduce autocorrelation between posterior samples. Thus, 2000 posterior samples were retained per parameter. Finally, we repeated this process 1000 times.

Evaluation of SBC results was based on graphical summaries recommended by Säilynoja, Bürkner, and Vehtari (2022). We report the evaluations in Figures B2 and B3 – these figures were produced using the *SBC* package in R (Kim, Moon, Modrák, & Säilynoja, 2023).

Our expectation is that the distribution of ranks for each parameter are uniformly distributed. When this is true, the histogram counts will often remain within the 95% simultaneous confidence bands – this expectation is met for all parameters with very few exceptions, see Figure B2. This suggests adequate calibration of all parameters.

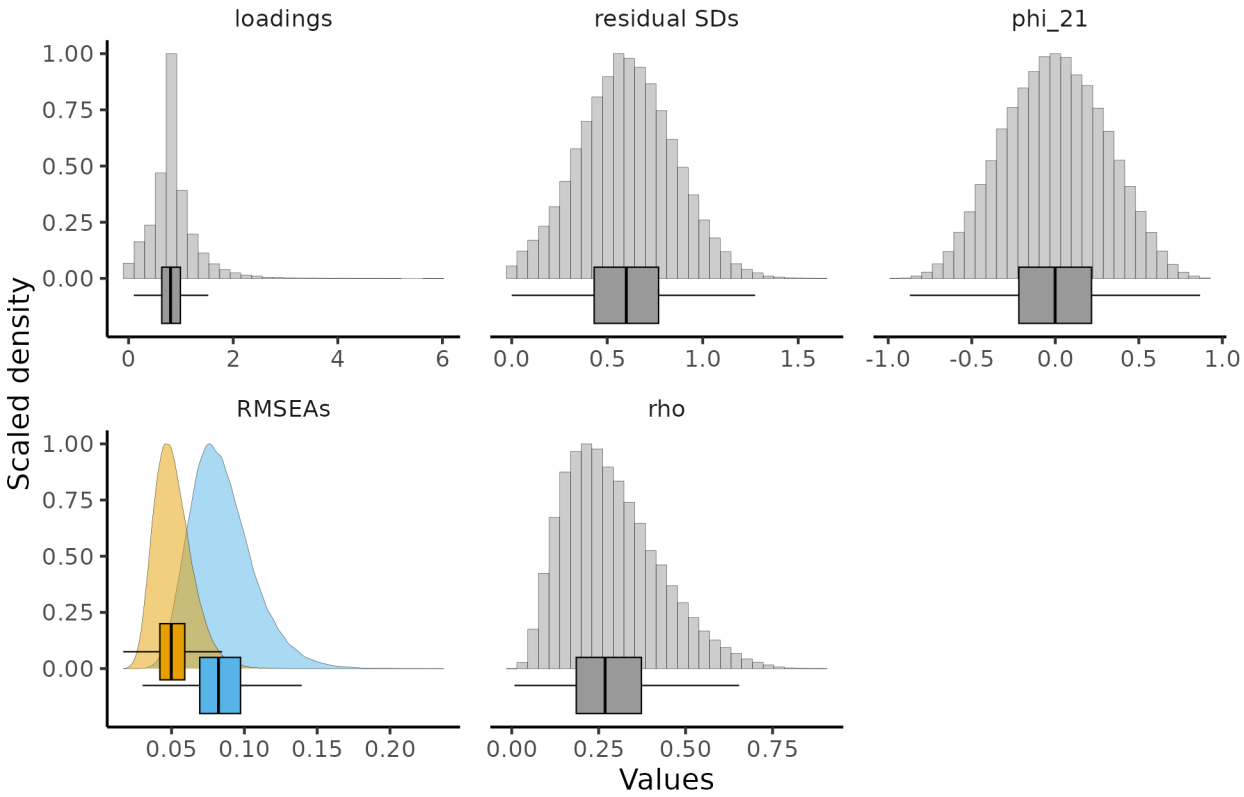
The evaluation via histogram is sensitive to the number of bins. Hence, we also assessed the empirical cumulative distribution function (ECDF) of the ranks. Precisely, we assessed the difference of the ECDF from the theoretical CDF of a uniform variable. When these differences are contained within the 95% simultaneous bands, parameters are adequately calibrated. This expectation is met for all parameters, Figure B3.

We also repeated the testing-based SBC evaluation procedures in Uanhoro (2023). The SBC ranks are first transformed to rankits: $q_i = (r_i + 0.5)(L+1)^{-1}$, where r_i are the ranks and $L = 2000$, the number of retained posterior samples. The standard normal quantile function was applied to the rankits. If the ranks were approximately uniform, then the result should be an approximately standard normal variable. The bias of the mean (difference from 0 based on the one-sample t_{999} test), bias of the variance (difference from 1 based on the one-sample χ^2_{999} test), and a χ^2_{1000} test of standard normality (Cook, Gelman, & Rubin, 2006) were then used to assess the standard normality expectation.[8] As shown in Figure B4, no parameter resulted in a statistically significant test suggesting calibration for all parameters.

[8] The degrees of freedom are based on the number replications, 1000.

Figure B1

Draws from prior distributions



There are two RMSEA distributions: between (lower values) and within (larger values).

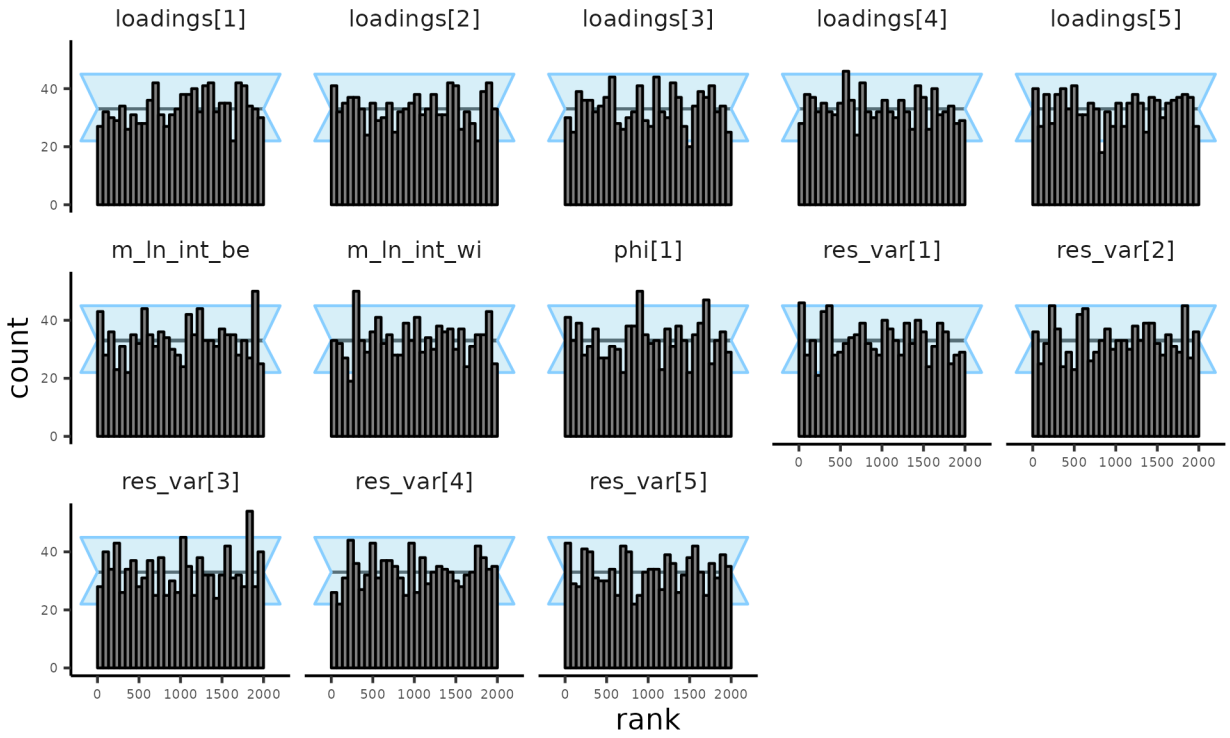
Appendix C
Simulation study of dependent correlation matrices

As mentioned in the Discussion section, we repeated the simulation study in the paper, but transformed the sample covariance matrices to correlation matrices prior to data analysis. Following from the expectation of an inverse-Wishart distribution, our model when applied to these data should return the following structured covariance matrix: $\boldsymbol{\Omega}(\boldsymbol{\theta})(m_1 - p - 1)m_1^{-1}$, where $m_1 = \varepsilon_1^{-2} + p - 1$ and $\varepsilon_1 = \varepsilon\sqrt{(1-\rho)}$. Hence, the bias and empirical coverage rate evaluations are adjusted to reflect this. We excluded conditions where $\rho = .75$ and $\varepsilon = 0.05$ as analysis runtimes for the three conditions ($c \in \{5, 15, 25\}$) were overly time consuming. Finally, we only ran 300 replications, down from 1000 replications in the original study.

Results are reported in Figures C1, C2 and C3. Most parameters had acceptable levels of bias ($< |10\%|$), apart from ε which was sometimes downwardly biased (especially when $\varepsilon = 0.05$). We believe this downward bias occurs because the process of converting a covariance matrix to a correlation matrix eliminates important variation, given the data generation process. Posterior standard deviations were often upwardly biased, especially for loading parameters. This suggests overly conservative inference, and resulted in higher than nominal coverage rates especially for loading parameters. Coverage for ε was always low given the downward parameter bias. And there were also periods of under-coverage for loading parameters at the combination of high values of ε and larger number of clusters. This under-coverage likely occurs even in the presence of

Figure B2

Histogram of SBC ranks – Digman (1997) application

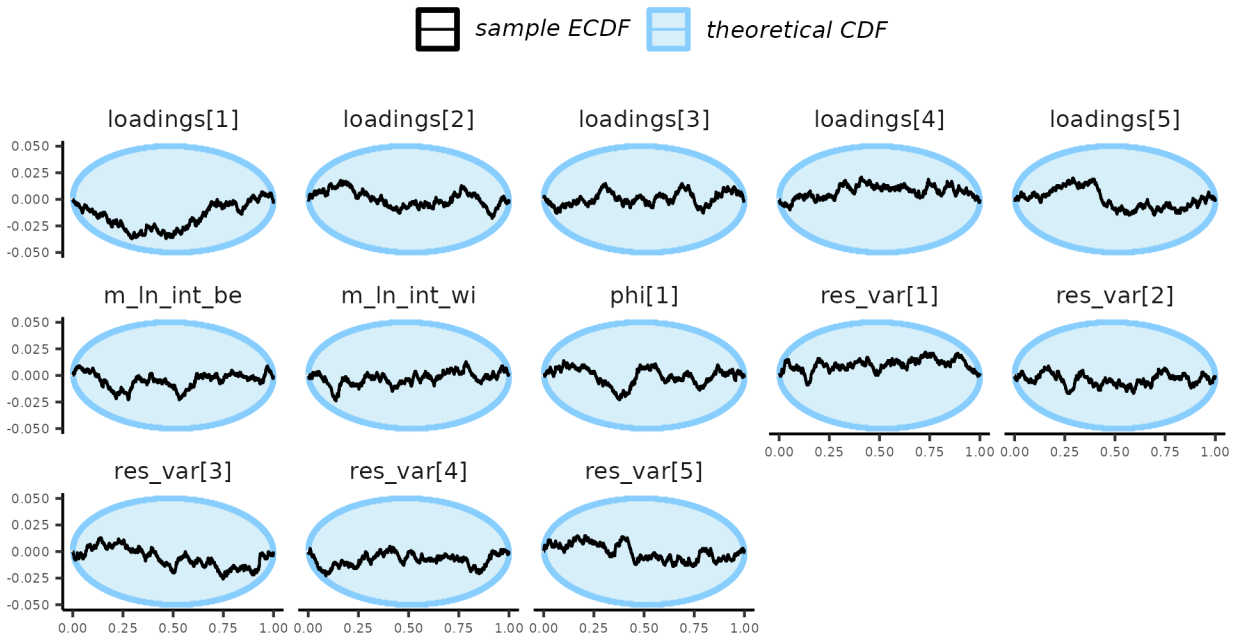


Note. m_ln_int_wi = m_1 and m_ln_int_be = m_2; phi = interfactor correlation; res_var = residual variances. Expectation is that the histogram counts are contained in the 95% simultaneous confidence bands.

overly wide posterior standard deviations because of the combination of some parameter bias and increased precision of posterior standard deviations at larger sample size. Finally, as with the original simulation study, there were problems estimating the dispersion parameters, see Figure C4.

Figure B3

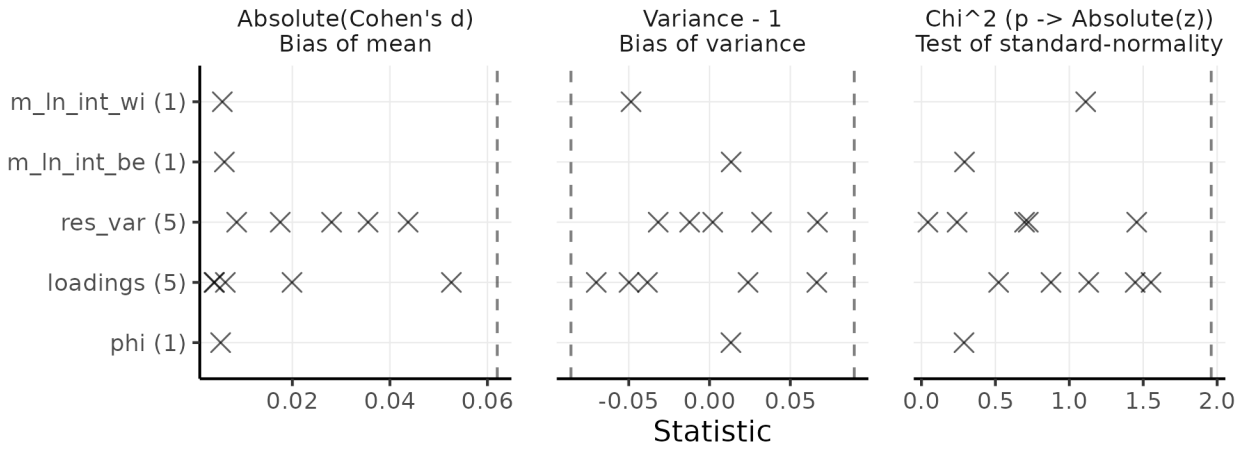
Empirical CDF check – Digman (1997) application



Note. m_ln_int_wi = m_1 and m_ln_int_be = m_2; phi = interfactor correlation; res_var = residual variances. Expectation is that the sample ECDF is contained within the 95% simultaneous confidence bands about the theoretical CDF.

Figure B4

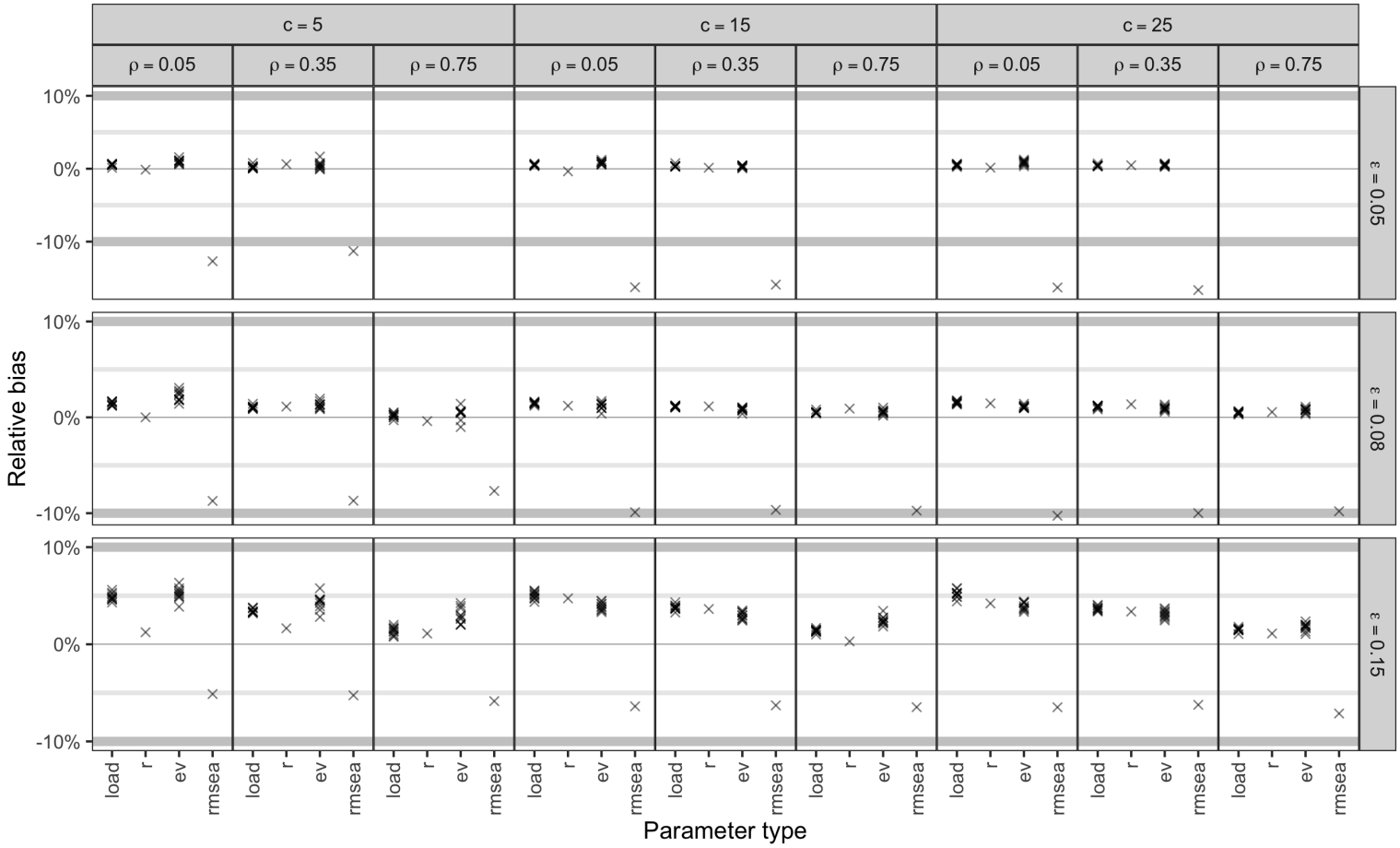
Additional SBC evaluation metrics – Digman (1997) application



Note. m_ln_int_wi = m_1 and m_ln_int_be = m_2; phi = interfactor correlation; res_var = residual variances. Number in parenthesis on y-axis is count of parameters. Vertical dashed lines are 95% confidence limits based on hypothesis tests; no estimate exceeded the limits suggesting adequate calibration for all parameters.

Figure C1

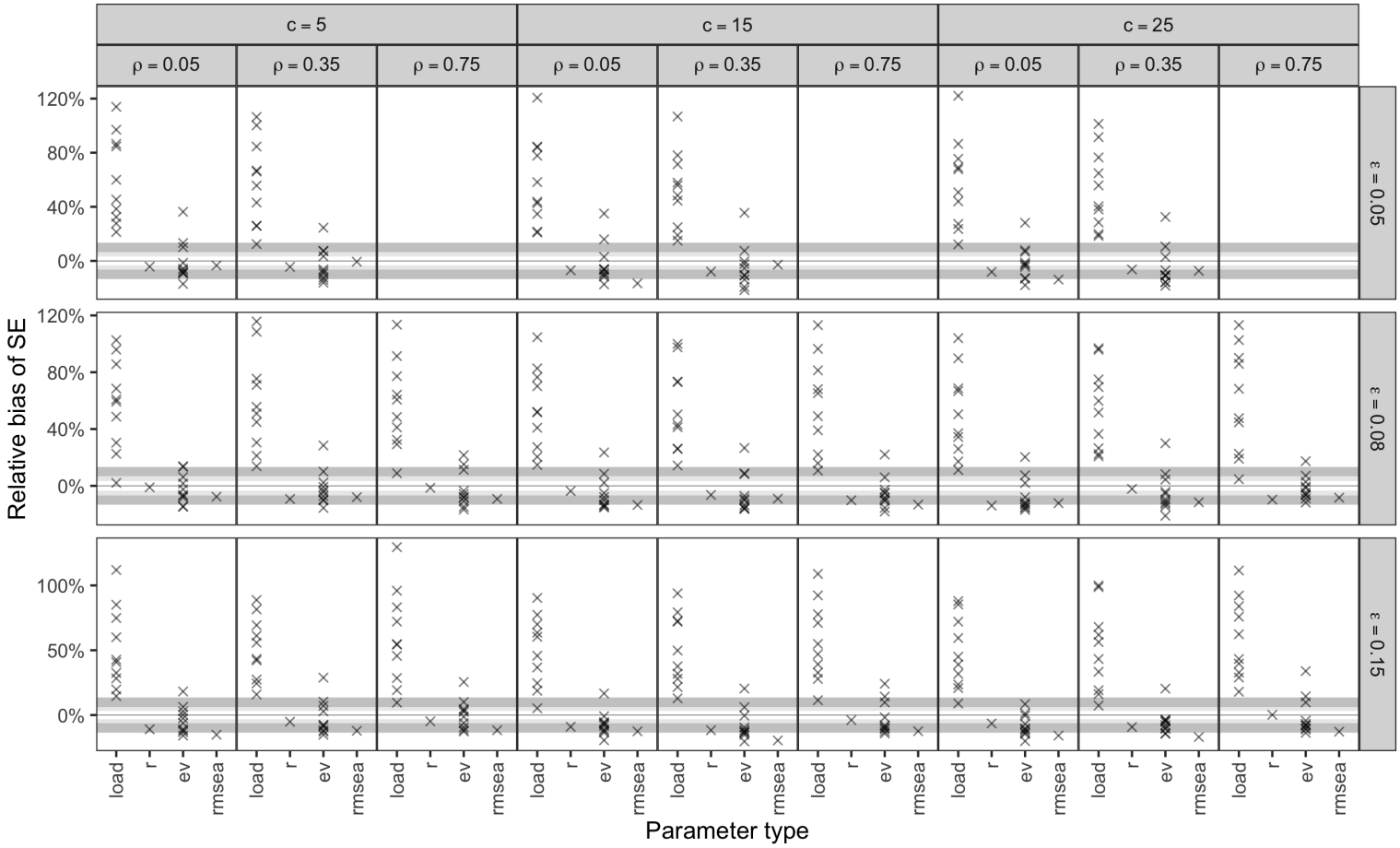
Relative bias of mean of posterior distribution for correlation simulation study



Note. load. = 10 loading estimates, r = inter-factor correlation, ev. = 10 error variance estimates.

Figure C2

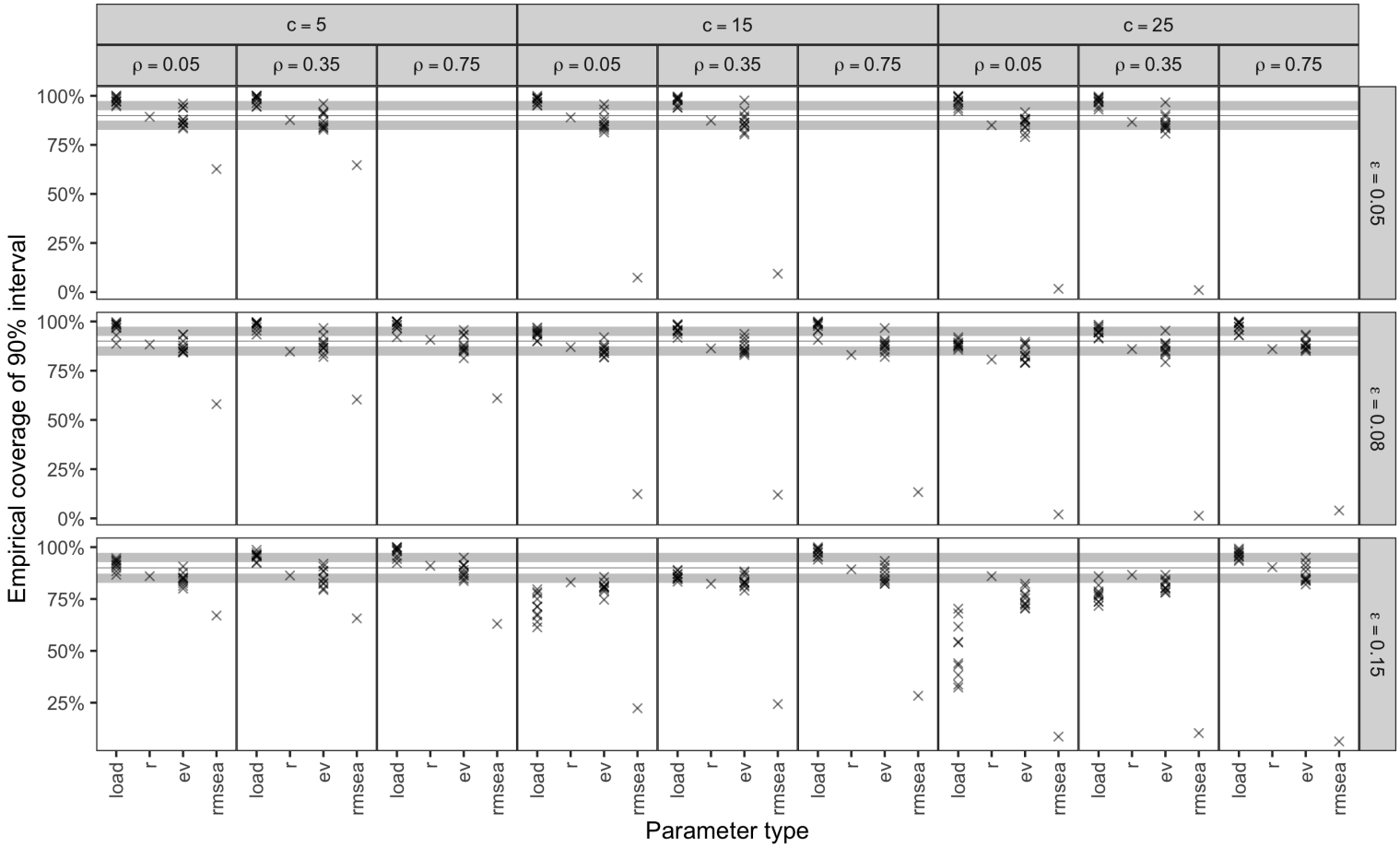
Relative bias of standard deviation of posterior distribution for correlation simulation study



Note. load. = 10 loading estimates, r = inter-factor correlation, ev. = 10 error variance estimates.

Figure C3

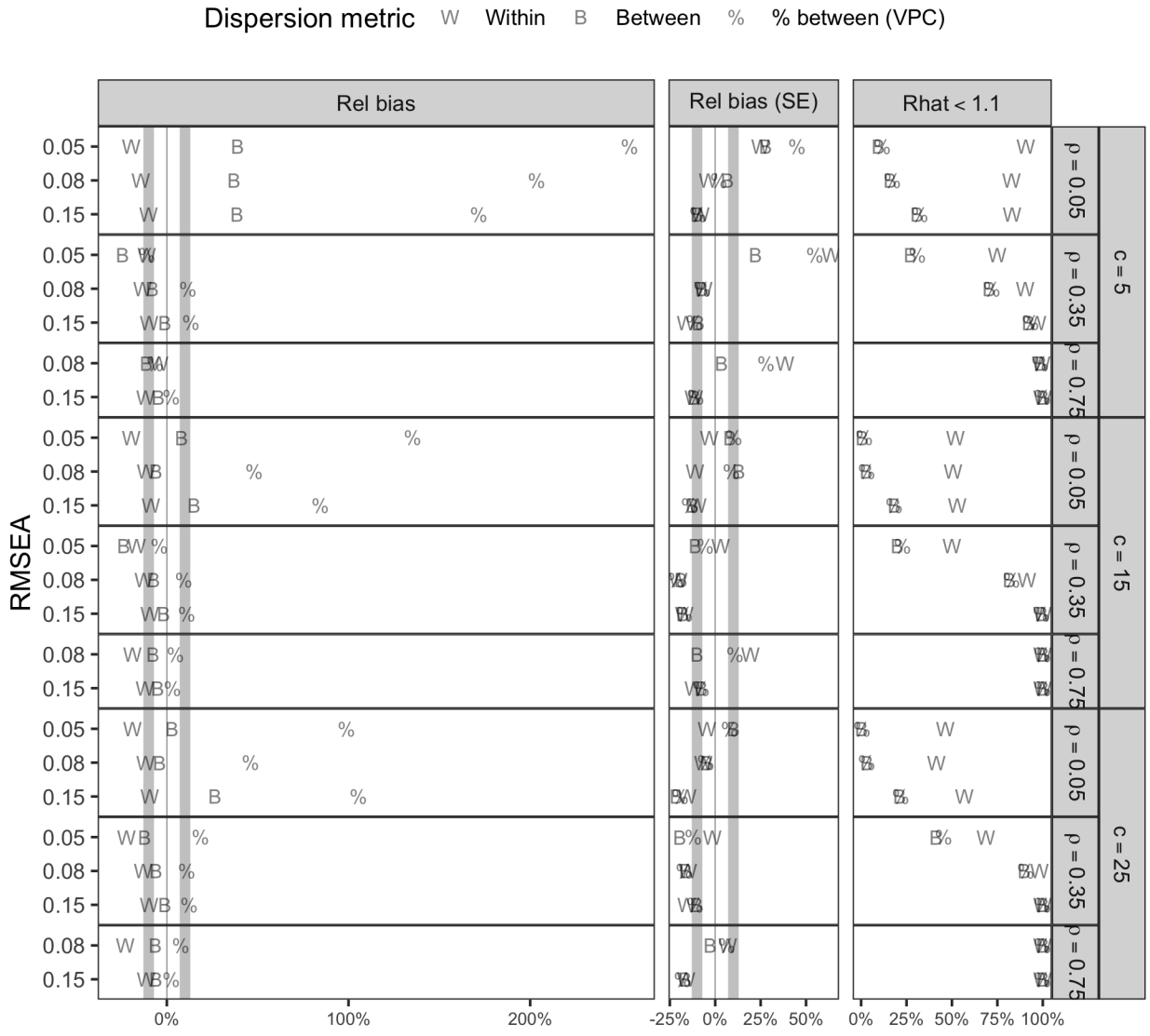
Empirical coverage rate of the 90% credible interval for correlation simulation study



Note. load. = 10 loading estimates, r = inter-factor correlation, ev. = 10 error variance estimates.

Figure C4

Recovery of dispersion parameters for correlation simulation study



Note. Within = ε_1, Between = ε_2, % between = ρ