Hierarchical covariance estimation approach to meta-analytic structural equation modeling

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

Abstract

We present a fully Bayesian approach to meta-analytic SEM based on hierarchical modeling of sample covariance matrices. The approach allows for flexible models that would not be identified under a traditional maximum likelihood approach. The approach allows for inclusion of moderators, produces a global fit index and permits investigation of local misspecification. Simulation-based calibration studies show that the Bayesian computation procedure produces valid inference for commonplace meta-analytic SEM applications. We demonstrate the approach with diverse data analysis examples and provide accompanying R code to support adoption and additional study of the approach. Finally, we lay out proposals that have the potential to extend the approach to accommodate the wide variety of analyses and data conditions that comprise meta-analytic SEM applications.

Structural equation modeling (SEM) is a popular statistical method for modeling covariance structures; for commonplace SEMs, the covariance matrix is sufficient for data analysis. Meta-analytic SEM (MASEM, Cheung & Chan, 2005; Viswesvaran & Ones, 1995) combines ideas from meta-analysis (Hedges & Olkin, 1985) and SEM to estimate and test covariance structures assumed to underlie multiple covariance matrices.

In this paper, we present a fully Bayesian approach to MASEM. The approach is based on hierarchical modeling of sample covariance matrices (Wu & Browne, 2015), returns a global fit index similar to the standardized root mean squared residual (Ogasawara, 2001), and permits investigation of local misspecification. Additionally, the approach allows for more flexible models than have been previously estimated in the MASEM literature. Prior to elaborating the approach, we briefly review major approaches to estimating meta-analytic SEMs. Across the paper, we assume study i in $1, \ldots, k$ studies has a $p \times p$ sample covariance matrix $\mathbf{S}_i$.

The Version of Record of this manuscript has been published and is available in Structural Equation Modeling: A Multidisciplinary Journal (2022) http://www.tandfonline.com/10.1080/10705511.2022.2142128

One major approach to estimating MASEMs is two-stage SEM (TSSEM, Cheung & Chan, 2005, 2009). In the first step, a pooled correlation matrix is estimated. Each $\mathbf{S}_i$ is decomposed thus: $\mathbf{R}_i = \mathbf{D}_i^{-1}\mathbf{S}_i\mathbf{D}_i^{-1}$, where $\mathbf{D}_i$ is a diagonal matrix with j^{th} diagonal element $s_{i(jj)}^{1/2}$, such that $\mathbf{R}_i$ is the correlation matrix. Under a fixed-effect TSSEM, all studies are assumed to have the same $\mathbf{R}_i$, i.e. $\mathbf{R}_1 = \ldots = \mathbf{R}_k$. Under a random-effects TSSEM, $\mathbf{R}_i$ is assumed to vary across studies. In a fixed-effect TSSEM, one can compute the degree of misspecification due to assuming the same $\mathbf{R}_i$ across studies. If the degree of misspecification is deemed unacceptable, then a random-effects TSSEM is preferred. Under a random-effects TSSEM, $\mathbf{r}_i$, the q-dimensional ($q = p \times (p-1)$) vector of elements below the main diagonal of $\mathbf{R}_i$, is modeled thus (Becker, 1992):

$$\mathbf{r}_i = \boldsymbol{\rho} + \mathbf{u}_i + \mathbf{e}_i,\ \mathbf{u}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Delta}),\ \mathbf{e}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi}_i) \tag{1}$$

where $\boldsymbol{\rho}$ is the pooled vector of correlation across studies, $\mathbf{u}_i$ are study specific deviations from the pooled estimates, and $\mathbf{e}_i$ are deviations resulting from sampling variation. As $\boldsymbol{\Delta} \to \mathbf{0}$, the fixed-effects model becomes more realistic for the data. $\boldsymbol{\Psi}_i$ is often assumed known (following equations 3–5 in Olkin & Finn, 1995), while $\boldsymbol{\rho}$ and $\boldsymbol{\Delta}$ may be estimated using maximum-likelihood (Cheung, 2014). Regardless of fixed-/random- effects approach, the resulting pooled correlations $\boldsymbol{\rho}$ are then used to estimate the hypothesized correlation structure $\boldsymbol{\rho}(\boldsymbol{\theta})$ with weighted least squares (WLS) discrepancy function (Browne, 1974). As an alternative to TSSEM, there is a one-stage maximum-likelihood method (Oort & Jak, 2016) – Yuan and Kano (2018) show that this one-stage method is asymptotically equivalent to the two-stage approach described above.

A competing approach is the one-stage parameter-based approach of Ke, Zhang, and Tong (2019). In this approach, the structural parameters of the hypothesized covariance structure are assumed to vary by study, $\boldsymbol{\rho}(\boldsymbol{\theta}_i)$. Each parameter j in study i, θ_{ji}, is hierarchically estimated: $\theta_{ji} \sim \mathcal{N}(\theta_j, \sigma_j)$, such that θ_j is the pooled estimate of a given parameter and σ_j^2 is the parameter's dispersion. Ke et al. (2019) take a Bayesian estimation approach to this model. Conceptually, this approach is akin to a multi-group SEM where all groups share the same structure but are allowed to have different parameter estimates, which are themselves hierarchically estimated. The stated benefit of this approach is the ability to directly model heterogeneity of structural parameters.

The maximum-likelihood estimation approach has recently been extended to account for

study characteristics or moderators (one-stage MASEM or OSMASEM, Jak & Cheung, 2020). In this model, any structural parameter (e.g. loadings, latent regression coefficients, path coefficients) can be the outcome of a regression equation with moderators as predictors. Hence, this approach shares some relation to parameter-based MASEM in that the structural parameters themselves are assumed to vary across studies.

In the next section of the paper, we lay out our method for MASEM. Afterward, we present some data analysis examples using data in the R metaSEM package (Cheung, 2015). Then we provide simulation results that show the method allows for valid inference and parameter recovery. And we conclude with discussion of the approach we present and some thoughts for further developing the approach. Finally, code for simulation studies and data analysis examples is available at `https://osf.io/rstzk/`.

Hierarchical covariance matrix (HCM) estimation

Under the assumption that the data in a study are multivariate normal, the sample covariance matrix is a Wishart matrix variate:

$$\mathbf{S} \sim \mathcal{W}_p \left(\frac{1}{n^*} \mathbf{\Sigma}, n^* \right), \tag{2}$$

where $n^* = \text{sample size} - 1$, $\mathbf{\Sigma}$ is the population covariance matrix underlying the study. As $n^* \to \infty$, $\mathbf{S} \to \mathbf{\Sigma}$, and the effect of sampling error is negligible. The population covariance matrix may be assumed to be inverse-Wishart (which is conjugate to Wishart):

$$\mathbf{\Sigma} \sim \mathcal{W}_p^{-1}(\mathbf{\Omega} \times m, m), \tag{3}$$

where $\mathbf{\Omega}$ 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 $m \to \infty$, $\mathbf{\Sigma} \to \mathbf{\Omega}$. Wu and Browne (2015, hereafter WB) assumed $\mathbf{\Omega}$ to be a structured covariance matrix, $\mathbf{\Omega}(\boldsymbol{\theta})$ with parameter vector, $\boldsymbol{\theta}$, such that differences between $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ are due to what WB (2015) term *adventitious error*.

Adventitious error refers to error resulting from differences between the general population for which a psychometric theory is hypothesized and the specific population from which observations

are made. Echoing an example in WB (2015), a group of researchers may develop an instrument designed to measure depression in U.S. adults. However, the team tests the instrument on adults on a hot summer day in Chicago. The difference between the structured covariance matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ and the population covariance matrix $\boldsymbol{\Sigma}$ in the Chicago study arises from the fact that observations were not made from a random sample. Given that most psychology research is done with non-random samples of the population, one should assume adventitious error as a given in applied contexts.

WB (2015) presented the model above in the context of a single study and suggested it may be useful for multi-group contexts. We agree and believe that this modeling approach holds promise for MASEM. It is reasonable to assume that differences in sample covariance matrices across studies result from different populations, with each of these populations being somewhat different from the generic population the hypothesized covariance structure holds for – if it holds. As we will revisit later, the size of m for a given study informs us of the extent to which the hypothesized covariance structure resembles the unobserved population covariance matrix underlying the observed sample covariance matrix.

The models in equations 2 and 3 form a hierarchical model for $\mathbf{S}$ – our primary interest is in obtaining $\boldsymbol{\Omega}$ not $\boldsymbol{\Sigma}$. Hence, one can integrate out $\boldsymbol{\Sigma}$, the resulting marginal distribution for $\mathbf{S}$ is a generalized matrix variate beta type II distribution (Roux & Becker, 1984, as cited in Wu & Browne, 2015):[1]

$$\mathbf{S} \sim \mathrm{GB}_p^{\mathrm{II}}\left(\frac{n^*}{2}, \frac{m}{2}, \frac{m}{n^*}\boldsymbol{\Omega}, \mathbf{0}_{p\times p}\right) \tag{4}$$

with log-likelihood:

$$\begin{aligned}\ln \mathcal{L} = {} & f(p, m+n^*) - f(p,m) - f(p,n^*) + \left(\frac{n^*-p-1}{2}\right)\ln|\mathbf{S}| \\ & + \left(\frac{m}{2}\right)\ln|\boldsymbol{\Omega}| - \left(\frac{n^*+m}{2}\right)\ln\left|\frac{m\boldsymbol{\Omega}+n^*\mathbf{S}}{m+n_i^*}\right|,\end{aligned} \tag{5}$$

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

[1] WB (2015) refer to this distribution as the second type of matrix variate beta distribution citing chapter 5 of Gupta and Nagar (1999). However, Gupta and Nagar (1999) in definition 5.2.4 include the term *generalized* to describe this distribution.

& Nagar, 1999, definition 1.4.2).

One key result in WB (2015) is that the quantity, $1/\sqrt{(m+p-1)}$, approximates the root mean square error of approximation (RMSEA, ε) from assuming the structured covariance matrix ($\boldsymbol{\Omega}(\boldsymbol{\theta})$) matches $\boldsymbol{\Sigma}$.

We now present an approach for fitting MASEMs that takes advantage of developments in WB (2015). We take a Bayesian approach to model estimation. To elaborate the method, we focus on the confirmatory factor analysis (CFA) model as it is the most common latent variable model in MASEM applications, however, the approach can be extended to the broad class of SEMs for which the sample covariance matrix is a sufficient statistic. Using MASEM classification, the model is a one-stage random-effects MASEM approach. The full Bayesian model is:

$$
\begin{aligned}
&\mathbf{S}_i \sim \mathrm{GB}_p^{\mathrm{II}}\left(\frac{n_i^*}{2}, \frac{m_i}{2}, \frac{m_i}{n_i^*}\boldsymbol{\Omega}(\boldsymbol{\theta}), \mathbf{0}_{p\times p}\right) \text{ for } i \in \{1,\ldots,k\}\\
&m_i = \exp(\mathbf{x}_i^{\intercal}\boldsymbol{\beta}) + p - 1,\ \beta_1 \sim t(3,0,5),\ \boldsymbol{\beta}\setminus\{\beta_1\} \sim t(3,0,2.5),\\
&\boldsymbol{\Omega}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} + \boldsymbol{\Psi},\\
&\boldsymbol{\lambda} \sim \mathcal{N}(0,\sigma_\lambda),\ \sigma_\lambda \sim t^+(3,0,1),\ \boldsymbol{\Phi} \sim \mathrm{LKJ}(1),\\
&\sqrt{\mathrm{diagonal}(\boldsymbol{\Theta})} \sim t^+(3,0,1),\ \boldsymbol{\rho}_{\boldsymbol{\Theta}} \sim \mathrm{Beta}(2,2)\times 2 - 1,\\
&\mathrm{diagonal}(\boldsymbol{\Psi}) = \mathbf{0}_p,\ \boldsymbol{\psi} \sim \mathcal{N}(0,\tau_\psi),\ \tau_\psi \sim t^+(3,0,1)
\end{aligned}
\tag{6}
$$

We explain each line in the model in the following paragraphs.

The degrees of freedom parameter for study i, m_i, is assumed to depend on the study i characteristics contained in row-vector, $\mathbf{x}_i^{\intercal}$.[2] The first element of this vector is constant with corresponding coefficient, β_1 – β_1 has a very wide t prior (location-scale notation). All other coefficients have weakly informative priors (Lemoine, 2019). This is a form of meta-regression i.e. study characteristics, often termed moderators, are assumed to influence the degree to which the unobserved population covariance matrix reflects the structured pooled structured covariance matrix, $\boldsymbol{\Omega}(\boldsymbol{\theta})$. The regression is on a shifted log-scale to ensure $m_i > p-1$. A positive coefficient of a moderator implies that studies with larger values of the given moderator have population covariance matrices that are closer to $\boldsymbol{\Omega}(\boldsymbol{\theta})$.

[2] We thank an anonymous reviewer for suggesting that the degrees of freedom be dependent on study characteristics.

$\boldsymbol{\Lambda}$ is the factor loading matrix, and $\boldsymbol{\lambda}$ is the vector of non-zero loadings assumed normal with scale hyperparameter σ_λ, which has its own half t-prior. The inter-factor correlation matrix ($\boldsymbol{\Phi}$) has an LKJ-prior (Lewandowski, Kurowicka, & Joe, 2009). $\boldsymbol{\Theta}$ is the *typical* residual covariance matrix – the square root of its diagonal elements, the item residual standard deviations, have a half t-prior. These priors are weakly informative default choices that should be suitable for items scaled such that their variances are close to 1, we encourage researchers to modify the priors for their particular applications.

In addition to $\boldsymbol{\Theta}$, $\boldsymbol{\Psi}$ also forms the residual covariance matrix (see "method 3" in Muthén & Asparouhov, 2012). $\boldsymbol{\Theta}$ is the usual residual covariance matrix, i.e. it contains hypothesized residual covariances. Off-diagonal elements are estimated via parameter expansion (e.g. Merkle & Rosseel, 2018), and its constituent correlations ($\boldsymbol{\rho}_{\boldsymbol{\Theta}}$) are assumed beta-distributed, with a boundary avoiding prior, rescaled to the $(-1, 1)$ interval.

$\boldsymbol{\Psi}$ is a full residual covariance matrix reflecting the fact that minor factors may induce small correlations between items. Estimating $\boldsymbol{\Psi}$ leads to model identification problems. Muthén and Asparouhov (2012) estimate $\boldsymbol{\Psi}$ by placing a 0-mean small-variance prior (precisely a highly informative inverse-Wishart prior) on $\boldsymbol{\psi}$ (= vechs($\boldsymbol{\Psi}$)), where vechs($\cdot$) is the strict half-vectorization function. Alternatively, we assume the diagonal of $\boldsymbol{\Psi}$ is $\mathbf{0}_p$, and borrow from standard hierarchical modeling and assume the scale of $\boldsymbol{\psi}$, τ_ψ, can be learned from the data. This similarly has the effect of regularizing/shrinking these elements towards 0. We place a weakly-informative hyper-prior on this scale, and a small value of τ_ψ suggests that weak residual covariances populate $\boldsymbol{\Psi}$.

As an alternative model to estimating a full residual covariance matrix, Muthén and Asparouhov (2012) also suggested estimating all cross-loadings using strong priors that shrink cross-loadings to zero. Both of these approaches (full residual covariance matrix and full loading matrix) allow for a model that better reflects patterns in data, as compared to models that impose local independence or simple structure. The model with the full loading matrix is substantively interesting as it allows for a meta-analytic *exploratory* SEM. However, in this paper, we opt for the full residual covariance matrix approach because this approach has a better chance of capturing patterns in data. Even when all cross-loadings are estimated, the factors in a model may fail to fully capture the relations between items. Additionally and as we will elaborate below, $\boldsymbol{\Psi}$ serves model diagnostic purposes.

Notably, $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is a covariance matrix regardless of whether $\mathbf{S}_i$ are correlation or covariance matrices, as opposed to TSSEM where the pooled matrix is often a correlation matrix.

Guidelines for model fitting and interpretation

The model in equation 6 is focused on estimating elements of the hypothesized structured covariance matrix, $\boldsymbol{\Omega}(\boldsymbol{\theta})$. We note that identifying $\boldsymbol{\Omega}(\boldsymbol{\theta})$ in equation 6 as the *hypothesized structured covariance matrix* is somewhat misleading. This is because $\boldsymbol{\Omega}(\boldsymbol{\theta})$ includes $\boldsymbol{\Psi}$ which is an unstructured residual covariance matrix containing the influences of minor factors.

Model diagnostics

One expectation of a credible hypothetical model would be that the correlations induced by minor factors are negligible. Hence, we rely on τ_ψ and $\boldsymbol{\Psi}$ for model diagnostics.

τ_ψ is the standard deviation of residual covariances; we standardize it to ease interpretation: $\tau'_\psi = \tau_\psi \Big/ \sqrt{\frac{1}{p(p-1)} \sum_{i=2}^{p} \sum_{j=1}^{i-1} \omega_{jj}^{0.5} \omega_{ii}^{0.5}}$, where $\omega_{ii/jj}$ is the i/j-th diagonal element of $\boldsymbol{\Omega}(\boldsymbol{\theta})$. Hence, τ'_ψ communicates the standard deviation of standardized residual covariances (SRCs). Substantively, τ'_ψ is similar to the standardized root mean squared residual (SRMR), and summarizes the typical size of standardized residual covariances. When τ'_ψ is under 0.05, then most SRCs will fall in the $(-0.1, 0.1)$ interval,[3] suggesting the SRCs are mostly small (Maydeu-Olivares, 2017).

It is also reasonable to examine the individual SRCs themselves. To do this, we standardize $\boldsymbol{\Psi}$: $\boldsymbol{\Psi}' = \mathbf{D}^{-1}\boldsymbol{\Psi}\mathbf{D}^{-1}$, where $\mathbf{D} = \text{diag-matrix}\left(\sqrt{\text{diagonal}(\boldsymbol{\Omega}(\boldsymbol{\theta}))}\right)$, such that $\boldsymbol{\Psi}'$ is the SRC matrix. A reasonable expectation would be that all SRCs have absolute values under 0.1 at a minimum, or preferably 0.05 (e.g. section 10 in Maydeu-Olivares, 2017). When all SRCs have absolute values under 0.05, one may regard $\boldsymbol{\Omega}(\boldsymbol{\theta})$ as practically being the hypothesized structured covariance matrix. An advantage of Bayesian estimation is that we may interpret the uncertainty of the SRCs without concerns about multiple comparisons. In situations where the SRCs are highly variable, there is insufficient information in the data to perform conclusive model diagnostics. When the SRCs are negligible and precisely estimated, we can conclude with high certainty that the correlations induced by minor factors are indeed trivial.

[3] Based on the empirical rule, 95% of values will fall within ≈ 2 standard deviations of the mean. The claim above follows from the prior under which the SRCs are assumed to have a 0-mean and τ'_ψ standard deviation.

In summary, τ'_ψ serves as an interpretable global fit index. Moreover, two models fit to the same data can be compared on the basis of the posterior distribution of their respective τ'_ψ values. And the elements of $\boldsymbol{\Psi}'$ can similarly be examined for closer investigation of the fit of the hypothesized structure to the pooled covariance structure, and has the potential to identify local misspecifications.

On the hypothesized structured covariance matrix

In the event that residual covariances are judged trivial such that $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is practically the hypothesized structured covariance matrix, then the theory behind the model structure may be assumed to underlie the population covariance matrices across the different studies. However, these population covariance matrices may be quite different from each other, and adventitious error offers a conceptual explanation for these differences. If the different studies were performed under identical conditions with identical populations, then the population covariance matrices would be identical and would follow directly from the hypothesized theory. However, given enough studies in the meta-analysis, this expectation is unrealistic. In summary, we see the question of the credibility of the hypothesized covariance matrix (as elaborated on in the preceding model diagnostics section) as separate from the dispersion between different studies (elaborated on in the next section). Even when the population covariance matrices are highly different from each other, the hypothesized structured covariance matrix may still be credible.

Interpreting the regression equation for m_i

The model incorporates a regression equation for m_i. Hence we can attempt to understand factors that lead to greater alignment between the hypothesized theory and available data. As already stated, a positive coefficient for a study factor means studies with higher levels of the given factor are closer to the hypothesized structure. However, the magnitude of m_i is not easily intelligible. Revisiting the relation between m and the RMSEA ($\varepsilon = (m+p-1)^{-1/2}$) from Wu and Browne (2015) allows for easier interpretations of coefficients. Hence, we can obtain study-level RMSEA values, $\varepsilon_i = (m_i + p - 1)^{-1/2}$, and an average RMSEA, $\bar{\varepsilon}$. Although one can interpret ε_i as a misfit of the hypothesized covariance structure to each study's population covariance matrix, we prefer to see the RMSEA values as communicating (to-be-expected) variation of the different

studies from the hypothesized structure. We next compute average marginal effects (AME, partial derivative for each predictor variable) to transform $\boldsymbol{\beta}$ in equation 6 to the scale of the RMSEA. For the j-th coefficient ($j > 1$), the AME of the corresponding moderator is:

$$-\beta_j \times \frac{1}{k} \sum_{i=1}^{k} \frac{e^{\mathbf{x}_i^\intercal \boldsymbol{\beta}}}{2(e^{\mathbf{x}_i^\intercal \boldsymbol{\beta}} + p - 1)^{\frac{3}{2}}}$$

This allows for direct estimation of the relation between moderators and the RMSEA.

Missing data handling

Since we focus on CFA MASEM applications in this paper, missing data are unlikely to be a problem, as it is common to report complete sample covariance matrices when performing CFA. However, we lay out a simple missing data strategy that should be adequate when the missing data mechanism is either missing completely at random (MCAR) or missing at random (MAR).

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, 2018a). Case ii usually occurs when only the covariance between predictors and outcomes are reported, as opposed to the full covariance matrix of all variables. 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 are imputed within the Bayesian model based on the posterior predictive distribution in equation 6. For any variables j and l in study i whose missing covariances are imputed, we increment their variance slightly by ϑ_{ji} and ϑ_{li} ($[\vartheta_{ji}, \vartheta_{li}] \sim \mathcal{N}^{+}(0, 1)$) ensuring that $\mathbf{S}_i$ remains a valid covariance matrix.

Implementation details

We provide R and Stan (Carpenter et al., 2017) code to estimate the model above. Stan is a statistical programming language for fitting models using Markov Chain Monte Carlo methods using the relatively efficient No-U-Turn sampler.

Data analysis examples

We now present some data analysis examples to demonstrate the hierarchical covariance matrix (HCM) approach we recommend. For Bayesian estimation, we draw 2000 posterior samples, retaining 1000 samples post-warmup across 4 chains, resulting in 4000 posterior samples per parameter. We assessed the model using sampler-agnostic (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2020) and sampler-specific (Betancourt, 2017) diagnostics, and all diagnostics were adequate – this is verifiable by re-running the shared code and traceplots are provided in appendix A.

As a comparison to the proposed model, we followed the standard TSSEM approach: estimate the fixed-effects pooled correlation matrix; if there is considerable error from assuming fixed correlations across studies (as identified by a first stage model with poor SEM fit indices),[4] follow a random-effects approach; otherwise complete the fixed-effects approach. We performed TSSEM using the metaSEM package in R. Although OSMASEM handles moderators like HCM, we do not include OSMASEM in our comparisons as the OSMASEM approach to handling moderators is fundamentally different.

The examples are diverse. Example 1 has two correlated factors and three moderator variables, example 2 has three correlated factors and notable residual covariances, while example 3 is a bifactor model with two moderator variables.

Example 1: Correlated factors model with moderators, Digman (1997)

The data are inter-factor correlation matrices of a five-factor model from 14 studies in Digman (1997), previously analyzed by Cheung (2014). The five variables are hypothesized to load onto two factors: *Alpha* and *Beta*. Alpha has three indicators: agreeableness (A), conscientiousness (C) and emotional stability (ES); Beta has two indicators: extroversion (E) and intellect (I).

We identified two moderators as predictors of m_i: the study population (children, $n = 4$; adolescents and young adults (AYAs), $n = 4$; adults, $n = 6$), and the study year (range = 1963 – 1994). We set adults as the reference group, and standardized study year.

We estimated the pooled structured covariance matrix using the HCM approach. τ'_{ψ} was

[4]The TSSEM literature does not provide precise cut-offs for SEM fit indices, though cut-offs in Hu and Bentler (1999) were cited by Cheung and Chan (2005).

Table 1

Parameter estimates and global fit indices for example 1, Digman (1997)

Parameter	HCM	TSSEM (WLS)
Fit indices		
τ'_ψ	0.062 [0.010, 0.162]	
Largest SRC	0.042 $[-0.055, 0.178]$	
χ^2(df)		7.82(4), $p = .098$
RMSEA		0.015 $[0.000, 0.030]$
SRMR / CFI / TLI		0.044 / .99 / .98
Moderator analysis predicting ε_i (AMEs)		
$\bar{\varepsilon}$	0.171 $[0.154, 0.190]$	
AYAs vs. Adults	-0.017 (0.031)	
Children vs. Adults	$+0.085$ (0.036)	
Study year	-0.009 (0.013)	
Factor loadings		
alpha $\rightarrow$ A	0.58 (0.096)	0.57 (0.052)
alpha $\rightarrow$ C	0.55 (0.098)	0.59 (0.053)
alpha $\rightarrow$ ES	0.69 (0.113)	0.76 (0.062)
beta $\rightarrow$ E	0.70 (0.158)	0.68 (0.076)
beta $\rightarrow$ I	0.60 (0.155)	0.64 (0.072)
Inter-factor correlations		
ρ(alpha, beta)	.35 (.105)	.38 (.047)
Error variances		
σ^2(A)	0.53 (0.110)	0.68
σ^2(C)	0.56 (0.110)	0.65
σ^2(ES)	0.43 (0.147)	0.42
σ^2(E)	0.39 (0.207)	0.54
σ^2(I)	0.54 (0.191)	0.59

Note. AYAs = adolescents & young adults. Showing parameter estimates and standard errors. RMSEA & $\bar{\varepsilon}$ intervals are 90%, residual interval is 95%. TSSEM error variances are computed not estimated within the model given that observed variable total variances were fixed to 1.

0.062, 95% credible interval (CI) [0.010, 0.162], suggesting a 95% range of SRCs ($\approx \pm 2\tau'_{\psi}$) exceeding the $(-0.1, 0.1)$ interval. However, the largest SRC was 0.042, 95% CI [-0.055, 0.178], hence the SRC estimates were all under 0.05. However, the CI for the largest SRC was notably large and several other 95% CI for SRCs exceeded the (-0.1, 0.1) range. Taken together, one is unable to conclude confidently that the degree of misspecification was indeed low, despite the small SRC point estimates.

We do not report the results of the regression on m_i, rather we report the results transformed to the RMSEA scale. The average RMSEA was 0.171, 90% CI [0.154, 0.190], suggesting considerable differences on average between the pooled covariance matrix and the population covariance matrices underlying the individual studies. Adjusted for study year, the RMSEA was on average higher by 0.085 points (SE = 0.036) for children samples as compared to adults. This is a considerable increase in RMSEA suggesting that the children samples had significantly larger variation from the pooled structure. Other moderators had imprecise estimates (95% CI included 0). Additional parameter estimates are reported in Table 1.

We also estimated the TSSEM fixed-effects pooled correlation matrix, however, there was considerable misfit of this matrix to the correlation matrices across studies: $\chi^2(130) = 1505$, $p < .001$, RMSEA = 0.182 (similar to the average RMSEA above) 90% CI [0.174, 0.190], SRMR = 0.162, CFI = .68, TLI = .66. Hence, we proceeded with a random-effects approach, and used the resulting pooled correlation matrix to estimate the hypothesized SEM – results in Table 1.

Estimates from both models were somewhat comparable though the HCM standard errors were consistently larger. The global fit indices for the TSSEM would be judged excellent by most evaluators and confidently $\big(p(\chi^2) > .05$ and narrow RMSEA interval$\big)$, differently from the conclusions based on the HCM analysis.

Example 2: Correlated factors with residual covariances

The data are the covariance matrices of 9 variables on work-related attitudes from 11 countries (*International Social Science Program: Work orientations, 1989*, 1992) – *issp89* dataset in the metaSEM package. The nine variables are hypothesized to load onto three factors: job *prospects*, job *nature*, and *time* demand. Job prospects had three indicators: job security (x1), income (x2), and advancement opportunity (x3); job nature had four indicators: interesting job

Table 2

Parameter estimates and global fit indices for example 2

Parameter	HCM	TSSEM (WLS)	HCM (Modified)
Fit indices			
τ'_ψ	0.072 [0.051, 0.100]		0.033 [0.017, 0.053]
Largest SRC	0.194 [0.045, 0.321]		−0.050 [−0.101, −0.007]
χ^2(df)		364(24), $p < .001$	
RMSEA		0.044 [0.041, 0.049]	
SRMR / CFI / TLI		0.066 / .86 / .79	
Moderator analysis predicting ε_i (AMEs)			
$\bar{\varepsilon}$	0.083 [0.078, 0.089]		0.084 [0.078, 0.089]
Factor loadings			
prospects → x1	0.47 (0.085)	0.51 (0.018)	0.46 (0.051)
prospects → x2	0.58 (0.091)	0.60 (0.018)	0.59 (0.051)
prospects → x3	0.65 (0.105)	0.69 (0.020)	0.66 (0.057)
nature → x4	0.63 (0.086)	0.72 (0.018)	0.72 (0.057)
nature → x5	0.51 (0.079)	0.53 (0.017)	0.52 (0.050)
nature → x6	0.59 (0.086)	0.50 (0.017)	0.45 (0.049)
nature → x7	0.52 (0.081)	0.42 (0.018)	0.37 (0.048)
time → x8	0.68 (0.232)	0.70 (0.093)	0.68 (0.170)
time → x9	0.28 (0.104)	0.33 (0.045)	0.33 (0.078)
Inter-factor correlations			
ρ(prospects, nature)	0.44 (0.086)	0.52 (0.019)	0.49 (0.055)
ρ(prospects, time)	0.43 (0.175)	0.43 (0.064)	0.42 (0.110)
ρ(nature, time)	0.42 (0.161)	0.38 (0.054)	0.32 (0.092)
Error covariances			
σ(x6, x7)			0.31 (0.048)
σ(x5, x8)			0.17 (0.052)
Error variances			
σ^2(x1)	0.82 (0.083)	0.83 (0.063)	0.82 (0.052)
σ^2(x2)	0.59 (0.106)	0.60 (0.036)	0.58 (0.058)
σ^2(x3)	0.65 (0.138)	0.65 (0.048)	0.64 (0.072)
σ^2(x4)	0.42 (0.107)	0.34 (0.037)	0.30 (0.077)
σ^2(x5)	0.62 (0.081)	0.69 (0.089)	0.62 (0.051)
σ^2(x6)	0.62 (0.099)	0.78 (0.073)	0.77 (0.049)
σ^2(x7)	0.58 (0.083)	0.73 (0.067)	0.71 (0.042)
σ^2(x8)	1.07 (0.350)	1.16 (0.139)	1.10 (0.268)
σ^2(x9)	0.96 (0.073)	0.99 (0.054)	0.94 (0.061)

Note. Showing parameter estimates and standard errors. RMSEA & $\bar{\varepsilon}$ intervals are 90%, residual interval is 95%.

(x4), independent work (x5), help other people (x6), and useful to society (x7); and *time* demand had two indicators flexible working hours (x8) and lots of leisure time (x9). We did not have any moderator information for these data.

We estimated the pooled structured covariance matrix using the HCM approach. τ'_{ψ} was 0.072, 95% CI [0.051, 0.100], suggesting a 95% range of SRCs exceeding the $(-0.1, 0.1)$ interval. There were two notably large SRCs: 0.194, 95% CI [0.045, 0.321] between x6 (help other people) and x7 (useful to society); and 0.104, 95% CI [0.021, 0.180] between x5 (independent work) and x8 (flexible working hours). It is reasonable on face value that these item pairs are correlated beyond the correlations induced by their factors. Hence, we decided to modify the model by purposely specifying both residual covariances (in the $\mathbf{\Theta}$ matrix). On adding both residual covariances to the model, τ'_{ψ} dropped to 0.033, 95% CI [0.017, 0.053]. We compared the posterior distributions of both τ'_{ψ} values to compare both models and there was a 99% chance that the degree of misspecification dropped from the original to the modified model. The largest SRC for the modified model was: -0.050, 95% CI [-0.101, -0.007] between x2 (income) and x7 (useful to society). Although it is again reasonable that these items are negatively beyond the model-implied correlation, we judged the magnitude of the SRC small enough to be considered a trivial misspecification. Hence, no SRC exceeded ± 0.05 and intervals were mostly contained in the (-0.1, 0.1) range. This suggests that the modified hypothesized structure in fact captures the dynamics underlying the items. One notable difference between the original and modified model is the drop in uncertainty about structural parameters (see standard errors in Table 2) after the largest SRCs are purposely modeled in the $\mathbf{\Theta}$ matrix. We now interpret other aspects of modified model.

The average RMSEA was 0.083, 90% interval [0.078, 0.089], suggesting some differences on average between the hypothesized covariance matrix and the population covariance matrices underlying the individual studies. Parameter estimates are reported in Table 2.

For the TSSEM approach, we directly meta-analyzed the sample covariance matrices. We estimated the fixed-effects pooled covariance matrix. The degree of misfit of this matrix to the covariance matrices across studies was greater than desirable: $\chi^2(450) = 2514$, $p < .001$, RMSEA = 0.084 (similar to the average RMSEA above), 90% CI [0.081, 0.087], SRMR = 0.122, CFI = .80, TLI = .82. Hence, we fit a pooled random-effect covariance matrix and used the resulting pooled covariance matrix to estimate the original hypothesized model – results in Table 2.

Point estimates from both models are somewhat comparable. Standard errors based on the HCM are notably larger. The global fit indices for the TSSEM would be judged acceptable based on the RMSEA and SRMR, and unacceptable based on the CFI and TLI.

Table 3

Parameter estimates and global fit indices for example 3, Norton, Cosco, Doyle, Done, and Sacker (2013)

Parameter	HCM		TSSEM (WLS)	
Fit indices				
τ'_ψ	0.022 [0.017, 0.027]			
Largest SRC	0.043 [0.014, 0.073]			
χ^2(df)			350(63), $p < .001$	
RMSEA			0.014 [0.013, 0.016]	
SRMR / CFI / TLI			0.022 / .98 / .98	
Moderator analysis predicting ε_i (AMEs)				
$\bar{\varepsilon}$	0.072 [0.069, 0.074]			
Non-patient vs. Patient	+0.021 (0.003)			
Study year	−0.005 (0.002)			
Factor loadings	General	Specific	General	Specific
Anx. → x1	0.63 (0.024)	0.17 (0.054)	0.68 (0.010)	0.13 (0.029)
Dep. → x2	0.44 (0.025)	0.48 (0.040)	0.42 (0.009)	0.50 (0.022)
Anx. → x3	0.58 (0.030)	0.38 (0.059)	0.62 (0.011)	0.37 (0.028)
Dep. → x4	0.49 (0.025)	0.43 (0.040)	0.48 (0.010)	0.46 (0.022)
Anx. → x5	0.66 (0.026)	0.22 (0.054)	0.71 (0.009)	0.17 (0.025)
Dep. 1 → x6	0.56 (0.026)	0.33 (0.042)	0.57 (0.010)	0.37 (0.019)
Anx. → x7	0.69 (0.028)	−0.20 (0.107)	0.73 (0.012)	−0.26 (0.046)
Dep. → x8	0.45 (0.024)	0.21 (0.039)	0.46 (0.008)	0.23 (0.016)
Anx. → x9	0.54 (0.029)	0.31 (0.055)	0.57 (0.011)	0.28 (0.030)
Dep. → x10	0.34 (0.024)	0.27 (0.037)	0.34 (0.010)	0.31 (0.018)
Anx. → x11	0.46 (0.024)	0.084 (0.056)	0.48 (0.010)	0.07 (0.028)
Dep. → x12	0.47 (0.025)	0.53 (0.041)	0.47 (0.010)	0.57 (0.022)
Anx. → x13	0.60 (0.031)	0.41 (0.063)	0.64 (0.011)	0.40 (0.032)
Dep. → x14	0.40 (0.023)	0.22 (0.037)	0.40 (0.009)	0.24 (0.016)

Note. Showing parameter estimates and standard errors. RMSEA & $\bar{\varepsilon}$ intervals are 90%, residual interval is 95%. Error variances not reported for brevity.

Example 3: Bifactor model with moderators, Norton et al. (2013)

For the final example, the data are the correlation matrices of 14 items from the Hospital Anxiety and Depression scale (HADS, Zigmond & Snaith, 1983) from 28 studies, collected and analyzed by Norton et al. (2013) and re-analyzed by Jak and Cheung (2018b). Of the models examined by Norton et al. (2013), we opted for the best-fitting model: a bifactor model with a general distress factor, with odd-numbered items loading onto an anxiety factor and even-numbered

items loading onto a distress factor – all factors were orthogonal.

We identified two moderators as predictors of m_i: the study population which we split similarly to Jak and Cheung (2020) (18 patient samples; 10 non-patient samples), and the study year (range = 2001 – 2012). We set the patient samples as the reference group, and standardized study year.

We estimated the pooled structured covariance matrix using the HCM approach. τ'_{ψ} was 0.022, 95% CI [0.017, 0.027], suggesting a 95% range of SRCs contained within the $(-0.05, 0.05)$ interval. The largest SRC was: 0.043, 95% CI [0.014, 0.073] and all SRC 95% CIs were in the (-0.1, 0.1) interval. Hence, we judged the effect of minor factors to be trivial, and the bifactor structure in fact captures the dynamics underlying the items.

The average RMSEA was 0.072, 90% interval [0.069, 0.074], suggesting some differences on average between the hypothesized covariance matrix and the population covariance matrices underlying the individual studies. Adjusted for study year, the RMSEA was on average higher by 0.021 points (SE = 0.003) for non-patient samples as compared to patient samples. This suggests that non-patients samples had significantly larger variation about the hypothesized bifactor structure. Study year had a relatively small but precisely estimated coefficient; over time, studies are on average less distant from the hypothesized structure. Additional parameter estimates are reported in Table 3.

For the TSSEM approach, we estimated the fixed-effects pooled correlation matrix. The degree of misfit of this matrix to the correlation matrices across studies was greater than desirable but not overly large: $\chi^2(2457) = 10400,\ p < .001$, RMSEA = 0.064, 90% CI [0.063, 0.066], SRMR = 0.098, CFI = .92, TLI = .91. We fit a pooled random-effect covariance matrix and used the resulting pooled covariance matrix to estimate the hypothesized model – results in Table 3. We note that estimation of the first-stage in the random-effect model took 30 minutes to converge. For comparison, HCM converged in under 150 seconds.[5]

Point estimates from both models are somewhat comparable. Standard errors based on the HCM are notably larger. The global fit indices for the TSSEM would be judged acceptable by most evaluators, similarly to the HCM results.

[5]For reference, the machine is a Ryzen 9 3900 3.1 GHz 12-core processor with 32 GB of RAM.

Simulation studies

We now present some simulation studies to evaluate the HCM approach to MASEM. Given that this is a Bayesian approach where priors do impact analysis, traditional standards for evaluating frequentist methods are less than adequate. Two basic properties of good frequentist estimators (in absolute terms as opposed to comparison amongst estimators) are: unbiasedness, which can be evaluated using bias; and adequate inference, which can be evaluated using a variety of methods including empirical error/coverage rates. With Bayesian estimation, the posterior distribution is a function of the prior and likelihood, such that the expectation of the posterior is unknown, except in the simplest of models. Hence, checking estimator bias against the parameter underlying the data can result in misleading conclusions. Moreover, that Bayesian estimators may be biased in the conventional sense can be a desirable feature of Bayesian modeling (e.g. Davis-Stober, Dana, & Rouder, 2018). Additionally, there is no expectation that Bayesian estimators maintain nominal error or coverage rates. The preceding comments make evaluation of Bayesian models challenging.

Building on the work of Cook, Gelman, and Rubin (2006), Talts, Betancourt, Simpson, Vehtari, and Gelman (2018) developed simulation-based calibration (SBC) to solve this problem. SBC simultaneously evaluates Bayesian algorithms (in our case: Stan implementation of NUTS) and model code (written by the authors) for the validity of the Bayesian inference applied to a particular model. This is particularly helpful since we provide model code that we hope researchers adopt in their MASEM applications. Moreover, SBC has been used to verify Bayesian SEM software (e.g. Merkle, Fitzsimmons, Uanhoro, & Goodrich, 2021).

With SBC, we simulate each population parameter, θ, from its prior distribution. The vector of parameters is used to simulate data according to a likelihood function. Bayesian analysis of the simulated data using the same prior and likelihood should result in parameter posterior distributions that match the original prior distributions. Across L iterations of a single Bayesian analysis and for each relevant parameter, we count the number of iterations where the posterior sample, θ_l, exceeds the simulated parameter, θ: $r = \sum_{l=1}^{L} \mathrm{I}(\theta_l > \theta)$, where $\mathrm{I}(\cdot)$ is the indicator function. The resulting value, r, will be an integer in the $[0, L]$ interval. We repeat the above process R times, simulating a different parameter value from the prior distribution each time.[6]

[6]This step differentiates SBC from traditional simulations where the population parameter is usually (but not always) kept fixed across replications (within a design condition).

Hence, we obtain r_i for $i \in \{1, \ldots, R\}$ ranks. When the Bayesian inference is valid, the distribution of these ranks, r_i, should be discrete uniform with bounds $[0, L]$.

SBC implementation details

We performed two SBC studies using the data structure from examples 1 and 3 to set up the structure of the problem, i.e. the number of indicators, moderators and assumed factor structure. We did not consider example 2 because the example has no moderator variables. We focused on calibration (or validity of inference) of the meta-regression coefficients for m_i, the residual covariances scale parameter τ_ψ, item residual variance parameters, loadings, inter-factor correlations and off-diagonal elements in $\boldsymbol{\Psi}$.

We modified the model in equation 6 to reduce the chance that invalid covariance matrices were simulated. The default priors we present may be adequate for real-world data but are capable of producing non positive semi-definite symmetric matrices (invalid covariance matrices) when they are the basis for model parameters. This leads to lack of calibration (e.g. Merkle et al., 2021). Hence we narrowed select model priors to ensure the prior choices led to valid covariance matrices. Additionally, we assumed loadings were non-negative which ensures the latent variables do not switch direction during estimation. All changes to priors are highlighted below:

$$
\begin{aligned}
& m_i = \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}) + p - 1, \;\; \beta_1 \sim \mathcal{N}(5, 0.5), \;\; \boldsymbol{\beta} \setminus \{\beta_1\} \sim \mathcal{N}(0, 0.75)\,, \\
& \boldsymbol{\Omega}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} + \boldsymbol{\Psi}, \\
& \boldsymbol{\lambda} \sim \mathcal{N}^+(0.75, \sigma_\lambda), \;\; \sigma_\lambda \sim \mathcal{N}^+(0, 0.5), \;\; \boldsymbol{\Phi} \sim \text{LKJ}(5)\,, \\
& \sqrt{\text{diagonal}(\boldsymbol{\Theta})} \sim \mathcal{N}^+(1, 0.25), \;\; \boldsymbol{\rho}_{\boldsymbol{\Theta}} \sim \text{Beta}(5, 5) \times 2 - 1\,, \\
& \text{diagonal}(\boldsymbol{\Psi}) = \mathbf{0}, \;\; \boldsymbol{\psi} \sim \mathcal{N}(0, \tau_\psi), \;\; \tau_\psi \sim \mathcal{N}^+(0.05, 0.0125)
\end{aligned}
\tag{7}
$$

For each SBC study, we repeated the process $R = 1000$ times, and requested 5000 posterior samples per parameter. We dropped the first 1000 samples as part of the warmup phase, and thinned the remaining 4000 samples by retaining every other iteration – thinning improves the validity of the SBC assessment metrics below (Talts et al., 2018). Hence, for each parameter, we retained $L = 2000$ posterior samples.

Evaluation metrics

Given the ranks, r_i, across R replications, we compute the empirical quantile for each replication i as rankits: $q_i = (r_i + 0.5)(L + 1)^{-1}$. If the ranks are discrete uniform with bounds, $[0, L]$ – as is the case under adequate inference – then the distribution of the quantiles, q_i, should be approximately continuous uniform with bounds $(0, 1)$. Hence, the standard normal quantile function applied to these quantiles, $\Phi^{-1}(q_i)$, should result in an approximately standard normal variable – approximate given the discreteness of the empirical quantiles (Talts et al., 2018). The standard normal transformation of the empirical quantiles forms the basis for evaluating the simulation results.

Bias of mean. For each parameter, the mean of the transformed variable, $\frac{1}{R}\sum_{i=1}^{R}\Phi^{-1}(q_i)$ over R replications should be 0, since the transformed variable is standard normal under valid inference. Any significant deviations of this mean from 0 suggest that estimation for the given parameter is biased. We compute the mean deviation from 0 in standardized units, such that it can be interpreted as a standardized mean difference, Cohen's d. For each parameter of interest, we assess bias using a one-sample t-test, where statistically significant bias occurs when the absolute Cohen's d exceeds: $\Phi^{-1}(.975)(R-1)^{-1/2} = 1.96 \times (1000-1)^{-1/2} = 0.062$. This test is akin to testing parameter bias in a more traditional simulation study.

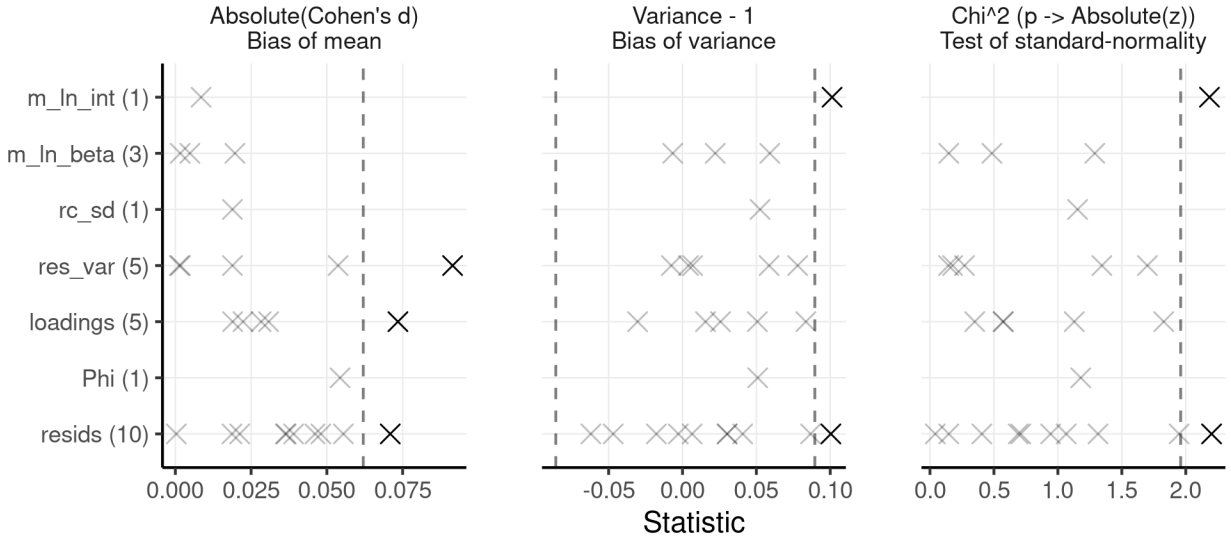
Bias of variance. For each parameter, the variance of the transformed variable, $\frac{1}{R}\sum_{i=1}^{R}\left[\Phi^{-1}(q_i)\right]^2 - \left[\frac{1}{R}\sum_{i=1}^{R}\Phi^{-1}(q_i)\right]^2$ over R replications should be 1, since the transformed variable is standard normal under valid inference. Any significant deviations of this variance from 1 suggest that posterior variance for the given parameter is biased leading to inadequate inference. For each parameter of interest, we assess bias of the variance using a one-sample χ^2-test of variance, where statistically significant bias occurs when sample variance of the transformed variable falls outside the bounds: $\left(\frac{\chi^2_{.025,R-1}}{R-1}, \frac{\chi^2_{.975,R-1}}{R-1}\right) = (0.914, 1.090)$. This test is akin to testing parameter standard error bias in a more traditional simulation study.

Test of standard-normality. The preceding metrics have tested how well the transformed variable matches the first two moments of a standard normal variable. The present test examines the standard-normality assumption in its entirety. For each parameter, if the transformed variable is standard normal, then the sum of squares, $\sum_{i=1}^{R}\left[\Phi^{-1}(q_i)\right]^2$ should be χ^2 with $R = 1000$ degrees

of freedom. A normality test based on the sum of squares is suitable because it is sensitive to extreme empirical quantiles (Cook et al., 2006). We evaluate the distribution function of χ^2_{1000} at the sum of squares, and if the resulting p-value is extreme ($p < .025$ or $p > .975$), then it is unlikely the transformed variable follows a standard normal distribution. Following Cook et al. (2006), we transform the p-values to z-statistics to more easily distinguish extreme p-values. Statistically significant deviations from normality occur when the absolute z-statistic exceeds $\Phi^{-1}(.975) = 1.96$.

Figure 1

SBC evaluation metrics for example 1 dataset: correlated factors, three moderators



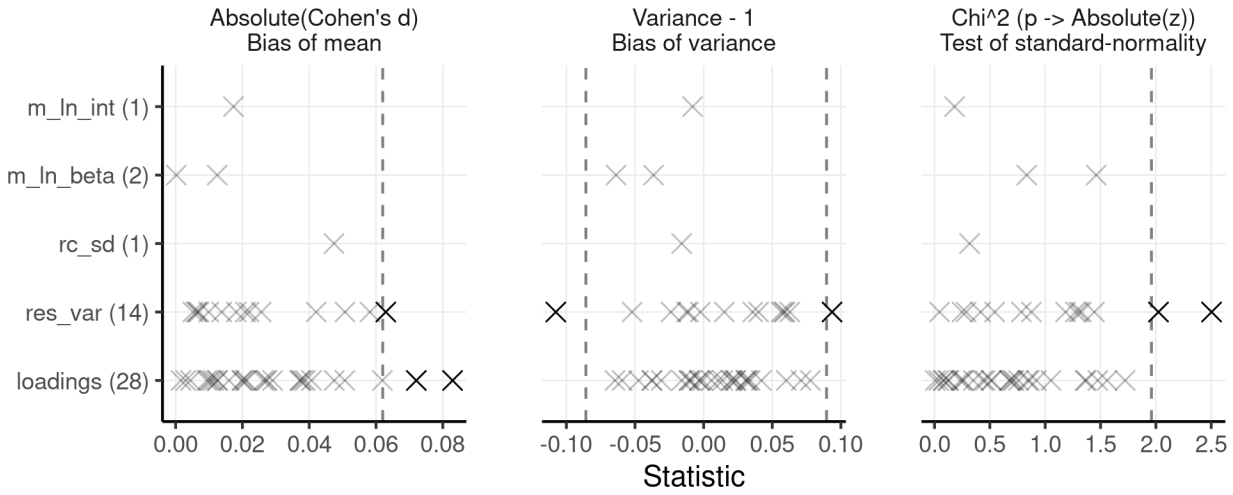
Note. m_ln_int: Moderator regression intercept; m_ln_beta: Moderator regression coefficients; rc_sd: τ_ψ; res_var: residual variance parameters; Phi: inter-factor correlations; resids: off-diagonal elements in $\mathbf{\Psi}$. Number in parenthesis on y-axis is count of parameters. Vertical dashed lines are limits defined in SBC evaluation metrics subsection; estimates exceeding limits (bolded points) are inadequate estimates.

SBC results

We discuss the SBC results based on both datasets together as the results are substantively identical. As shown in Figures 1 and 2, most parameters had mean and variance bias that were not statistically significant, providing some evidence that parameter estimation was both unbiased and inference was valid. Similarly, most parameters did not have statistically significant deviations from standard-normality providing further evidence of adequate estimation and inference. For the handful of parameters that had inadequate estimates, there was no pattern to them, i.e. inadequate results did not cluster by parameter type.

Figure 2

SBC evaluation metrics for example 3 dataset: bifactor model, two moderators



Note. m_ln_int: Moderator regression intercept; m_ln_beta: Moderator regression coefficients; rc_sd: τ_ψ; res_var: residual variance parameters. Number in parenthesis on y-axis is count of parameters. Vertical dashed lines are limits defined in SBC evaluation metrics subsection; estimates exceeding limits (bolded points) are inadequate estimates. Off diagonal-elements ($n = 91$) in $\mathbf{\Psi}$ are not shown as they clutter the plot. However, 10, 2 and 3 out of 91 parameters had inadequate mean bias, variance bias, and standard-normality tests respectively.

Additionally, the distribution for all parameter ranks were uniform for both SBC studies (see Figures B1 and B2 in Appendix). In conclusion, the SBC results suggest the proposed approach produces valid Bayesian inference for typical meta-analytic confirmatory factor problems.

Discussion

In this paper, we have presented a one-stage Bayesian method for meta-analytic SEM, based on a hierarchically estimated pooled structured covariance matrix. The approach includes a global fit index – which can be evaluated by itself or used for comparing models – and permits investigation of local misspecification. We have provided code to estimate the model, and simulation-based calibration suggests the provided code provides valid Bayesian inference. And we have demonstrated the method for meta-analytic confirmatory factor examples, comparing results to the commonplace two-stage MASEM approach as implemented in the metaSEM package.

The HCM approach is based on the concept of adventitious error. Adventitious error is error due to differences between the population for which a hypothesized SEM holds, and the real-world population from which data were collected. These differences arise from non-random

sampling of study participants. For this reason, we should expect a discrepancy between the pooled covariance matrix which exists at the level of the theoretical population, and the population covariance matrix underlying individual studies. The average RMSEA describes the discrepancy between the hypothesized SEM and the population covariance matrix underlying the different studies. And the regression equation for the RMSEA allows us identify populations for which the hypothesized structure is most likely to hold on average.

We formulated the HCM approach within a Bayesian framework. By adopting a Bayesian framework, we were able to model misspecification simultaneously with structural parameters. Both misspecification and structural parameters are estimated at the level of the theoretical population underlying the different populations in the individual studies. Only after it is confirmed that the degree of this misspecification is low can the hypothesized covariance structure be considered credible. Modeling misspecification simultaneously with structural parameters represents an advancement in MASEM, similarly to the work of WB (2015) in single-study SEMs. Importantly, uncertainty due to model misspecification is reflected via increased uncertainty about structural parameters – this can be seen in the data analysis examples, as TSSEM does not account for uncertainty due to model misspecification.

We now compare the proposed HCM approach to extant MASEM methods. Compared to TSSEM, HCM is one-stage and implemented within a Bayesian framework. Additionally, the Bayesian foundation of HCM reduces the accessibility of the approach. However, there are two practical advantages of HCM. Although model runtime may typically be slower than TSSEM, the reverse is increasingly true as the size of the covariance matrix increases. For example, the random-effect TSSEM in example 3 (bifactor model with 14 items across 28 correlation matrices) took 30 minutes to converge on our test machine, while HCM converged in under 2 minutes.[7] Second, HCM improves over TSSEM by allowing for the inclusion of moderators. In that regard, HCM is more similar to OSMASEM which permits moderators. However, HCM and OSMASEM address the moderators differently. OSMASEM assumes moderators may influence all structural parameters resulting in a regression equation for each structural parameter. This is likely most interpretable or useful in a path analytic MASEM contexts. Alternatively, HCM assumes a single hypothesized structure, and moderators explain the discrepancy from this hypothesized structure. We believe

[7] OSMASEM as implemented in the metaSEM package took 17 minutes, no moderators in the model.

this is more interpretable or useful in the confirmatory factor analytic contexts we focused on. Additionally, as discussed by the authors, OSMASEM has convergence problems when there is more than one moderator unlike the HCM approach. Finally, we hope to systematically compare extant MASEM methods to HCM in the future.

With regard to limitations, one clear case where the HCM approach we recommend will be inadequate is when the MASEM application includes both long and short forms of an instrument. Usually, short form instruments include indicators chosen for their strong relation to the latent variable(s) of interest (Widaman, Little, Preacher, & Sawalani, 2011), hence missing indicators are missing because of their likely smaller correlations with other items on the scale – a case of variables missing not at random, which has been shown to bias MASEM methods (Furlow & Beretvas, 2005).

Future directions

Two-stage approach. An alternative to the current one-stage proposal is a two-stage approach. In the first step, one hierarchically estimates a pooled unstructured covariance matrix and its asymptotic covariance matrix (see theoreom 5.3.20 in Gupta & Nagar, 1999) as opposed to the structured covariance matrix – this step may also account for moderators. In the second step, the unstructured matrix and its asymptotic covariance are used to estimate the hypothesized SEM using robust WLS estimation (Muthén, du Toit, & Spisic, 1997). The advantage of this two-stage approach is purely practical: the modeler can use their usual SEM software and maintain their modeling traditions.

Handling dependent sample covariance matrices. Sample covariance matrices may be dependent, e.g. multiple covariance matrices from the same study. Ignoring this source of uncertainty can be a problem (Wilson, Polanin, & Lipsey, 2016). The hierarchical covariance estimation approach we lay out can be extended to account for such data by extending equation 6:

$$\mathbf{S}_{ij} \sim \text{GB}_p^{\text{II}}\left(\frac{n_i^*}{2}, \frac{m_i}{2}, \frac{m_i}{n_i^*}\boldsymbol{\Sigma}_{j[i]}, \mathbf{0}_{p\times p}\right) \text{ for } i \in \{1, \ldots, k\}$$

$$o_j\boldsymbol{\Sigma}_j \sim \mathcal{W}_p(\boldsymbol{\Omega}(\boldsymbol{\theta}), o_j) \text{ for } j \in \{1, \ldots, c\}$$

where $\boldsymbol{\Sigma}_j$ is the unstructured population covariance matrix which varies by cluster j. This proposal

simply exploits conjugate pairings (Wishart → inverse-Wishart → Wishart). Additionally, such an approach would allow cluster-level predictors which influence the dispersion parameter, o_j, that controls the distance between the cluster covariance matrices and the hypothesized structured matrix, $\boldsymbol{\Omega}(\boldsymbol{\theta})$.

We believe that investigation of the above proposals will extend the current approach to accommodate the variety of analyses and data conditions that comprise MASEM.

Declarations

Data Availability Statement. All code for simulation studies, data analyses and Stan scripts are available at `https://osf.io/rstzk/`.

References

Becker, B. J. (1992, December). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*, *17*(4), 341–362. doi: 10.3102/10769986017004341

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

Browne, M. W. (1974, January). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*(1), 1–24. doi: 10.10520/AJA0038271X_175

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

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

Davis-Stober, C. P., Dana, J., & Rouder, J. N. (2018, November). Estimation accuracy in the psychological sciences. *PLOS ONE*, *13*(11), e0207239. (Publisher: Public Library of Science) doi: 10.1371/JOURNAL.PONE.0207239

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

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.

Hu, L., & Bentler, P. M. (1999, January). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi: 10.1080/10705519909540118

International Social Science Program: Work orientations, 1989. (1992). Inter-university Consortium for Political and Social Research.

Jak, S., & Cheung, M. W.-L. (2018a, 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. (2018b, August). Testing moderator hypotheses in meta-analytic structural equation modeling using subgroup analysis. *Behavior Research Methods*, *50*(4), 1359–1373. doi: 10.3758/s13428-018-1046-3

Jak, S., & Cheung, M. W. L. (2020). Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological Methods*, *25*(4), 430–455. (Place: US Publisher: American Psychological Association) 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

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

Lewandowski, D., Kurowicka, D., & Joe, H. (2009, October). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. doi:

10.1016/J.JMVA.2009.04.008

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533–558. doi: 10.1007/s11336-016-9552-7

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

Merkle, E. C., & Rosseel, Y. (2018, June). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. doi: 10.18637/jss.v085.i04

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. doi: 10.1037/a0026802

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997, November). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Retrieved from `https://www.statmodel.com/download/Article_075.pdf`

Norton, S., Cosco, T., Doyle, F., Done, J., & Sacker, A. (2013, January). The hospital anxiety and depression scale: A meta confirmatory factor analysis. *Journal of Psychosomatic Research*, *74*(1), 74–81. doi: 10.1016/j.jpsychores.2012.10.010

Ogasawara, H. (2001, September). Standard errors of fit indices using residuals in structural equation modeling. *Psychometrika*, *66*(3), 421–436. doi: 10.1007/BF02294443

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

Roux, J. J. J., & Becker, P. J. (1984). *On prior inverted Wishart distribution.* (Tech. Rep.). Pretoria, South Africa: Department of Statistics and Operations Research, University of South Africa.

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

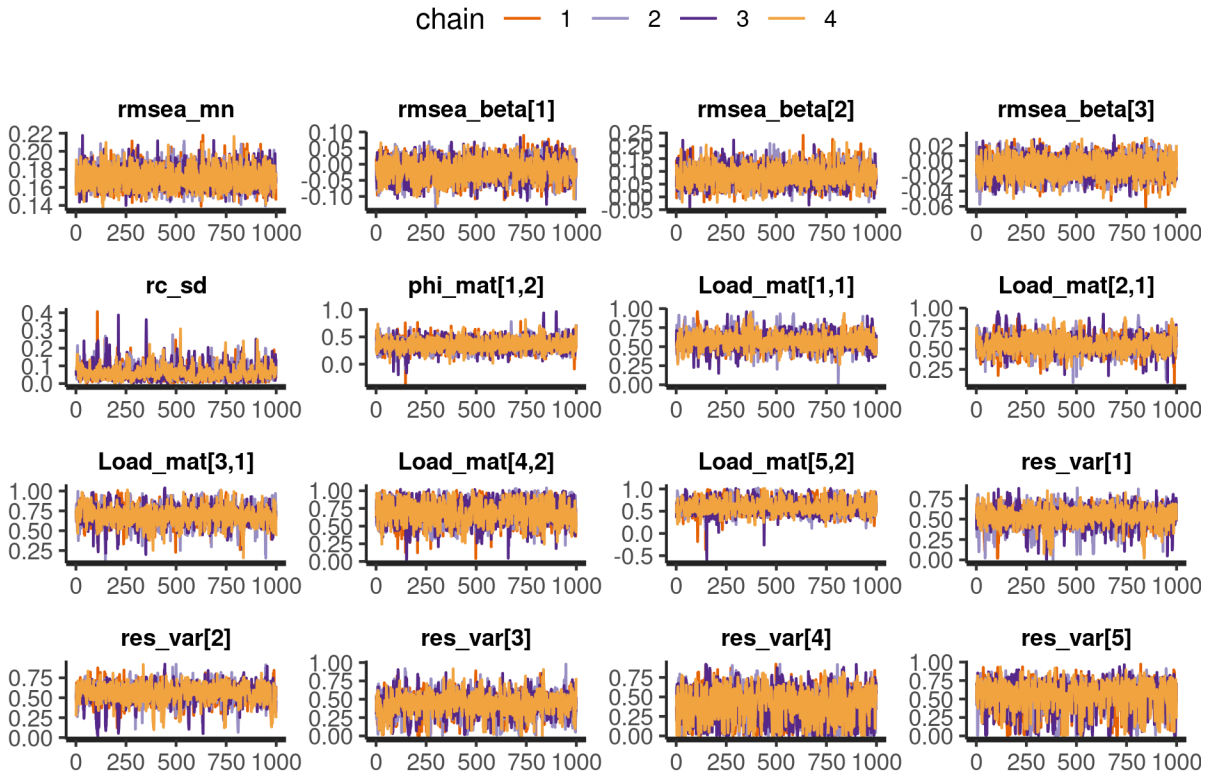
Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*, 1–28. doi: 10.1214/20-BA1221

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

Figure A1

Traceplot for τ'_{ψ}, average RMSEA, RMSEA coefficients and structural parameters from Digman (1997) example.



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

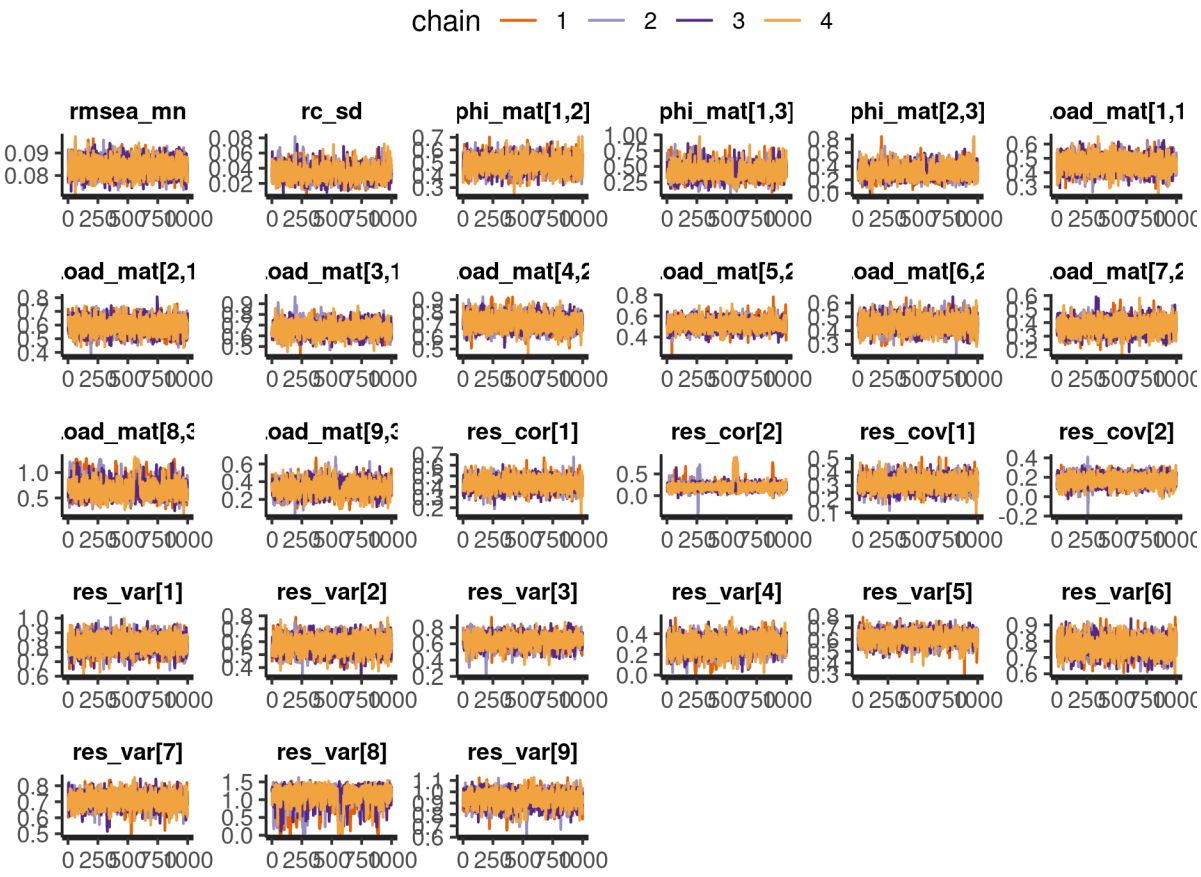
Yuan, K.-H., & Kano, Y. (2018, December). Meta-analytical SEM: Equivalence between maximum likelihood and generalized least squares. *Journal of Educational and Behavioral Statistics*, *43*(6), 693–720. doi: 10.3102/1076998618787799

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, *67*(6), 361–370.

Appendix A

Figure A2

Traceplot for τ'_{ψ}, average RMSEA and structural parameters from example 2 modified model.



Hierarchical covariance matrix diagnostic traceplots

Appendix B

Hierarchical covariance matrix SBC histograms

Figure A3

Traceplot for τ'_{ψ}, average RMSEA, RMSEA coefficients and structural parameters from Norton et al. (2013) example.

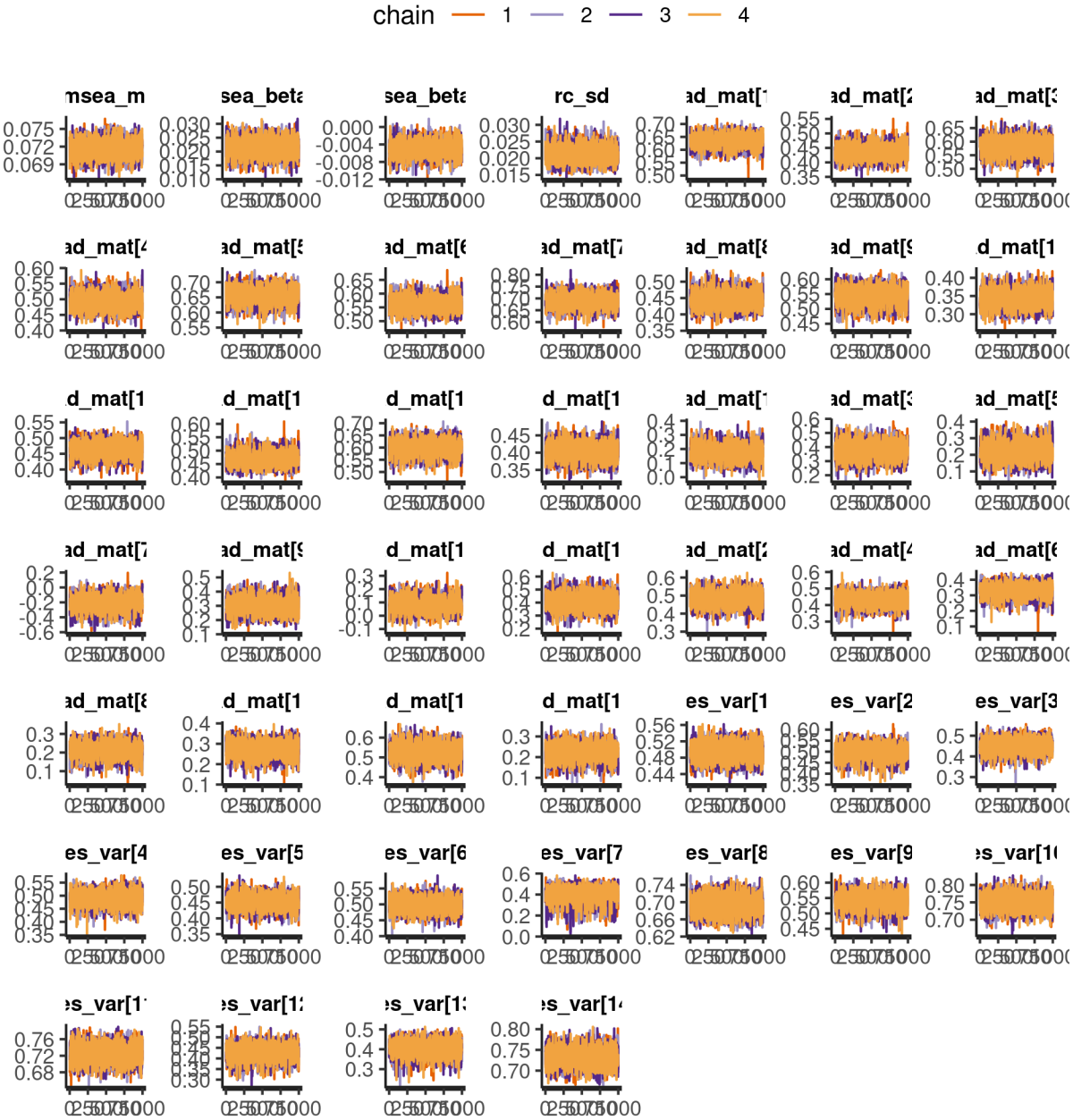
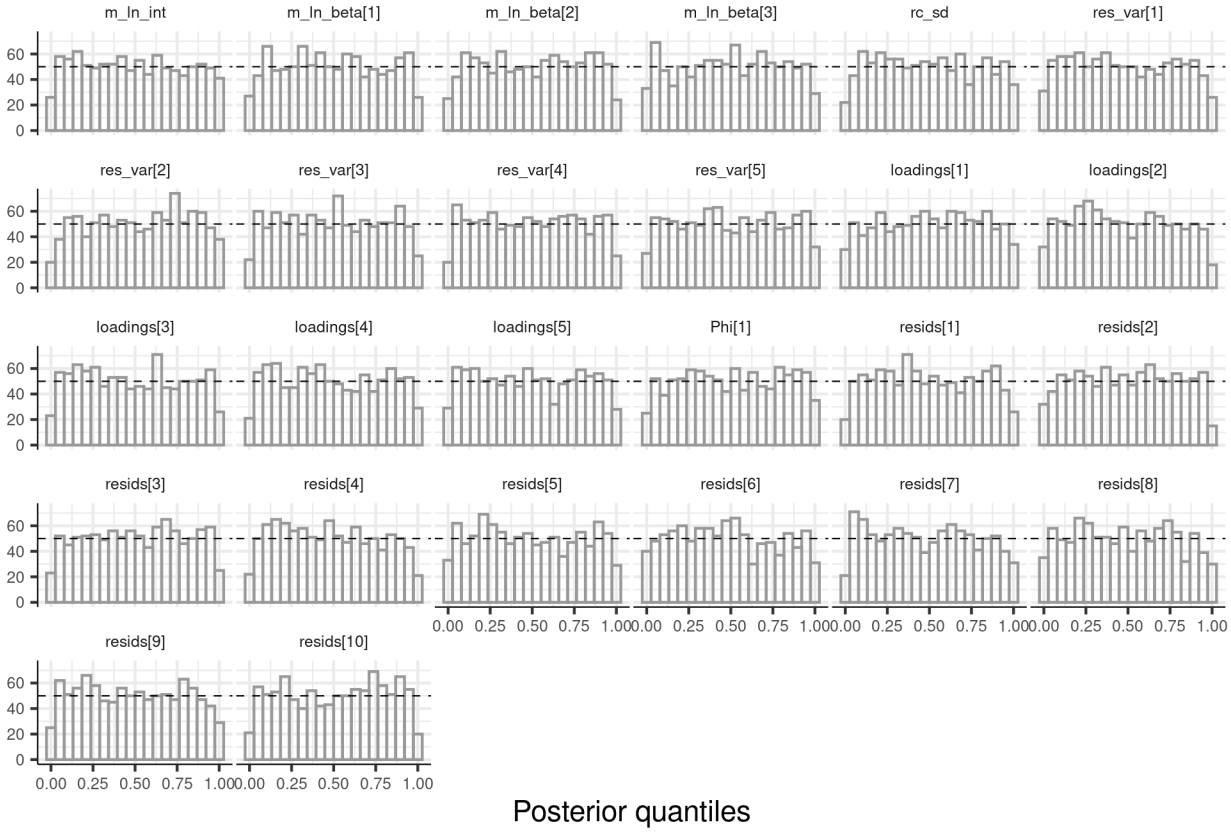


Figure B1

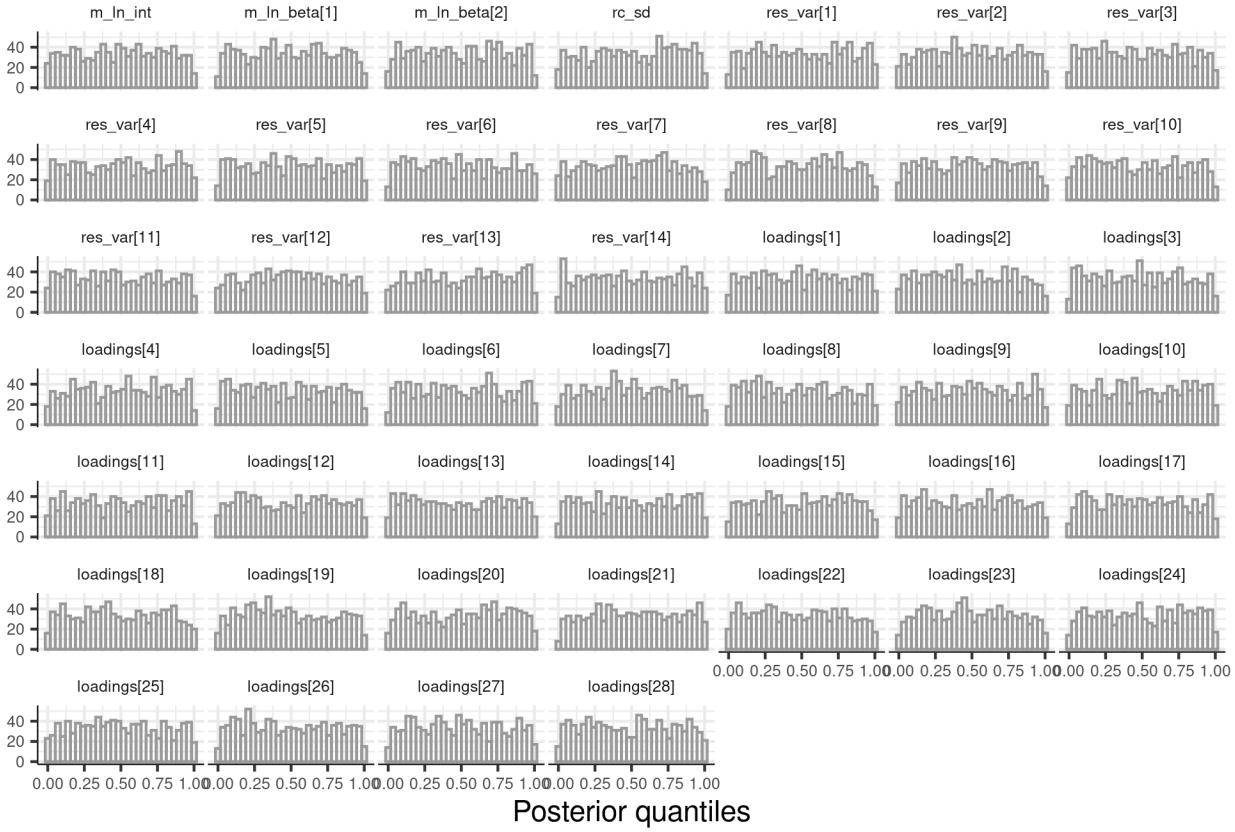
SBC histogram for example 1 dataset: correlated factors, three moderators



Note. Uniform histograms suggest the Bayesian sampler is adequate for parameters; deviations from uniformity suggest the sampler is not properly calibrated.

Figure B2

SBC histogram for example 3 dataset: bifactor model, two moderators



Note. Uniform histograms suggest the Bayesian sampler is adequate for parameters; deviations from uniformity suggest the sampler is not properly calibrated.