STRUCTURAL EQUATION MODELING
2025, VOL. 32, NO. 2
ISSN: 1070-5511 | E-ISSN: 1532-8007

Assessing Heterogeneity of Correlation Matrices in Misspecified Meta-Analytic Structural Equation Models

Christian Bloszies and Tobias Koch

Friedrich Schiller University Jena

ABSTRACT

Meta-analytic structural equation modeling (MASEM) techniques are increasingly common tools to synthesize data across multiple studies. One popular approach is two-step MASEM, where study correlation matrices are pooled in a first stage using either a fixed- or random-effects model, to then fit one or multiple structural equation models onto the pooled correlation matrix in a second stage. In a simulation study, we examined the performance of different fit criteria and resulting parameter estimates under both random- and fixed-effects pooling when fitting a three-factor CFA model to study populations that were partly misspecified. We discuss benefits and issues when using a random-effects model in this scenario and discuss future research directions regarding correlation matrix heterogeneity when using MASEM methods.

KEYWORDS

Fixed-effects; heterogeneity; meta-analysis; random-effects; structural equation modeling

Meta-analytic structural equation modeling (MASEM) uses various statistical techniques to combine structural equation modeling (SEM) with meta-analytic approaches. It can be used to synthesize correlation or covariance matrices across different studies to then fit structural equation models on the pooled correlation or covariance matrix. In that way, MASEM is extending the advantages of structural equation modeling (e.g., flexibility of modeling structure, modeling multiple outcomes, estimation of latent variables, assessment of measurement error) beyond single datasets, allowing for more powerful and generalizable applications. Common MASEM approaches involve two analysis stages. In the first stage, correlation or covariance matrices from different studies are pooled to create an average correlation matrix. This matrix is then used to fit structural equation models in the second analysis stage (Cheung, 2015).

When analyzing data using MASEM in this way, a researcher has to decide between using a fixed- or a random-effects model. Fixed-effects models assume that studies share a common population effect size, or that all study-level influences causing effect size variation are known and can be statistically adjusted for, while random-effects models allow variation of effect sizes across studies (Valentine et al., 2022). A random-effects model is commonly recommended as the assumption of common effect sizes is often conceived as too strict and unrealistic when combining studies of different origin, design and quality (Borenstein et al., 2010). Despite their limitations, fixed-effects models can still be applied meaningfully to meta-analytic research. Even with effect size heterogeneity, fixed-effects models estimate a common effect shared by different individual study populations. This estimated common effect has been shown to have good statistical properties even without assuming effect size homogeneity (Rice et al., 2018).

Apart from the decision between a fixed-effects and a random-effects model in stage one based on theoretical grounds, researchers can also use study-level moderators to account for heterogeneity between different studies (Jak & Cheung, 2018b). When moderator variables are available, the data can be split up into subgroups in which specific MASEM assumptions (i.e., homogeneity for fixed-effects) do hold. By accounting for these moderators, fixed-effects models can then be applied to subsets of the original data. A more recent approach is to conduct MASEM analyses in one single step, which has been implemented in an easy to access web application (Jak et al., 2021). This allows a direct and user-friendly way to do MASEM analyses with the benefit of using continuous study-level variables (Jak & Cheung, 2018b). Results from both the one-step and two-step approaches are similar in many cases (Jak & Cheung, 2020) with one-step MASEM providing a flexible framework to test continuous moderator hypotheses, while two-step MASEM allows the assessment of correlation matrix heterogeneity within a dedicated and separate analysis step.

We focus on the two-step approach implemented in the `metaSEM` package available for *R* (Cheung, 2015). At the first stage, individual study correlation matrices are pooled using multiple group SEM. The resulting average (or pooled) correlation matrix is then used to fit a specific structural equation model in stage two. Homogeneity of

individual study correlation matrices, corresponding to equality constraints between studies, can be assessed by goodness-of-fit indices available in `metaSEM` alongside the pooled stage one correlation matrix.

Fit indices used in MASEM analyses differ in their assumptions and implications (Jak, 2015). The chi-squared test assumes that homogeneity strictly holds for the whole population. This null model of homogeneity can be tested with a chi-squared test and its associated p value. When fitting a large SEM (i.e., many manifest variables) to comparatively small sample sizes, the chi-squared statistic is known to be positively biased, resulting in type I error inflation. This is known as the model size effect (Shi et al., 2019). Another potential issue when assessing heterogeneity in stage one with the chi-squared statistic is rejecting the strict null hypothesis of homogeneity even under negligible amounts of heterogeneity or misspecification, so researchers might want to supplement these results with other fit indices also available in the `metaSEM` package instead of only relying on one single statistical test.

Indices of approximate fit like the root mean squared error of approximation (RMSEA; see Browne & Cudeck, 1992) relax the strict assumption of homogeneity in the whole population to only an approximate fit of the fixed effect model. In that sense, we assume homogeneity to not exactly hold when relying on such approximative indices, even when using a fixed-effects model. Instead, we assume that the population is sufficiently close to the conditions of a fixed-effects analysis being met. It is the researcher's decision which amount of heterogeneity and its associated amount of parameter bias are still acceptable while using a fixed-effects model. The RMSEA is evaluating the model misfit per degrees of freedom. Within the `OpenMx` functions used by the `metaSEM` package, the RMSEA is calculated as

$$\widehat{RMSEA} = \sqrt{\frac{\max\left(\left(\chi^2_T - df_T\right)/(n-1), 0\right)}{df_T}} \tag{1}$$

where χ^2_T is the chi-square statistic value of the target model, n is the overall sample size and df_T are the degrees of freedom of the target model.

The standardized root mean square residual (SRMR) quantifies the average discrepancy between the model-implied correlation matrix and the sample correlation matrix. The closer the candidate model corresponds to the true data generating model, the closer the model-implied correlation matrix will be to the sample correlation matrix. The SRMR for a covariance structure analysis is calculated as

$$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left(\frac{s_{ij}-\hat{\sigma}_{ij}}{\sqrt{s_{ii}s_{jj}}}\right)^2}{\frac{p(p+1)}{2}}} \tag{2}$$

where s_{ij} and $\hat{\sigma}_{ij}$ are elements of the sample covariance matrix and the model-implied covariance matrix, respectively.

The comparative fit index (CFI) quantifies the proportion reduction when comparing the chi-square values of the baseline model and a target model and is bounded between 0 and 1. A value bigger than .95 is considered indicating good fit (Hu & Bentler, 1999). The CFI is calculated as

$$CFI = 1 - \frac{\max\left(\left(\chi^2_T - df_T\right), 0\right)}{\max\left(\left(\chi^2_T - df_T\right), \left(\chi^2_B - df_B\right), 0\right)} \tag{3}$$

for the chi-square test statistics and degrees of freedom of the baseline model (χ^2_B, df_B) and a target model (χ^2_T, df_T). The magnitude of CFI values is related to observed correlations, with lower correlations leading to a lower CFI value, thus preferring the baseline model. For that reason, previous research does not expect the CFI to be helpful in assessing stage one heterogeneity (Jak, 2015).

While there is a substantial amount of research on fit indices in SEM applications (Hu & Bentler, 1999), simulation studies evaluating their performance in specific MASEM settings are comparatively sparse. Previous simulation research has evaluated use cases such as the general performance of fixed-effects MASEM with missing correlation coefficients (Jak & Cheung, 2018a) and a general comparison of one- and two-step MASEM approaches using a two-factor CFA model (Jak & Cheung, 2020). Our study adds to existing research by focusing on additional heterogeneity caused by misspecified models that might be missed by researchers when analyzing their data.

In our study, different population models holding for two subpopulations cause additional variation in study correlation matrices and factor loadings. In absence of a moderator explaining these differences, there is a practical choice between using a fixed-effects model and treating possible violations of assumptions as negligible on one hand and using a random-effects model to estimate variability on the other. The present study evaluated the performance of stage one fit indices for assessing correlation matrix heterogeneity and parameter bias resulting from both a fixed- or random-effects pooled correlation matrix for testing a structural equation model in stage two.

Methods

Simulation Design

Baseline Model

All datasets were generated from a three-factor CFA model with three indicators per latent factor (Figure 1). Main loadings were set to (0.4, 0.5, 0.6) for each factor, with latent factors having a correlation of 0.3. These loadings represent an order of magnitude regularly encountered in psychological studies, e.g., personality research (Church & Burke, 1994; Haynes et al., 2000).

Sample Sizes

We varied the number of studies in the simulated meta-analyses N_S = (12, 24, 48) as well as individual sample size (i.e., number of observations) in each study N_O = (125, 250,

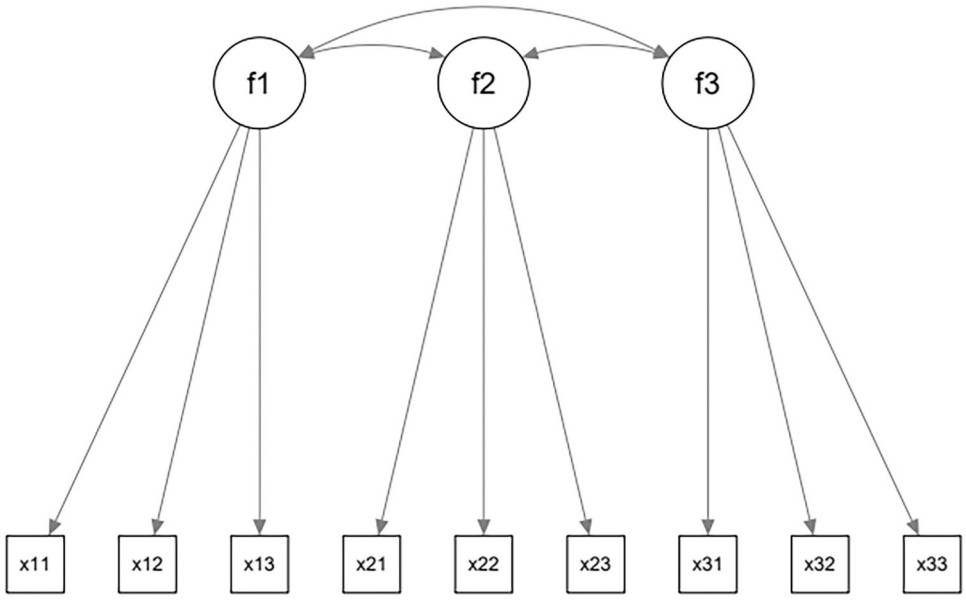

Figure 1. Simulation model. *Note*. Baseline CFA model with three factors (f1,f2,f3) and three indicators each. Cross-loading conditions: one (f1 to x32), two (f1 to x32, f2 to x12), three (f1 to x32, f2 to x12, f3 to x12).

500) to provide insight into typical use cases and more optimistic cases where larger samples are available.

Heterogeneity

Baseline effect size heterogeneity was simulated for each individual study's underlying population by sampling factor loadings from a normal distribution centered around the population grand mean with a standard deviation of 0.05. The resulting study population correlation matrices were used to simulate correlation matrices for individual studies.

In addition to this underlying effect size variability, we added heterogeneity on a model specification level by generating a percentage of datasets within a meta-analysis from another population model with one, two or three cross-loadings added (all of size 0.3). The percentage of datasets within a meta-analysis being generated from the baseline model was set to two levels (50%, 75%). We also simulated a baseline condition without cross-loadings (i.e., one single population model with no misspecifications).

Analysis

Stage One

All analyses were performed using R Statistical Software (V4.3.0; R Core Team, 2023). Simulated correlation matrices were pooled and analyzed using both the fixed-effects and random-effects settings of the `tssem1()` function from the `metaSEM` package (V1.3.1; Cheung, 2015). The assumption of correlation matrix homogeneity after fitting a stage one fixed-effects model was assessed with a chi-square statistic testing the strict null hypothesis of correlation matrix homogeneity, as well as commonly used fit indices (RMSEA, SRMR, CFI). We used common guidelines and rules of thumb as cutoffs for both the chi-square test and fit indices. Although this heuristic approach has been criticized (Hu & Bentler, 1999; Nye & Drasgow, 2011; Schermelleh-Engel et al., 2003), rules of thumb remain common practice in applied SEM research and should be evaluated accordingly.

Stage Two

For the second TSSEM stage, a correlated three-factor model without any misspecification (i.e., no cross-loading misspecifications) was fitted to the pooled correlation matrices from stage one using the `tssem2()` function from the `metaSEM` package. This corresponds to fitting an "ideal" correlated three-factor model to data with (partly) misspecified population models (i.e., cross-loadings in a proportion of simulated studies). In that way we could evaluate the extent to which the parameters in the TSSEM are biased when this heterogeneity of the correlation matrices remains undetected, and an "incorrect" model is fitted to a subset of studies. We want to emphasize that the term "bias" in this type of simulation study needs to be interpreted with great caution, because it fully depends on both the data generating model and the specific analysis model. We calculated the relative parameter estimation bias (peb_p) for the parameter estimates of all manifest indicator loadings as

$$peb_p = \frac{|M_p - e_p|}{e_p} \tag{4}$$

where M_p denotes the average of the stage two parameter estimates over all replications and e_p is the true population value. Note that true population value refers to parameters in a correlated three-factor model without misspecification (i.e., no cross-loadings). However, M_p is computed based on the simulated data containing correctly and misspecified population models. The relative parameter estimation bias should not exceed 10% for any parameter included in the model (Koch, 2013; Muthén & Muthén, 2002).

Similarly, we calculated the relative factor correlation bias (peb_c) for each factor combination as

$$peb_c = \frac{|M_c - e_c|}{e_c} \quad (4)$$

where M_c denotes the average of the stage two factor correlations over all replications and e_c is the true population value for a given factor correlation.

Results

Stage One – Fixed-Effects Fit Indices

Cross-Loading Misspecification

p Values. The p values for the chi-square test of stage one heterogeneity were sensitive to the number of studies, sample sizes and the number of cross-loadings (Figure 2). Using 0.8 as a cutoff for an acceptable rejection rate (corresponding to 80% statistical power), the chi-square test performed reliably for the three-crossloading condition. Performance for two cross-loadings was good for at least 24 studies. With only one cross-loading, the chi-squared test did not reject homogeneity reliably for 12 studies and lower sample sizes.

RMSEA. When using 0.08 as a conservative cutoff for adequate fit, RMSEA values were consistently below the threshold for all conditions. With the stricter 0.05 cutoff, only the three cross-loadings condition showed notable differences between sample sizes. Overall, RMSEA values did not indicate poor fit for cross-loading misspecifications (Figure 3).

SRMR. The SRMR values were indicating good fit using a cutoff at 0.08 for all conditions except the lowest sample size of 125. While both higher misspecification levels and percentage of misspecified models were associated with higher SRMR values, the changes were not substantial when using common rules of thumb. Overall, the SRMR worked mostly as a test for sample size regardless of the underlying data (Figure A1).

CFI. The CFI rejected homogeneity reliably for three cross-loadings (Figure 4). Both low and medium misspecification showed a counterintuitive pattern of model fit rejection increasing with lower sample sizes. This is consistent with the small sample size effect (Shi et al., 2019), where chi-squared values (which are used in the CFI calculation) can be positively biased when fitting relatively big models to relatively small samples. Compared to the chi-squared test, the CFI did react less strict to violations of homogeneity while still showing a pattern of indicating poorer fit when more misspecifications were present.

Baseline Condition

p Values. For the baseline condition with no misspecification, the chi-squared test did not reliably reject correlation matrix heterogeneity even though the underlying baseline heterogeneity was still present (Figure 5). Differing from the misspecification conditions, the baseline condition clearly showed the small sample size effect for all conditions.

RMSEA. Using an RMSEA cutoff of 0.05, all conditions were marked as having good fit with only baseline heterogeneity present.

SRMR. Similar to conditions with cross-loadings, SRMR values were predominantly sensitive to sample size in the baseline heterogeneity condition (Figure A2). Model fit was generally considered bad for a sample size of 125 and good for sample sizes of 250 or 500.

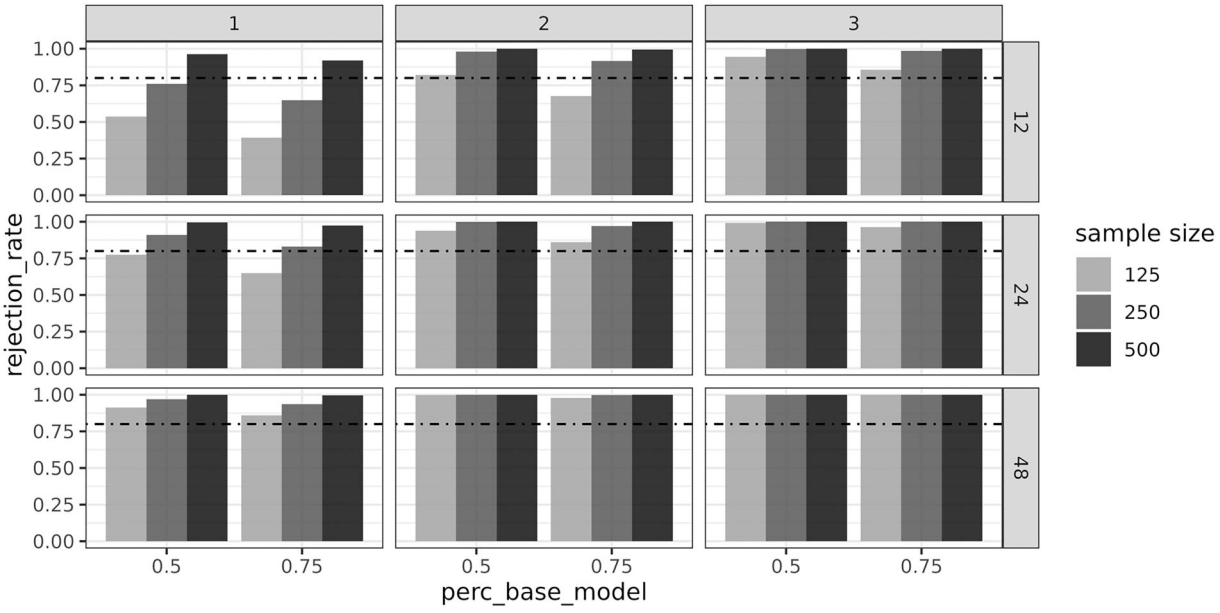

Figure 2. Chi-square test rejection rates for stage one correlation matrix homogeneity with cross-loading misspecifications. *Note*. Bars show the homogeneity rejection rate for the chi-square test of stage one homogeneity with cross-loading misspecifications. Proportion of studies simulated from the baseline model indicated on the *x*-axis. Dashed lines indicate a rejection rate of 80%. Rows indicate number of studies. Columns indicate the number of cross-loadings.

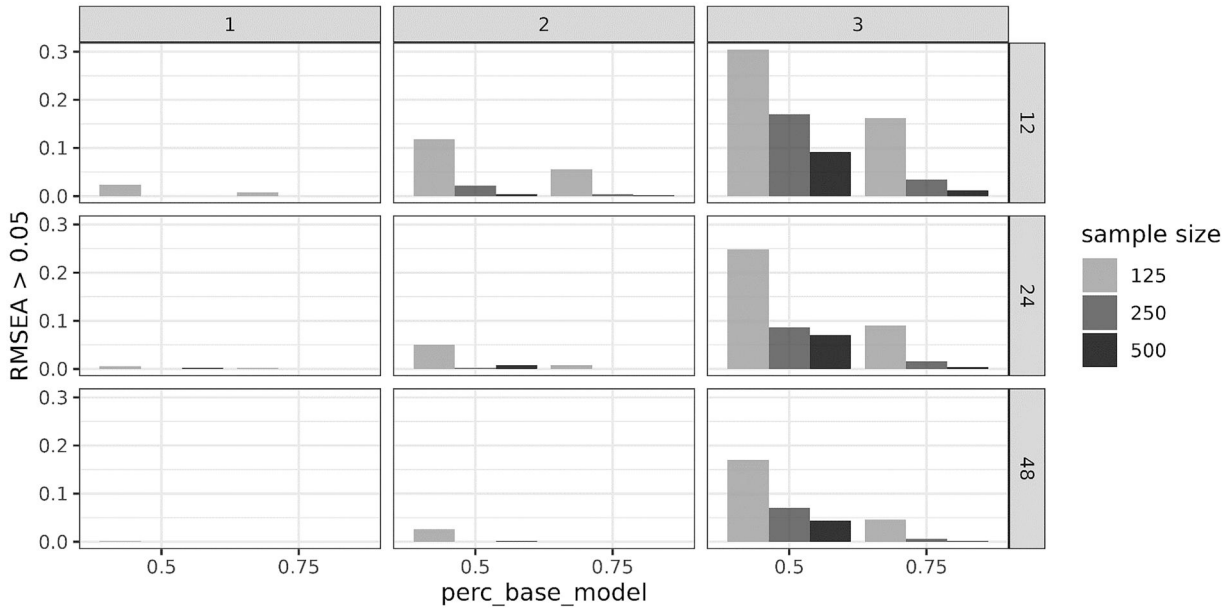

Figure 3. Proportion of RMSEA values >0.05 when assessing stage one homogeneity with cross-loading misspecifications. *Note.* Bars show the proportion of RMSEA values >0.05 (rejection of good fit) when assessing stage one homogeneity. Proportion of studies simulated from the baseline model indicated on the *x*-axis. Rows indicate number of studies. Columns indicate the number of cross-loadings.

CFI. Using CFI values to assess stage one baseline heterogeneity in the baseline condition (no cross-loading misspecification) did not lead to reliable rejection of homogeneity (Figure A3).

Overall, correlation matrix homogeneity was not reliably rejected in the baseline condition (no cross-loading misspecifications). In this situation, fixed-effects pooling would be considered acceptable even with underlying baseline heterogeneity according to commonly used rules of thumb. This shows that in the presence of baseline heterogeneity, additional heterogeneity caused by structural misspecification can alter stage one homogeneity test results in meaningful ways.

Stage Two

Comparison of Fixed-Effects and Random-Effects Pooling

In general, peb values were expectedly higher for indicators with a cross-loading (Figure A4). Within misspecification conditions, the proportion of misspecified models was the biggest contributor to peb increase, while sample size and number of studies were comparatively negligible. For indicators without a cross-loading, no peb values exceeded the 0.1 threshold, indicating acceptable parameter bias across all conditions.

Using random-effects pooling in stage one did lead to decreased parameter bias across all simulation conditions for both factor loadings and factor correlations. The benefits of random-effects pooling increased with higher sample sizes for both factor loading estimates (Figure 6) and particularly factor correlations (Figure 7).

Random-effects pooling did also lead to lower rejection rates for the chi-squared test of model fit, particularly for a low number of studies (Figure 8). Other fit indices (RMSEA, SRMR, CFI) generally indicated good model fit across all conditions for both random- and fixed-effects pooling.

Discussion

In this article, we discuss the performance of stage one correlation matrix homogeneity indices for meta-analytic SEM using the `metaSEM` package in R. We consider cross-loadings as structural misspecifications leading to additional correlation matrix heterogeneity that ideally should be detected by stage one fit indices at least in more severe cases. The simulation illustrates a scenario where homogeneity does not hold in the baseline condition, with further heterogeneity added in the cross-loading conditions for a number of studies. The issue of misspecified models is common when using CFA and SEM techniques and extends beyond psychological research (Bagozzi & Yi, 2012; Baumgartner & Homburg, 1996; Jarvis et al., 2003). Such scenarios can lead to different possible sources of stage two parameter bias:

1. Stage one heterogeneity not being modelled due to using a fixed-effects model
2. Misspecified analysis model used in the stage two analysis

Baseline heterogeneity without additional misspecification did not lead to significant bias of loadings or factor correlations even when using a fixed-effects analysis. Although we agree with the general advise to use a random-effects model when possible, with a correctly specified stage two model (or acceptable misspecification severity), fixed-effects pooling can be a practical alternative if a random-effects model cannot be fitted to the data and heterogeneity is not too severe.

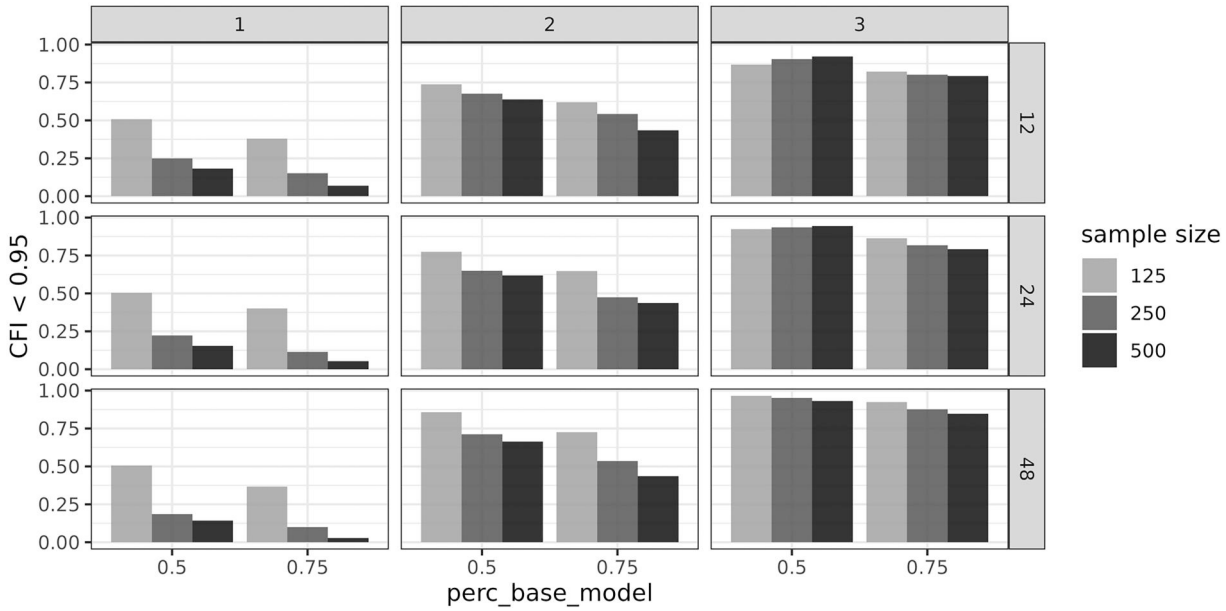

Figure 4. Proportion of CFI values <0.95 (indicating bad fit) when assessing stage one homogeneity with cross-loading misspecifications. *Note*. Bars show the proportion of CFI values <0.95 when assessing stage one homogeneity. Proportion of studies simulated from the baseline model is indicated on the *x*-axis Rows indicate number of studies. Columns indicate the number of cross-loadings.

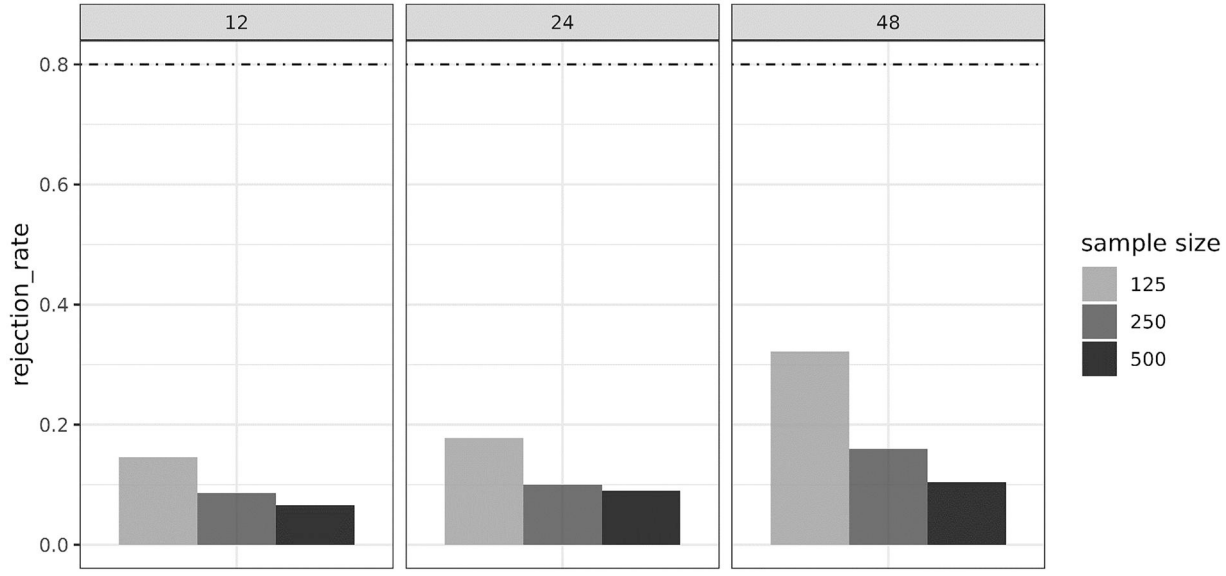

Figure 5. Chi-square test rejection rates for stage one correlation matrix homogeneity without mis-specified models (baseline condition). *Note*. Bars show the rejection rates for the stage one homogeneity chi-square test without misspecifications (baseline). Dashed lines indicate a rejection rate of 80%. Rows indicate number of studies. Columns indicate the number of studies.

The second source of heterogeneity (structurally different population models in stage one) was present in all misspecification conditions due to studies being simulated from two subpopulations with different population models (baseline without cross-loadings vs. misspecified with cross-loadings). Analyzing subgroups is a core concept in meta-analysis (Borenstein & Higgins, 2013) with well-known disadvantages such as lack of statistical power (Cuijpers et al., 2021). In our simulation, correctly assigning studies to their respective subgroups would have resulted in unbiased estimates within subgroups. Without the assignment process being explicitly modeled, parameters were necessarily biased for the pooled stage two data. Nevertheless, stage one random-effects pooling could model at least parts of the additional correlation matrix heterogeneity caused by structural misspecification, leading to reduced stage two factor loading and correlation bias. If sample sizes and the number of studies allow successful fit of a random-effects model in stage one, it seems to be preferable to a fixed-effects model even with only negligible heterogeneity present. This is in line with the general recommendation to use a random-effects model as a default when conducting a meta-analysis and only use fixed-effects models for substantive reasons (Schmidt et al., 2009). Random-effects pooling did not lead to severe convergence issues in our simulation, with non-converged stage one random-effects pooling rate not exceeding 5–6% even in the most severe misspecification conditions (Figure A5).

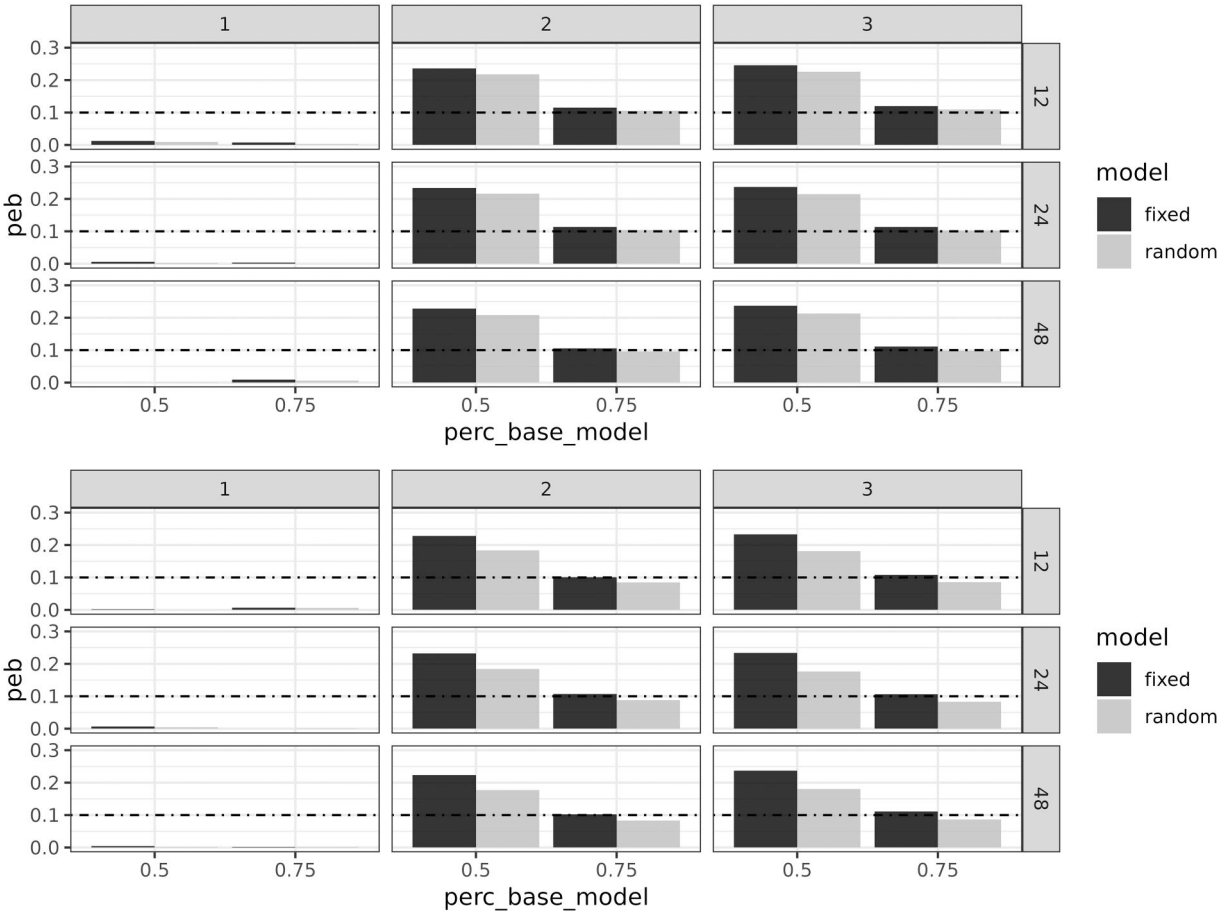

Figure 6. Comparison of stage two parameter bias for for fixed- and random-effects stage one correlation matrix pooling. *Note*. Top: sample size = 125. Bottom: sample size = 500. Bars show the stage two parameter estimation bias for indicator x12 when using fixed- or random-effects pooling in stage one. Proportion of studies simulated from the baseline model indicated on the *x*-axis. Dashed lines indicate a peb of 10%. Rows indicate number of studies. Columns indicate the number of cross-loadings. Indicator x12 has no cross-loading in the left column.

While baseline heterogeneity was not reliably detected by fit indices, our simulation showed that both the chi-squared test of stage one correlation matrix homogeneity and to a lesser degree the stage one CFI values were sensitive to heterogeneity caused by cross-loading misspecifications in a proportion of simulated studies. The sensitivity of these indices to sample sizes can be advantageous, but potentially lead to overly strict rejections of acceptable heterogeneity for larger samples.

The widely followed recommendation of using random-effects pooling whenever feasible seems to be appealing at first glance, and we generally agree. But our study also highlights certain limitations that deserve more attention in discussions about how to approach meta-SEM analyses. Applied researchers should be mindful of the difference between heterogeneity caused by different population models across studies and "baseline" heterogeneity in a random-effects sense. Our simulation shows that random-effects pooling can improve parameter estimates and model fit even if the stage two analysis model is wrongly specified for 50% of all studies. In one sense, this can be desirable if unbiased estimations of factor loadings and correlations are the main research goal. There are also no general rules for "unacceptable" model misspecifications, and structural differences like minor cross-loadings might be ignorable when synthesizing multiple studies.

In stage one, heterogeneity caused by structural misspecification for a proportion of studies did lead to higher rejection rates for correlation matrix homogeneity when using the chi-squared test and the CFI. Since these fit indices and tests cannot distinguish between different sources of heterogeneity (misspecified models vs. baseline heterogeneity), researchers might miss the possibility of structurally different subpopulations causing heterogeneity in stage one, and focus on the more general concept of random-effects heterogeneity instead. Heterogeneity issues can appear "solved" by using random-effects pooling due to stage two improvements even though the more severe factor is structural misspecification.

In stage two analysis, random-effects pooling can lead to overconfidence in a misspecified model when more severe and unknown misspecifications are present for a number of studies. Researchers might not know the subgroup assignment process, explanatory variables are incomplete or unavailable, or there are disagreements in the literature on which model is best suited for the data at hand. For example, in the context of bifactor models, researchers often choose a classical orthogonal bifactor model (Holzinger &

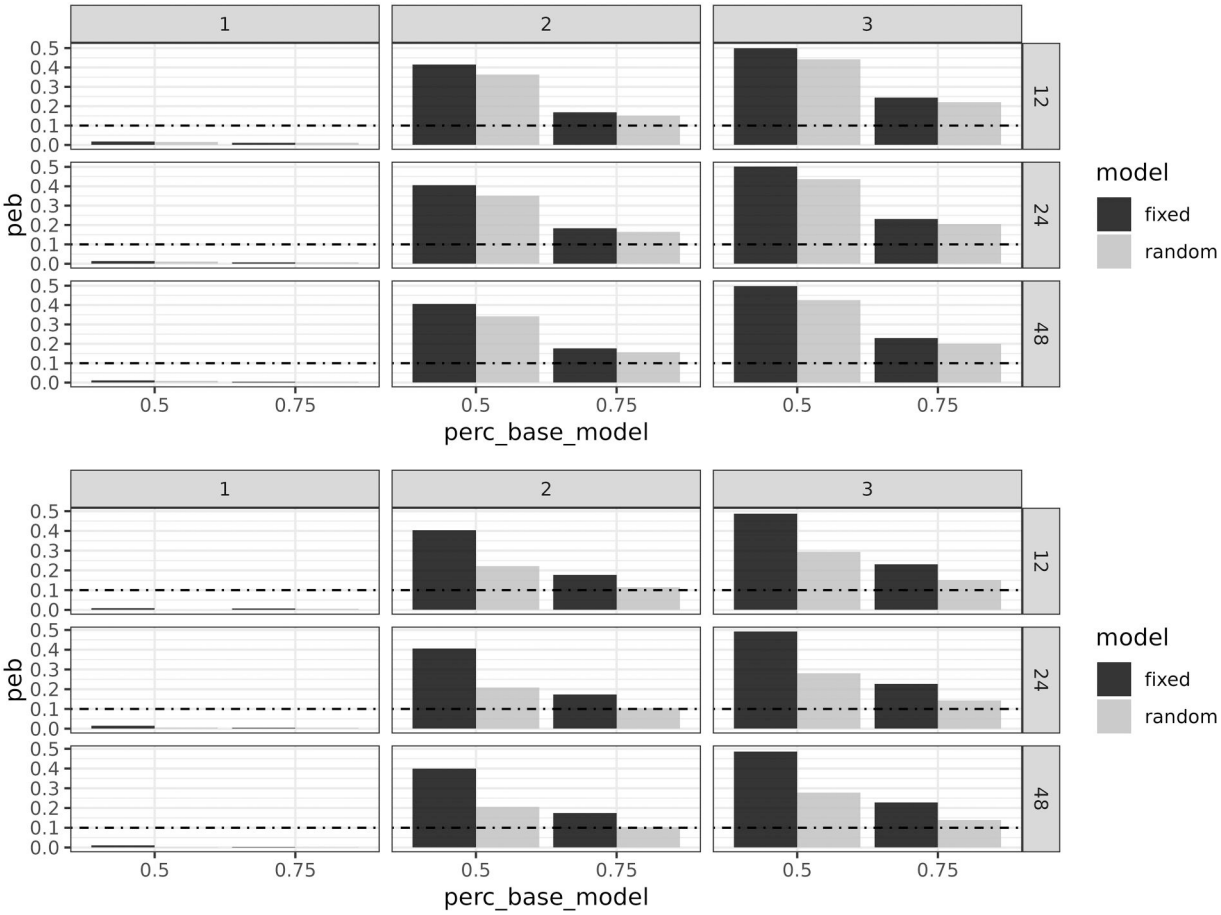

Figure 7. Comparison of stage two factor correlation bias for fixed- and random-effects stage one correlation matrix pooling. *Note*. Top: sample size = 125. Bottom: sample size = 500. Bars show the stage two bias for the correlation between factors 1 and 2. Proportion of studies simulated from the baseline model indicated on the *x*-axis. Dashed lines indicate a peb of 10%. Rows indicate number of studies. Columns indicate the number of cross-loadings. Indicator x12 has no cross-loading in the left column.

Swineford, 1937) by default without acknowledging the existence of different and presumably more appropriate bifactor models (Eid et al., 2017; Koch & Eid, 2024). In these situations, a random-effects model can mask relevant structural misspecifications. Researchers might not consider other candidate models or think about the possibility of structurally different subgroups before the pooling stage because they are satisfied with the stage two performance of their (partly misspecified) pooled stage two model. This issue might be negligible for minor misspecifications. It can be argued that a single small cross-loading does not change a measurement model in a meaningful way. But with a larger number of more severe misspecifications, the meaning of factors and indicators can change substantially.

Beyond the scope of this study, ensemble-based approaches could be used to compare different sets of fit indices, cut-off values and their resulting stage two estimation performance, conditional on decision criteria used in stage one. If stage one fit indices show differential sensitivity to different types of misspecifications, they can as indicators for researchers to decide whether correlation-based pooling is a sensible analysis step.

Future research can also focus on alternative measures of fit, such as dynamic fit indices (McNeish & Wolf, 2023), entropy-based fit measures (Golino et al., 2024), as well as Bayesian alternatives that are potentially avoiding issues related to fitting complex models to smaller samples (Garnier-Villarreal & Jorgensen, 2020). In this context, predictive fit methods and information criteria like the widely applicable information criterion (WAIC) and Leave-one-out-cross-validation (LOO-CV) might be useful. These methods can potentially be used to evaluate individual study correlation matrices before using meta-SEM pooling to detect correlation matrix heterogeneity and facilitate analyses of more homogenous subgroups by filtering out "surprising" individual studies based on their influence on correlation matrix homogeneity.

When potential moderator variables are available, researchers can include them into their two-stage MASEM analysis if they are dichotomous, or instead use a one-step approach if they are continuous (Jak et al., 2021). A pragmatic approach to deal with correlation matrix heterogeneity is to use parameter-based MASEM. Since these approaches first fit candidate models to all available studies, and only estimated effect sizes get pooled in a later stage, model misspecification can be evaluated for individual studies. These approaches come with their own disadvantages, such as not being as robust when data is missing compared to correlation-based approaches (Cheung & Cheung, 2016).

Another promising approach is trying to identify relatively homogenous correlation matrix subgroups with

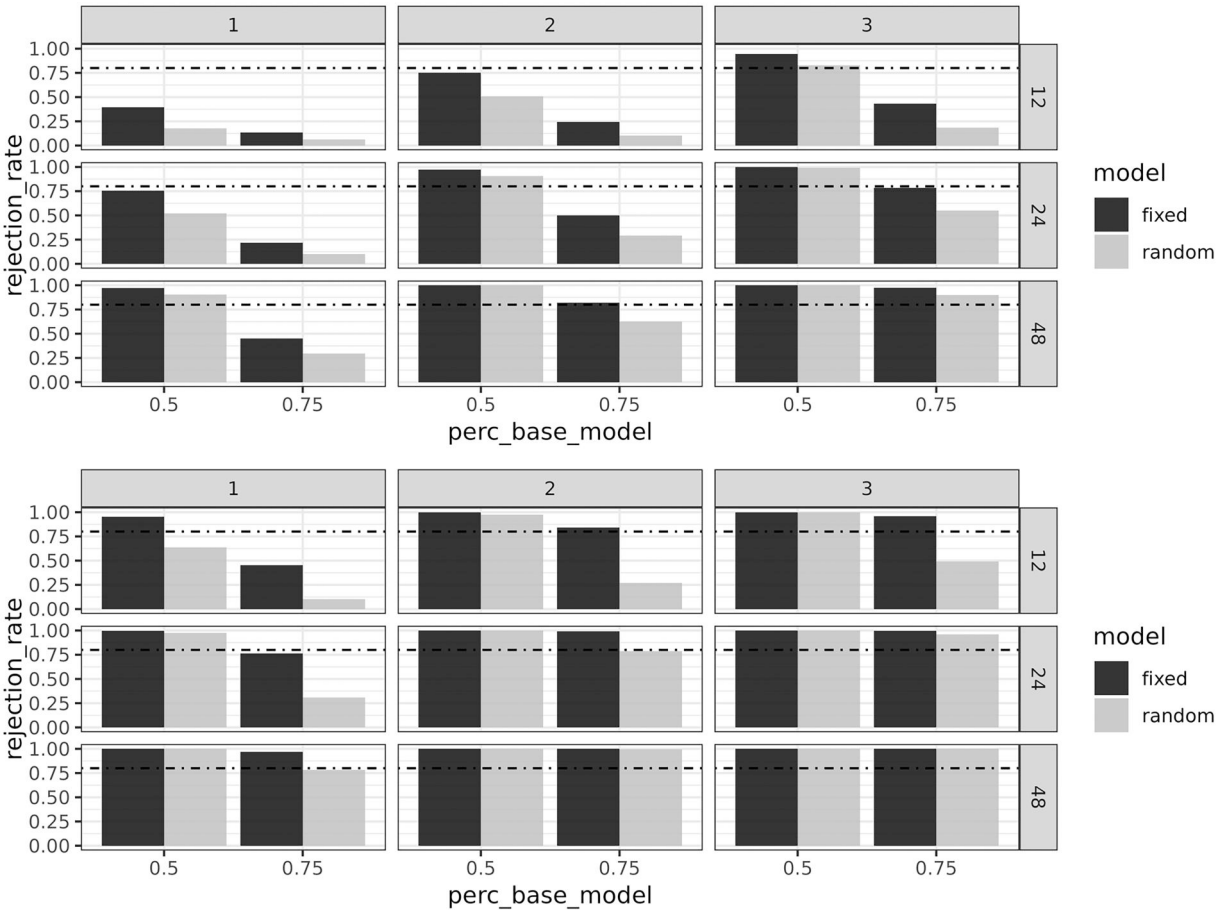

Figure 8. Comparison of the stage two chi-square test rejection rates for sample sizes of 125 (top) and 500 (bottom). *Note*. Bars show the model fit rejection rate for the chi-square test when assessing stage two model fit. Proportion of studies simulated from the baseline model indicated on the *x*-axis. Dashed lines indicate a rejection rate of 80%. Rows indicate number of studies. Columns indicate the number of cross-loadings.

explorative methods before conducting MASEM. Cluster analysis has been shown to be a promising approach (Cheung & Chan, 2005). Recent research mostly focused on practical application (Schroeders et al., 2021; Steinmetz et al., 2020), leaving potential for future theoretical and methodological developments. It can be particularly useful to further investigate exploratory approaches to identify homogenous study subpopulations for scenarios with no available moderator variables or clear theories about data generating population models. Aside from comparing different clustering approaches, future research can also further compare different distance measures to quantify correlation matrix heterogeneity, both in a general sense and related to different specific models commonly used in psychological research.

With the limitations of stage one fit indices in mind, we advise practical researchers to make theory-informed decisions between stage one fixed and random-effects pooling. If the number of studies and sample sizes is large enough to allow reliable model convergence, random-effects pooling can reduce parameter and correlation bias with partially misspecified structural equation models as the main heterogeneity source. An important and sometimes overlooked issue is that stage one fit indices and subsequent stage two pooling can mask indications of structurally different subgroups and lead to overconfidence in a singular stage two analysis model by reducing factor loading and correlation bias. Researchers should keep the issue of model subgroup misspecification in mind when stage one indices reject homogeneity and be careful about equating good stage two fit with their analysis model being adequate for all pooled studies.

Disclosure statement

The authors report there are no competing interests to declare.

References

Bagozzi, R. P., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, *40*, 8–34. https://doi.org/10.1007/s11747-011-0278-x

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational & Psychological Measurement*, *79*, 310–334. https://doi.org/10.1177/0013164418783530

Steinmetz, H., Bosnjak, M., & Isidor, R. (2020). Meta-analytische Strukturgleichungsmodelle: Potenziale und Grenzen illustriert an einem Beispiel aus der Organisationspsychologie. *Psychologische Rundschau*, *71*, 111–118. https://doi.org/10.1026/0033-3042/a000483

Valentine, J. C., Cheung, M. W.-L., Smith, E. J., Alexander, O., Hatton, J. M., Hong, R. Y., Huckaby, L. T., Patton, S. C., Pössel, P., & Seely, H. D. (2022). A primer on meta-analytic structural equation modeling: The case of depression. *Prevention Science: The Official Journal of the Society for Prevention Research*, *23*, 346–365. https://doi.org/10.1007/s11121-021-01298-5

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, *13*, 139–161. https://doi.org/10.1016/0167-8116(95)00038-0

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111. https://doi.org/10.1002/jrsm.12

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science: The Official Journal of the Society for Prevention Research*, *14*, 134–143. https://doi.org/10.1007/s11121-013-0377-7

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258. https://doi.org/10.1177/0049124192021002005

Cheung, M. W. -L. (2015). *Meta-analysis: A structural equation modeling approach* (1. Aufl.). Wiley. https://doi.org/10.1002/9781118957813

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*, 140–155. https://doi.org/10.1002/jrsm.1166

Cheung, M. W.-L. (2014). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, 1521. https://doi.org/10.3389/fpsyg.2014.01521

Cheung, M. W.-L., & Chan, W. (2005). Classifying correlation matrices into relatively homogeneous subgroups: A cluster analytic approach. *Educational & Psychological Measurement*, *65*, 954–979. https://doi.org/10.1177/0013164404273946

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality & Social Psychology*, *66*, 93–114. https://doi.org/10.1037/0022-3514.66.1.93

Cuijpers, P., Griffin, J. W., & Furukawa, T. A. (2021). The lack of statistical power of subgroup analyses in meta-analyses: A cautionary note. *Epidemiology & Psychiatric Sciences*, *30*, e78. https://doi.org/10.1017/S2045796021000664

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*, 541–562. https://doi.org/10.1037/met0000083

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*, 46–70. https://doi.org/10.1037/met0000224

Golino, H., Jiménez, M., Garrido, L. E., & Christensen, A. P. (2024). *Generalized total entropy fit index: A new fit index to compare bifactor and correlated factor structures in SEM and network psychometrics*. Advance online publication. https://doi.org/10.31234/osf.io/5g3hb

Haynes, C. A., Miles, J. N. V., & Clements, K. (2000). A confirmatory factor analysis of two models of sensation seeking. *Personality & Individual Differences*, *29*, 823–839. https://doi.org/10.1016/S0191-8869(99)00235-4

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, *2*, 41–54. https://doi.org/10.1007/BF02287965

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

Jak, S. (2015). *Meta-analytic structural equation modelling* (1st ed.). Springer International Publishing: Imprint: Springer. https://doi.org/10.1007/978-3-319-27174-3

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

Jak, S., & Cheung, M. W.-L. (2018b). Testing moderator hypotheses in meta-analytic structural equation modeling using subgroup analysis. *Behavior Research Methods*, *50*, 1359–1373. https://doi.org/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*, 430–455. https://doi.org/10.1037/met0000245

Jak, S., Li, H., Kolbe, L., De Jonge, H., & Cheung, M. W. -L. (2021). Meta-analytic structural equation modeling made easy: A tutorial and web application for one-stage MASEM. *Research Synthesis Methods*, *12*, 590–606. https://doi.org/10.1002/jrsm.1498

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*, 199–218. https://doi.org/10.1086/376806

Koch, T. (2013). *Multilevel structural equation modeling of multitrait-multimethod-multioccasion data* [Dissertation]. Erziehungswissenschaft und Psychologie, Freie Universität Berlin. http://dx.doi.org/10.17169/refubium-11340

Koch, T., & Eid, M. (2024). Augmented bifactor models and bifactor-(S-1) models are identical. A comment on Zhang, Luo, Zhang, Sun & Zhang (2023). *Structural Equation Modeling: A Multidisciplinary Journal*. Advance online publication. https://doi.org/10.1080/10705511.2024.2339387

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*, 61–88. https://doi.org/10.1037/met0000425

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, *14*, 548–570. https://doi.org/10.1177/1094428110368562

Rice, K., Higgins, J. P. T., & Lumley, T. (2018). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*, 205–227. https://doi.org/10.1111/rssa.12275

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *The British Journal of Mathematical & Statistical Psychology*, *62*, 97–128. https://doi.org/10.1348/000711007X255327

Schroeders, U., Kubera, F. R., & Gnambs, T. (2021). *The structure of the Toronto Alexithymia Scale (TAS-20): A meta-analytic confirmatory factor analysis*. [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/s7eny

Appendix A

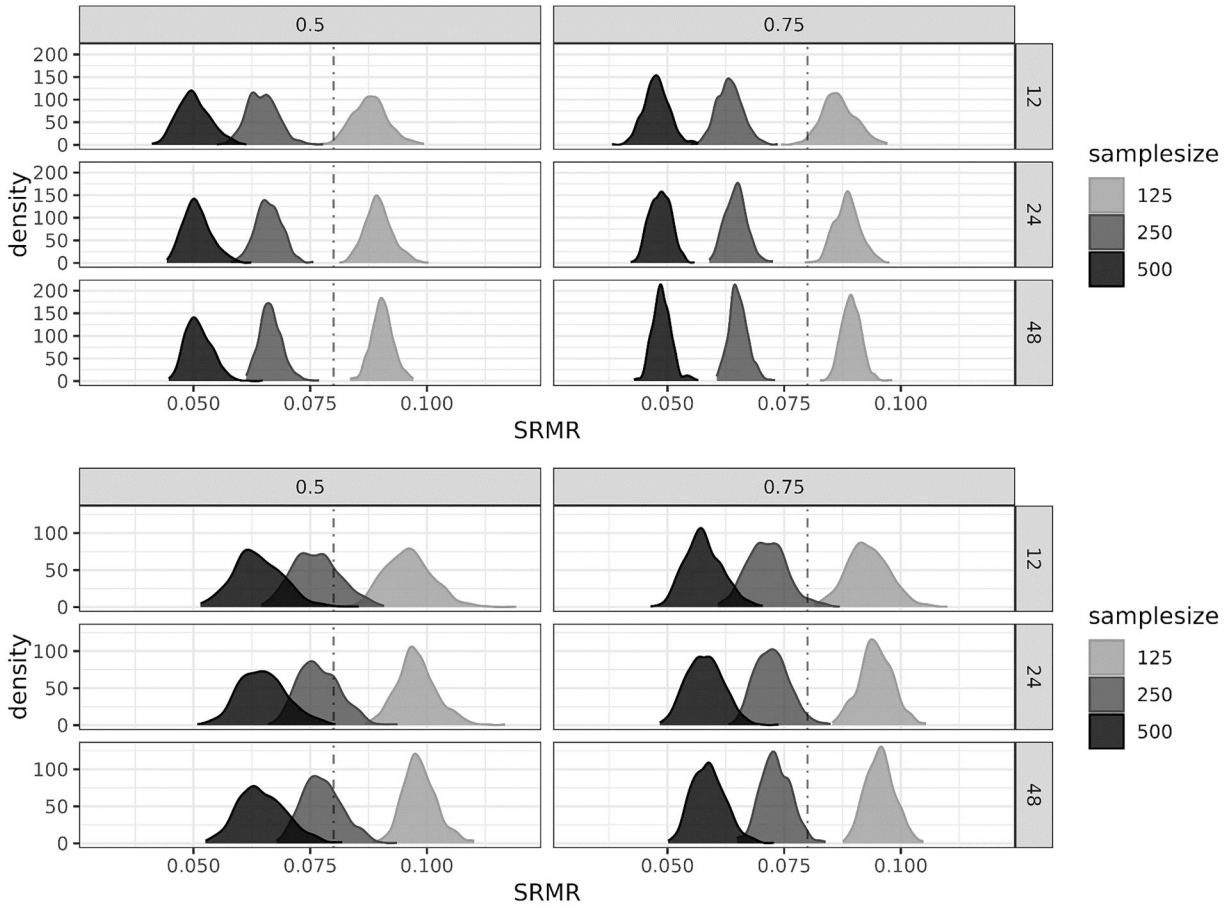

Figure A1. Distribution of SRMR values when assessing stage one homogeneity for one cross-loading (top) and three cross-loadings (bottom). *Note*. Histograms show the distribution of SRMR values when assessing stage one homogeneity. Dashed lines indicate an SRMR cutoff value of 0.08. Rows indicate number of studies. Columns indicate baseline model proportion.

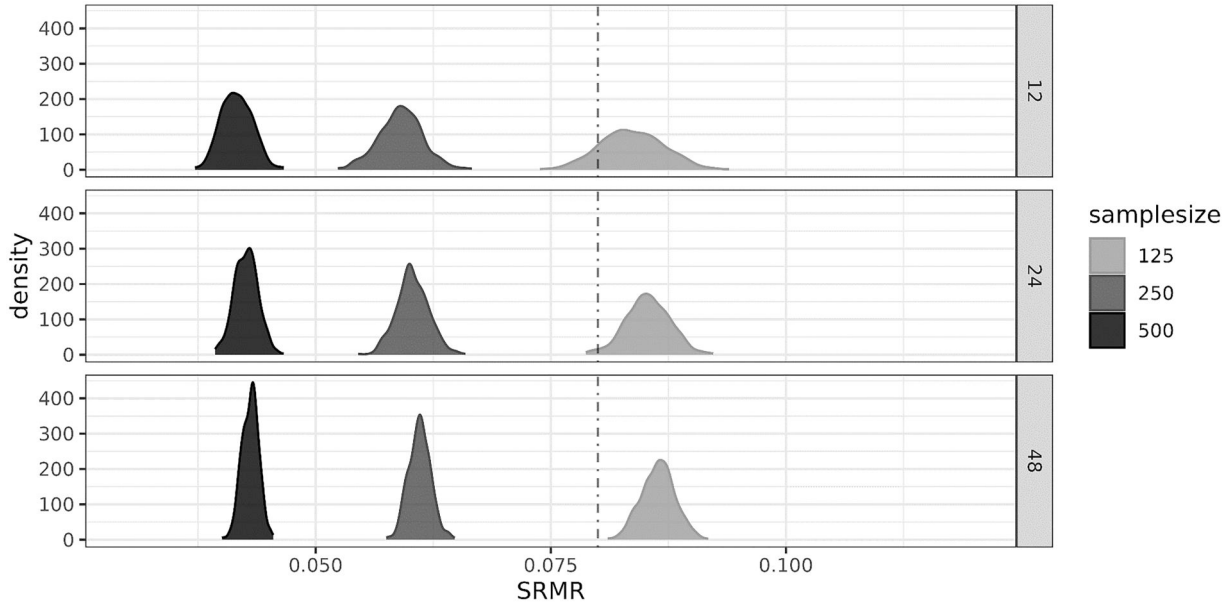

Figure A2. Distribution of SRMR values when assessing stage one homogeneity for the baseline condition (no misspecification). *Note*. Histograms show the distribution of SRMR values for assessing stage one homogeneity without misspecification. Dashed lines indicate an SRMR cutoff value of 0.08. Rows indicate number of studies.

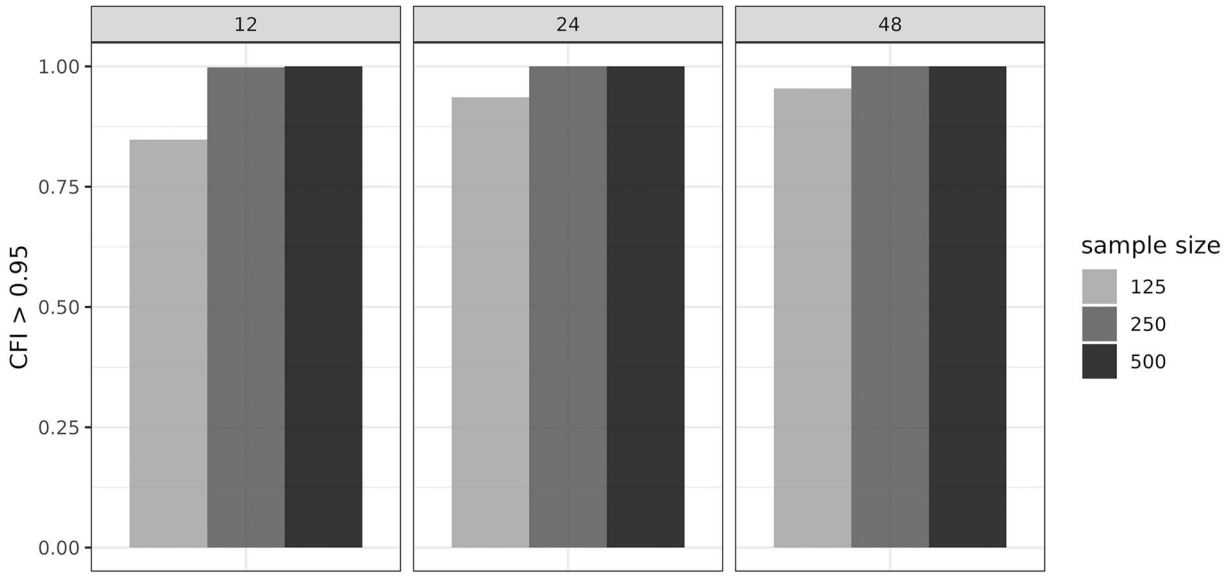

Figure A3. Proportion of CFI values >0.95 when assessing stage one homogeneity for the baseline condition (no misspecification). *Note*. Bars show the proportion of CFI values >0.95 (i.e. good fit) when assessing stage one homogeneity. Columns indicate number of studies.

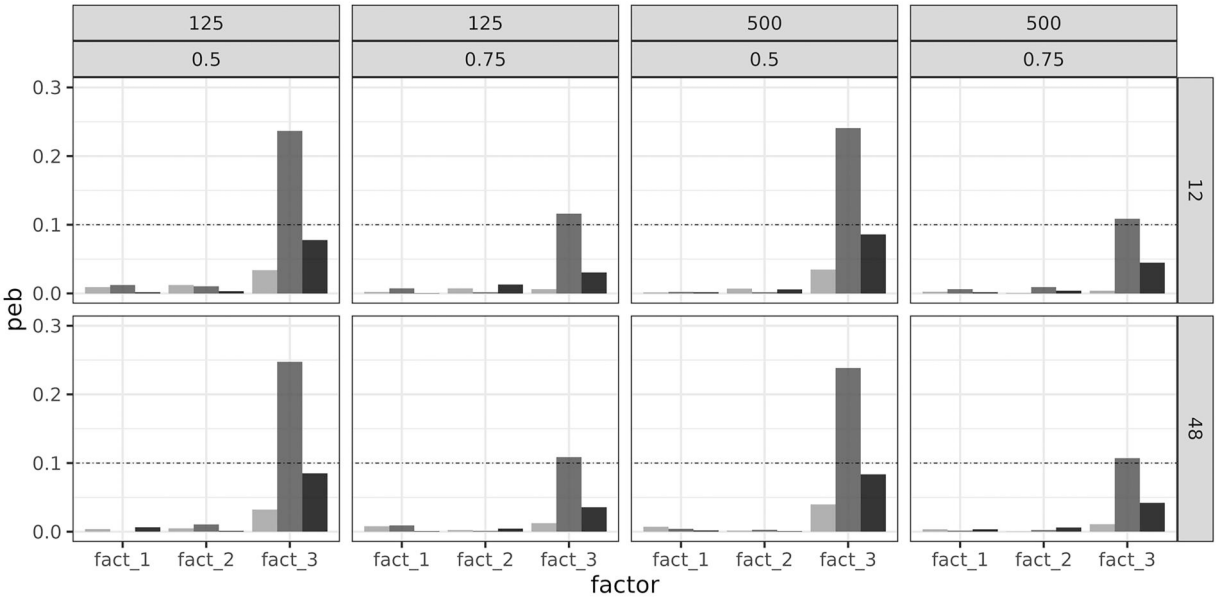

Figure A4. Stage two parameter estimation bias for one cross-loading. *Note*. Bars show the parameter estimation bias for all nine indicators. Factors indicated on the *x*-axis. Rows indicate number of studies. Columns indicate sample size (top) and baseline model proportion (bottom). Dashed lines indicate a bias of 10%.

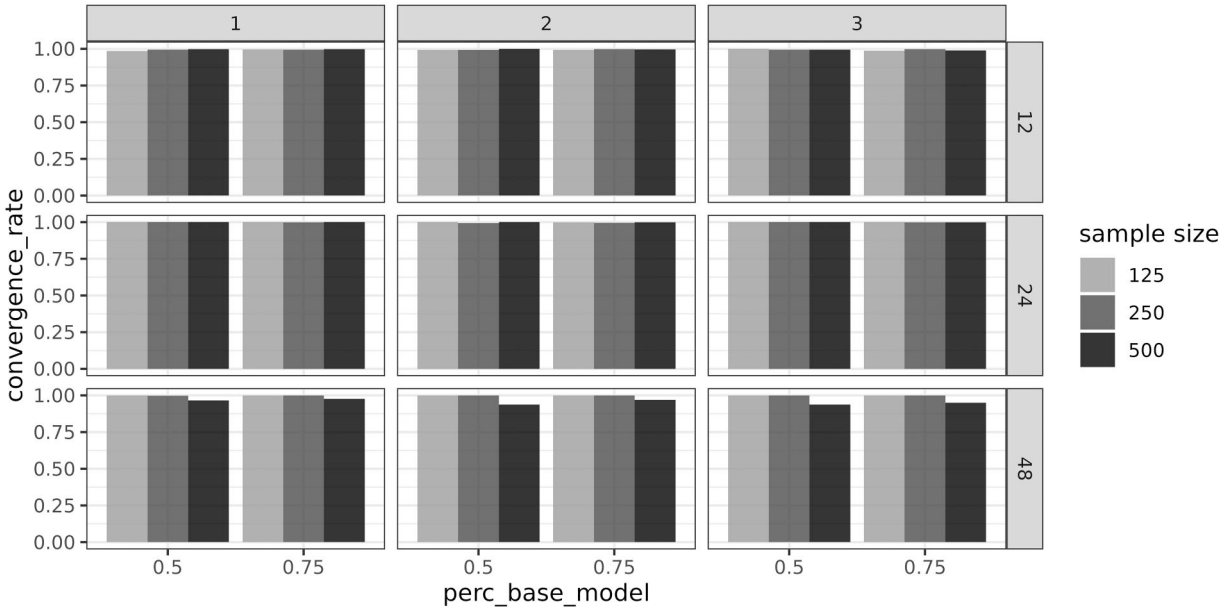

Figure A5. Stage one random-effects pooling convergence rates with cross-loading misspecifications. *Note*. Bars show the stage one random-effects model convergence rate. Proportion of studies simulated from the baseline model indicated on the *x*-axis. Rows indicate number of studies. Columns indicate the number of cross-loadings.