Historical measurement information can be used to improve estimation of structural parameters in structural equation modelling with small samples

James Ohisei Uanhoro[1] and Olushola O. Soyoye[2]

[1]Research, Measurement & Statistics, Department of Educational Psychology, University of North Texas

[2]Educational Statistics and Research Methods, Department of Education & Human Development, University of Delaware

Author Note

The Version of Record of this manuscript, when published, will be available in Educational and Psychological Measurement

https://doi.org/10.1177/00131644251330851

Abstract

This study investigates the incorporation of historical measurement information into structural equation models (SEM) with small samples to enhance the estimation of structural parameters. Given the availability of published factor analysis results with loading estimates and standard errors for popular scales, researchers may use this historical information as informative priors in Bayesian SEM. We focus on estimating the correlation between two constructs using Bayesian SEM after generating data with significant bias in the Pearson correlation of their sum scores due to measurement error. Our findings indicate that incorporating historical information on measurement parameters as priors can improve the accuracy of correlation estimates, mainly when the true correlation is small – a common scenario in psychological research. Priors derived from meta-analytic estimates were especially effective, providing high accuracy and acceptable coverage. However, when the true correlation is large, weakly informative priors on all parameters yield the best results. These results suggest leveraging historical measurement information in Bayesian SEM can enhance structural parameter estimation.

Keywords: Bayesian SEM, priors, small samples, historical information, weakly informative priors, informative priors

Historical measurement information can be used to improve estimation of structural parameters in structural equation modelling with small samples

This study applies Bayesian structural equation modelling (SEM) to small samples, specifically in scenarios where historical information from prior factor analyses is readily available to create informative priors. Scale validation studies frequently report factor analysis results. Researchers can then utilize published factor analysis results as informative priors for the measurement component in their SEMs. Practising researchers are unlikely to do measurement or latent-variable modelling with small samples. When such small samples are analyzed, researchers are usually interested in the structural relations between variables (e.g., correlations between sum scores). Therefore, our primary goal is to determine whether incorporating historical information about the measurement component can enhance the estimation of structural parameters in new data.

We focus on SEM because SEM is one of the most common approaches for modelling with difficult-to-measure variables. SEM rests on the common-factor model, which posits a latent (hidden) variable (scale construct) as the cause of observed indicators (items within the scale). In SEM, researchers then use observed items to identify constructs within a structural model that returns the relations between the constructs. Under the assumption that the common-factor model is correct (Rhemtulla, van Bork, & Borsboom, 2020), the importance of accounting for measurement error using SEM is clear (e.g. Bollen, 1989; Cole & Preacher, 2014; Westfall & Yarkoni, 2016).

A significant challenge in implementing SEM in practice is their large sample requirement. SEM can yield poor results with small samples. The first hurdle is model convergence: the more complex the model, the less likely there is sufficient information in the data to estimate the model (e.g. Boomsma, 1985; Nevitt & Hancock, 2004). The second hurdle is accuracy: parameter estimates, although unbiased, may vary a lot across samples such that results from a given sample are highly misleading.

As an example, Savalei (2019) studied the accuracy (captured by the mean squared

error) of regression coefficients in the context of a mediation dynamic between latent variables (with unidimensional scales) for sample sizes under 200. She used the same data generation process (DGP) as Ledgerwood and Shrout (2011) and Cole and Preacher (2014). The strategies she studied were SEM and sum score path analysis with and without reliability-based correction for attenuation. The alternatives to SEM yielded more accurate estimates for small samples, with the reliability-based corrections being quite promising for inference (assessed via the confidence interval coverage) when reliability was assumed known a priori and thus fixed.

Although simplification strategies as above are viable for certain DGPs, there are contexts where explicit measurement error modelling is necessary, e.g. specific nuisance factors in an essentially unidimensional scale, one rationale for the bifactor model. When measurement error modelling is needed, and the data do not contain sufficient information to yield accurate results, Bayesian methods are known for producing adequate inference. Given the correct choice of prior and likelihood, Bayesian inference is calibrated and does not rely on large-sample approximations. With this in mind, SEM researchers have recommended Bayesian methods for SEM with small samples (e.g. Hox, Schoot, & Matthijsse, 2012; Lee & Song, 2004; Scheines, Hoijtink, & Boomsma, 1999). When applied to small samples, precisely samples with relatively little information given model complexity, it is essential that prior information is used to supplement the information in the data (e.g. McNeish, 2016; Smid, McNeish, Miočević, & van de Schoot, 2020; Smid & Winter, 2020).

As an example, Ulitzsch, Lüdtke, and Robitzsch (2023) studied the same DGP as Savalei (2019) and compared several frequentist methods including the fixed reliability approach recommended by Savalei (2019), and constrained maximum likelihood where the parameter space is restricted such that the resulting SEMs converge (F. Chen, Bollen, Paxton, Curran, & Kirby, 2001), with Bayesian methods using informative priors.[1]

[1] Ulitzsch et al. (2023) referred to their priors as *weakly informative*. However, that description stretched

Ulitzsch et al. (2023) found that constrained ML and the Bayesian methods they studied had conservative or acceptable type I error rates, with the Bayesian methods exhibiting higher accuracy. And, as expected, the more informative the prior (especially when the prior reflected the DGP), the more accurate the parameter estimates for small samples – Smid and Rosseel (2020) reached similar conclusions with a smaller simulation study.

These studies are examples that demonstrate the unsurprising conclusion that informative priors are necessary for Bayesian SEM (BSEM) with small samples to be (i) helpful in mitigating concerns associated with standard SEM and (ii) competitive with alternative approaches for handling small samples. As already mentioned, the results of the published factor analysis of scales serve as a source of prior information for the measurement component of SEMs. Our goal is to examine whether researchers can use this historical information to improve the estimation of structural parameters when measurement error modelling is necessary to obtain accurate estimates of structural parameters.

Before elaborating on the current study, we note that an alternative approach to SEM with small samples is regularization methods, such as penalized maximum likelihood (Jacobucci, Grimm, & McArdle, 2016) and its Bayesian analogues (Jacobucci & Grimm, 2018). From a traditional perspective, regularization methods add a penalty for model complexity, such that analysis of data that are not so informative relative to a maximal model returns parameters that either imply a much simpler model or weakened relations within the maximal model. From a Bayesian perspective, regularization methods invoke prior distributions that attempt to restrict substantive parameters to a prespecified

---

the typical operationalization of the concept. Weakly informative priors attempt to restrict the range of parameters to reasonable values based on relatively limited subject matter expertise (Lemoine, 2019; McElreath, 2020). For example, a weakly informative prior on a standardized correlation would not posit a direction for the coefficient and would be highly sceptical of values above one as such values, though possible, are unlikely. An example is $\mathcal{N}(0, \sigma = 0.5)$, which has approximately 95% of the distribution in the $(-1, 1)$ interval. Ulitzsch et al. (2023) characterized priors on correlations with the true correlation as the location parameter of the distribution as 'weakly informative'. Prior knowledge of the true relationship between variables would likely come from substantial subject matter expertise. In particular, Smid and Rosseel (2020) examined similar priors in their study and termed them 'informative' priors.

location; this location usually but not necessarily implies a null result. The prior scale may be set manually (e.g. B. O. Muthén & Asparouhov, 2012) or is learned from the data (e.g. J. Chen, Guo, Zhang, & Pan, 2021). Given the suitability of these methods for working with small samples, we will also incorporate regularized Bayesian SEM into the current study.

Also relevant to the current study is the literature on constructing prior distributions from historical data. One approach, *Bayesian synthesis* or augmented data-dependent priors (Marcoulides, 2017), is classic Bayesian updating, whereby the posterior distribution from a previous study serves as the prior for the subsequent analysis in the sequence, and so on.

Another class of methods known as power priors (Ibrahim & Chen, 2000; Ibrahim, Chen, Gwon, & Chen, 2015) relies on historical data, assuming that historical and present data do not arise from the same population. These methods do this by weighting the historical results. The weight ranges from 0: without reliance on historical information to 1: Bayesian updating of historical results. The weight can either be determined a priori or given its hyper-prior. Finch (2024) studied power priors in the context of structural parameter estimation within Bayesian SEMs and found that their application improved parameter accuracy relative to other methods for incorporating historical information.

A related idea is commensurate priors (Hobbs, Carlin, Mandrekar, & Sargent, 2011). For any parameter, the point estimate and variance from a previous study serve as the expected value and variance of the prior distribution. However, the inverse of a commensurability parameter additionally inflates the variance. A low value of commensurability implies significant differences between current and historical data.

Another approach is Bayesian dynamic borrowing (Kaplan, Chen, Yavuz, & Lyu, 2023; Viele et al., 2014) where the choice to consider historical data information depends on the similarity of the current and historical data. A variant of Bayesian dynamic borrowing is hierarchical modelling, where the modeller treats parameters of interest as

draws from a distribution, and parameters from historical studies constitute the other draws from this distribution.

We now describe the current study, highlighting aspects that distinguish it from the published literature on this issue.

The current study

The current study contains two simulation studies. We take a different approach to data generation from previous research on using SEM with small samples. Below, we highlight these differences.

First, we focus on the situation with substantial prior information: we study the scenario where a researcher is interested in estimating the relation between two constructs; both constructs have well-established scales with published factor analysis results. The factor analysis results provide prior information for the measurement component of the researcher's SEM. The simulation study by Finch (2024) is similar to ours in that he compared methods of incorporating historical data in Bayesian SEM via simulation. However, our study is different in some ways: (i) we focus on small samples; (ii) we assume no historical information for the parameter of interest. Another study by Finch and Miller (2019) is similar to ours in its focus on estimating structural parameters in small-sample Bayesian MIMIC models. Specifically, our study is close to the uninformative prior on regression coefficients + informative prior on loading conditions within their study. One limitation of their study was the choice of bias as a primary outcome; when studying small samples, we believe accuracy to be more important than bias (Davis-Stober, Dana, & Rouder, 2018). We also explore our question in significantly more depth than they do – our study 2 explores moderation of choice between methods, such that our analysis provides more room for nuance.

Second, we focus on the situation where measurement error modelling is necessary. Reducing latent variable models to observed variable models can be a viable small sample strategy for regression with truly unidimensional scales (e.g. Rosseel, 2020; Savalei, 2019;

Smid & Rosseel, 2020; Ulitzsch et al., 2023). Instead, we assume a bifactor structure for both constructs with method effects that similarly affect the constructs. Methods that ignore measurement error modelling will have significant bias, so they are unlikely to be accurate even with small samples.

A couple of additional features are unique to the current study.

Highly realistic DGPs. We increase the realism of the DGPs in the study; the researcher's population never matches previously studied populations. We do this by randomly varying both structured and unstructured parameters across samples (e.g. Uanhoro, 2024; Wu & Browne, 2015), whether historical or current. The effect of this choice is that historical results analyzed using standard SEM methods yield inferences that are only accurate for the specific population of the historical study.

Simulation design and analysis. In Study 2, we identify several potential design factors that can affect the relative performance of different methods. We assume distributions for each of these factors and sample from these distributions within each iteration (e.g. Walters, Hoffman, & Templin, 2018). This choice allows us to study how these factors can affect the relative performance of different methods. We then use a combination of importance sampling (Owen, 2013), generalized additive models (Wood, 2017) and trees (Hothorn, Hornik, & Zeileis, 2006) to compare the relative performance of the methods, including moderation of their relative performance by design factors.

In the next section, we report on simulation study 1, where we test several approaches to incorporating historical measurement information. Afterwards, we report on simulation study 2, in which we test a subset of methods from study 1, compare them against each other, and assess moderation by design factors. Study 1 helps reduce a variety of potential approaches to a subset of approaches. Study 2 provides guidance on when to use specific methods. Then, we end with discussion points.

All code for simulation studies, data analyses and Stan scripts are available at `https://osf.io/rk45m/`. The OSF repository also contains a data demonstration file with

an applied example.

Simulation study 1

Data generation

We assumed the following data generation process for an 18×18 population covariance matrix for 18 items in study i, $\boldsymbol{\Sigma}_i$:

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda_i \Phi \Lambda_i}^{\mathsf{T}} + \mathbf{D}_i \mathbf{R}_i \mathbf{D}_i, \ \boldsymbol{\Phi} = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & .5 \\ 0 & 0 & .5 & 1 \end{pmatrix} \tag{1}$$

The loadings randomly varied between the samples ($\boldsymbol{\Lambda}_i$), and the error standard deviations ($\text{diag}(\mathbf{D}_i)$) were calculated such that $\text{diag}(\boldsymbol{\Sigma}_i) = \mathbf{1}_{18}$. Furthermore, the 18 items had residual correlations ($\mathbf{R}_i$) according to a random process. The interfactor correlation matrix ($\boldsymbol{\Phi}$) shows four factors underlying the 18 items. The first two factors are substantive factors correlated at ρ, and the other factors are substantively similar method-effect factors correlated at 0.5.

We first describe the random process for $\boldsymbol{\Lambda}_i$:

$$\begin{aligned}
&\boldsymbol{\Lambda}_i = [\, \boldsymbol{\lambda}_{1i} \ \boldsymbol{\lambda}_{2i} \ \boldsymbol{\lambda}_{3i} \ \boldsymbol{\lambda}_{4i} \,], \\
&\boldsymbol{\lambda}_{1i} = \Big[h(1), \ldots, h(9), \underbrace{0, \ldots, 0}_{9 \text{ times}}\Big]', \ \boldsymbol{\lambda}_{2i} = \Big[\underbrace{0, \ldots, 0}_{9 \text{ times}}, h(1), \ldots, h(9)\Big]', \\
&\boldsymbol{\lambda}_{3i} = \Big[\underbrace{0, \ldots, 0}_{5 \text{ times}}, \mathbf{u}_{1i}, \underbrace{0, \ldots, 0}_{9 \text{ times}}\Big]', \ \boldsymbol{\lambda}_{4i} = \Big[\underbrace{0, \ldots, 0}_{14 \text{ times}}, \mathbf{u}_{2i}\Big]',
\end{aligned}$$

where $h(x)$ is a function with three steps:

$$\begin{aligned}
&\text{Step 1: } t = .7 - (x-1)(.7-.4)/8 \\
&\text{Step 2: } r \sim \mathcal{N}(t, \sigma_\lambda) \\
&\text{Step 3: } y = \begin{cases} .2 & \text{if } r < .2 \\ .9 & \text{if } r > .9 \\ r & \text{otherwise} \end{cases} \\
&\text{Step 4: Return } y
\end{aligned} \tag{2}$$

and $\mathbf{u}_{\cdot i} = \{u_{ji}\}_{j=1}^{4}, \ u_{ji} \sim \text{Uniform}(.3, .4)$

The first two factors are reflected in the first nine and second nine items, respectively. In practice, the first nine items represent one scale, and the second nine items represent a second scale. The final four items in each scale reflect the last two factors. The items within each scale are ordered by decreasing relation to their substantive factors, with equidistant loadings on average ranging from 0.7 to 0.4. The specific loading for an item on its substantive factor in study i is a random draw from a normal distribution with a location depending on its position within its scale and a scale depending on σ_λ, winsorized at .2 and .9. Hence, if $\sigma_\lambda = 0$, the same loadings are identical in all studies. In contrast, large values of σ_λ mean that the same loadings vary more in all studies. The loadings of the last four items on each scale on their method effect factors fall in the (.3,.4) interval; these loadings are small but large enough to lead to poor estimation of ρ when the modeller ignores both method effect factors. Given the average loading values, the average reliability coefficient ω for either scale was 0.762.

$\mathbf{R}_i$ is random-misspecification error (Uanhoro, 2024; Wu & Browne, 2015) and reflects the specificity of the measurement instance that causes hypothesized models to be misspecified in practice. Another way to view $\mathbf{R}_i$ is as a random residual network structure (Epskamp, Rhemtulla, & Borsboom, 2017). When a measurement instance produces too different patterns from the hypothesized structure, we reject the hypothesized structure for the population under study. In this study, residual correlations were LKJ-distributed (Lewandowski, Kurowicka, & Joe, 2009): $\mathbf{R}_i \sim \text{LKJ}(\eta_i)$. The marginal distribution for error correlations is a symmetric beta distribution rescaled to the (-1, 1) interval. Varying η_i allows us to vary the scale of the correlations. We set η_i such that the standard deviation of standardized residual covariances (τ, akin to the CRMR) was .04. Hence, models with the correct theoretical structure would still show misspecification; however, a CRMR of .04 would be considered acceptable by most evaluators.

Given the data generation process above, we varied the following factors in the simulation study:

1. $n \in \{30, 70, 150\}$ (3 levels): sample size of the current study.

2. whether or not ρ was 0 (2 levels): For nonzero ρ, we set ρ to -.5, -.3, and -.2 for $n = \{30, 70, 150\}$, respectively. The values are close to the minimum detectable correlations under a one-tailed power analysis (power = 80%) without accounting for measurement error.

3. $\sigma_\lambda \in \{.05, .1\}$ (2 levels): the degree of difference in the loadings between samples. We examined measurement invariance analysis at both levels of σ_λ. At $\sigma_\lambda = .05$, measurement invariance analysis suggests medium levels of non-invariance, $\text{RMSEA}_D \approx .06$ when assessing two groups on either scale for metric invariance.[2] At $\sigma_\lambda = .10$, measurement invariance analysis suggests relatively high non-invariance, $\text{RMSEA}_D \approx .10$ when assessing two groups on either scale for metric invariance. See Figure A1 in the appendix for the RMSEA_D distribution.

Hence, this study had 12 design conditions ($3 \times 2 \times 2$). To represent the results of the factor analysis that existed before the current study, we generated 12 historical datasets. Note that the substantive correlation, ρ, was identical across all population covariance matrices. Within any simulation iteration, these historical datasets had different population covariance matrices as described in Equations 1 and 2. The datasets were:

- 2 large-sample studies ($n = 2000$): The first study reported factor analysis results (parameter estimates and standard errors) for scale one, and the second reported factor analysis results for scale 2. Both studies accounted for the substantive and method effect factors and provided precise priors for measurement parameters different from the true measurement parameters underlying the current study since $\sigma_\lambda > 0$.

[2] RMSEA_D is the RMSEA difference for nested model comparisons (Savalei, Brace, & Fouladi, 2023).

- 2 very large-sample studies ($n = 10000$): Both studies were similar to those above but with much more participants. Compared to the first two studies, the parameters in these studies are much more precise.

- 8 medium-sample studies ($350 \leq n \leq 480$ about 95% of the time): The first four studies were about the first scale, and the second four were about the second scale. The median sample size of these studies was 400. We generated the exact sample sizes from a shifted log-normal distribution: $n = 300 + \lfloor s \rfloor,\ s \sim \text{log-normal}(\ln(100), .3)$.

Note that no historical study includes both scales, so there is no prior information on the relationship between both constructs. We considered the historical studies to exist in three separate universes:

1. *2k*: The researcher has two results from large-sample factor analysis ($n = 2,000$), one per scale.

2. *10k*: The researcher has two results from very large-sample factor analysis ($n = 10,000$), one per scale.

3. *Meta*: The researcher has eight results from typical sample factor analysis (median $n = 400$), four per scale.

Data analysis

For the analysis of historical data, we fit accurate bifactor models separately by scale since the data for each scale came from a different study. We set maximum likelihood (ML) as the estimator, matching the default in commonplace SEM software.

For the present small sample data analysis, we assumed a three-factor structure for all measurement error models by combining both method effects into a single factor loaded on items 5–9 and 15–18. We considered several analysis strategies, classifying them as considering or ignoring historical information.

Before reviewing these strategies, all methods estimated the structural relation of interest as the correlation between the two substantive factors. Additionally, all Bayesian analyses assumed an LKJ prior ($\eta = 1$) on the interfactor correlation matrix. Given that the interfactor correlation matrix has three variables, this translates into a weakly informative marginal beta(1.5, 1.5) prior rescaled to the (-1, 1) interval on the correlation of interest. Finally, whenever we reference weakly informative priors on other parameters, we mean standard normal priors on loadings and Student $t^+(3, 0, 1)$ priors on residual standard deviations.

Analytical strategies that considered historical information

The strategies that considered historical information were:

1. Bayesian SEM with **informative priors** on loadings based on historical data and weakly informative priors on other parameters (3 models): we had three sets of informative priors corresponding to the three universes above: 2k, 10k, and Meta. For the 2k and 10k universes, we used the estimates and standard errors from historical studies as the mean and scale for the normal distribution priors for loadings. For the *meta* universe, we performed a random-effect meta-analysis (Hedges & Olkin, 1985) for each element of the loading matrix using the restricted ML estimator (Viechtbauer, 2005). The input information was the estimate and standard error for each loading across the studies that converged out of the four studies – there were sometimes non-convergent solutions. The output information we used as the prior was the meta-analytic average and standard error for each loading, so there were 26 (2×9 substantive loading + 2×4 method-effect loadings) meta-analytic averages and 26 standard errors. The 2k and 10k informative priors are incorrect (since $\sigma_\lambda > 0$) and may be misleadingly precise. The priors based on a meta-analysis of four typical-size studies should be less misleading, but four studies are too few to fully capture the variation inherent to the DGP.

2. Bayesian SEM with **commensurate priors** on loadings and weakly informative priors on other parameters (3 models): This strategy is very similar to the previous strategy and is similarly an informative prior, but with one significant difference: it includes a commensurability parameter ($1/\tau$) to acknowledge differences between historical studies and current data, such that the degree of borrowing is dynamic. For some generic loading, λ_1, we assumed: $\lambda_1 \sim \mathcal{N}\left(\lambda_1^{(j)}, \sqrt{\text{var}(\lambda_1^{(j)}) + \tau^2}\right)$, where j corresponds to the study index among historical studies, such that the prior for each loading in the *meta* universe had four factors – one from each study. We assumed that the commensurability parameter was invariant across loadings, assuming: $\tau \sim \mathcal{N}^{+}(0, 0.1)$.

3. SEM with *ML* estimation with **loadings fixed** to values from previous studies (3 models): Hence, there was a fixed set of loadings based on the 2k studies, another based on the 10k studies, and another based on the 26 meta-analytic averages from the *meta* universe. This approach dramatically reduces model complexity but may yield misleading inference for two reasons: (i) it does not propagate the uncertainty from the historical estimates, and (ii) it assumes historical estimates are correct for the current data – this is incorrect since $\sigma_\lambda > 0$.

Analytical strategies that ignored historical information

The strategies that ignored historical information were:

1. SEM with **ML** estimation (1 model): This approach will be inefficient given the model complexity and small sample sizes.

2. Bayesian SEM with **weakly informative priors** (1 model): Bayesian SEM assuming weakly informative priors on all parameters. This strategy increases the chance that the measurement parameters behave well. However, this strategy may not provide sufficient information to accurately estimate the measurement component parameters, potentially impacting the structural parameter of interest.

3. Bayesian SEM with **regularized estimation of nonzero loadings** and weakly informative priors on other parameters (1 model): we assumed the distribution of all nonzero loadings within a factor was normal with mean zero and a scale that varies by factor, that is, $\lambda_{kl} \sim \mathcal{N}(0, s_l)$, for item k on factor l where s_l is itself learned from the data: $s_l \sim \mathcal{N}^+(0, 1)$. This specification of the normal prior is akin to ridge regularization of loadings (Hsiang, 1975). Regularization of loadings may stabilize the measurement component of the model, potentially improving the accuracy of the structural parameter.

4. **Linear regression** (1 model): Sum the scores from the current study by scale, standardize the sum scores, and compute the regression coefficient using the usual t test for inference. This method should be highly biased given the correlated method effects in the DGP.

In total, we ran 13 different models for the current data: three Bayesian SEMs with informative priors (Inf-2k/10k/Meta); three Bayesian SEMs with commensurate priors (Com-2k/10k/Meta); three ML SEMs with fixed loadings (Fix-2k/10k/Meta)l; one ML SEM (ML-None); one Bayesian SEM with weakly informative priors (BSEM-Weak); one Bayesian SEM with regularized estimation of loadings (BSEM-Reg); and one linear regression (LR), $3 + 3 + 3 + 1 + 1 + 1 + 1 = 13$.

Bayesian strategies we do not use

There are additional Bayesian methods for accounting for historical information that we have not considered. We did not include them for the following reasons: inadequate assumptions, complexity of implementation, or similarity to already included strategies.

Power priors (Ibrahim et al., 2015). Estimating the weight in the power prior can be challenging when the weight is not known a priori. More recent methods are either time-consuming (e.g. Duan, Ye, & Smith, 2006) or require both the original data (which we assume to be unavailable in practice) and the current data to assess the degree of similarity

between both datasets (e.g. Pan, Yuan, & Xia, 2017; Shi, Li, & Liu, 2023). For this reason, we do not include the standard power prior in this study. However, when historical information is a normal distribution with a known mean and variance, the power prior corresponds to scaling the variance by the inverse of the prior weight (Ibrahim & Chen, 2000). This feature of the power prior connects it to the commensurate prior, as applied in this study. Since our prior distributions are assumed to be normal with known mean and variance, we consider the commensurate prior equivalent to a power prior with the weight learned from the data. Estimating the commensurability parameter can also be difficult, but it can be placed on the same scale as the parameter of interest, slightly easing the specification of its hyper-prior.

Bayesian synthesis (Marcoulides, 2017). Our current study captures historical information as factor loading estimates and standard errors, which are assumed to be normal distribution parameters. Under these conditions, the resulting prior from Bayesian synthesis would be equivalent to the prior obtained via a fixed-effects meta-analysis if the initial prior in the Bayesian synthesis is uninformative. Our study focuses on the situation where the current and historical data do not arise from the same population, so we do not consider Bayesian synthesis in this study.

Test-then-pool (Viele et al., 2014). In this approach, we test the similarity of the current data to the historical samples. Historical samples similar to the current data would then be considered prior information for the current analysis. Our current scenario would require fitting the CFA separately for each scale using current data. Next, we test the equality of the loadings between the current data and each historical set of loadings. One may conduct this equality test via pairwise differences in loadings under the assumption of known or estimated variances, followed by a test of whether the average gap is 0. For historical datasets where we fail to reject the null, we may use their historical information as priors for the current data analysis. For historical datasets where we reject the null, we do not use their information as priors for the current data. In our study, we

generated the data to be different between samples, such that *test-then-pool*, when adequately powered (due to large-sample historical data), would often discard historical information. The approach is also laborious and would require model convergence for the current data analysis before the eventual Bayesian model. For these reasons, we do not consider this strategy.

Evaluation metrics

We wanted to estimate the correlation between the two substantive factors ρ. We monitored two primary metrics: root mean square error (RMSE) to assess the accuracy of parameter estimates (mean of posterior samples for Bayesian models) and coverage of the 90% confidence (or credible) interval (CI) to assess inference.

To better understand the differences in accuracy between methods, we also tracked the bias for $\rho = 0$, the relative bias for $\rho \neq 0$ and the standard deviation of the parameter estimates. For example, an unbiased estimator may have such high variance that it is overall less accurate (higher RMSE). In this situation, the preference is for the more accurate estimator; we only examine bias and standard deviation to understand differences in accuracy.

To better understand the differences in inference quality, we also evaluated the bias-adjusted coverage of the 90% CI. When coverage is inadequate but bias-adjusted coverage is adequate, the inadequate coverage is due to bias rather than under- or overestimating parameter uncertainty.

Within each design condition, we divided the RMSE of each analytic approach by the minimum RMSE within the condition and then subtracted one from the result. The lowest RMSE method would have a value of 0; we considered methods with values below 5% as practically comparable to the best method and under 10% competitive. We considered coverage in the (87.5%, 92.5%) interval ideal and coverage in the (85%, 95%) interval adequate. We view under-coverage more negatively than over-coverage, especially

when $\rho = 0$ since under-coverage implies increased type I error.[3] For relative bias, we considered values in ±5% ideal and values in ±10% adequate.

We also assessed power when $\rho \neq 0$ with preference for methods that maintained adequate coverage and had high power.

Simulation conditions and convergence assessment

We repeated each condition with 2000 replications. We estimated ML SEM with the lavaan package in R (Rosseel, 2012). We estimated the Bayesian models with the minorbsem package in R (based on Stan, Uanhoro, 2023), except for the Bayesian SEMs with commensurate priors and regularized estimation for which we wrote custom Stan code (Carpenter et al., 2017). For Bayesian modelling, we used three chains, requested 2000 posterior samples per chain, and retained the final 1000 samples for inference. For ML SEM models (both prior and present data), we tracked model convergence assessment using `lavaan::lavInspect(model, "converged")`. For Bayesian models, we tracked the convergence of the coefficient of interest, ρ, using the improved statistic $\widehat{R}$ (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2020), preferring $\widehat{R} \leq 1.05$.

Results

Convergence assessment. For the ML models, the convergence problems were rare, occurring for typical sample studies within the *meta* universe about 1.5% of the time, and the current data analysis less than 1% of the time. For the Bayesian models, $\widehat{R}$ for ρ was greater than 1.05 less than 0.1% of the time, except for the models with commensurate priors where $\widehat{R}$ for ρ exceeded 1.05 about 1.4% of the time.[4] In summary, convergence was rarely a problem in the current study.

---

[3] One could argue that for small samples, type-II error is more severe than type-I error since small sample results are likely to be considered exploratory by most evaluators, and when exploring, it is reasonable to entertain false positives. Ultimately, the context should determine which error is more concerning.

[4] The convergence problem for the models with commensurate priors was concentrated in the 2k and 10k models when fit to data with a sample size of 30 and $\rho = -.5$. For this subset of conditions, $\widehat{R} \geq 1.05$ occurred about 4–5% of the time.

Results for $\rho = 0$

Strategies that relied on historical information were consistently more accurate than comparable strategies that did not; see the top panel of Figure 1. The most accurate model was linear regression with models based on informative or commensurate priors a close second. ML with fixed-loadings (ML-Fix) provided no advantage over standard ML (ML-None). These results suggest that researchers can use historical measurement information within Bayesian SEM to improve structural parameter estimation.

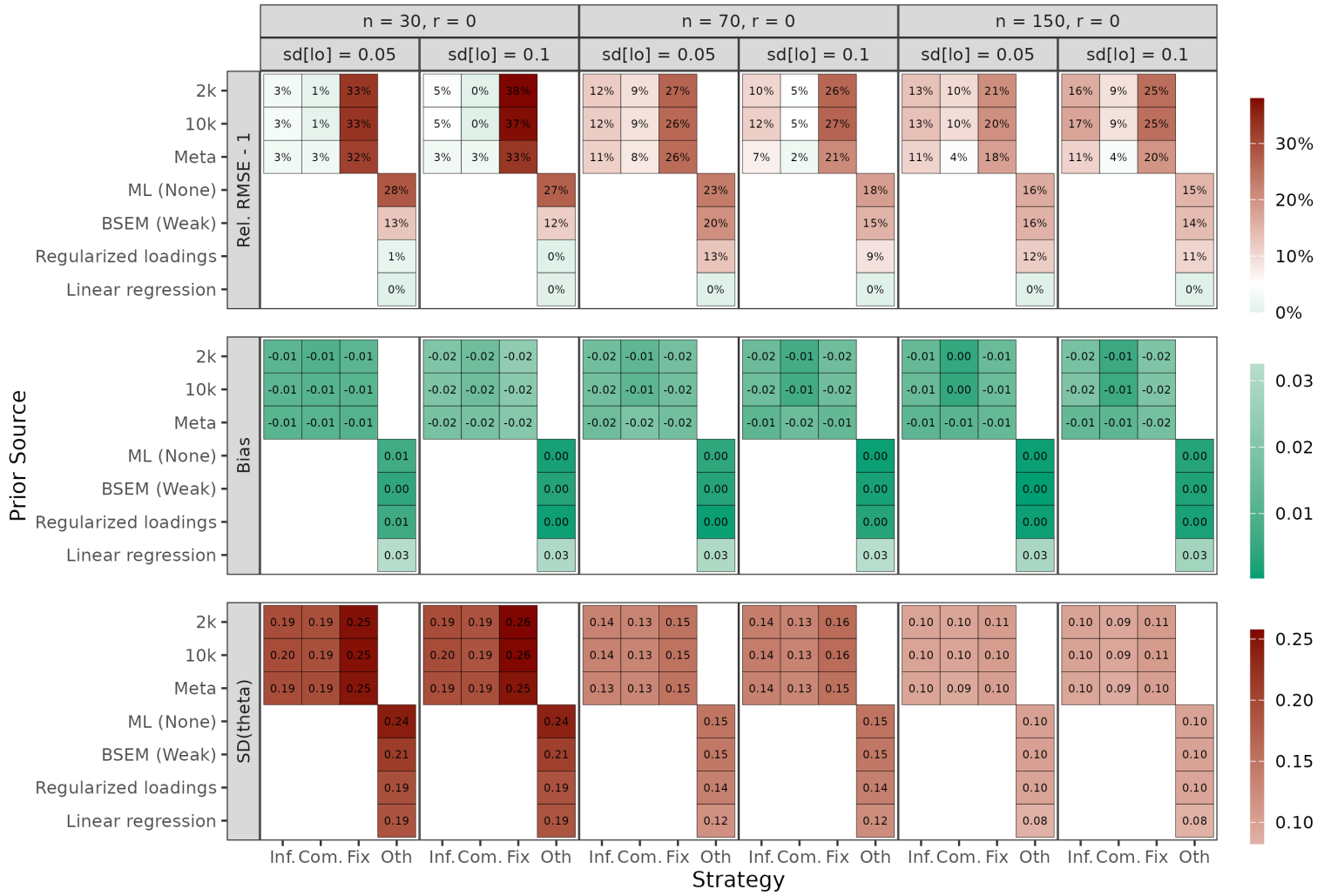
ML-Fix and ML-None were the least efficient approaches and were never competitive with the most accurate models (top panel of Figure 1), likely due to their relatively high variance; see bottom panel of Figure 1. For the smallest sample size ($n = 30$), all methods except both ML methods and BSEM-Weak were competitive in accuracy with the most accurate method. At other sample sizes ($n \geq 70$), only models with commensurate priors were consistently competitive with LR. Models with informative priors based on the *meta* studies and BSEM-Reg were sometimes competitive.

Regarding inference, all methods except ML had intervals that included ρ at rates close to or exceeding the nominal coverage rate. ML methods are often undercovered ρ, with undercoverage being more severe for smaller samples.

Given the results in this section, the ML methods are the least appealing option when the true structural parameter is zero. Informative priors based on the *meta* studies, commensurate prior approaches, regularized loadings, and linear regression had acceptable inference, and these methods were sometimes competitive with each other in terms of accuracy.

Figure 1

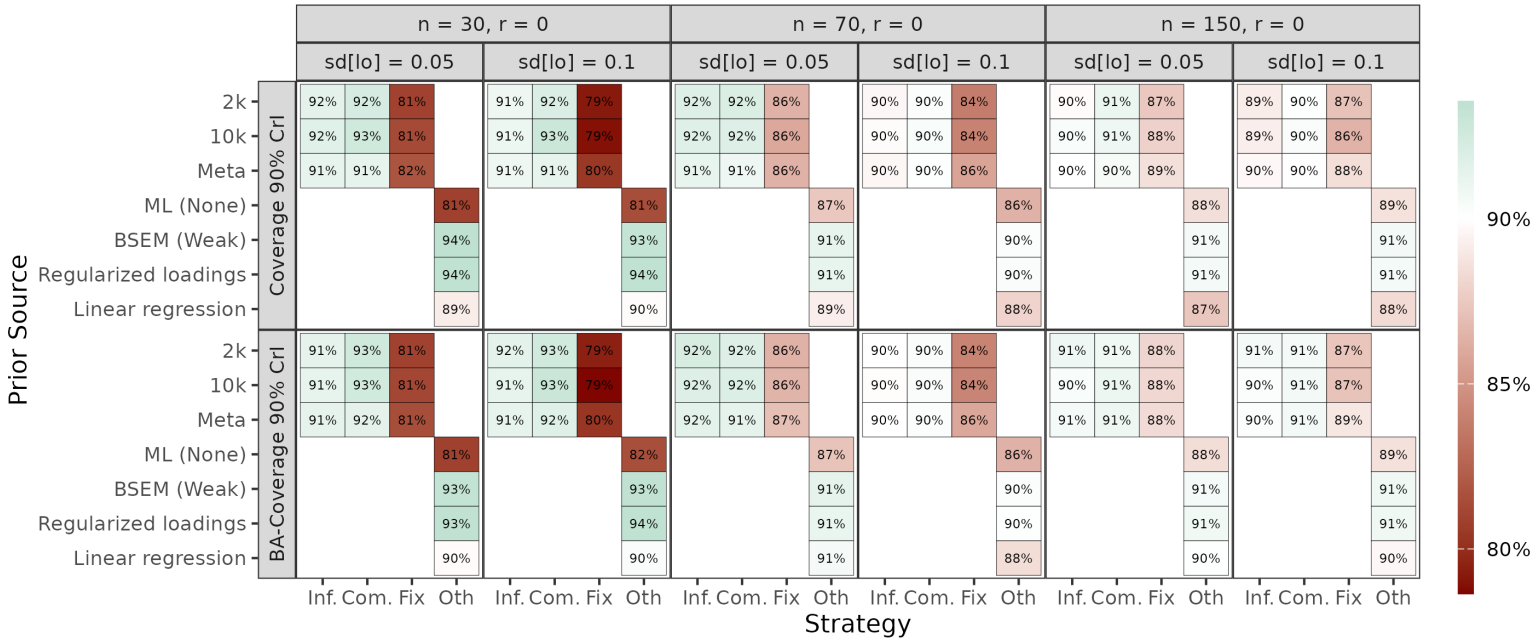
Simulation study 1: Accuracy of parameter estimates when $\rho = 0$



Note. Inf.: Informative priors based on historical data; Com.: Commensurate priors; Fix: ML SEM with fixed loadings. For models under the *Oth* strategy: N = ML (None); W = BSEM (Weak); R = Regularized loadings; L = linear regression. sd[lo]: σ_λ

Figure 2

Simulation study 1: Inference metrics when $\rho = 0$



Note. Inf.: Informative priors based on historical data; Com.: Commensurate priors; Fix: ML SEM with fixed loadings. For models under the *Oth* strategy: N = ML (None); W = BSEM (Weak); R = Regularized loadings; L = linear regression. sd[lo]: σ_λ

Results for $\rho \neq 0$

When $n = 30$ and $\rho = -0.5$, the three methods that depended on historical information – BSEM-Inf, BSEM-Com, and ML-Fix – were the most accurate approaches, while BSEM-Weak was competitive; see the top-left panels of Figure 3. BSEM-Inf and BSEM-Com had high bias and low variance, whereas ML-Fix had low bias and high variance. For larger samples ($n \geq 70$), BSEM-Inf and BSEM-Com remained the most accurate approaches, while BSEM-Weak and BSEM-Reg were competitive.

Regarding inference, most methods had acceptable coverage across conditions except for ML-Fix and ML-None in the smallest samples ($n = 30$), and LR at all sample sizes; see the top panel of Figure 4. The bias-adjusted coverage of LR (middle panel) was adequate, implying that its inadequate coverage was due to its high bias.

When $n = 30$, BSEM-Com based on the *meta* universe and BSEM-Inf methods were the most powerful among methods that maintained nominal coverage. The ML methods had high power, but their coverage was inadequate, suggesting that modellers cannot rely on these methods for inference. For larger samples, the same methods continued to be the most powerful. Their power was higher than other methods that maintained nominal coverage, such as BSEM-Weak and BSEM-Reg.

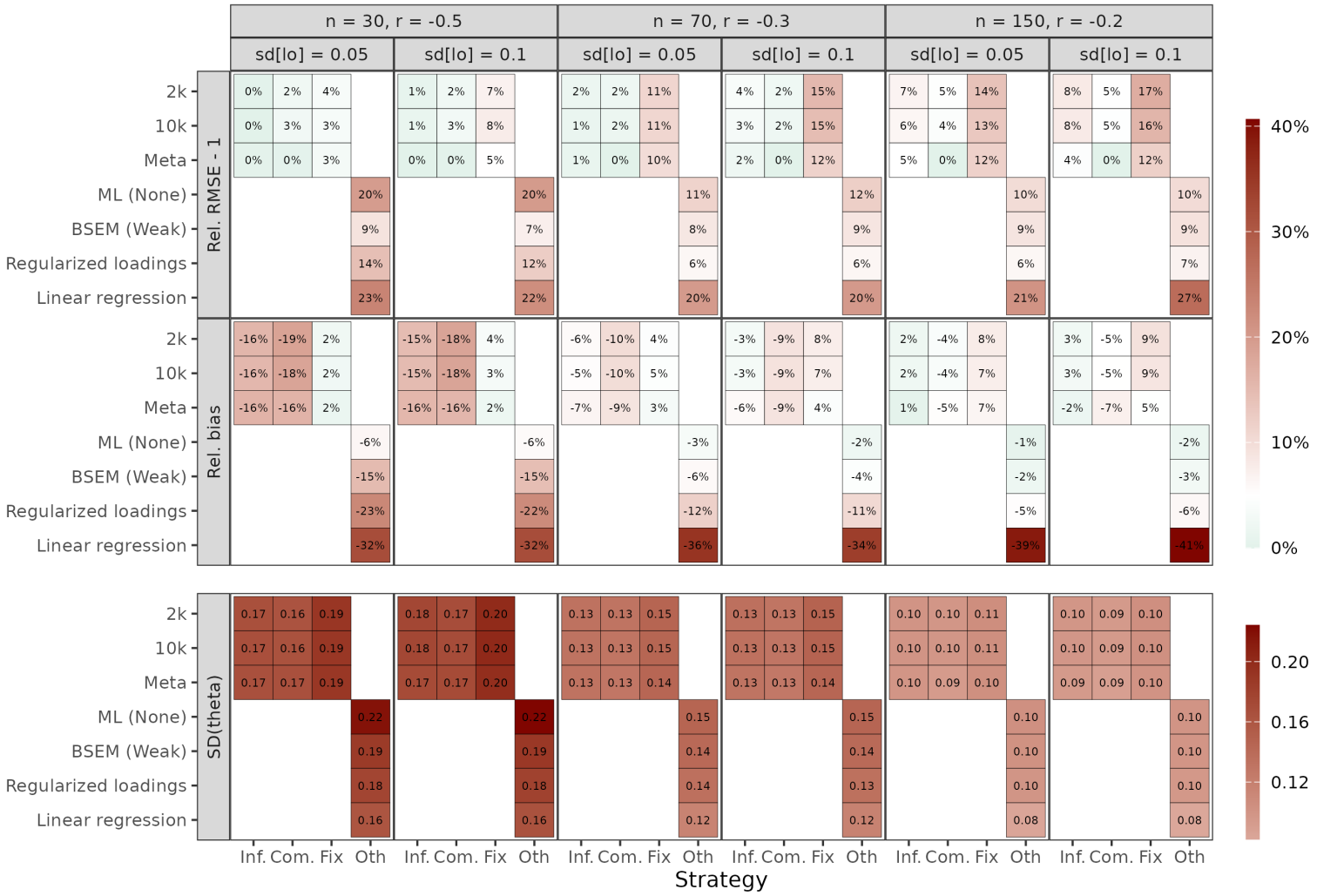
Summary of results

Across design conditions, the informative and commensurate prior approaches were often adequate, with priors based on the *meta* universe sometimes performing better than priors based on the 2k and 10k universes – see Appendix B for further elaboration of this point. ML methods that fixed the loadings to historical results performed well for the smallest samples in terms of accuracy, though their coverage was inadequate.

Among the methods that ignored historical information, standard ML was always a poor choice. Linear regression was the worst choice when $\rho \neq 0$. BSEM with weakly informative priors on loadings had mediocre performance, and regularizing loadings did not provide any benefit over weakly informative priors on loadings.

Figure 3

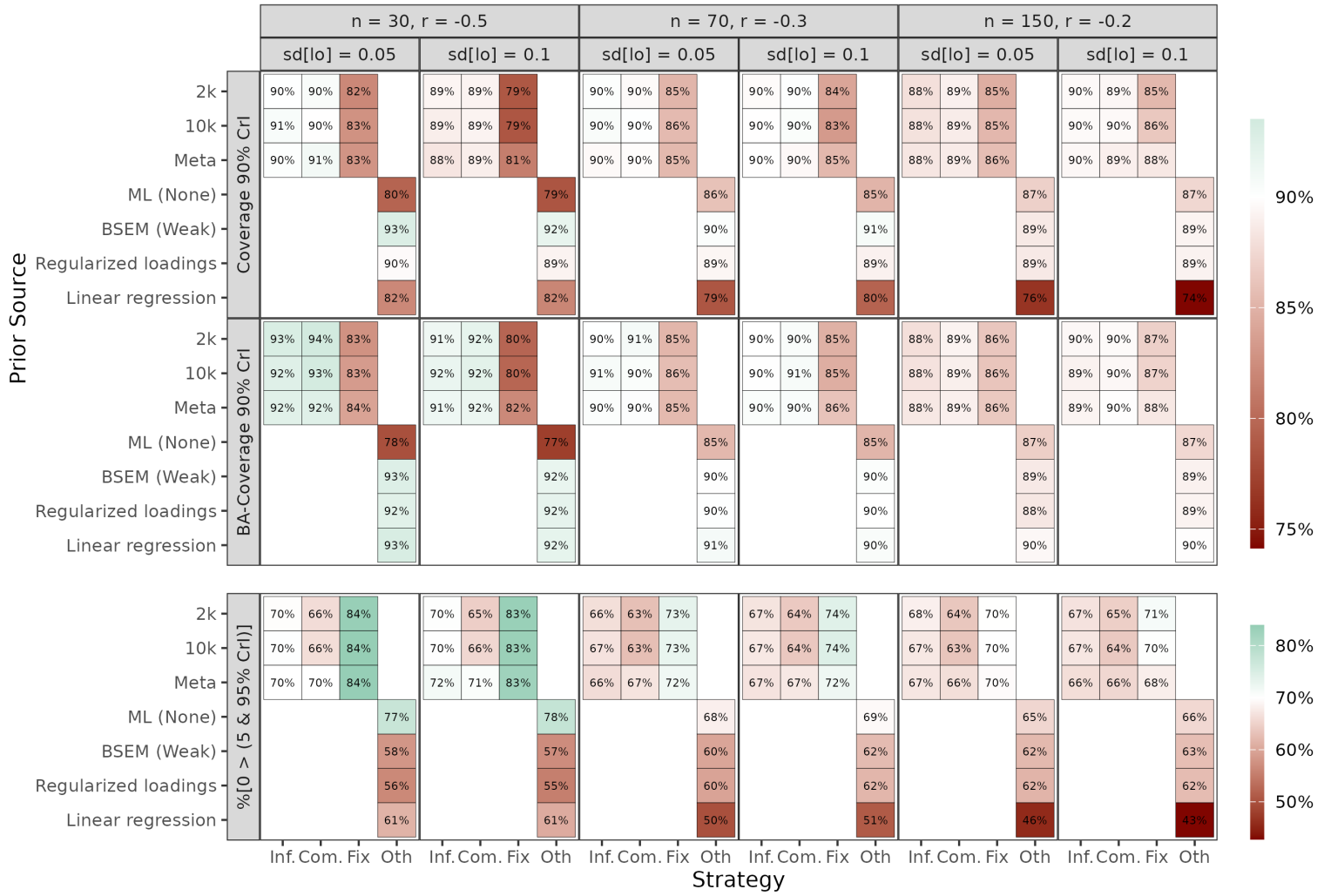
Simulation study 1: Accuracy of parameter estimates when $\rho \neq 0$



Note. Inf.: Informative priors based on historical data; Com.: Commensurate priors; Fix: ML SEM with fixed loadings. For models under the *Oth* strategy: N = ML (None); W = BSEM (Weak); R = Regularized loadings; L = linear regression. sd[lo]: σ_λ

Figure 4

Simulation study 1: Inference metrics when $\rho \neq 0$



Note. Inf.: Informative priors based on historical data; Com.: Commensurate priors; Fix: ML SEM with fixed loadings. For models under the *Oth* strategy: N = ML (None); W = BSEM (Weak); R = Regularized loadings; L = linear regression. sd[lo]: σ_λ

Simulation study 2

For study 2, we retained five models from study 1 to study how the design factors affected their performance. The five models were: BSEM with informative (*BS-Inf*) and commensurate priors (*BS-Com*) as they were often adequate; BSEM with weakly informative (*BS-Weak*) as a mediocre default Bayesian model; standard ML (*ML-None*) as a typical default model despite its poor performance; and ML with fixed loadings (*ML-Fix*) as a cheap easy-to-implement option with potentially good performance for the smallest samples when $\rho \neq 0$.

For historical data, we operated only in the *meta* universe since this resulted in the best-performing models in study 1. We sampled the sample size for historical studies from: $300 + \lfloor s \rfloor,\ s \sim \text{log-normal}(\ln(100), \ln(5)/\Phi^{-1}(.99))$. Hence, the median sample size for historical studies was 400, and there was a 99% chance that the sample size for historical studies was less than 800, $300 + 5 \times 100$.

We retained the same data generation process from Study 1 but treated the design factors as random variables.

Simulation design conditions

We draw the levels of continuous design factors from either continuous or discrete uniform distributions. This approach enables us to study the behaviour of the methods across the selected interval of each design factor. We note that the selected distributions do not accurately reflect the distribution of these factors in applied studies. For example, we sample the correlation between substantive factors from $\text{Unif}(-0.75, 0.75)$. High correlations, such as 0.7, are less common than small correlations closer to zero (e.g. Bosco, Aguinis, Singh, Field, & Pierce, 2015; Kraft, 2020; Lovakov & Agadullina, 2021). As a result, the average behaviour of any method may not be particularly informative if its behaviour varies as a function of the design factor. However, this design choice allows us to understand better how the method's behaviour depends on the design factor. We report the distribution of design factors in Table 1.

Table 1
Simulation study 2: Distribution of design factors

Design Factor	Distribution
Correlation between substantive factors	$\rho \sim \mathcal{U}(-0.75, 0.75)$
Correlation between method factors	$\varrho \sim \mathcal{U}(0.25, 0.75)$
Difference in loadings across studies	$\sigma_\lambda \sim \mathcal{U}(0.01, 0.15)$
Number of historical studies for factor 1	$k_1 \sim \mathcal{U}\{3, 8\}$
Number of historical studies for factor 2	$k_2^* \sim k_1 + \mathcal{U}\{-2, 0\},\ k_2 = \max(k_2^*, 2)$
Random misspecification error	$\tau \sim \mathcal{U}(.01, .09)$
Number of items per substantive factor	$p \sim \mathcal{U}\{7, 15\}$
Sample size of current study	$n \sim \mathcal{U}\{30, 150\}$

The number of items reflecting the method effect within each scale was under half p. Specifically, it was the quotient of $(p-1)/2$. For example, when p was 7 or 8, three items per scale reflected each method effect factor.

We simulated 10,000 iterations, sampling the levels of the design conditions from their distributions.

Analysis methods

We were primarily interested in how the five methods compared in accuracy and how well they maintained their nominal coverage, as assessed using the 90% confidence interval.

Primary analysis of accuracy

We measured accuracy using the mean absolute error (MAE). The MAE for method m was $MAE_m = \frac{1}{10000}\sum_{r=1}^{10000}|\hat{\rho}_{rm} - \rho_r|$ where r is the design replication index. We performed 10 paired z-tests to compare the five methods to each other and computed the standardized mean difference based on the difference in absolute errors between methods.

A limitation of this analysis is that the distributions of simulation conditions do not reflect the actual distributions of parameters in applied research. As a result, the comparisons may not generalize well to typical scenarios in practice. To address this, we implemented a re-weighting strategy to increase the generalizability of the results.

Among the design factors, the distribution of absolute correlations between

variables in psychology is well-studied. Bosco et al. (2015) analyzed 147,328 correlations from two major psychology journals from 1980 to 2010. In Table 2 of their study, Bosco et al. (2015) report the following quantiles for absolute correlations, rounded to two decimal places: 20% (.05), 25% (.07), 33% (.09), 40% (.12), 50% (.16), 60% (.21), 67% (.26), 75% (.32), 80% (.36).

To approximate this distribution, we used the *rrisk* R package (Belgorodski, Greiner, Tolksdorf, & Schueller, 2017), which supports fitting quantiles to known distributions. Of ten candidate distributions (e.g., gamma, normal, Weibull), we identified beta, exponential, and truncated normal distributions that closely matched the quantiles. For instance, a beta distribution with shape parameters 0.87 and 3.25 matched all quantiles except the 75th percentile, which differed by only 0.01 points. Given the plausibility of the beta distribution for absolute correlations, we adopted it as our target distribution.

The beta(0.87, 3.25) distribution has a mean of 0.21, a standard deviation of 0.18, and is right-skewed. These characteristics suggest smaller correlations than those sampled from $\mathcal{U}(-0.75, 0.75)$. However, if we had sampled directly from beta(0.87, 3.25), we would have had limited data on large correlations due to their low frequency, reducing our ability to evaluate the performance of the five methods for larger correlations.

Importance sampling (Owen, 2013) is useful in this regard as it allows us to identify the appropriate weights (*likelihood ratios*) for estimates from each iteration given the *target* distribution (beta(0.87, 3.25)) and the sampled or *proposal* distribution $\mathcal{U}(0, 0.75)$ – the magnitude not the sign of the true correlations is the relevant design factor. We computed *likelihood ratios* (importance weights) as: $(p(|\rho_r|; 0.87, 3.25)/q(|\rho_r|; 0, .75))$ where p and q are the beta and uniform densities respectively. We used these weights to re-weight the pairwise differences in absolute errors between methods. We report both weighted and unweighted results to provide a complete picture of the comparisons.

Finally, we adjusted p-values for multiple comparisons separately within the unadjusted and adjusted analyses. The adjustment method was the Benjamini-Yuketieli

method (Benjamini & Yekutieli, 2001), which attempts to control the false discovery rate.

Primary analysis of coverage

We computed the coverage of the 90% CI for each method's iteration and then computed the average coverage across iterations. For the adjusted analysis, we re-weighted each binary variable using the likelihood ratios computed above before taking the average.

Additional analysis of accuracy and coverage

We wanted to assess how design factors impacted the accuracy of each method. To do this, we computed whether each estimated correlation, $\hat{\rho}_{rm}$, was within one standard error of the true correlation ρ_r. We performed this check on the Fisher z-scale, which has the advantage of being variance-stabilizing, i.e. the standard error of z_r, $(n_r - 3)^{-\frac{1}{2}}$, is symmetric across all possible values of ρ_r. Based on the empirical rule and from a frequentist perspective, the ideal method would achieve an average of about 68.3% on this indicator. This metric is similar to coverage, but the interval in this assessment is about the population parameter instead of the sample estimate. We modelled this indicator with the design factors as predictors. The predictors were: $|\rho|, n, \varrho, \sigma_\lambda, p, k_1, k_1 - k_2, \tau$; see Table 1 for the names of the design factors.

We used conditional inference trees (Hothorn et al., 2006) to identify potential interactions between design factors. For each tree, we required that the p-value governing a split be less than .001 (i.e. $\alpha = .001$) for parsimony.

Next, we used generalized additive logistic models (Wood, 2017) – one per method – to model continuous but nonlinear relations between design factors and the performance of each method. We specified continuous design factors with thin-plate regression splines (Wood, 2003). We adjusted p-values within each model for multiple comparisons using the Benjamini-Yuketieli method (Benjamini & Yekutieli, 2001).

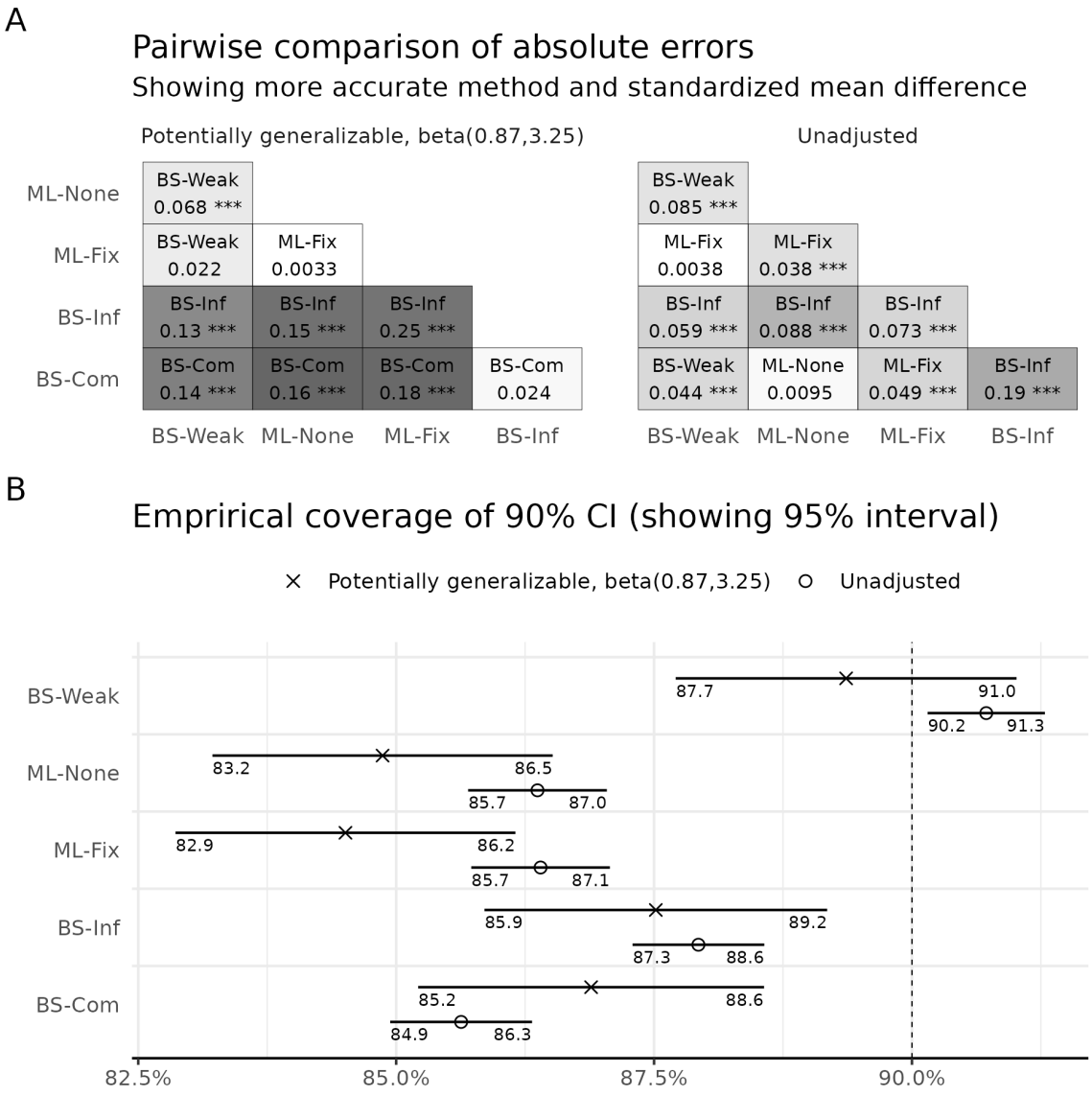
Finally, we used plots of the predicted probability of adequate performance as a function of statistically significant ($p < .05$) design factors based on the additive models to describe the behaviour of the methods. When the predicted probability plot did not

include ρ, the predictions were re-weighted using likelihood ratios to ensure that the average relation level reflected the assumed distribution of ρ based on the prior literature.

Coverage. We applied the same methods used in the additional accuracy analysis to the indicator of each method's coverage within each iteration.

Figure 5

Simulation study 2 primary results



Note. Panel A: Pairwise differences in MAE. ' ' $p > .05$, '*' $p < .05$, '**' $p < .01$, '***' $p < .001$ Panel B: Coverage of 90% CI. ML: Maximum-likelihood; BS: BSEM; Inf: Informative priors; Com: Commensurate priors; Fix: Fixed loadings.

Results

Primary analysis of accuracy

Results of the primary accuracy analyses are in panel A of Figure 5. We first focus on the re-weighted analysis (top-left panel). The method with the highest accuracy overall was BS-Com, which had a lower MAE than all other methods on average. Its accuracy was comparable to BS-Inf; their MAE differed by 0.024 standardized mean difference (SMD), which was not statistically significant, $p = .056$. The worst choice was ML-None, which had a higher MAE on average than all other methods, though the second worst choice, ML-Fix, only had a slightly lower MAE by 0.003 SMD. BS-Weak was the middle-performing method; it performed worse than BS-Inf and BS-Com and better than ML-Fix and ML-None. These results are informative for choosing between the methods without specific contextual factors.

In the unadjusted analysis, BS-Inf performed best, followed by ML-Fix. BS-Weak performed middle, followed by ML-None. BS-Com performed worst, with a 0.0095 SMD difference compared to the second-worst method, ML-None.

Re-weighting by likelihood ratios down-weights large and up-weights small correlations since small correlations are more common in the literature. These results suggest that the relative performance of BS-Com and ML-Fix, two methods incorporating historical information, improves when the true substantive correlation (ρ) is smaller.

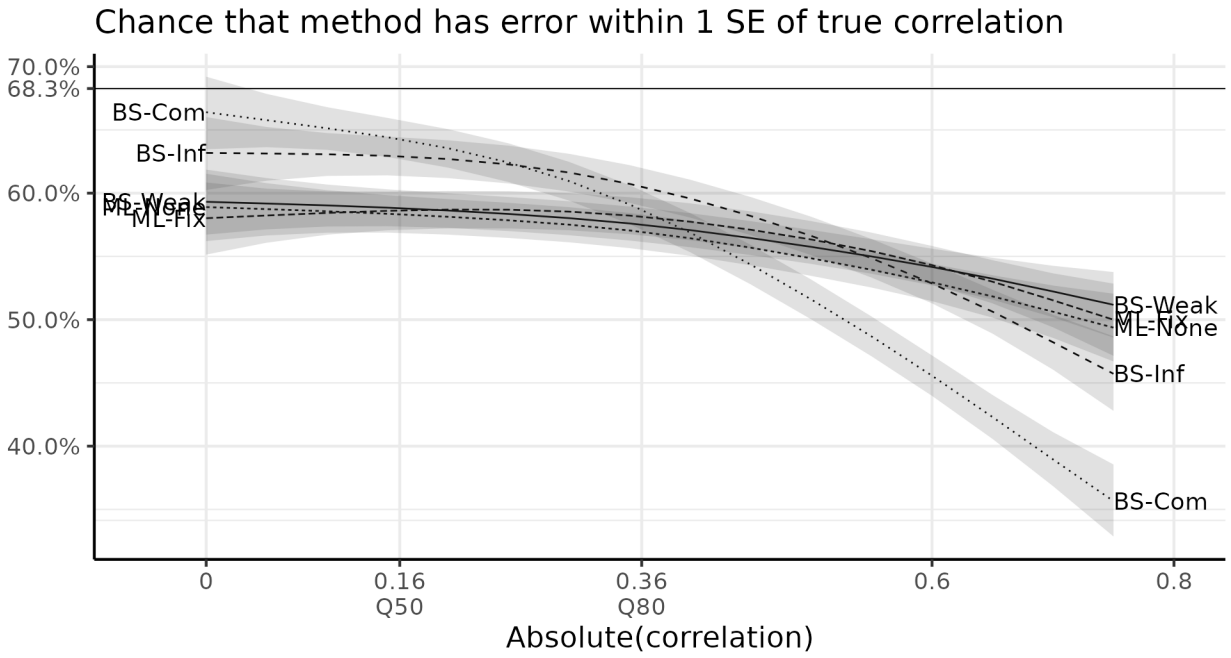
Additional analysis of accuracy

The trees did not identify any interaction that significantly altered the interpretation of the marginal effects of the design factors. So, we focus on the results from the additive models.

The absolute size of the correlation, $|\rho|$, affected the relative accuracy of all five methods; see Figure 6. BS-Inf and BS-Com were most affected by this design factor. When the true correlation is small or less than the 80th percentile of correlations, both methods were most accurate, with about 60 – 68% chance of estimates falling within 1 SE of the true

Figure 6

Simulation study 2: Method accuracy as a function of the correlation between factors



Note. 0.16 and 0.36 are the 50th and 80th percentiles of absolute correlations respectively identified by Bosco et al. (2015). The shaded areas around each line are 95% confidence intervals of prediction.

correlation. Other methods averaged at about 60% on this metric. When correlations were larger, the other methods performed better than BS-Inf and BS-Com. The performance of BS-Inf was still comparable to other methods, but the performance of BS-Com heavily deteriorated, dropping to about 40% of estimates falling within 1 SE of the true correlation.

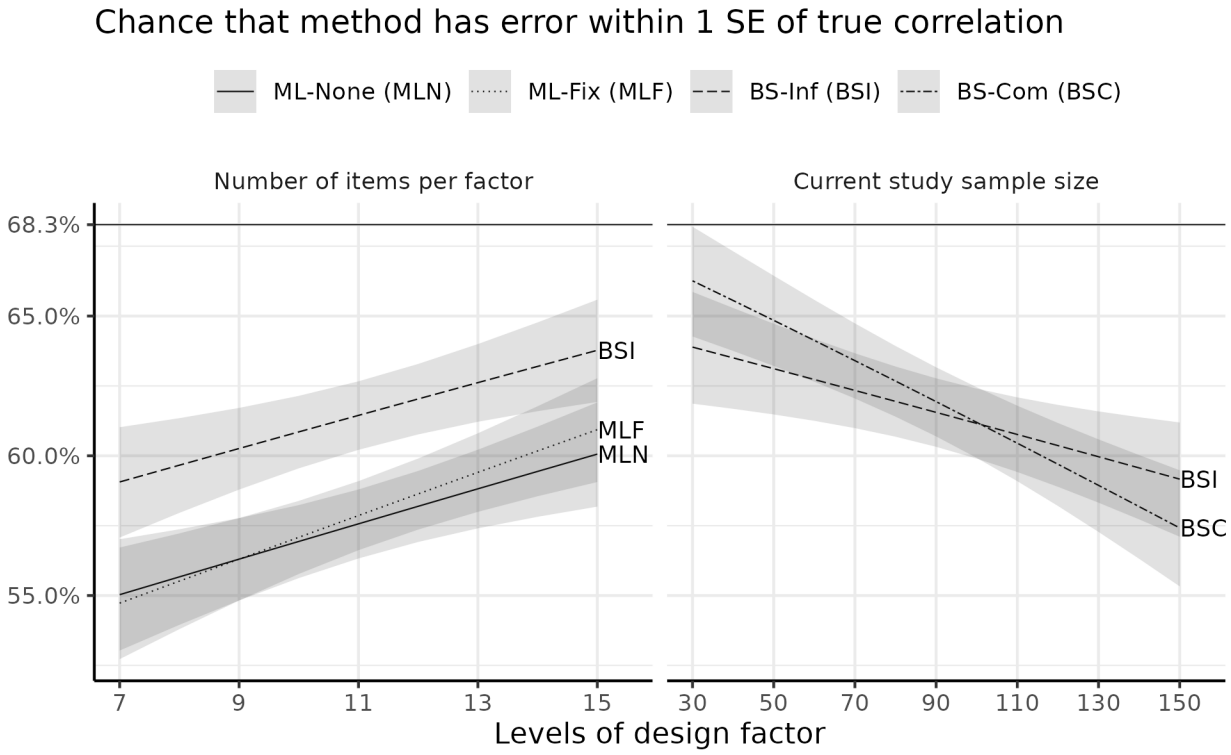
Finally, the relative accuracy of BS-Inf, ML-Fix and ML-None improved with more items per factor; see the left panel of Figure 7. The relative accuracy of BS-Inf and BS-Com decreased with larger sample sizes; see the right panel of Figure 7.

Primary analysis of coverage

As shown in panel B of Figure 5, only BS-Weak had a coverage rate statistically or practically indistinguishable from the nominal rate. BS-Inf had a coverage rate of at least 86%, which most evaluators would consider acceptable. The middle-performing method was BS-Com, whose coverage was at least 85%. The coverage of the ML methods was less than adequate in the adjusted analysis: at least 83%, with estimates just under 85%. In

Figure 7

Simulation study 2: Method accuracy as a function of additional factors



the unadjusted analysis, both ML methods had more acceptable coverage. This result suggests that the coverage of the ML methods is poorer for smaller correlations.

Additional analysis of coverage

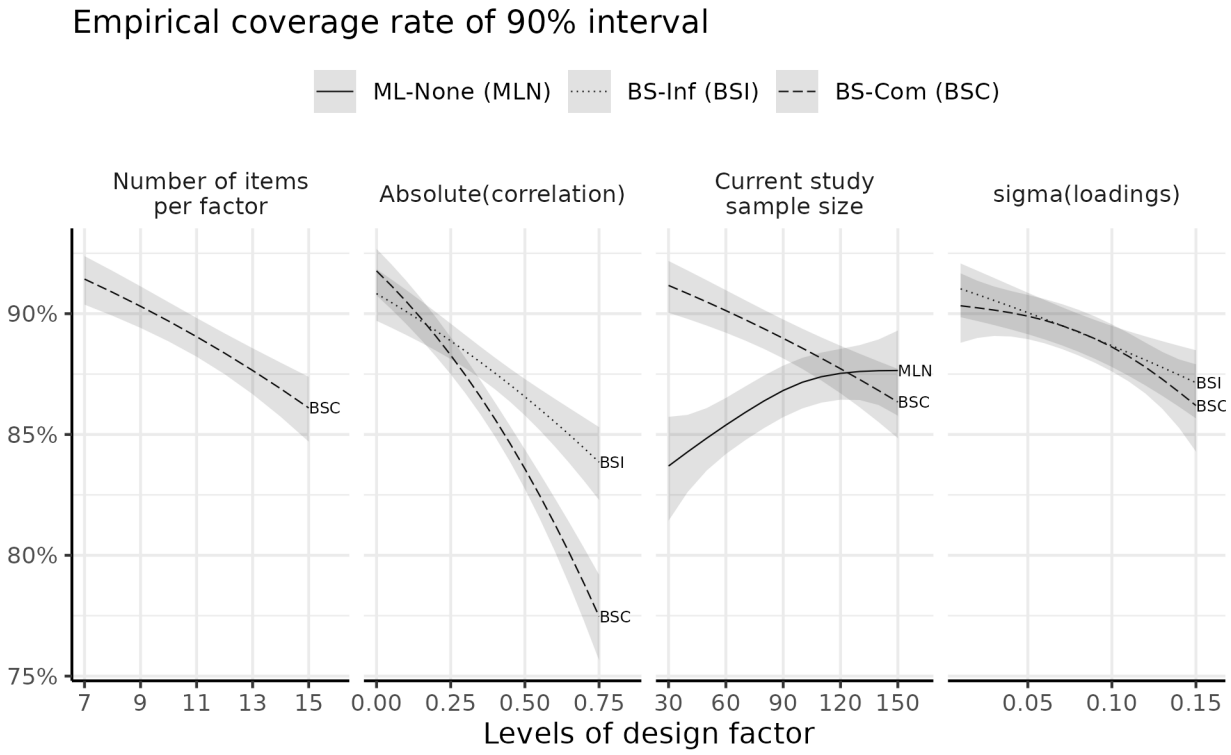
Similar to the additional analysis of accuracy, the trees did not identify any interaction that significantly altered the interpretation of the marginal effects of the design factors. So, we focus on the results from the additive models; see Figure 8.

Number of items per factor, p. Increasing p worsened the coverage of BS-Com significantly, though the coverage drop due to this factor did not result in coverage below 85%.

Absolute value of the correlation, $|\rho|$. Increasing $|\rho|$ negatively impacted the coverage of BS-Com and BS-Inf. At very high $|\rho|$, the coverage of BS-Inf was about 83%, while the coverage of BS-Com was about 78%.

Figure 8

Simulation study 2: Method coverage as a function of design factors



Current study sample size, n. Increasing n improved the coverage of ML-None from under 85% when $n = 30$ to about 87.5% when $n \geq 100$. However, increasing n led to lower coverage for BS-Com, dropping to about 86% when $n = 150$.

Heterogeneity of loadings, σ_λ. Increasing σ_λ across studies led to a drop in coverage for BS-Inf and BS-Com, though the resulting drop in coverage did not drop coverage below 85%.

Summary. In summary, the more troubling drops in coverage ($< 85\%$) were in the coverage of ML-None for the smallest samples ($n \lessapprox 50$), the coverage of BS-Inf for extreme correlations ($|\rho| \gtrapprox .6$), and the coverage of BS-Com for large correlations, $|\rho| \gtrapprox .4$. Also, the coverage of BS-Com was affected by four factors, suggesting that this method had fragile frequentist inference properties.

Discussion

Summary of findings

Our study aimed to determine whether incorporating historical measurement models into SEM with small samples could improve the estimation of structural parameters. This question is pertinent because, for widely used scales, it is often possible to find published factor analysis results that include loading estimates and standard errors. Researchers can use this historical information to establish informative priors for the measurement parameters in a new BSEM. We focused on the problem of estimating the correlation or standardized correlation between two constructs. We generated data to introduce significant bias in the Pearson correlation between the scale scores, thus necessitating measurement error modelling. In addition, we incorporated randomness into the data generation process to ensure that no two studies had identical population covariance matrices. Our findings indicate that by using historical information about the measurement parameters in the form of priors on loading, researchers may improve the accuracy of the correlation of interest.

Priors based on random-effects meta-analysis will often be optimal. The improvement in accuracy is notable when the true correlation is small, the most common situation in psychological research (Bosco et al., 2015; Kraft, 2020; Lovakov & Agadullina, 2021). Furthermore, informative priors based on meta-analysis will produce results with acceptable coverage, making it the optimal choice. Attempting to discount the historical nature of the data using a commensurability parameter did not provide any benefit over informative priors based on meta-analysis for the small samples we studied. The exception to the advice above is studies focused on large correlations. When the true correlation is large, BSEM with weakly informative priors on all parameters was the best choice with high relative accuracy and adequate coverage.

Do not fix loadings. Under our simulation conditions, which we consider realistic due to heterogeneity in structural and unstructured parameters, there is no good reason to

fix loadings based on results from historical studies. When such historical information is available, researchers should use this information to specify informative priors in a Bayesian model. Suppose only loading estimates are available, and there is no way to obtain standard error information. In that case, using weakly informative priors is likely better than fixing loadings, especially considering the poor frequentist inferential properties of fixing loadings.

Design factors without identifiable impact. We could not identify the impact of several design factors in Study 2. These were the correlation between the method factors, the number of historical studies, and the degree of random misspecification error. The degree of random misspecification error likely had no identifiable effects because these effects are more notable for large samples (see Table 1 in Satorra, 2015). Supplementary analysis of Study 1 (in Appendix B) suggests having more historical studies is better than having only one historical study. So, it may be the case that the range of this variable in Study 2 was insufficient to identify its effects. Finally, the correlation between method effects reflects model misspecification in the current data analysis since the current data analysis always assumes the method factors are identical (e.g. Maydeu-Olivares, 2017). This misspecification did not impact the performance of any of the methods.

Limitations

We now review some limitations of this study, some of which may inform future research efforts.

First, the results of this study only apply to the limited situations where:

1. There is no significant prior information on structural parameters;
2. There is significant prior information on measurement parameters and
3. The prior information on the measurement parameters suggests measurement error modelling is necessary.

When points 1 and 2 are correct, but point 3 is false, linear regression with sum scores is probably sufficient. Linear regression will yield lower variance estimates than estimates from latent variable models. If measurement error modelling is unnecessary, the overall accuracy of linear regression may be higher than that of latent variable approaches since the drop in variance may be less than the increase in bias. When point 1 is false, researchers can markedly improve accuracy by supplying prior information directly on the structural parameters of interest (e.g. Finch & Miller, 2019; McNeish, 2016; Smid et al., 2020; Smid & Rosseel, 2020; Smid & Winter, 2020).

Second, we did not correctly meta-analyze the prior information. The results of the historical factor analysis are better meta-analyzed using multivariate meta-analytic confirmatory factor analysis methods (e.g. Cheung, 2014; Cheung & Chan, 2009) that account for the correlation between parameters within the historical studies. However, these methods require original historical data or sufficient statistics (correlation and covariances). Hence, we would recommend applying multivariate meta-analytic methods that allow the same analysis across datasets when the historical data (or sufficient statistics) are available – the availability of historical data should be more common in the future, given open science efforts that encourage transparent reporting and data sharing.

Third, we explored only one method to account for the historical nature of prior information that dynamically discounts such information. Commensurate priors performed less favourably than using historical information directly without accounting for its historical nature. This finding is helpful because researchers can implement the recommended approaches in popular Bayesian SEM software such as Mplus (L. K. Muthén & Muthén, 1998–2017) and blavaan (Merkle, Fitzsimmons, Uanhoro, & Goodrich, 2021). However, as noted in the introduction, there are other methods for accounting for the historical nature of past results. We believe that methods that discount historical information based on the degree of similarity between historical and current data (e.g. Pan et al., 2017; Shi et al., 2023) may be more promising than the approach we employed. As

with the second limitation above, such methods should be easier to implement in the future as more researchers share their study data.

Finally, we operationalized the estimation of structural parameters in this study as a correlation between two substantive factors. These results may not generalize to other structural configurations, such as mediation models. Although we do not believe that altering the dynamic of the structural parameter would change the findings presented here, a future study should evaluate this claim to confirm its validity.

Disclosure Statement

There are no relevant financial or non-financial competing interests to report.

References

Belgorodski, N., Greiner, M., Tolksdorf, K., & Schueller, K. (2017). rriskDistributions: Fitting distributions to given data or known quantiles [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rriskDistributions` (R package version 2.1.2)

Benjamini, Y., & Yekutieli, D. (2001, August). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. doi: 10.1214/aos/1013699998

Bollen, K. A. (1989). *Structural equations with latent variables*. Hoboken, NJ, USA: John Wiley & Sons, Inc. (Pages: 514) doi: 10.1002/9781118619179

Boomsma, A. (1985, June). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242. doi: 10.1007/BF02294248

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015, March). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. doi: 10.1037/a0038047

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

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. (2001, May). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*(4), 468–508. doi: 10.1177/0049124101029004003

Chen, J., Guo, Z., Zhang, L., & Pan, J. (2021). A partially confirmatory approach to scale development with the Bayesian lasso. *Psychological Methods*, *26*, 210–235. doi: 10.1037/met0000293

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., & 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

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. doi: 10.1037/a0033805

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

Duan, Y., Ye, K., & Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, *17*(1), 95–106. doi: 10.1002/env.752

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017, December). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927. doi: 10.1007/s11336-017-9557-x

Finch, W. H. (2024, March). A comparison of methods for synthesizing results from previous research to obtain priors for Bayesian structural equation modeling. *Psych*, *6*(1), 45–88. doi: 10.3390/psych6010004

Finch, W. H., & Miller, J. (2019, July). The use of incorrect informative priors in the estimation of MIMIC model parameters with small sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 497–508. doi: 10.1080/10705511.2018.1553111

Hedges, L. V., & Olkin, I. (1985, January). Random effects models for effect sizes. In L. V. Hedges & I. Olkin (Eds.), *Statistical Methods for Meta-Analysis* (pp. 189–203).

San Diego: Academic Press. doi: 10.1016/B978-0-08-057065-5.50014-2

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., & Sargent, D. J. (2011, September). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, *67*(3), 1047–1056. doi: 10.1111/j.1541-0420.2011.01564.x

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. doi: 10.1198/106186006X133933

Hox, J. J. C. M., Schoot, R. v. d., & Matthijsse, S. (2012, July). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*(2), 87–93. (Number: 2) doi: 10.18148/srm/2012.v6i2.5033

Hsiang, T. C. (1975, November). A Bayesian view on ridge regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(4), 267–268. Retrieved from `http://www.jstor.org.proxy.lib.ohio-state.edu/stable/2987923` (Publisher: [Royal Statistical Society, Wiley]) doi: 10.2307/2987923

Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., & Chen, F. (2015, December). The power prior: Theory and applications. *Statistics in medicine*, *34*(28), 3724–3749. doi: 10.1002/sim.6728

Jacobucci, R., & Grimm, K. J. (2018, July). Comparison of frequentist and bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 639–649. doi: 10.1080/10705511.2017.1410822

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, *23*(4), 555–566. doi: 10.1080/10705511.2016.1154793

Kaplan, D., Chen, J., Yavuz, S., & Lyu, W. (2023, March). Bayesian dynamic borrowing

of historical information with applications to the analysis of large-scale assessments. *Psychometrika*, *88*(1), 1–30. doi: 10.1007/s11336-022-09869-3

Kraft, M. A. (2020, May). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. doi: 10.3102/0013189X20912798

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, *101*(6), 1174–1188. doi: 10.1037/a0024776

Lee, S.-Y., & Song, X.-Y. (2004, October). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*(4), 653–686. doi: 10.1207/s15327906mbr3904_4

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

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. doi: 10.1002/ejsp.2752

Marcoulides, K. M. (2017). *A Bayesian synthesis approach to data fusion using augmented data-dependent priors* (Tech. Rep.). Arizona State University. Retrieved 2024-03-23, from `https://keep.lib.asu.edu/items/155625`

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

McElreath, R. (2020). *Statistical rethinking : A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press/Taylor & Francis. (Pages: 594)

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling*, *23*(5). doi: 10.1080/10705511.2016.1186549

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

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, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (Eighth ed.). Los Angeles, CA: Muthén & Muthén.

Nevitt, J., & Hancock, G. R. (2004, July). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439–478. doi: 10.1207/S15327906MBR3903_3

Owen, A. B. (2013). Importance sampling. In *Monte Carlo theory, methods and examples* (p. 46). Retrieved 2024-12-02, from `https://artowen.su.domains/mc/Ch-var-is.pdf`

Pan, H., Yuan, Y., & Xia, J. (2017, November). A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *66*(5), 979–996. doi: 10.1111/rssc.12204

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. doi: 10.1037/met0000220

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–20. doi: 10.18637/jss.v048.i02

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. van de Schoot & M. Miocević (Eds.), *Small Sample Size Solutions: A Guide for Applied*

Researchers and Practitioners (pp. 226–269). Taylor & Francis. doi: 10.4324/9780429273872

Satorra, A. (2015, September). A comment on a paper by H. Wu and M. W. Browne (2014). *Psychometrika*, *80*(3), 613–618. doi: 10.1007/s11336-015-9455-z

Savalei, V. (2019, June). A comparison of several approaches for controlling measurement error in small samples. *Psychological Methods*, *24*(3), 352–370. doi: 10.1037/met0000181

Savalei, V., Brace, J. C., & Fouladi, R. T. (2023). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*. doi: 10.1037/met0000537

Scheines, R., Hoijtink, H., & Boomsma, A. (1999, March). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37–52. doi: 10.1007/BF02294318

Shi, Y., Li, W., & Liu, G. F. (2023, March). A novel power prior approach for borrowing historical control data in clinical trials. *Statistical Methods in Medical Research*, *32*(3), 493–508. doi: 10.1177/09622802221146309

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020, January). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161. doi: 10.1080/10705511.2019.1577140

Smid, S. C., & Rosseel, Y. (2020). SEM with small samples: Two-step modeling and factor score regression versus Bayesian estimation with informative priors. In R. van de Schoot & M. Miocević (Eds.), *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners* (pp. 239–254). Taylor & Francis. doi: 10.4324/9780429273872

Smid, S. C., & Winter, S. D. (2020, December). Dangers of the defaults: A tutorial on the impact of default priors when using Bayesian SEM with small samples. *Frontiers in*

Psychology, *11*, 611963. doi: 10.3389/fpsyg.2020.611963

Uanhoro, J. O. (2023, June). minorbsem: An R package for structural equation models that account for the influence of minor factors. *Journal of Open Source Software*, *8*(86), 5292. doi: 10.21105/joss.05292

Uanhoro, J. O. (2024, April). Modeling misspecification as a parameter in Bayesian structural equation models. *Educational and Psychological Measurement*, *84*(2), 245 – 270. doi: 10.1177/00131644231165306

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling—A comparison of constrained maximum likelihood, Bayesian estimation, and fixed reliability approaches. *Psychological Methods*, *28*(3), 527–557. doi: 10.1037/met0000435

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

Viechtbauer, W. (2005, September). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. doi: 10.3102/10769986030003261

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., ... Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, *13*(1), 41–54. doi: 10.1002/pst.1589

Walters, R. W., Hoffman, L., & Templin, J. (2018, May). The power to detect and predict individual differences in intra-individual variability using the mixed-effects location-scale model. *Multivariate Behavioral Research*, *53*(3), 360–374. doi: 10.1080/00273171.2018.1449628

Westfall, J., & Yarkoni, T. (2016, March). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*(3), e0152719. doi: 10.1371/journal.pone.0152719

Wood, S. N. (2003, August). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 95–114. doi: 10.1111/1467-9868.00374

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Boca Raton, FL: CRC Press. (Pages: 476)

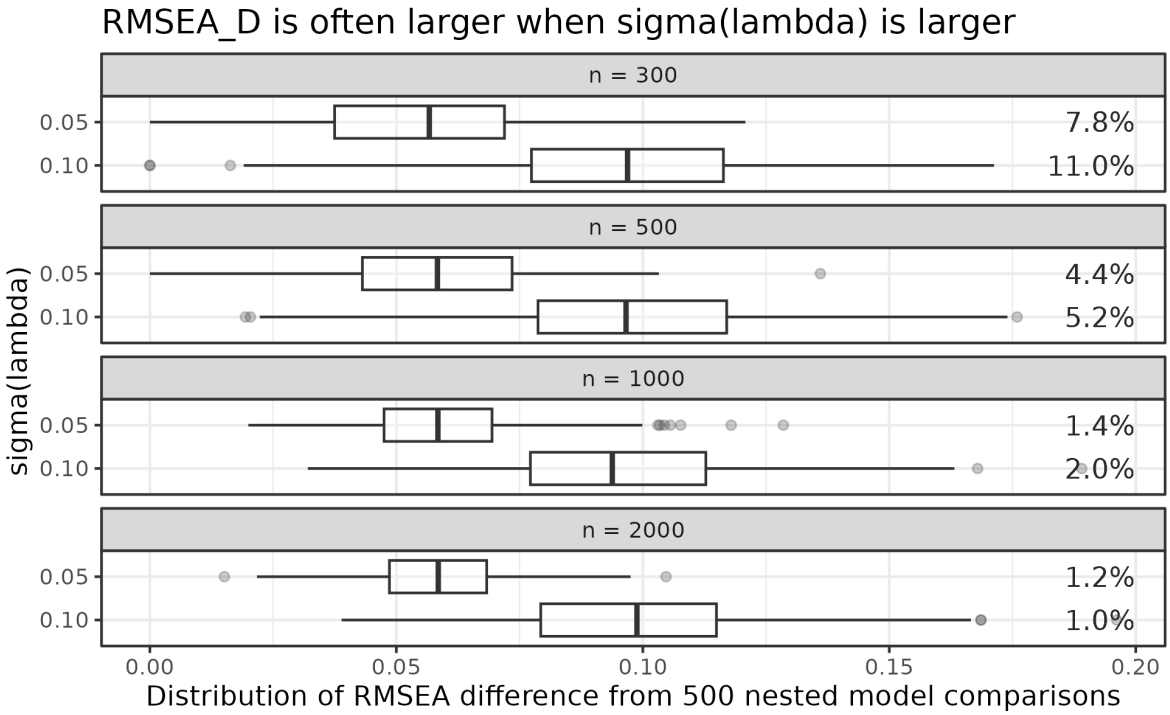
Wu, H., & Browne, M. W. (2015, September). Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika*, *80*(3), 571–600. doi: 10.1007/s11336-015-9451-3

Appendix A

Additional simulation results

Figure A1

Simulation study 1: Distribution of $RMSEA_D$



Note. Percentages on the right end of the figure are the proportion of comparisons that contained at least one non-convergent model – we could not compute RMSEA_D in these cases. The proportion of such cases is increasingly negligible for larger samples.

Appendix B

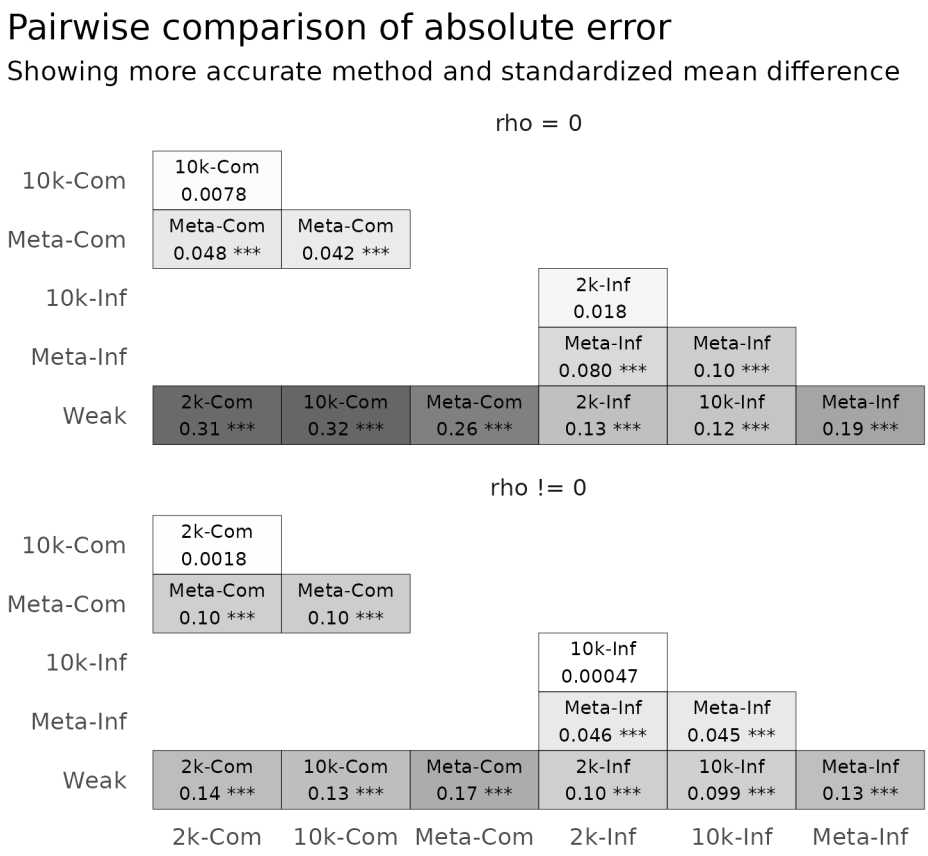
Study 1: Regression analyses of accuracy

We were interested in quantifying the differences in accuracy between the 2k, 10k, and Meta universes. Based on the preceding visual inspection of trends in the data, the results from the *meta* universe appeared more accurate. We focus on the estimates from six BSEM models with informative and commensurate priors and BSEM with weakly informative priors to quantify the differences between the universes. We did not include the ML models with fixed loadings as these models do not incorporate uncertainty about historical information into the present analysis. For this reason, the ML models with fixed loadings will dampen differences between the different universes. We included BSEM with weakly informative priors as a default model. We computed pairwise differences in absolute errors as in simulation study 2 and used the same reporting standards; however, we did not adjust the current analysis with importance weights. We analyze the data separately by whether $\rho \neq 0$.

As shown in Figure B1, the meta-universe produced between 0.04 and 0.10 standardized mean difference reductions in MAE compared to the 2k and 10k methods. The 2k and 10k were not statistically significantly different from each other. This effect is relatively minor but points to a gain from pooling results from multiple studies instead of relying on a single large study. It is also of note that one cannot distinguish the 2k and 10k results in accuracy. This finding implies that increasingly narrowing the priors on the measurement parameters incorrectly (since $\sigma_\lambda > 0$) did not impact the accuracy of the structural parameter estimates. Finally, the methods that relied on historical information were more accurate than BSEM with weakly informative priors.

Figure B1

Simulation study 1 regression analysis of MAE



Note. The top panel shows results for $\rho = 0$, the bottom shows results for when $\rho \neq 0$. ' ' $p > .05$, '*' $p < .05$, '**' $p < .01$, '***' $p < .001$.