MULTIVARIATE BEHAVIORAL RESEARCH
2026, VOL. 61, NO. 4, 716–734

OPEN ACCESS

Multivariate Location-Scale Models for Meta-Analysis

Katrin Jansen and Steffen Nestler

Department of Psychology, University of Münster, Münster, Germany

ABSTRACT

Often, primary studies that are pooled in a meta-analysis provide information on several outcomes of interest. Multivariate meta-analysis allows to analyze these outcomes simultaneously and model their relationship, and in addition can be more efficient than separate, univariate meta-analyses. However, standard multivariate meta-analysis models typically assume that the between-study variances and correlations are constant across studies. While it is possible to relax this assumption of constant heterogeneity by using location-scale models in univariate meta-analysis, extensions to the multivariate case have not yet been proposed. Here, we fill this gap by describing a location-scale model for the multivariate setting where both the between-study variances of the different outcomes and the correlations between them can depend on covariates. We examine its performance in a simulation study, where we compare univariate and bivariate location-scale models and different estimation methods. In addition, we show how to apply this model to data from a meta-analysis on the effects of motivational reading instruction on reading achievement and motivation. We discuss the implications of our findings for further research on meta-analysis of multiple outcomes and provide recommendations for the use of multivariate location-scale meta-analysis in applications.

KEYWORDS

Multivariate meta-analysis; location-scale model; heterogeneity; mixed-effects models

Introduction

Meta-analyses often reveal that there is substantial variability in the outcomes which are summarized. Typically, a large part of this variability is systematic variability that goes beyond sampling error, also known as heterogeneity. A central goal of meta-analysis is to identify reasons for this heterogeneity. This can be accomplished by conducting a meta-regression, which examines whether the mean of an outcome depends on covariates. However, a limitation of standard meta-regression models is that they assume residual heterogeneity, that is, heterogeneity which is not explained by the covariates, to be constant across studies (homoscedastic). To relax this assumption, location-scale models have been proposed, which allow not only the mean of the outcome (location), but also the between-study variance (scale) to depend on covariates (Viechtbauer & López-López, 2022). Location-scale meta-analysis provides a means of examining in which situations effects are more or less heterogeneous, and can thus provide important insights into meta-analytic data that cannot be obtained with conventional univariate meta-analysis methods. Moreover, allowing for heterogeneous between-study variances can improve statistical inference in certain situations (Rodriguez et al., 2023).

Thus far, location-scale models have only been considered for the univariate setting. However, meta-analyses often examine several outcomes, for example, the effects of student misbehavior on different dimensions of teacher burnout (Aloe et al., 2014), the effects of coaching on verbal and mathematical subtests of the Scholastic Aptitude Test (Kalaian & Raudenbush, 1996), or the sensitivity and specificity of a diagnostic test (Reitsma et al., 2005). Different meta-analytic models have been proposed for analyzing multiple outcomes simultaneously (Fernández-Castilla et al., 2021). Here, we focus on multivariate meta-analysis, which assumes that different outcomes are correlated both within- and between-studies, and in addition, allows to separately estimate the mean and the between-study variance of each outcome. Multivariate meta-analysis has several advantages over separate, univariate meta-analyses: First, it allows to account for the correlation

© 2026 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

of different outcomes within a study. Second, it allows to quantify the relationship of different outcomes between studies. Finally, multivariate meta-analysis can yield more efficient parameter estimates than separate, univariate meta-analyses (Riley, Abrams, Lambert, et al., 2007; Riley, Abrams, Sutton, et al., 2007). Although it appears desirable to examine the impact of covariates on heterogeneity also in multivariate meta-analysis, extensions of the location-scale model to this setting have not yet been proposed. In this article, we close this gap, because we think that multivariate location-scale meta-analysis models allow to answer research questions beyond those covered by their univariate counterparts and because they can be more efficient compared to these models.

With regard to the first point, while univariate location-scale meta-analysis allows to examine the impact of one or several (continuous or categorical) study-level covariates on the mean (location) and between-study variance (scale) of *one* outcome, its multivariate extension can be used to simultaneously assess how study-level covariates affect the means of *multiple* outcomes, their between-study variances, *and* their correlations. In consequence, multivariate location-scale meta-analysis allows to examine between-study differences in variability while correctly accounting for dependencies between outcomes. Beyond that, it allows to *model* between-study differences in correlations between outcomes. When there are more than two outcomes, modeling their correlations is challenging because the variance-covariance matrix must be positive semidefinite. We show how this can be achieved by using a multivariate generalization of the Fisher's Z-transformation (Archakov & Hansen, 2021), which is order-invariant and equivalent to the univariate Fisher's Z-transformation for the bivariate case.

Second, multivariate location-scale meta-analysis may have statistical advantages over univariate location-scale meta-analysis. Research comparing multivariate and univariate meta-analysis showed that estimates from multivariate meta-analysis have better statistical properties than their equivalents from univariate meta-analysis, resulting from the "borrowing of strength" across outcomes (Jackson et al., 2011). In particular, estimates of pooled effects from a multivariate approach are expected to have a higher precision than their univariate counterparts, with efficiency gains being most pronounced when within- or between-study correlations are large, and when some studies do not report estimates for all outcomes of interest (i.e., some effect sizes are missing; Riley, Abrams, Lambert, et al., 2007; Riley, Abrams, Sutton, et al., 2007; Jackson et al., 2011). Furthermore, it has also been suggested that borrowing of strength results in a more efficient estimation of the between-study variances in multivariate meta-analysis (Jackson et al., 2011). We expect that these findings transfer directly to multivariate and univariate location-scale meta-analysis, that is, that gains in efficiency and power can be achieved when a multivariate is chosen over a univariate approach. Investigating the magnitude of these gains is especially relevant because simulations have shown that the power of tests for the scale coefficients is often small for the univariate location-scale model (Blázquez-Rincón et al., 2025). Gains in efficiency and power can also be achieved by means of parameter restrictions (e.g., restricting the parameters for the between-study variances of different outcomes to be equal). An advantage of the multivariate approach is that such parameter restrictions can be easily implemented and tested.

Given these potential advantages of multivariate location-scale meta-analysis, an extension of the existing univariate model to settings with two or more outcomes seems much needed. After reviewing multivariate meta-analysis and meta-regression to establish the notation used in this article, we introduce a bivariate location-scale model which allows to examine the effect of study-level covariates on the means, between-study variances, and the correlation of two outcomes. We then show how this model can be extended to the general multivariate setting. We also describe how this model can be estimated using maximum likelihood (ML) and Bayesian methods. Afterwards, we report the results of a simulation study that we conducted to systematically compare the performance of bivariate location-scale meta-analysis to that of univariate location-scale meta-analysis. We then use the model to reanalyze an example of a recent meta-analysis. In the final section, we discuss theoretical and practical implications of our results and suggest avenues for future research.

Multivariate meta-analysis

Suppose that we have aggregate data from k studies, where each study i ($i = 1, \ldots, k$) provides estimates for p different outcomes, which can for instance represent effect sizes obtained from different study groups, or effect sizes for different outcomes measured in the same group. We denote the $p \times 1$-vector containing the outcomes from study i by $\mathbf{y_i}$. Following multivariate meta-analysis, we assume that $\mathbf{y_i}$ is multivariate normally distributed:

$$\mathbf{y_i}|\boldsymbol{\mu_i} \sim N(\boldsymbol{\mu_i}, \mathbf{S_i}), \tag{1}$$

where $\boldsymbol{\mu_i}$ is a $p \times 1$-vector of the true outcomes for the ith study and $\mathbf{S_i}$ is the within-study variance-covariance matrix of the ith study (i.e., the variance-covariance matrix of the outcomes' sampling errors). Following standard multivariate meta-analysis, we assume that the variances and covariances in this matrix have been estimated and are treated as known.

For the true outcomes, we also assume that they are normally distributed

$$\boldsymbol{\mu_i} \sim N(\boldsymbol{\mu}, \Sigma), \tag{2}$$

with the $p \times 1$-vector of the true outcomes' expectation $\boldsymbol{\mu}$ and the between-study variance-covariance matrix Σ. The marginal distribution of $\mathbf{y_i}$ is then given by

$$\mathbf{y_i} \sim N(\boldsymbol{\mu}, \mathbf{S_i} + \Sigma). \tag{3}$$

The diagonal entries of Σ refer to the variances of the true outcomes across studies, which we denote by $\boldsymbol{\tau^2} = \mathrm{diag}(\Sigma) = (\tau_1^2, ..., \tau_p^2)$. The off-diagonal entries refer to the covariances of the true outcomes. For two outcomes j and l, the covariance can be written as $\rho_{jl}\sqrt{\tau_j^2\tau_l^2}$, where ρ_{jl} is the correlation of the true outcomes j and l ($j \in (1, ..., p)$ and $l \in (1, ..., p)$, $j \neq l$) across studies. In multivariate meta-analysis, Σ is assumed to be constant across studies (Jackson et al., 2011). Later, we describe how this assumption can be relaxed by using covariates to model the between-study variances and correlations.

Multivariate meta-regression

Covariates for $\boldsymbol{\mu}$ can be included in multivariate meta-analysis by replacing $\boldsymbol{\mu}$ by $\mathbf{X_i}\boldsymbol{\beta}$, where $\mathbf{X_i}$ is a $p \times q$-design matrix containing the covariates for the ith study, and $\boldsymbol{\beta}$ is a $q \times 1$-vector of regression coefficients. The p different outcomes can be regressed on the same or different covariates. When different covariates are used, we model the mean of the jth outcome by $\mathbf{x_{ij}}\boldsymbol{\beta_j}$, where $\mathbf{x_{ij}}$ is a $1 \times q_j$-covariate vector and $\boldsymbol{\beta_j}$ is a $q_j \times 1$-vector of regression coefficients. We then write

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} & \cdots & \boldsymbol{\beta_p} \end{pmatrix}^T \tag{4}$$

and

$$\boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{x_{i1}} & 0 & \cdots & 0 \\ 0 & \boldsymbol{x_{i2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x_{ip}} \end{pmatrix}, \tag{5}$$

as suggested by White et al. (2012). It is also possible to use the same covariates for all outcomes (i.e., by restricting $\mathbf{x_{ij}}$ to be equal for all outcomes). In addition, regression coefficients can be restricted to be the same for all outcomes by restricting $\mathbf{x_{ij}}$ and $\boldsymbol{\beta_j}$ to be equal for all outcomes.

Multivariate location-scale meta-analysis

In the following, we first describe how to extend a bivariate meta-regression model to a location-scale model ($p = 2$). We then extend this model to situations with more than two outcomes ($p > 2$). After that, we present different ways of exploring heterogeneity with these models, before providing details on their estimation.

The bivariate location-scale model

To obtain a model for bivariate location-scale meta-analysis, we first relax the assumption that Σ is constant by modeling the elements of the between-study variance-covariance matrix as dependent on study-level covariates. Hence, we replace Σ by Σ_i in Eqs. (2) and (3), where

$$\Sigma_i = \begin{pmatrix} \tau_{1i}^2 & \rho_i\tau_{1i}\tau_{2i} \\ \rho_i\tau_{1i}\tau_{2i} & \tau_{2i}^2 \end{pmatrix} \tag{6}$$

For the between-study variances $\boldsymbol{\tau_i^2} = \mathrm{diag}(\Sigma_i)$, we write

$$\boldsymbol{\tau_i^2} = \exp(\mathbf{Z_i}\boldsymbol{\gamma}), \tag{7}$$

where $\mathbf{Z_i}$ is a $2 \times s$-design matrix containing the covariates for the ith study and $\boldsymbol{\gamma}$ is an $s \times 1$-vector of regression coefficients. We use the exponential function here to ensure that the values predicted for the between-study variance are positive. The covariates for the different elements in $\boldsymbol{\tau_i^2}$ can either be different (this is achieved by specifying $\boldsymbol{\gamma}$ and $\mathbf{Z_i}$ analogous to Eqs. (4) and (5)) or the same. Furthermore, we can either use the same or different covariates to model the mean and variance of each outcome. The former is achieved by setting $\mathbf{Z_i}$ equal to $\mathbf{X_i}$.

We model the between-study correlation using

$$\rho_i = \tanh(\mathbf{r_i}\boldsymbol{\delta}), \tag{8}$$

where $\mathbf{r_i}$ is a $1 \times r$-vector containing the covariates for the ith study and $\boldsymbol{\delta}$ is an $r \times 1$-vector of regression coefficients. Here, we use the tanh-function (i.e., the hyperbolic tangent, also the inverse of the Fisher's Z-transformation) to achieve that the predicted correlations lie between -1 and 1.

Extending the model to more than two outcomes

If we want to model the between-study variances of more than two outcomes, we use the same regression model as proposed for two outcomes (cf. Eq. (7)). However, modeling the correlations with covariates is more demanding than in the bivariate case, because one needs to ensure that $\boldsymbol{\Sigma_i}$ is positive semidefinite for all i. Simply using the regression model given in Eq. (8) for all $p(p-1)/2$ correlations can give rise to improper solutions because the bounds for a (Fisher's Z-transformed) correlation in a multivariate model depend on the other correlations in a nonlinear way (Olkin, 1981). To ensure that we obtain positive semidefinite variance-covariance matrices, we model the elements of an unconstrained $p(p-1)/2 \times 1$-vector $\boldsymbol{\psi}_i$, and then use a multivariate transformation to map this vector to a correlation matrix $\mathbf{P_i}$.

Specifically, we model $\boldsymbol{\psi}_i$ using:

$$\boldsymbol{\psi_i} = \mathbf{R_i}\boldsymbol{\delta}, \tag{9}$$

where $\mathbf{R_i}$ is a $p(p-1)/2 \times r$-matrix that contains covariates, and $\boldsymbol{\delta}$ is an $r \times 1$-vector of regression coefficients. Then, we transform $\boldsymbol{\psi_i}$ with a function g to obtain a correlation matrix, that is $\mathbf{P_i} = g(\boldsymbol{\psi_i})$, where g is a multivariate extension of the inverse Fisher's Z-transformation described by Archakov and Hansen (2021). More specifically, the function g is based on the conception that the elements in $\boldsymbol{\psi_i}$ are the lower diagonal elements of the matrix logarithm of $\mathbf{P_i}$. In practice, numerical methods are required to compute it. Advantages of this approach are that any vector $\boldsymbol{\psi_i}$ maps to a *unique* correlation matrix $\mathbf{P_i}$ (and vice-versa). In addition, it is invariant to the ordering of outcomes, and it is equivalent to the univariate Fisher's Z-transformation in the bivariate case. We provide more details on this transformation in the appendix.

The same or different covariates can be used to model the different elements of $\boldsymbol{\psi}_i$. Note, however, that when using different covariates, the value of a covariate for a specific element of $\boldsymbol{\psi}_i$ may not only influence one, but several correlations. This is because g does not map the single elements of $\boldsymbol{\psi}_i$ onto the single elements of the vector of correlation coefficients, but rather, the whole vector $\boldsymbol{\psi}_i$ is mapped onto a correlation matrix $\mathbf{P_i}$.

When $p > 2$ and some studies only report a subset of the outcomes that are of interest, i.e., when there are missing values, there are two possible approaches to obtaining $\mathbf{P}_{i_i}$ (and thus, $\boldsymbol{\Sigma_i}$), which are detailed in the appendix.

Exploring heterogeneity

Based on the estimated coefficients of the multivariate location-scale model, we can obtain a predicted variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_i$ for each of the k studies in our meta-analysis. Similarly, we can obtain a predicted variance-covariance matrix for a hypothetical new study by choosing values for the covariates in $\mathbf{Z_i}$ and $\mathbf{R_i}$. This allows us to study how, for instance, the value of the (residual) between-study variance of a certain outcome depends on the level of one or several covariates.

In order to facilitate assessing the amount of heterogeneity, meta-analyses often report additional statistics such as I^2 (Higgins & Thompson, 2002) or prediction intervals (PIs; Higgins et al., 2009). For I^2, different suggestions have been made regarding its extension to multivariate meta-analysis (see, e.g., Jackson et al., 2012; White, 2011). Some of these suggestions involve estimating a single I^2 for a model with multiple outcomes, which does not seem to align well with the flexibility of multivariate location-scale modeling. We therefore propose to rely on the definition of I^2 that is also used in meta-regression (e.g., within the metafor package; Viechtbauer, 2010), because it can easily be extended in a way that allows to estimate I^2 per outcome and in addition allows to consider the influence of covariates in the scale part of the model ($\mathbf{Z_i}$) on the value of I^2. Specifically, let $\mathbf{X}$ be $(\mathbf{X_1} \cdots \mathbf{X_k})^T$, and let $\mathbf{S}$ be a block-diagonal matrix which contains the within-study variance covariance matrices $\mathbf{S_1}, \ldots, \mathbf{S_k}$. Then, we can estimate I^2_{ij} (i.e., I^2 for outcome j conditional on the values of the covariates in study i) by

$$I^2_{ij} = 100\% \cdot \frac{\hat{\tau}^2_{ij}}{\hat{\tau}^2_{ij} + \tilde{s}}, \tag{10}$$

where $\hat{\tau}^2_{ij}$ is the between-study variance of outcome j predicted for study i using Eq. (7), and

$$\tilde{s} = \frac{\sum_{i=1}^{k} p_i - q}{\mathrm{tr}\left(\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{S}^{-1}\right)} \tag{11}$$

is an estimate of the "average" within-study variance across outcomes. If the outcomes have different scales, we recommend estimating $\tilde{s}$ separately per outcome.

A $100 \cdot (1-\alpha)\%$ prediction interval for outcome j conditional on the values of the covariates in study i can be obtained as

$$\hat{\mu}_{ij} \pm z_{(1-\alpha/2)}\sqrt{SE(\hat{\mu}_{ij})^2 + \exp\left(\mathbf{z_{ij}}\hat{\boldsymbol{\gamma}}_j\right)}, \tag{12}$$

where $\hat{\mu}_{ij} = \mathbf{x_{ij}}\hat{\boldsymbol{\beta}}_j$, $z_{(1-\alpha/2)}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution, $SE(\hat{\mu}_{ij})$ is the

standard error of the prediction for outcome j in study i, and $\mathbf{z_{ij}}$ and $\boldsymbol{\gamma}_j$ are defined analogously to $\mathbf{x_{ij}}$ and $\boldsymbol{\beta}_j$.

Estimation

ML estimation

The log-likelihood function for the multivariate location-scale model is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = -\frac{kp}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{k}\left(\log(|\mathbf{M_i}|) + (\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta})^T\mathbf{W_i}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta})\right), \tag{13}$$

where $\mathbf{M_i} = \mathbf{S_i} + \boldsymbol{\Sigma_i}$ and $\mathbf{W_i} = \mathbf{M_i}^{-1}$. Note that the only difference of this log-likelihood function to the log-likelihood function of standard multivariate meta-regression is that $\boldsymbol{\Sigma_i}$ is not constant across studies, but defined in terms of Eqs. (7) and (9). Equation (13) can be easily adapted to situations where any of the outcomes are missing at random by using a full information ML approach (Nam et al., 2003).

To obtain the ML estimates, Eq. (13) can be maximized simultaneously over $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ using a numerical optimization routine. Alternatively, one can estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{k}\mathbf{X_i}^T\mathbf{W_i}\mathbf{X_i}\right)^{-1}\sum_{i=1}^{k}\mathbf{X_i}^T\mathbf{W_i}\mathbf{y_i}, \tag{14}$$

and after substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in Eq. (13), maximize the resulting (profile) log-likelihood $l(\boldsymbol{\gamma}, \boldsymbol{\delta}; \hat{\boldsymbol{\beta}})$ over $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Since ML estimation is known to yield biased estimates of between-study variances, we use restricted ML (REML) estimation here. In this case, the restricted log-likelihood that is maximized is given by:

$$l_{REML}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) + \frac{q}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{k}\left(\log(|\mathbf{X_i}^T\mathbf{X_i}|) - \log(|\mathbf{X_i}^T\mathbf{W_i}\mathbf{X_i}|)\right). \tag{15}$$

Just like in ML estimation, one can replace $\boldsymbol{\beta}$ by Eq. (14) and maximize $l_{REML}(\boldsymbol{\gamma}, \boldsymbol{\delta}; \hat{\boldsymbol{\beta}})$ over $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$.

Standard errors for the elements of $\hat{\boldsymbol{\beta}}$ can be obtained by computing

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{k}\mathbf{X_i}^T\mathbf{W_i}\mathbf{X_i}\right)^{-1}. \tag{16}$$

For the elements in $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$, we compute standard errors by taking the inverse of the (numerically obtained) Hessian of the negative (restricted) log-likelihood function. Confidence intervals and tests for each parameter can then be obtained based on a normal approximation.

Bayesian estimation

When the number of studies or the sample sizes within studies are small, ML and REML estimation may fail to converge or yield estimates of the correlation which are close to or on the boundary of the parameter space (i.e., −1 or 1, see Riley, Abrams, Sutton, et al., 2007). Hence, we also consider Bayesian estimation, which can help to stabilize estimates by incorporating additional information through prior distributions. The goal of Bayesian estimation is to obtain an estimate of the joint posterior distribution of the parameters. This is achieved by utilizing the fact that the posterior distribution is proportional to the likelihood times the prior distribution of the parameters:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}), \tag{17}$$

where the likelihood L is defined by (exponentiating) Eq. (13). When defining the prior distribution $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$, we assume that the parameters are independent:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{j=1}^{q}p(\beta_j) \cdot \prod_{k=1}^{s}p(\gamma_k) \cdot \prod_{l=1}^{r}p(\delta_l), \tag{18}$$

where the products are over the elements of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, respectively. For the parameters in $\boldsymbol{\beta}$, we consider uninformative normal priors with zero means and variances of 10,000. Because there was no previous research on prior distributions for multivariate location-scale models, we studied the performance of different sets of prior distributions for the variance and correlation parameters. Based on results from pre-simulations,[1] we use $N(-2.3, 2^2)$ priors for variance parameters that are intercepts. The mean of this distribution is equal to the (log-transformed) average estimate of τ^2 obtained from a large data set of between-study variance estimates reported in meta-analyses of standardized mean differences (Van Erp et al., 2017). For the remaining variance parameters, we use $N(0, 2^2)$ priors. For the correlation parameters, we use $N(0, 0.7^2)$ priors. It is important to consider how informative priors for variance and correlation parameters (that are on the logarithmic or Fisher's Z-scale, respectively) are when transformed to the

[1] In these pre-simulations, we considered normal priors with different variances, as well as half-normal priors or inverse-gamma priors for the intercepts of the variance parameters and a uniform prior for the correlation intercept.

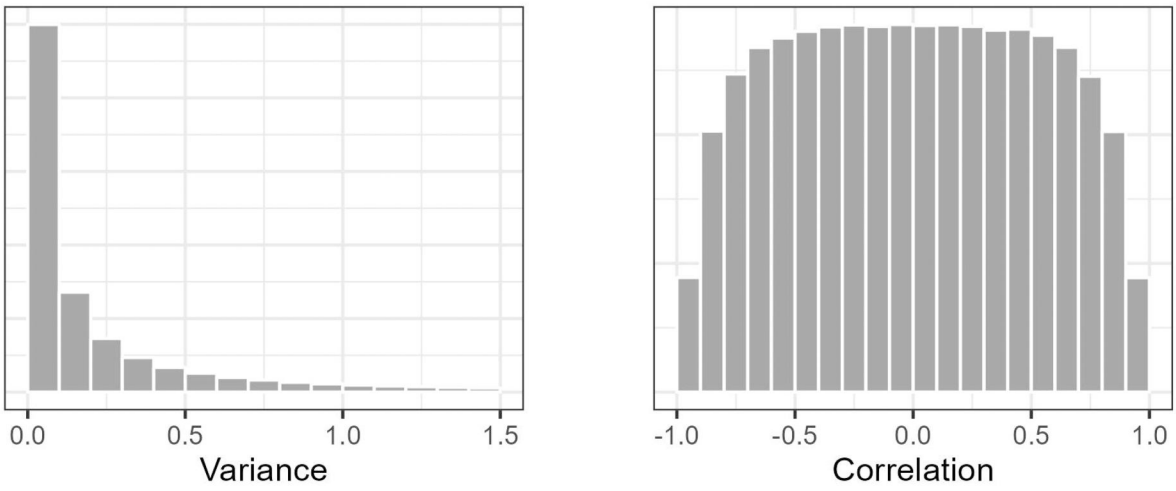


Figure 1. Prior distributions on the variance and correlation scale.
Note. The plot on the left-hand side shows a histogram of 1,000,000 values that were generated from a $N(-2.3, 2^2)$ distribution and transformed by taking the exponential. The plot on the right-hand side shows a histogram of 1,000,000 values that were generated from a $N(0, 0.7^2)$ distribution and transformed by applying the inverse Fisher's Z-transformation.

original scale. Figure 1 illustrates the $N(-2.3, 2^2)$ and the $N(0, 0.7^2)$ prior on the variance and correlation scale, respectively. When choosing priors, one should take into account that the wider the (normal) prior for the correlation parameters, the more weight is allocated to extreme correlations (i.e., close to 1 or -1). As our pre-simulations showed, wider priors for the correlation parameters lead to a larger variability of the estimates, and wider credible intervals.

Monte Carlo simulation

In order to systematically examine the performance of multivariate location-scale meta-analysis in different settings and compare it to that of its univariate counterpart, we conducted a simulation study in which we considered different scenarios. In all scenarios, we generated meta-analytic data with two outcomes,[2] and kept the means of these outcomes constant (at zero). We varied the sample sizes within studies, the effect of a dichotomous predictor on the between-study variances, and the effect of the same dichotomous predictor on the between-study correlation. In addition, we considered conditions in which there were missings in one of the outcomes. We generated 1,000 replications per scenario.

[2]We ran an additional, less extensive simulation study with three outcomes that included two of the scenarios of the main simulation study to test our code. The results of this simulation were similar to those obtained for two outcomes.

Data generation

In each simulation replication, we generated data for k studies, where k was either 30, 50, or 100. We selected these numbers based on a review by Tipton et al. (2019), who found that the average number of studies included in meta-regression analyses ranged from 29 to 99 across different journals publishing meta-analyses on psychological, educational, or medical topics. For each study, data for two outcomes were generated, which are denoted by outcome A and outcome B in the following. In addition, the level of a dichotomous predictor p_i was sampled from a $\text{Ber}(0.5)$ distribution for each study. The overall means of both outcomes (μ^A and μ^B) were both set to zero, that is, no predictor was included in the location model. The design matrix for the between-study variances was defined as

$$\mathbf{Z_i} = \begin{pmatrix} 1 & p_i & 0 & 0 \\ 0 & 0 & 1 & p_i \end{pmatrix}, \tag{19}$$

and the design matrix for the correlation between the true outcomes was

$$\mathbf{R_i} = \begin{pmatrix} 1 & p_i \end{pmatrix}. \tag{20}$$

The true outcomes (μ_i^A, μ_i^B) of each study were sampled from a bivariate normal distribution with zero means and variance-covariance matrix as defined by Eqs. (7) and (8). Across simulation conditions, we varied the values of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. The values of these parameters are given in Table 1. For $\boldsymbol{\gamma}$, the first condition resulted in a between-study variance of 0.05 regardless of the value of p_i, while for the second

Table 1. Values of the simulation parameters.

Parameter	Values
k	30, 50, 100
γ	$(\log(0.05), 0, \log(0.05), 0)^T$, $(\log(0.05), \log(3), \log(0.05), \log(3))^T$,
δ	$(\text{atanh}(0.3), 0)^T$, $(\text{atanh}(0.3), \text{atanh}(0.7) - \text{atanh}(0.3))^T$
n	"small" (10–30), "medium" (30–60), "large" (90–120)

condition, the between-study variance was 0.05 for $p_i = 0$ and 0.15 for $p_i = 1$. For $\boldsymbol{\delta}$, the first condition implied a (medium-sized) correlation of 0.3 regardless of the value of p_i and the second condition implied a correlation of 0.3 for $p_i = 0$ and a correlation of 0.7 for $p_i = 1$. We chose to simulate a fairly large difference in correlations in the second condition in order to be better able to detect any potential effects of this manipulation.

For each study, the sample size n was generated from a uniform distribution. We varied the average sample size and used uniform distributions with different bounds to generate small, medium-sized and large samples. The average sample sizes were 20, 45, and 105. The numbers chosen for small and medium sample sizes are representative of meta-analyses in clinical psychology (cf., e.g., Rubio-Aparicio et al., 2018). In education or medicine, sample sizes tend to be larger (cf., e.g., Ahn et al., 2012; Davey et al., 2011), so we also included conditions with a larger average sample size.

Raw data were simulated for a control and a treatment group: For the treatment group, n data points were simulated from a bivariate normal distribution with mean (μ_i^A, μ_i^B) and variance-covariance matrix

$$\begin{pmatrix} 1 & r_i \\ r_i & 1 \end{pmatrix}, \tag{21}$$

where r_i is the within-study correlation which was sampled from a $\text{Unif}(0.1, 0.5)$ distribution such that on average, there was a medium-sized correlation. For the control group, n data points were simulated from a bivariate normal distribution with mean $(0, 0)$ and variance-covariance matrix as given by (21).

The values of the simulation parameters are summarized in Table 1. Fully crossing these values resulted in 36 simulation conditions. Furthermore, each condition was replicated with an additional manipulation: In these 36 additional conditions, 50% of the observed values of outcome B were deleted at random before conducting the analyses, in order to reflect scenarios in which some studies only report information on one outcome.

Models and estimation

For each of the generated data sets, we fitted bivariate location-scale models as well as univariate location-scale models per outcome using REML and Bayesian estimation. Models were fitted in R (R Core Team, 2022). In all models, the dichotomous predictor was only used to model the between-study variances (and the correlation, in bivariate location-scale models), while it was not included in the location part.

For REML estimation, we used our own implementation, for which the code can be downloaded from the OSF directory associated with this manuscript. Bayesian location-scale models were fitted using Stan (Stan Development Team, 2024) with cmdstanr (Gabry et al., 2024). For each model, four chains were sampled, with 5,000 iterations per chain. In addition, burn-in phases of 2,500 iterations were used. These values were determined in pre-simulations and usually resulted in reasonable effective sample sizes (ESS). Posterior medians were used as point estimates throughout.

Performance measures

We first evaluated the estimability of all models in each simulation condition. For models estimated with REML, estimability was defined in terms of the percentage of trials in which both point estimates and standard errors were obtained for all parameters. For models estimated using Bayesian methods, estimability was defined as the percentage of trials in which the ESS for all parameters was at least 400.[3] The remaining performance measures were calculated based on the simulation replications in which estimability was achieved.

Each model's performance was assessed by computing median bias, a robust version of the root mean squared error (RMSE), and coverage (of confidence or credible intervals) for all parameters. In addition, we computed the power of tests for those variance parameters indicating a between-condition difference in variances (i.e., γ_2 and γ_4), as well as for the parameter indicating a between-condition difference in correlations (i.e., δ_2). Let θ denote the true value of the parameter under consideration, $\hat{\theta}$ its estimator, and $\hat{\theta}_i$ its estimate in the ith replication. Median bias was computed as $\overline{\theta} - \theta$, where $\overline{\theta}$ was the median of the

[3] An ESS of 400 per parameter is recommended as the minimum ESS required to obtain accurate estimates, while an ESS of 1,000 is considered optimal (see, e.g. Zitzmann & Hecht, 2019). In applications, we therefore recommend to aim for an ESS of at least 1,000. Larger ESS can be obtained by increasing the number of iterations per chain.

Table 2. Point and interval estimation of the mean parameters in conditions with 50 studies and missings on the second outcome.

			REML estimation			Bayesian estimation		
Parameter	Measure	Model	small *n*	moderate *n*	large *n*	small *n*	moderate *n*	large *n*
μ^A	Median Bias	Biv	−0.002	−0.002	−0.002	−0.003	−0.003	−0.001
	Median Bias	Uni	−0.003	−0.002	−0.002	−0.003	−0.002	−0.001
	Robust RMSE	Biv	0.057	0.050	0.041	0.056	0.049	0.041
	Robust RMSE	Uni	0.056	0.050	0.042	0.056	0.050	0.042
	Coverage	Biv	0.950	0.922	0.930	0.950	0.933	0.940
	Coverage	Uni	0.949	0.929	0.934	0.953	0.937	0.942
μ^B	Median Bias	Biv	−0.001	0.002	0.003	−0.001	0.004	0.003
	Median Bias	Uni	0.001	0.005	0.001	−0.000	0.005	0.002
	Robust RMSE	Biv	0.079	0.069	0.056	0.079	0.067	0.055
	Robust RMSE	Uni	0.081	0.070	0.056	0.081	0.067	0.056
	Coverage	Biv	0.925	0.916	0.910	0.931	0.939	0.946
	Coverage	Uni	0.925	0.930	0.930	0.936	0.945	0.953

Note. Biv = bivariate model, Uni = univariate model, *n* = sample size.

estimates of all available simulation replications. The robust RMSE was obtained as

$$\text{RMSE}(\hat{\theta}) = \sqrt{(\overline{\theta} - \theta)^2 + \text{MAD}(\hat{\theta})^2}, \tag{22}$$

where MAD is the median absolute deviation calculated as $\text{MAD}(\hat{\theta}) = 1.4826 \times \text{median}(|\hat{\theta}_i - \overline{\theta}|)$ (Salditt & Nestler, 2023). The robust RMSE was used in order to avoid that results were affected by outliers, which often occurred in the estimation of variance parameters, especially when using univariate location-scale models. Coverage was calculated as the percentage of simulation replications in which the 95% Wald-type confidence interval (for models estimated with REML) or the 95% credible interval (for Bayesian models) covered the true value of the parameter. Finally, power was evaluated as the percentage of simulation replications in which the confidence or credible interval did not cover zero.

For both the bivariate and the univariate location-scale model, we expect performance to be best (indicated by a low bias, a small robust RMSE, a close-to-nominal coverage, and a high power) for large numbers of studies, large samples, and no missings. In line with previous research on multivariate meta-analysis, we anticipate the greatest performance advantage for the bivariate location-scale model in conditions with missings (cf., e.g., Riley, Abrams, Lambert, et al., 2007). In particular, we expect the robust RMSEs of the bivariate model to be notably smaller than those of the separate univariate models, and we also expect a higher power in these conditions. With regard to the remaining conditions, we also expect performance of the bivariate model to be slightly better than that of its univariate counterpart, albeit differences between the two may be small (Jackson et al., 2011).

Results of the simulation study

Estimability

For univariate and bivariate location-scale models estimated with REML, estimability ranged from 75.4 – 100% and 71.1 – 100%, respectively, across simulation conditions. Estimability was lower in conditions with smaller numbers of studies, smaller sample sizes, and missings. With Bayesian estimation, full estimability was achieved in all conditions, indicating a sufficient number of iterations per chain.

Point and interval estimation

In the following, we describe the performance of the models regarding point and interval estimation for the mean parameters μ^A and μ^B, the variance parameters γ_1 to γ_4 and the correlation parameters δ_1 and δ_2. We only show the results for 50 studies here. For 30 studies, the results were similar. For 100 studies, performance was notably better for all models (in particular, the biases and robust RMSEs were smaller). Moreover, for mean and variance parameters, we only show the results for conditions with missings on the second outcome, as the results for the parameters associated with the first outcome (μ^A, γ_1 and γ_2) were similar in conditions with and without missings. Whenever appropriate, we summarize results across simulation conditions with different values of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Figures showing the results for all performance measures, parameters and simulation conditions are available in the Supplement.

Mean parameters μ^A and μ^B

Table 2 shows the median bias, the robust RMSE, and coverage for the mean parameters in conditions with 50 studies and missings on the second outcome averaged across the different values of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Results are grouped according to the model (bivariate

Table 3. Point and interval estimation of the variance parameters in conditions with 50 studies and missings on the second outcome.

			REML estimation						Bayesian estimation					
			$\gamma_2 = \gamma_4 = 0$			$\gamma_2 = \gamma_4 = \log(3)$			$\gamma_2 = \gamma_4 = 0$			$\gamma_2 = \gamma_4 = \log(3)$		
Parameter	Measure	Model	small n	moderate n	large n	small n	moderate n	large n	small n	moderate n	large n	small n	moderate n	large n
γ_1	Median Bias	Biv	−0.206	−0.071	−0.064	−0.172	−0.078	−0.061	−0.598	−0.192	−0.049	−0.255	−0.068	0.004
	Median Bias	Uni	−0.180	−0.095	−0.067	−0.248	−0.097	−0.066	−0.545	−0.169	−0.049	−0.226	−0.058	−0.004
	Robust RMSE	Biv	0.968	0.579	0.437	0.909	0.580	0.431	0.953	0.604	0.427	0.664	0.512	0.407
	Robust RMSE	Uni	0.933	0.605	0.438	1.054	0.603	0.437	0.918	0.588	0.422	0.654	0.508	0.406
	Coverage	Biv	0.973	0.976	0.982	0.972	0.975	0.980	0.983	0.964	0.952	0.990	0.978	0.954
	Coverage	Uni	0.970	0.976	0.983	0.969	0.977	0.979	0.984	0.965	0.956	0.990	0.979	0.955
γ_2	Median Bias	Biv	0.013	−0.021	−0.019	−0.014	−0.003	−0.003	−0.548	−0.198	−0.069	−0.344	−0.090	−0.079
	Median Bias	Uni	−0.192	−0.004	−0.015	0.102	0.028	0.004	−0.503	−0.167	−0.049	−0.312	−0.068	−0.071
	Robust RMSE	Biv	1.436	0.849	0.593	1.120	0.704	0.545	0.956	0.817	0.577	0.897	0.619	0.510
	Robust RMSE	Uni	1.665	0.871	0.601	1.284	0.714	0.558	0.955	0.806	0.575	0.892	0.613	0.519
	Coverage	Biv	0.999	0.998	0.965	0.987	0.979	0.952	0.993	0.975	0.948	0.982	0.973	0.946
	Coverage	Uni	0.998	0.995	0.965	0.985	0.980	0.952	0.991	0.977	0.951	0.979	0.973	0.951
γ_3	Median Bias	Biv	−0.057	−0.057	−0.078	−0.036	−0.045	−0.072	−0.540	−0.286	−0.078	−0.206	−0.039	0.039
	Median Bias	Uni	−0.064	−0.083	−0.098	−0.118	−0.073	−0.080	−0.475	−0.229	−0.083	−0.151	−0.016	0.026
	Robust RMSE	Biv	1.123	0.710	0.609	1.094	0.712	0.611	0.928	0.780	0.578	0.746	0.595	0.527
	Robust RMSE	Uni	1.225	0.774	0.620	1.304	0.770	0.599	0.895	0.739	0.574	0.721	0.592	0.532
	Coverage	Biv	0.962	0.976	0.979	0.957	0.976	0.979	0.996	0.977	0.956	0.996	0.979	0.969
	Coverage	Uni	0.959	0.976	0.983	0.960	0.979	0.987	0.995	0.978	0.968	0.993	0.981	0.973
γ_4	Median Bias	Biv	−0.193	−0.070	−0.033	−0.237	−0.014	−0.010	−0.641	−0.364	−0.134	−0.861	−0.238	−0.156
	Median Bias	Uni	−0.504	−0.085	−0.043	−0.126	0.037	−0.011	−0.605	−0.308	−0.109	−0.781	−0.219	−0.160
	Robust RMSE	Biv	1.786	1.242	0.899	1.544	1.055	0.785	0.890	0.974	0.847	1.288	0.843	0.699
	Robust RMSE	Uni	2.653	1.343	0.925	1.734	1.071	0.807	0.877	0.969	0.824	1.262	0.820	0.706
	Coverage	Biv	1.000	0.999	0.985	0.991	0.986	0.967	1.000	0.990	0.968	0.987	0.977	0.957
	Coverage	Uni	0.999	0.998	0.992	0.995	0.989	0.973	1.000	0.989	0.975	0.988	0.977	0.961

Note. Biv = bivariate model, Uni = univariate model, n = sample size.

vs. univariate), the estimation method (REML vs. Bayesian), and the sample size (small vs. moderate vs. large). Bias was generally small for all models and in all conditions. Robust RMSEs decreased as the sample size increased, and were similar for univariate and bivariate models as well as for REML and Bayesian estimation. Especially for REML estimation, coverage was slightly below the nominal level, unless for μ^A when sample sizes were small.

Variance parameters γ_1, γ_2, γ_3 and γ_4

Table 3 shows the results for the variance parameters in conditions with 50 studies and missings on the second outcome averaged across the different values of $\boldsymbol{\delta}$. Median bias decreased for larger sample sizes. For the intercept parameters γ_1 and γ_3, a negative bias was observed even for large samples, while estimation of the slope γ_2 and γ_4 was fairly accurate. When REML estimation was used, the bivariate model often yielded a smaller bias compared to the univariate model when samples were small. The reverse was true for Bayesian estimation, which in general yielded larger biases for smaller sample sizes. Bias regarding the slope of the variance model was usually stronger for the outcome in which there were missings (i.e., for γ_4 compared to γ_2).

Robust RMSEs decreased considerably for larger samples, but were relatively large in all conditions. When REML estimation was used, the bivariate model usually had a smaller RMSE than the univariate model, in particular when sample sizes were small or moderate, and more so for the parameters associated with the outcome on which there were missings (i.e., γ_3 and γ_4). Again, the reverse was observed for Bayesian estimation. However, robust RMSEs obtained with Bayesian estimation were often smaller than those obtained with REML estimation.

Coverages were mostly similar for univariate and bivariate models, and tended to be above the nominal level for both estimation methods. They were especially large for the slope parameters γ_2 and γ_4 when the slope was truly zero, indicating that both Wald-type confidence intervals and Bayesian credible intervals were too conservative in this situation. In particular for Bayesian estimation, coverage tended to decrease toward the nominal level for larger sample sizes.

In summary, these results show that variance parameters can be estimated fairly well by bivariate (but also univariate) location-scale models. In terms of bias, Bayesian estimation performed significantly worse than REML estimation when sample sizes were small or moderate, probably due to the impact of prior distributions. However, Bayesian estimation typically yielded smaller RMSEs. When REML estimation was used, the bivariate model outperformed the univariate model in terms of point estimation when sample sizes were small or moderate.

Table 4. Point and interval estimation of the correlation parameters in conditions with 50 studies and moderate or large sample sizes.

Simulation conditions				Parameter: δ_1						Parameter: δ_2					
				REML estimation			Bayesian estimation			REML estimation			Bayesian estimation		
Missings	n	$\gamma_2\ (=\gamma_4)$	δ_2	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage
no	moderate	0	0	0.029	0.464	1.000	−0.079	0.307	0.981	−0.036	0.824	0.998	0.045	0.311	1.000
			0.56	0.013	0.458	0.998	0.051	0.300	0.988	0.290	1.316	0.996	−0.186	0.337	0.999
		log (3)	0	0.025	0.437	0.999	−0.057	0.268	0.989	−0.053	0.569	0.985	0.004	0.287	0.991
			0.56	0.029	0.472	1.000	0.089	0.284	0.988	0.009	0.654	0.990	−0.124	0.320	0.989
	large	0	0	0.010	0.328	0.975	−0.036	0.253	0.968	0.013	0.453	0.966	0.044	0.324	0.982
			0.56	0.009	0.325	0.974	0.039	0.253	0.965	0.082	0.532	0.971	−0.055	0.313	0.979
		log (3)	0	0.013	0.330	0.976	−0.028	0.248	0.977	−0.004	0.404	0.951	0.019	0.288	0.980
			0.56	0.014	0.328	0.976	0.046	0.251	0.971	0.031	0.424	0.961	−0.091	0.303	0.971
yes	moderate	0	0	0.039	0.678	1.000	−0.114	0.393	0.985	−0.035	1.666	1.000	0.044	0.253	1.000
			0.56	0.033	0.720	1.000	0.006	0.372	0.989	0.595	3.373	1.000	−0.307	0.386	1.000
		log (3)	0	0.036	0.653	1.000	−0.089	0.330	0.990	−0.049	0.883	0.998	0.044	0.281	0.999
			0.56	0.031	0.707	1.000	0.084	0.331	0.992	0.129	1.126	0.999	−0.175	0.313	0.998
	large	0	0	0.041	0.505	0.987	−0.053	0.310	0.981	−0.025	0.804	0.990	0.041	0.344	0.995
			0.56	0.045	0.504	0.988	0.058	0.328	0.979	0.131	0.957	0.990	−0.170	0.368	0.996
		log (3)	0	0.040	0.514	0.990	−0.039	0.294	0.990	−0.044	0.689	0.962	0.022	0.328	0.992
			0.56	0.042	0.511	0.987	0.079	0.307	0.983	0.021	0.703	0.984	−0.167	0.366	0.987

Note. Bias = median bias, RMSE = robust RMSE, n = sample size.

Correlation parameters δ_1 and δ_2

Table 4 summarizes the performance of the bivariate model regarding point and interval estimation of the correlation parameters. Results are displayed separately for conditions with and without missings, different sample sizes, different values of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, and different estimation methods. Because results were generally unsatisfactory when samples were small (i.e., biases and RMSEs were very large), we only show the results for conditions with moderate and large samples here, and note that correlation estimates should not be taken at face value when all studies in a bivariate location-scale meta-analysis are small.

Median biases were smaller for larger samples and in conditions without missings, and estimation was fairly accurate in these conditions. For δ_2, large positive biases were observed with REML estimation when γ_2 and γ_4 were zero, δ_2 was not zero and sample sizes were moderate, especially when there were missings. Bayesian estimation, in constrast, tended to show a somewhat less pronounced negative bias in these situations.

Robust RMSEs decreased for larger samples and were smallest when there were no missings. Bayesian estimation often yielded lower RMSEs than REML estimation. Coverage was above the nominal level in all conditions and for both estimation methods.

To summarize, large numbers of studies and sample sizes are required to achieve a good performance with respect to the estimation of correlation parameters. In terms of point estimation of the correlation parameters, Bayesian estimation showed some advantages over REML estimation, especially in terms of efficiency. However, both Wald-type confidence and Bayesian credible intervals for these parameters tended to be too conservative.

Power of Wald-Tests for differences in variances and correlations

Results concerning the power of the tests for the hypothesis $\gamma_4 \neq 0$ are presented in Figure 2. Results for γ_2 were omitted because they were similar to those for γ_4 in conditions without missings. Results were aggregated across conditions with different values of $\boldsymbol{\delta}$, and are displayed separately for different numbers of studies, different sample sizes, and estimation methods. Power was larger for larger sample sizes and numbers of studies, but achieved a satisfactory level only in conditions without missings and large numbers of studies as well as large sample sizes. In general, power was smaller when there were missings. Results were similar for univariate and bivariate models.

Figure 3 shows the results regarding power of the test for the hypothesis $\delta_2 \neq 0$ for conditions with and without missings, different numbers of studies, different sample sizes, different values of $\boldsymbol{\gamma}$, and different estimation methods. In general, power was low, in particular in conditions with missings, small numbers of studies, small and moderate sample sizes, and when γ_2 and γ_4 were zero. The largest power observed was 0.531 for REML estimation for 100 studies, no missings, large sample sizes, and $\gamma_2 = \gamma_4 = \log(3)$.

Taken together, these results underscore the need for a large number of studies with large samples to achieve satisfactory power for testing hypotheses concerning variance and, in particular, correlation

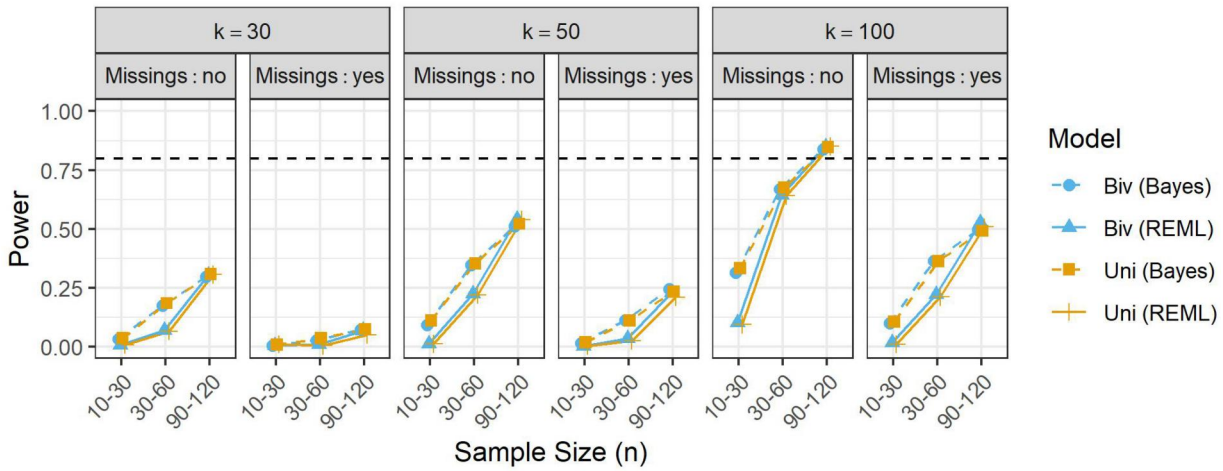


Figure 2. Power of Wald-tests for variance differences across different simulation conditions.
Note. Nested horizontal panels depict the results for different numbers of studies *k*, and conditions for which there are no missings (Missings: no) or 50% missings on the second outcome (Missings: yes). Different colors depict the results for different models (Uni = univariate model, Biv = bivariate model), different linetypes depict different estimation methods (REML, Bayes). The dashed horizontal line indicates a power of .80.

parameters in bivariate location-scale models. This is even more true in situations where some of the outcomes are missing in a significant number of studies.

Illustrative example

In this section, we apply bivariate location-scale meta-analysis to data from a recent meta-analysis by McBreen and Savage (2021), that examined the effect of motivational reading instruction on reading achievement (RA) and reading motivation (RM). Effects were quantified in terms of the standardized mean difference. In the original analysis, RA and RM were analyzed separately, and both outcomes were found to vary significantly between studies. The meta-analysis included 49 studies, of which 43 reported an effect size for RA, and 28 reported an effect size for RM. In total, 22 studies reported effect sizes for both RA and RM. Here, we explore whether bivariate location-scale meta-analysis can yield additional insights, especially regarding the question whether study-level covariates have an impact on the between-study variances and the between-study correlation of RA and RM.

We consider two potential covariates: The first covariate is the study design, that is, whether or not the study was a randomized controlled trial (RCT). It is well established that effects in non-RCT studies can differ from those in RCTs, because they may fail to control for important confounders. We believe that it is also of interest whether effects are more heterogeneous in non-RCT studies than in RCTs, because if they were, this could indicate that heterogeneity arises from between-study differences that are more pronounced in less controlled settings. With bivariate location-scale meta-analysis, we can also examine whether the effects for RA and RM are more or less correlated in non-RCTs vs. RCTs.

The second covariate is the standard error of the respective effect. If it affects the mean, this can indicate publication bias (as in Egger's regression test for publication bias; Egger et al., 1997). With bivariate location-scale meta-analysis, we can also examine whether RA or RM effects are more or less heterogeneous for studies with a larger standard error (i.e., for studies with smaller samples). Finally, we can examine whether the average standard error affects the correlation of RA and RM outcomes.

All models that are presented in the following were fitted using REML estimation. likelihood-ratio tests (LRTs) were used for model comparisons. When the models compared differed in terms of their fixed effects (i.e., how the mean was modeled), LRTs were computed based on the results of the same models refitted with ML estimation. Table 5 gives an overview of the predictors considered in the different models and the results of the model comparisons.

We assumed a sampling correlation of 0.5 between RA and RM effects, but conducted additional sensitivity analyses with sampling correlations of 0.3 and 0.8. We found that model comparisons as well as estimates of mean and variance parameters were robust to the value of the sampling correlation. Estimates of the correlation parameters changed when the sampling correlation was varied, and we describe these changes in footnotes. The R code required to fit all models described below can be obtained from the OSF repository associated with this paper (https://osf.io/4ey2w). There, we also show how

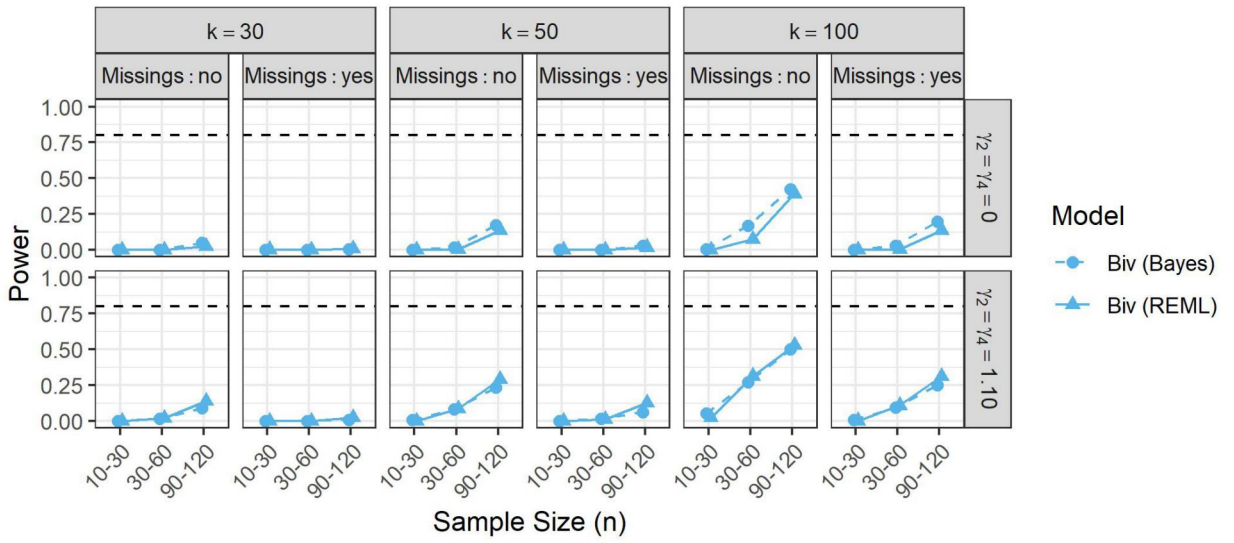


Figure 3. Power of Wald-tests for correlation differences across different simulation conditions.
Note. Nested horizontal panels depict the results for different numbers of studies *k*, and conditions for which there are no missings (Missings: no) or 50% missings on the second outcome (Missings: yes). Vertical panels depict the results for conditions that differ in terms of the value of γ_2 and γ_4. Different linetypes depict different estimation methods (REML, Bayes). The dashed horizontal line indicates a power of .80.

to estimate the final model (see below) using Bayesian estimation.

Before examining the impact of the two covariates on the parameters in the variance-covariance matrix, we aimed to establish a model for the mean, assuming intercept-only models for the variance of each outcome and their correlation. We do this because we want to examine how *residual* heterogeneity, i.e., heterogeneity that cannot be explained, is affected by the available covariates.

Modeling the mean

We first compared a model including only the outcome type (RA vs. RM) as a predictor for the mean (model 1) and a model considering study design as an additional predictor (with separate coefficients for each outcome, model 2). Specifically, for model 1, the mean effect sizes for RA and RM were modeled as

$$\begin{pmatrix} \mu_i^{RA} \\ \mu_i^{RM} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \tag{23}$$

and for model 2, the regression equations were

$$\begin{pmatrix} \mu_i^{RA} \\ \mu_i^{RM} \end{pmatrix} = \begin{pmatrix} 1 & d_i & 0 & 0 \\ 0 & 0 & 1 & d_i \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}. \tag{24}$$

where d_i is a dummy variable indicating whether a study was an RCT ($d_i = 1$) or not ($d_i = 0$). The between-study variance-covariance matrix was constant across studies for both model 1 and model 2:

$$\Sigma_i = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}. \tag{25}$$

A LRT indicated that model 2 did not exhibit a significantly better fit (cf. Table 5).

We then considered a model where the effect of each outcome was moderated by its standard error (model 3). The regression equations for this model were

$$\begin{pmatrix} \mu_i^{RA} \\ \mu_i^{RM} \end{pmatrix} = \begin{pmatrix} 1 & 10 \cdot se_i^{RA} & 0 & 0 \\ 0 & 0 & 1 & 10 \cdot se_i^{RM} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \tag{26}$$

where se_i^{RA} and se_i^{RM} are the standard errors associated with the study-specific effect sizes for RA and RM, respectively. These standard errors were rescaled by multiplying them by 10 in order to facilitate interpretation. Model 3 yielded a significantly better fit than model 1, cf. Table 5. The expected effects for RA and RM for a standard error of zero implied by model 3 were $\hat{\beta}_1 = -0.06$ (95% CI: $[-0.16; 0.04]$) and $\hat{\beta}_3 = 0.01$ (95% CI: $[-0.22; 0.23]$), respectively, indicating that according to our model, RA and RM effects would essentially disappear for an infinitely large precision. The expected change of the effect per 0.1

Table 5. Models considered in the analysis of the illustrative example.

Model	Predictors mean	Predictors variance	Predictors correlation	Restrictions	LRT
1	–	–	–	–	
2	study design	–	–	–	*vs. Model 1*: $\chi^2(2) = 2.99$, $p = .224$
3	SE	–	–	–	*vs. Model 1*: $\chi^2(2) = 37.28$, $p < .001$
4	SE	study design	study design	–	*vs. Model 3*: $\chi^2(3) = 4.93$, $p = .177$
5	SE	SE	average SE	–	*vs. Model 3*: $\chi^2(3) = 22.72$, $p < .001$
6	SE	SE	–	equal variances	*vs. Model 5*: $\chi^2(3) = 2.16$, $p = .541$

Note. Unless otherwise stated in the column *Restrictions*, separate means and between-study variances were estimated for the different outcome types (reading achievement and reading motivation). SE = standard error, LRT = likelihood-Ratio Test.

increase in the standard error was $\hat{\beta}_2 = 0.28$ for RA (95% CI: $[0.21; 0.34]$) and $\hat{\beta}_4 = 0.19$ for RM (95% CI: $[0.06; 0.31]$). The positive slopes indicate that both outcomes might be affected by publication bias. Descriptively, the slope is larger for RA, which is in line with the finding by McBreen and Savage (2021) that a funnel plot for RA showed stronger signs of publication bias than a funnel plot for RM.

Modeling the variance-covariance matrix

We proceeded with model 3, and additionally considered the impact of the study design on residual heterogeneity. We included the predictor in the model for the variance and the correlation simultaneously to assess whether it had any impact on the between-study distribution (model 4). Accordingly, the model-implied variance-covariance matrix varied across studies, with

$$\begin{pmatrix} \log(\tau^2_{i1}) \\ \log(\tau^2_{i2}) \end{pmatrix} = \begin{pmatrix} 1 & d_i & 0 & 0 \\ 0 & 0 & 1 & d_i \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix} \quad (27)$$

and

$$\operatorname{atanh}(\rho_i) = \begin{pmatrix} 1 & d_i \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}. \quad (28)$$

However, a LRT indicated that model 4 did not fit significantly better than model 3 (cf. Table 5). In a fifth model, we explored whether the standard error affected residual heterogeneity, by fitting a model that included the (rescaled) standard error as a predictor for the mean as well as the residual between-study variance of each outcome. The average (rescaled) standard error of both outcomes ($\overline{se}_i$) was included as a predictor for the correlation. Hence, the regression equations for the variances were

$$\begin{pmatrix} \log(\tau^2_{i1}) \\ \log(\tau^2_{i2}) \end{pmatrix} = \begin{pmatrix} 1 & 10 \cdot se_i^{RA} & 0 & 0 \\ 0 & 0 & 1 & 10 \cdot se_i^{RM} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix}, \quad (29)$$

and for the correlation, the equation was

$$\operatorname{atanh}(\rho_i) = \begin{pmatrix} 1 & 10 \cdot \overline{se}_i \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}. \quad (30)$$

According to a LRT, model 5 yielded a significantly better fit than model 3, cf. Table 5. Estimates for the fixed effects were similar to those obtained with model 3 (results not shown). The expected between-study variances for a standard error of zero were $\exp(\hat{\gamma}_1) = 0.001$ and $\exp(\hat{\gamma}_3) = 0.003$ for RA and RM, respectively. The expected increase in the logarithm of the between-study variance per 0.1 increase in the standard error was $\hat{\gamma}_2 = 1.56$ (95% CI: $[0.63, 2.49]$) for RA, and $\hat{\gamma}_4 = 1.55$ (95% CI: $[0.69; 2.41]$) for RM. The expected correlation for a standard error of zero was $\tanh(\hat{\delta}_1) = -0.06$, with a slope of $\hat{\delta}_2 = 0.09$ for the Fisher's Z-transformed correlation (95% CI: $[-0.84; 1.01]$).[4] These results suggest that it might be justified to assume equal effects of the standard error on the variances of RA and RM, and that the correlation might be constant across different values of the average standard error, i.e., $\operatorname{atanh}(\rho_i) = \delta$.

When we fitted a model with these restrictions to the data (model 6), we found that it did not yield a significantly worse fit compared to the model without these constraints (cf. Table 5). Figure 4 illustrates the results for this final model by showing the estimates of the mean for each outcome, as well as the

[4]For a sampling correlation of 0.3 (0.8), the expected correlation for a standard error of zero was 0.27 (−0.53), with a slope of 0.02, 95% CI: $[-0.93, 0.97]$ (0.21, 95% CI: $[-0.87, 1.29]$) for the Fisher's Z-transformed correlation.

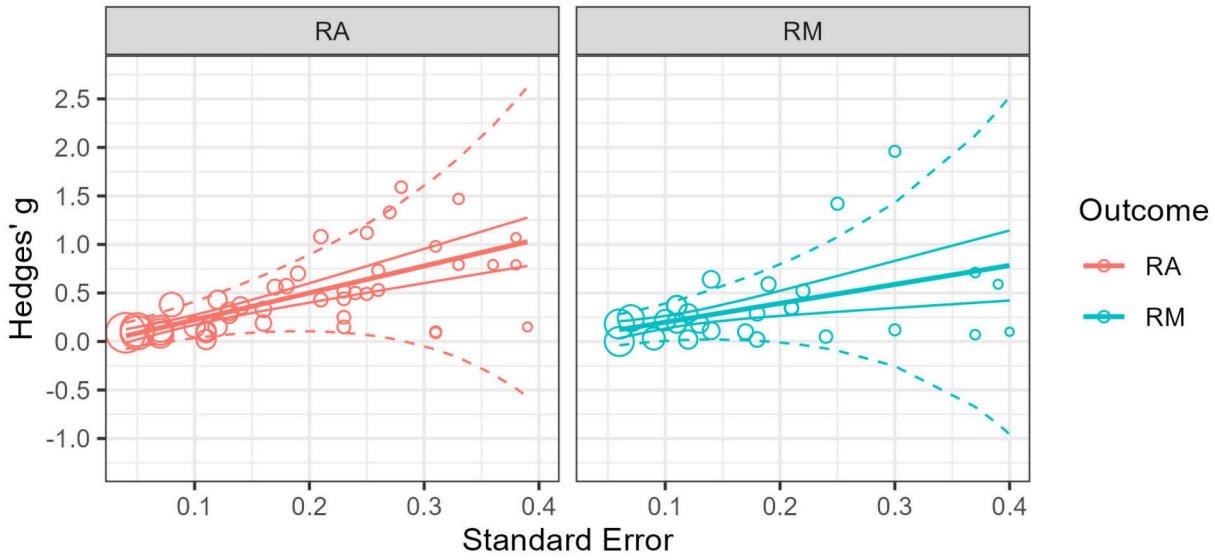


Figure 4. Results of a bivariate location-scale model with the standard error as a predictor for the mean and the variance, with equal effects on the between-study variances of reading achievement (RA) and reading motivation (RM).
Note. Circles represent the observed effects, scaled in proportion to their inverse variance. The thick solid line depicts the estimated mean effect for the respective standard error. The thin solid line depicts the 95% CI around this effect. The dashed line shows the 95% prediction interval.

Table 6. Comparison of univariate and bivariate location-scale meta-analysis for the illustrative example.

	Mean parameters		Variance parameters	
	$\hat{\beta}_0$ (95%− CI)	$\hat{\beta}_{se}$ (95%− CI)	$\hat{\gamma}_0$ (95%− CI)	$\hat{\gamma}_{se}$ (95%− CI)
			Outcome: RA	
Univariate	−0.04 [−0.12, 0.04]	0.26 [0.19, 0.34]	−6.94 [−9.41, −4.48]	1.59 [0.65, 2.53]
Bivariate (Model 5)	−0.05 [−0.13, 0.03]	0.27 [0.20, 0.35]	−6.86 [−9.29, −4.43]	1.56 [0.63, 2.49]
Bivariate (Model 6)	−0.05 [−0.15, 0.04]	0.28 [0.19, 0.36]	−6.26 [−7.73, −4.79]	1.49 [0.90, 2.09]
			Outcome: RM	
Univariate	−0.01 [−0.18, 0.16]	0.21 [0.06, 0.35]	−5.85 [−7.84, −3.87]	1.52 [0.67, 2.37]
Bivariate (Model 5)	0.004 [−0.16, 0.17]	0.19 [0.05, 0.33]	−5.94 [−7.93, −3.94]	1.55 [0.69, 2.41]
Bivariate (Model 6)	0.004 [−0.14, 0.15]	0.19 [0.07, 0.32]	−6.26 [−7.73, −4.79]	1.49 [0.90, 2.09]

Note. $\hat{\beta}_0$ = estimated mean for the respective outcome for a standard error of zero, $\hat{\beta}_{se}$ = estimated increase in the mean for the respective outcome per 0.1 increase of the standard error, $\hat{\gamma}_0$ = estimated logarithmic between-study variance for the respective outcome for a standard error of zero, $\hat{\gamma}_{se}$ = estimated increase in the logarithmic between-study variance for the respective outcome per 0.1 increase of the standard error.

associated confidence interval and prediction interval for different values of the standard error, along with the observed effects.

Comparison with univariate location-scale meta-analysis

To further illustrate the advantages of the bivariate model, we compared its results to those obtained with separate univariate models for RA and RM. Table 6 shows the results of univariate location-scale models in which the standard error was included as a covariate for both the mean and the variance, along with the results of model 5 and model 6. Overall, the results obtained with the univariate model were similar to those from model 5. Due to the restriction that the variance parameters are equal for RA and RM, their estimates obtained from model 6 were different. Importantly, in addition to slightly different estimates, the confidence intervals for the variance parameters obtained with this model were considerably narrower than those obtained with the univariate model or model 5 (cf. Table 6). This illustrates a potential advantage of the bivariate over the univariate location-scale model: Within this framework, parameter restrictions can be tested while simultaneously accounting for dependencies between outcomes. Such parameter restrictions can result in notable gains in precision and power and hence, they should be considered in multivariate (location-scale) meta-analysis when they are meaningful and do not impair model fit.

Discussion

In this article, we have proposed an extension of location-scale meta-analysis to the multivariate setting, which allows to examine the impact of categorical or continuous covariates on the between-study variances of different outcomes as well as on the correlations between outcomes. We systematically evaluated the model's performance in a simulation study, and illustrated its usefulness by applying it to an example examining the effects of motivational reading instruction on reading achievement and reading motivation.

In our illustrative example, we found that the between-study variances of the effects of motivational reading instruction on reading achievement and motivation increased as precision decreased. Combined with the results for the mean structure, this suggests that small studies show larger and more heterogeneous effects for both outcomes. This result is consistent with previous studies examining the impact of sample sizes on heterogeneity in meta-analyses (IntHout et al., 2015; Viechtbauer & López-López, 2022). IntHout et al. (2015) suggested that this finding could arise from different participant characteristics in larger vs. smaller studies, as well as from lower quality standards in smaller studies. We neither found an effect of the study design on the variances, nor did any variable show an impact on the correlation. It is possible that these null findings were due to the small number of studies. As our simulation study showed, the power of tests for variance parameters is small even for moderate numbers of studies (e.g., 50), and the power of tests for correlation parameters is even smaller.

The results of the simulation study partly supported the contention that multivariate location-scale meta-analysis may benefit from "borrowing strength" between outcomes, and in consequence, have statistical advantages over univariate location-scale meta-analysis. Specifically, bivariate location-scale models estimated with REML sometimes yielded less biased and more efficient estimates for the variance parameters than their univariate counterpart when sample sizes were small or when there were missing effect sizes for one of the outcomes. We emphasize that this does not indicate that multivariate location-scale models are a panacea under these circumstances, because we found biases and RMSEs to be largest and power to be insufficiently small for the same conditions. That said, the small performance differences between univariate and multivariate location-scale models were also partly related to the use of robust performance metrics (i.e., median bias and the robust RMSE). Inspecting the simulation results at the level of replications showed that the univariate model produced more outliers (e.g., large negative estimates of the intercepts of the variance models, or large biases in the slopes of the variance models in both directions). Importantly, we did not identify cases in which the multivariate location-scale model yielded a significantly worse performance than its univariate counterpart, and hence, conclude that it should be preferred in location-scale meta-analyses of two (and perhaps, more) outcomes, given it can be estimated for the data at hand.

Coverage rates for the variance and correlation parameters were often somewhat above the nominal level, and were accompanied by a low power for detecting differences in variances or correlations between the levels of a dichotomous predictor. Previous research has shown that profile confidence intervals and likelihood-ratio tests often perform better than their Wald-type equivalents for variance parameters (Viechtbauer, 2007a, 2007b). However, a recent simulation study examining the univariate location-scale model (Blázquez-Rincón et al., 2025) found no clear performance advantage of profile confidence intervals. We reached a similar conclusion based on an additional simulation performed to compare profile and Wald-type confidence intervals for a subset of the conditions considered in our main simulation (for more information, please refer to the OSF repository of this paper). Therefore, further research on interval estimation for variance (and correlation) parameters in location-scale models is required.

When using the Bayesian multivariate location-scale model in applications, it is important to assess the sensitivity of the results to the prior distributions by varying them. We found that for the illustrative example, using wider priors (i.e., priors with a larger standard deviation) for the variance parameters led to slight changes of their point and interval estimates, while using a wider prior for the correlation parameter notably increased the width of its credible interval. It is well known that the choice of prior distributions for (co)variance parameters can have a notable influence on their estimation (Lambert et al., 2005). Against this background, more research on sensible choices for prior distributions in location-scale models is needed.

In order to keep our simulation study feasible, we had to limit the number of scenarios that we examined. Most importantly, the simulation conditions that we considered were restricted to settings with a categorical (dichotomous) predictor. As our illustrative

example has shown, it may be appealing to examine the effect of a continuous predictor (e.g., the standard error) on the between-study variances of the outcomes. Furthermore, a simulation study on the univariate location-scale model has shown that convergence rates and coverages can be affected by the predictor type (Blázquez-Rincón et al., 2025). Therefore, scenarios with continuous predictors should also be given consideration in future simulation studies. In addition, more extensive simulations of the model for three or more outcomes are needed, especially because this model involves an additional step (recovering the correlation matrix from the parameter estimates using numerical methods).

We expect that in practice, the estimation of multivariate location-scale models will become harder the larger the number of outcomes. To give an example, for a model with five outcomes, one could model five means, five variances, and ten correlations, which can quickly result in a (too) large number of parameters to be estimated. In such situations, it may be necessary to impose restrictions, such as considering equal coefficients for certain or all correlation (or variance) models, or resorting to intercept-only models for the correlations (see McShane & Böckenholt, 2023, for a more detailed account on potential restrictions on heterogeneity parameters in meta-analyses). In this regard, it would be interesting to consider whether techniques such as Lasso or Ridge could successfully be used to reduce the parameters in the variance or correlation models (McNeish, 2015; Scharf & Nestler, 2019). In addition to the estimation, the interpretation of (high-dimensional) multivariate location-scale models may be challenging. To this end, we recommend using graphical methods for assessing the impact of covariates on parameters of interest.

When applying the proposed approaches, it is important to note that some practical limitations of multivariate meta-analysis also apply to multivariate location-scale meta-analysis. One limitation is that there is often no information on the covariances of sampling errors, such that they need to be replaced by a "guestimate." If this "guestimate" differs markedly from the true covariance, this can result in severe biases in the estimate of the between-study correlation (Ishak et al., 2008; Wei & Higgins, 2013). In multivariate location-scale meta-analysis, imputing the wrong within-study correlation may further result in erroneous conclusions regarding the effect of covariates on the correlation and hence, extra caution is required when choosing plausible values for the within-study correlations when these effects are of interest.

Here, we considered an extension of the location-scale model to multivariate meta-analysis, with a focus on the bivariate case. The model that we have suggested can be applied to any type of effect size (i.e., not only to standardized mean differences). When means and standard deviations of two outcomes are available separately for a treatment and a control group, it can be interesting to consider a quadrivariate version of our model which allows to separately model heterogeneity per group and outcome (cf. McShane & Böckenholt, 2020, 2023; Van Aert, 2022). In addition, there are further meta-analytic models for multiple outcomes to which location-scale modeling could be extended, such as hierarchical models (e.g., three-level models, see Fernández-Castilla et al., 2021), extensions thereof (e.g., the Correlated and Hierarchical Effects (CHE) model; Pustejovsky & Tipton, 2022), or combinations of hierarchical and multivariate models (McShane & Böckenholt, 2018). It seems warranted to extend location-scale modeling in these directions because the multivariate model presented here is not suited for situations in which studies report more than one effect size per outcome, or in which it is not possible to assign effect sizes to distinct outcome types.

To conclude, multivariate location-scale meta-analysis provides a new and potentially useful tool for modeling complex meta-analytic data structures. It allows to answer new substantive research questions and, in certain situations, has statistical advantages over univariate location-scale meta-analysis. Future research should try to identify more powerful methods to detect effects of covariates on variances and correlations, and consider extending location-scale modeling to further meta-analytic models for studies with multiple outcomes.

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, *82*(4), 436–476. https://doi.org/10.3102/0034654312458162

Aloe, A. M., Shisler, S. M., Norris, B. D., Nickerson, A. B., & Rinker, T. W. (2014). A multivariate meta-analysis of student misbehavior and teacher burnout. *Educational Research Review*, *12*, 30–44. https://doi.org/10.1016/j.edurev.2014.05.003

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

Blázquez-Rincón, D., López-López, J. A., & Viechtbauer, W. (2025). Performance of location-scale models in meta-analysis: A simulation study. *Behavior Research Methods*, *57*(4), 118. https://doi.org/10.3758/s13428-025-02622-5

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*(1), 160. https://doi.org/10.1186/1471-2288-11-160

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fernández-Castilla, B., Aloe, A. M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van Den Noortgate, W. (2021). Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behavior Research Methods*, *53*(2), 702–717. https://doi.org/10.3758/s13428-020-01459-4

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). Cmdstanr: R Interface to 'CmdStan'. Manual.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

IntHout, J., Ioannidis, J. P., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: A meta-meta-analysis. *Journal of Clinical Epidemiology*, *68*(8), 860–869. https://doi.org/10.1016/j.jclinepi.2015.03.017

Ishak, K. J., Platt, R. W., Joseph, L., & Hanley, J. A. (2008). Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine*, *27*(5), 670–686. https://doi.org/10.1002/sim.2913

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, *30*(20), 2481–2498. https://doi.org/10.1002/sim.4172

Jackson, D., White, I. R., & Riley, R. D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine*, *31*(29), 3805–3820. https://doi.org/10.1002/sim.5453

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*(3), 227–235. https://doi.org/10.1037/1082-989X.1.3.227

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*(15), 2401–2428. https://doi.org/10.1002/sim.2112

McBreen, M., & Savage, R. (2021). The impact of motivational reading instruction on the reading achievement and motivation of students: A systematic review and meta-analysis. *Educational Psychology Review*, *33*(3), 1125–1163. https://doi.org/10.1007/s10648-020-09584-4

McNeish, D. M. (2015). Using Lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. https://doi.org/10.1080/00273171.2015.1036965

McShane, B. B., & Böckenholt, U. (2018). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*(1), 255–271. https://doi.org/10.1007/s11336-017-9571-z

McShane, B. B., & Böckenholt, U. (2020). Enriching meta-analytic models of summary data: A thought experiment and case study. *Advances in Methods and Practices in Psychological Science*, *3*(1), 81–93. https://doi.org/10.1177/2515245919884304

McShane, B. B., & Böckenholt, U. (2023). Multilevel multivariate meta-analysis made easy: An introduction to MLMVmeta. *Behavior Research Methods*, *55*(5), 2367–2386. https://doi.org/10.3758/s13428-022-01892-7

Nam, I.-S., Mengersen, K., & Garthwaite, P. (2003). Multivariate meta-analysis. *Statistics in Medicine*, *22*(14), 2309–2333. https://doi.org/10.1002/sim.1410

Olkin, I. (1981). Range restrictions for product-moment correlation matrices. *Psychometrika*, *46*(4), 469–472. https://doi.org/10.1007/BF02293804

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

R Core Team. (2022). *R: A language and environment for statistical computing*. https://www.R-project.org/

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*(10), 982–990. https://doi.org/10.1016/j.jclinepi.2005.02.022

Riley, R. D., Abrams, K. R., Lambert, P. C., Sutton, A. J., & Thompson, J. R. (2007). An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, *26*(1), 78–97. https://doi.org/10.1002/sim.2524

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., & Thompson, J. R. (2007). Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, *7*(1), 3. https://doi.org/10.1186/1471-2288-7-3

Rodriguez, J. E., Williams, D. R., & Bürkner, P.-C. (2023). Heterogeneous heterogeneity by default: Testing categorical moderators in mixed-effects meta-analysis. *The British Journal of Mathematical and Statistical Psychology*, *76*(2), 402–433. https://doi.org/10.1111/bmsp.12299

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2018). A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behavior Research Methods*, *50*(5), 2057–2073. https://doi.org/10.3758/s13428-017-0973-8

Salditt, M., & Nestler, S. (2023). Parametric and nonparametric propensity score estimation in multilevel observational studies. *Statistics in Medicine*, *42*(23), 4147–4176. https://doi.org/10.1002/sim.9852

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 576–590. https://doi.org/10.1080/10705511.2018.1558060

Stan Development Team. (2024). *Stan modeling language users guide and reference manual, version 2.35*. https://mc-stan.org

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

Van Aert, R. C. M. (2022). Analyzing data of a multilab replication project with individual participant data meta-analysis: A tutorial. *Zeitschrift Für Psychologie*, *230*(1), 60–72. https://doi.org/10.1027/2151-2604/a000483

Van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data*, *5*(1), 4. https://doi.org/10.5334/jopd.33

Viechtbauer, W. (2007a). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*(1), 37–52. https://doi.org/10.1002/sim.2514

Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *The British Journal of Mathematical and Statistical Psychology*, *60*(1), 29–60. https://doi.org/10.1348/000711005X64042

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & López-López, J. A. (2022). Location-scale models for meta-analysis. *Research Synthesis Methods*, *13*(6), 697–715. https://doi.org/10.1002/jrsm.1562

Wei, Y., & Higgins, J. P. (2013). Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, *32*(7), 1191–1205. https://doi.org/10.1002/sim.5679

White, I. R. (2011). Multivariate random-effects meta-regression: Updates to mvmeta. *The Stata Journal*, *11*(2), 255–270. https://doi.org/10.1177/1536867X1101100206

White, I. R., Barrett, J. K., Jackson, D., & Higgins, J. P. T. (2012). Consistency and inconsistency in network meta-analysis: Model estimation using multivariate meta-regression. *Research Synthesis Methods*, *3*(2), 111–125. https://doi.org/10.1002/jrsm.1045

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 646–661. https://doi.org/10.1080/10705511.2018.1545232

Appendix A. Transformation of the correlation matrix for more than two outcomes

In the following, we describe the multivariate generalization of the Fisher's Z-transformation proposed by Archakov and Hansen (2021):

Let $\mathbf{P_i}$ be the $p \times p$ correlation matrix of the outcomes in study *i*. To obtain the multivariate Fisher's Z-transformation of $\mathbf{P}_i$, we first compute its matrix logarithm, $\Psi_i = \log(\mathbf{P_i})$ which is defined by the relationship $\exp(\Psi_i) = \mathbf{P_i}$. The matrix Ψ_i is a symmetric $p \times p$ matrix. We then define a vector $\boldsymbol{\psi}_i$, which contains the elements below the main diagonal of Ψ_i. If we denote the operation of vectorizing the elements below the main diagonal of a matrix by vecl, we can write the multivariate Fisher's Z-transformation as

$$f(\mathbf{P_i}) = \text{vecl}(\log(\mathbf{P_i})). \tag{A1}$$

The vector $\boldsymbol{\psi}_i$ is a $p(p-1)/2 \times 1$ real vector, that is, its elements lie between $-\infty$ and ∞.

We now describe how to backtransform a real vector $\boldsymbol{\psi}_i$ to a correlation matrix $\mathbf{P_i}$. This requires numerical methods. We first define a symmetric matrix Ψ_i^*, which contains the elements of $\boldsymbol{\psi}_i$ above and below its main diagonal. We denote the vectorization of the main diagonal of Ψ_i^* by $\boldsymbol{x}^*$. To obtain $\mathbf{P_i}$, we need to find the vector $\boldsymbol{x}^*$ for which $\exp(\Psi_i^*)$ is a correlation matrix. In other words, we need to find the values for the main diagonal of Ψ_i^* for which its matrix exponential has the following characteristics: (1) Its diagonal is a vector of ones and (2) the elements below and above the main diagonal lie between -1 and 1. In the following, we denote the vector $\boldsymbol{x}^*$ for which $\exp(\Psi_i^*)$ is a correlation matrix by $\boldsymbol{x}$. Archakov and Hansen (2021) showed that the following sequence converges to $\boldsymbol{x}$:

$$\boldsymbol{x}^*_{(m+1)} = \boldsymbol{x}^*_{(m)} - \log(\text{diag}(\exp(\Psi^*_{i(m)}))) \qquad (m = 0, 1, 2, \ldots) \tag{A2}$$

where diag describes the operation of vectorizing the main diagonal of a matrix. Hence, we can backtransform $\boldsymbol{\psi}_i$ to $\mathbf{P_i}$ by (1) choosing an arbitrary vector $\boldsymbol{x}^*_{(0)}$, (2) computing the sequence described in Eq. (A2) until convergence, and (3) computing $\mathbf{P_i} = \exp(\Psi_i^*)$ for the final $\boldsymbol{x}^*$.

Appendix B. Approaches to obtaining Σ_i when $p>2$ in scenarios with missings

A situation that is frequently encountered in multivariate meta-analyses is that some studies only report a subset of the outcomes of interest. That is, if we are interested in outcomes A, B, and C, some studies may report all these outcomes, while others may report only two (e.g., only outcomes A and C), or only one outcome (e.g., outcome C).

At the same time, information on study characteristics that serve as covariates for the means, the variances, and the correlations may or may not be available in each study. Here, we assume that covariate values are available for each study, irrespective of its reported outcomes (i.e., studies without covariate information are handled by listwise deletion, or covariate information is imputed using a valid method). In this situation, there are two approaches to obtaining Σ_i during estimation: One approach that retains covariate information of missing outcomes, and one that omits this information. Note that Σ_i is composed of $\tau_i^2 = \exp(\boldsymbol{Z}_i\boldsymbol{\gamma})$ and $\mathbf{P}_i$. This matrix is obtained by computing $\boldsymbol{\psi}_i = \boldsymbol{R}_i\boldsymbol{\delta}$. Furthermore, given all outcomes and all covariate information are available, $\boldsymbol{Z}_i$ has p rows (corresponding to the number of variance terms), and $\boldsymbol{R}_i$ has $p(p-1)/2$ rows (corresponding to the number of correlations).

In the first approach, we compute Σ_i as if all outcomes were available, and then extract the rows and columns corresponding to those outcomes that were reported. For example, assume that $p = 3$ and outcomes are denoted by A, B, and C. For a study that only reports outcomes A and C, we set up $\boldsymbol{Z}_i$ and $\boldsymbol{R}_i$ as if all outcomes were available (i.e., with three rows each) and, based on the current estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, compute the full vectors τ_i^2 and $\boldsymbol{\psi}_i$. From those, we obtain the full variance-covariance matrix Σ_i using the formulas described in this paper. To obtain the variance-covariance matrix for those outcomes that were reported, we then extract the rows and columns of Σ_i corresponding to outcomes A and C. An advantage of this approach is that for two studies that have equal values on all covariates, the (marginal) variance-covariance matrices for the subset of outcomes that is reported in both studies will also be equal.

In the second approach, we extract the rows in $\boldsymbol{Z}_i$ and $\boldsymbol{R}_i$ corresponding to the available outcomes. That is, if there are $p = 3$ outcomes A, B, and C, but only outcomes A and C were reported, we only keep the rows of $\boldsymbol{Z}_i$ that pertain to these outcomes. Similarly, we keep only the rows in $\boldsymbol{R}_i$ pertaining to the correlation of outcomes A and C. Based on these restricted design matrices (with two rows and one row, respectively), we compute $\boldsymbol{\tau}_i$ and $\boldsymbol{\delta}_i$ (which, in the given example, would have two and one elements, respectively). We then compute the 2×2 matrix Σ_i using the formulas described in this paper. In contrast to the first approach, this approach can also be used when covariate information is only available for the observed outcomes. However, a disadvantage of this approach is that for two studies with equal values on all covariates, the (marginal) variance-covariance matrices for the subset of outcomes that is reported in both studies may differ, and this may cause bias especially in meta-analyses with many missings. We conducted a simulation study to investigate whether and if so the extent to which the two approaches lead to different estimates.

Specifically, in this simulation, we generated meta-analytic data for three outcomes A, B, and C, with a binary predictor whose value was generated from a $\text{Ber}(0.5)$ distribution for each study. We considered two sets of parameter values: In both sets, we set $\boldsymbol{\gamma} = (-3, 0.69, -2.3, 0.69, -1.61, 0.69)^T$, so that the variances for the three outcomes were 0.05 (0.1), 0.1 (0.2), and 0.2 (0.4) when the binary predictor was zero (one). In the first set, we set $\boldsymbol{\delta} = (0.51, 0, 0.23, 0, 0.38, 0)^T$, so that correlations between outcomes A and B, A and C, and B and C were 0.5, 0.3, and 0.4, respectively. In the second set, we set $\boldsymbol{\delta} = (0.51, 0.27, 0.23, 015, 0.38, 0.21)^T$ so that the correlations were 0.5 (0.7), 0.3 (0.5), and 0.4 (0.6) when the binary predictor was zero (one). Within studies, outcomes were uncorrelated. For each set of parameter values, we generated meta-analytic data sets with $1,000$ studies, using the same data-generating mechanism as in the main simulation study with a large sample size within studies.

For each data set, we considered nine missingness scenarios by fully crossing the factors "missing outcome" (levels: A, B, C) and "percentage of missing values" (levels: 25%, 50%, 75%). For each of these scenarios, we randomly chose the studies for which the respective effect size was missing. We then estimated two trivariate location-scale meta-analysis models, one according to the first approach and one according to the second approach, including the predictor in all variance- and correlation equations. Figures that show the estimates of the models for both parameter sets and all missingness scenarios are available at this paper's OSF repository. They show that estimates for variance were unaffected by the approach used for handling missing outcomes, while correlation parameter estimates were similar for both approaches, especially when the percentage of missing values was below 75%. Altogether thus, this shows that the bias induced by using the second approach (omitting covariate information) will likely be small as long as the percentage of missing values is not large.