Received: 11 March 2022 | Revised: 17 August 2023 | Accepted: 31 August 2023

DOI: 10.1002/jrsm.1673

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Accuracy and precision of fixed and random effects in meta-analyses of randomized control trials for continuous outcomes

Timo Gnambs[1] | **Ulrich Schroeders**[2]

[1]Leibniz Institute for Educational Trajectories, Bamberg, Germany

[2]University of Kassel, Kassel, Germany

Correspondence
Timo Gnambs, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany.
Email: timo.gnambs@lifbi.de

Abstract

Meta-analyses of treatment effects in randomized control trials are often faced with the problem of missing information required to calculate effect sizes and their sampling variances. Particularly, correlations between pre- and posttest scores are frequently not available. As an ad-hoc solution, researchers impute a constant value for the missing correlation. As an alternative, we propose adopting a multivariate meta-regression approach that models independent group effect sizes and accounts for the dependency structure using robust variance estimation or three-level modeling. A comprehensive simulation study mimicking realistic conditions of meta-analyses in clinical and educational psychology suggested that imputing a fixed correlation 0.8 or adopting a multivariate meta-regression with robust variance estimation work well for estimating the pooled effect but lead to slightly distorted between-study heterogeneity estimates. In contrast, three-level meta-regressions resulted in largely unbiased fixed effects but more inconsistent prediction intervals. Based on these results recommendations for meta-analytic practice and future meta-analytic developments are provided.

KEYWORDS

effect size, meta-analysis, missing value, randomized control trial, robust variance estimation

What is already known

- Randomized control trials (RCT) estimate treatment effects by comparing the change between pre- and posttest in an intervention group to the change in a control group.
- For the calculation of RCT effects often a constant value is imputed for missing pre-post correlations.

What is new

- Meta-analyses with imputed pre-post correlations and multivariate approaches that allow pooling RCT effects with missing pre-post correlations result in largely unbiased point and interval estimates of fixed effects, albeit three-level

meta-analyses exhibit a slight overcoverage of the confidence intervals and substantial undercoverage rates of the prediction intervals.
- As compared to univariate meta-analyses of posttest effect sizes, meta-analyses of RCT effects were not affected by posttest variance heterogeneity or attrition bias.

Potential impact
- The choice of the meta-analytic model has a negligible impact on the pooled effects in RCTs when pre-post correlations are missing.
- For practical applications, it is recommended to conduct univariate meta-analyses that impute a fixed value of 0.8 for the missing correlation or, alternatively, adopt a multivariate meta-regression model with robust variance estimation.

1 | INTRODUCTION

Randomized control trials (RCT) are often considered the gold standard to infer scientific evidence from empirical data,[1,2] thus, informing decisions in health care and education policy. The focus of RCTs is on treatment or intervention effects that compare the difference in the average change of an outcome between two measurement occasions (pretest vs. posttest) for two randomly assembled groups (treatment vs. control group). Treatment effects are inferred if the average change in the treatment group that has received the intervention of interest (e.g., a novel therapy) between the two measurements is significantly larger (or smaller) as compared to the average change in the control group that received no intervention (e.g., a placebo) or an alternative intervention (e.g., an established therapy). Properly designed RCTs strengthen causal attributions of observed changes to intervention effects because they can account for three potential sources of bias, that is, time effects, selection effects, and time-selection interaction effects.[2,3] Thus, RCTs allow controlling for natural changes taking place between a pretest and posttest (e.g., maturation, fatigue) that are not caused by the intervention. If time effects are ignored, natural changes might be erroneously interpreted as treatment effects, despite the treatment having no relevant effect on the outcome. Simpler designs such as pre-post designs without a control group typically cannot control for these time effects. Selection effects can occur if non-random groups are compared because treatment and control groups exhibit important differences in the outcome at the pretest. If selection effects are ignored, posttest differences between groups might be erroneously interpreted as treatment effects although they merely reflect preexisting differences between groups. In RCTs, effective randomization to treatment and control groups typically controls for known and unknown confounders and, thus, ensures that groups at the pretest are comparable. Still, differential attrition between pre- and posttest can lead to nonequivalent groups at posttest, simply because response rates depend differentially on the measured outcome for the two groups. Simpler designs such as posttest comparisons between treatment and control groups that do not acknowledge pretest information often cannot control for the effects of these time-selection interactions. Finally, in within-subjects designs such as RCTs, each individual can be considered their own control which increases the power of statistical tests and the precision of inferences.[3] Therefore, RCTs are often considered the best practice for studying causal relationships in prevention and intervention research.[1,2]

Both in clinical and educational research, meta-analyses of RCTs are often considered the most reliable evidence for intervention efficacy, particularly in areas with a limited number of participants per trial or conflicting evidence. Consequently, these meta-analyses not only receive a lot of attention from the scientific community but are also used by stakeholders that base their decisions on scientific evidence. A prominent example is a recent discussion on the efficacy and safety of *umifenovir* for the treatment of the coronavirus disease which has initially been advocated as an effective treatment but turned out ineffective in a quantitative meta-analysis of the available RCTs.[4] Although combining the raw data of multiple studies in individual participant meta-analyses is preferable from a methodological point of view,[5] most psychological studies do not provide the respective raw data.[6,7] Particularly, in clinical research often legal restrictions or ethical considerations prevent sharing the raw data (see Reference [8], for a potential remedy). Therefore, meta-analyses of summary statistics are the only viable solution in many situations.

Because reporting practices in psychology and other behavioral sciences often do not follow prevalent

recommendations,[9] necessary information to adequately aggregate meta-analytic results is often missing. The current manuscript evaluates different strategies for meta-analyses of RCTs with a focus on situations when information to calculate the sampling variances of the effect sizes is missing. To this end, we propose to respecify the traditional univariate meta-analysis as a multivariate model that acknowledges dependent effects using robust variance estimation[10,11] or a three-level meta-analysis.[12,13] We present a comprehensive simulation study that contrasts these approaches under different realistic conditions to derive recommendations for future meta-analytic practice.

2 | META-ANALYSES OF STANDARDIZED MEAN DIFFERENCES IN RCTs

In the following, we summarize the prevalent method of synthesizing RCT effect sizes in meta-analytic research. Moreover, we will emphasize shortcomings in this approach that make its application infeasible in many situations.

2.1 | The RCT effect size

The conventional effect size for RCTs with continuous outcomes is the difference in the standardized mean change between the pretest and posttest for the treatment and control groups. Let us assume that the pretest and posttest scores for the metric outcome in both groups (T = treatment group, C = control group) follow a bivariate normal distribution in the population with means $\mu_{T,pre}$ and $\mu_{C,pre}$ at the pretest and $\mu_{T,post}$ and $\mu_{C,post}$ at the posttest. If we further assume a common variance σ^2 for both groups at the time points and a common correlation ρ between pre- and posttest scores, then the standardized mean change in the population is given by $\delta_g = (\mu_{g,post} - \mu_{g,pre})/\sigma$ in each group $g = \{T, C\}$. The RCT effect size for the difference in the standardized mean change is

$$\Delta = \delta_T - \delta_C = \frac{\left(\mu_{T,post} - \mu_{T,pre}\right) - \left(\mu_{C,post} - \mu_{C,pre}\right)}{\sigma} \quad (1)$$

with the corresponding sample estimate for Δ as

$$\widehat{\Delta} = c(df) \cdot \frac{\left(M_{T,post} - M_{T,pre}\right) - \left(M_{C,post} - M_{C,pre}\right)}{SD_{pre}}. \quad (2)$$

In (2), M_{pre} and M_{post} are the pretest and posttest means in the two groups, while SD_{pre} is the pooled pretest standard deviation

$$\widehat{\sigma} = SD_{pre} = \sqrt{\frac{(n_T - 1) \cdot SD_{T,pre}^2 - (n_C - 1) \cdot SD_{C,pre}^2}{n_T + n_C - 2}} \quad (3)$$

given the pretest standard deviations $SD_{T,pre}$ and $SD_{C,pre}$ and respective sample sizes n_T and n_C. Although different estimators for $\widehat{\sigma}$ have been proposed that either use independent estimates $\widehat{\sigma}_g$ for both groups[14] or also incorporate the posttest variance,[15] simulation research suggests that the pooled pretest SDs result in the most precise estimates of the sampling variances of $\widehat{\Delta}$.[16] Finally, $c(df)$ is a bias adjustment function to correct for a small sample bias with degrees of freedom (df) of $n_T + n_C$-2 and the gamma function $\Gamma(x)$[17,18]:

$$c(df) = \sqrt{\frac{2}{df}} \cdot \frac{\Gamma\left(\frac{df}{2}\right)}{\Gamma[(df-1)/2]} \approx 1 - \frac{3}{4 \cdot df - 1} \quad (4)$$

The asymptotic distribution of $\widehat{\Delta}$ in (2) has been derived as[16]

$$Var\left(\widehat{\Delta}\right) = c(df)^2 \cdot 2 \cdot (1-\rho) \cdot \left(\frac{n_T + n_C}{n_T \cdot n_C}\right) \cdot \left(\frac{n_T + n_C - 2}{n_T + n_C - 4}\right) \cdot \left(1 + \frac{n_T \cdot n_C}{n_T + n_C} \cdot \frac{\Delta^2}{2 \cdot (1-\rho)}\right) - \Delta^2. \quad (5)$$

2.2 | The random-effect meta-analytic model

If RCT effect size estimates are available from multiple samples, meta-analytic methods can be used to combine them to infer an average true effect $\widehat{\Delta}$ across samples. Consider K samples to contribute effect sizes for the meta-analysis. Let $\widehat{\Delta}_k$ denote the effect size estimate of Δ_k from the kth sample with $k \in \{1, \ldots, K\}$ and $Var\left(\widehat{\Delta}_k\right) = v_k$ as the corresponding sampling variance. Then, the univariate random effect model can be written as a multilevel model, such that[13]

$$\widehat{\Delta}_k = \Delta + u_k + e_k$$

$$u_k \sim N\left(0; \tau^2\right)$$

$$e_k \sim N(0; v_k)$$

$$Cov(u_i,u_j) = Cov(e_s,e_t) = Cov(u_i,e_s) = 0; \forall (i \neq j \wedge s \neq t) \text{ with } i,j,s,t \in \{1,...,K\} \quad (6)$$

where the sampling errors $e_k = \widehat{\Delta}_k - \Delta_k$ are assumed to be uncorrelated with known variance v_k.u_k represents the deviation of a sample-specific true effect from the average true effect and τ^2 gives the heterogeneity estimate of the distribution of the true effects. If $u_k = 0$ for all samples, (6) simplifies to a fixed-effect model. Because the assumption of no between-sample heterogeneity is rarely tenable in practice,[19] we will focus on the random-effect model. Although different estimators have been suggested for the random effects variance τ^2, restricted maximum likelihood (REML) has shown the most promising results for continuous outcomes under different conditions (see Reference [20] for a review). The average true effect Δ in (6) is typically derived as a weighted least square estimate given by $\widehat{\Delta} = \sum_k \left(w_k \cdot \widehat{\Delta}_k\right) / \sum_k w_k$ with $w_k = 1/\left(v_k + \widehat{\tau}^2\right)$.[21]

2.3 | Unresolved challenges in RCT meta-analyses

Morris[16] advocated the use of an effect size in RCT meta-analyses which is based on the pooled pretest standard deviation (see Reference [3]) because it "provides an unbiased estimate of the population effect size" (p. 24) and has a smaller sampling variance than competing estimates. However, the sampling variance estimator in (5) relies on the pre-post correlation. While means and standard deviations of the pretest and posttest scores and sample sizes are routinely reported in scientific publications, the correlation between pretest and posttest scores is seldom found. In fact, it is not uncommon that not a single primary study included in a meta-analysis informs about the respective correlation.[22] Therefore, these correlations are often imputed by a constant value such as 0.70,[23] 0.60,[24] or 0.50, thus, mimicking an independent groups design.[3] However, empirical effect size distributions of pre-post correlations in different fields highlight that these correlations can vary substantially depending on the domain and the studied effect.[25,26] For example, Taylor and colleagues[26] found pooled pre-post correlations for different types of training effects that varied between 0.43 and 0.82. Thus, using a specific value for the unknown pre-post correlation might be misleading and reduce the efficiency of the effect size estimator in (6) (see Reference [27] for similar concerns in the context of multivariate meta-analyses). Even if pre-post correlations are available from primary studies, it might not be advisable to use sample estimates for the population correlation ρ required in (5), because especially in small samples with less than 250 participants that dominate RCT research,[28] sample correlations are highly variable and a poor estimate of the population value.[29]

3 | A MULTIVARIATE META-REGRESSION APPROACH FOR RCTs

To overcome the problem of missing pre-post correlations, we propose modeling the RCT effect as independent group effect sizes in a meta-analytic regression framework. To do so, the RCT effect size in (2) is restructured as

$$\widehat{\Delta} = c(df) \cdot \frac{\left(M_{T,post} - M_{C,post}\right) - \left(M_{T,pre} - M_{C,pre}\right)}{SD_{pre}}$$

$$= c(df) \cdot \frac{\left(M_{T,post} - M_{C,post}\right)}{SD_{pre}} - c(df) \cdot \frac{\left(M_{T,pre} - M_{C,pre}\right)}{SD_{pre}}$$

$$= \widehat{\delta}_{post} - \widehat{\delta}_{pre} \quad (7)$$

and, thus, expressed as the difference between two independent group effect sizes for the pretest and the posttest. Then, the sampling variances of the effect size $\widehat{\delta}_t$ at each measurement occasion t ($0 =$ pretest, $1 =$ posttest) do not rely on the pre-post correlation, but correspond to (5) when setting ρ to 0.5.[17] The difference in (7) can be formalized in a multivariate meta-regression model[30] where each sample contributes two effect sizes ($\widehat{\delta}_{pre}$ and $\widehat{\delta}_{post}$) as

$$\widehat{\delta}_{kt} = \beta_0 + \beta_1 \cdot t + u_{kt} + e_{kt}$$

$$u_{kt} \sim N\left(0; \mathrm{T}^2\right)$$

$$e_{kt} \sim N(0; v_{kt})$$

$$Cov\left(u_{im}, u_{jn}\right) = 0; \forall (i \neq j) \text{ with } i,j \in \{1,...,K\} \text{ and } m,n \in \{0,1\}$$

$$Cov\left(u_{im}, e_{jn}\right) = 0 \text{ with } i,j \in \{1,...,K\} \text{ and } m,n \in \{0,1\}$$

$$Cov\left(e_{im}, e_{jn}\right) = 0; \forall (i \neq j) \text{ with } i,j \in \{1,...,K\} \text{ and } m,n \in \{0,1\}$$

$$Cov(e_{k0}, e_{k1}) = \rho_k \cdot v_{k0}^{0.5} \cdot v_{k1}^{0.5} \text{ with } k \in \{1,...,K\} \quad (8)$$

with $\widehat{\delta}_{kt}$ as the independent group effect size in sample k at measurement occasion t ($0 =$ pretest, $1 =$ posttest) and e_{kt} as the sampling error residual with known variance $v_{kt} = Var\left(\widehat{\delta}_{kt}\right)$. The regression coefficient β_1 represents

the difference in effect sizes between the pretest and posttest and, thus, estimates the average true effect Δ across samples as in (6), while the intercept β_0 represents the mean difference at the pretest (i.e., a selection effect).

The time-specific effects u_{kt} within a sample are jointly distributed, in the most general case, with an unstructured variance–covariance matrix $T^2 = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}$, thus, assuming different between-study heterogeneities at pre- and posttest. Then, $\tau_0^2 + \tau_1^2 - 2 \cdot \tau_{01}$ corresponds to the total between-study heterogeneity for the RCT effect size, that is, τ^2 in (6).[1] However, because the variance–covariance structure of the time-specific effects is rather complex, more restrictive specifications might be more appropriate in practice. For example, properly designed RCTs should result in no group differences at the pretest because respondents are randomized to the treatment and control groups (see Reference [31] for an overview of different approaches). Therefore, if the assumption of no (or negligible) pretest imbalance seems justified (which often is the case, see References [28,32]), β_0 as well as τ_0^2 and τ_{01} in T^2 could be constrained to 0; this would reduce the total between-study heterogeneity to $\tau^2 = \tau_1^2$. Moreover, because RCT meta-analyses often include rather few primary studies with small samples,[33] the covariance in T^2 might be sometimes practically non-identifiable,[34,35] thus, requiring modeling independent variances at pre- and posttest. In practice, proper constraints on the variance structure can be identified by comparing models with different random effects structures, for example, using likelihood ratio tests.[36]

Because each sample contributes two effect sizes to the multivariate meta-analytic model in (8), the $\widehat{\delta}_{k\cdot}$ are no longer independent but exhibit an unknown within-sample correlation ρ_k. However, information on the exact value of ρ_k is not essential because dependent effects can be acknowledged using robust variance estimation (RVE[10,11]) or extending (8) to a three-level model (TLM[12,13]) which do not rely on a known correlation.

3.1 | Robust variance estimation

RVE corrects the standard errors of the model parameters estimated in (8) without requiring information on the exact variance–covariance structure for the dependent effect sizes within a sample. This is achieved by specifying a "working model" for the dependence structure such as using a common correlation ρ_k of 0.8 within samples.[10] Then, estimates of the study-specific covariance matrices are empirically derived using weighted least squares as the products of the regression residuals (see Reference [11], and the Appendix A). Although each study-specific estimate might be rather imprecise, their average across multiple samples is sufficiently precise when the number of samples is large. However, small-sample corrections can be incorporated into the variance–covariance estimator to yield approximately unbiased standard errors when the number of studies is small.[37,38] Importantly, although the robust standard errors are unbiased even if the "working model" is not correctly specified, the precision of the resulting estimates increases the closer the true dependency structure is approximated.

3.2 | Three-level meta-analysis

TLM extends the model in (8) by an additional random effect u_k. Thus, the total random variance is split into two components, the between-sample variation u_k and the residual within-sample variation u_{kt} in true effects:

$$\widehat{\delta}_{kt} = \beta_0 + \beta_1 \cdot t + u_{kt} + u_k + e_{kt}$$

$$u_{kt} \sim N\left(0; \tau_w^2\right)$$

$$u_k \sim N\left(0; \tau_b^2\right)$$

$$e_{kt} \sim N\left(0; v_{kt}\right)$$

$$Cov\left(u_{im}, u_{jn}\right) = 0; \forall (i \neq j) \text{ with } i,j \in \{1,...,K\} \text{ and } m,n \in \{0,1\}$$

$$Cov\left(u_{im}, u_{in}\right) = 0; \forall (m \neq n) \text{ with } i \in \{1,...,K\} \text{ and } m,n \in \{0,1\}$$

$$Cov\left(u_{im}, u_j\right) = 0 \text{ with } i,j \in \{1,...,K\} \text{ and } m \in \{0,1\}$$

$$Cov\left(u_{im}, e_{jn}\right) = 0 \text{ with } i,j \in \{1,...,K\} \text{ and } m,n \in \{0,1\}$$

$$Cov\left(u_i, e_{jn}\right) = 0 \text{ with } i,j \in \{1,...,K\} \text{ and } n \in \{0,1\}$$

$$Cov\left(e_{im}, e_{jn}\right) = 0; \forall (i \neq j \vee m \neq n) \text{ with } i, j \in \{1,...K\} \text{ and } m,n \in \{01\} \quad (9)$$

Then, the total between-study heterogeneity for the RCT effect size, that is, τ^2 in (6), corresponds to $2 \cdot \tau_w^2$.[2] Moreover, the TLM in (9) assumes independent sampling errors e_{kt}. Thus, instead of specifying correlated errors, the TLM models correlated true effects within samples. This assumption is clearly violated when multiple effect

sizes are based on the same sample. Additionally, the TLM implies a similar degree of between-sample heterogeneity for all effect sizes.[39] Although these assumptions might not be tenable in empirical applications, Van den Noortgate and colleagues[13,40] showed that the three-level random effect structure can account for within-sample dependencies reasonably well when the number of effect sizes per sample are small and the random variances are large. Therefore, models with multiple random effects are increasingly used in applied meta-analytic research (see Reference [41] for a review). However, TLMs tend to suffer from convergence issues when the number of samples is small and show increased parameter bias when pooling outcome-specific effect sizes.[42]

4 | OBJECTIVES AND RESEARCH QUESTIONS

As outlined above, different approaches are currently in use to pool RCT effect sizes across multiple samples. Hitherto, little is known to what degree and under which conditions the use of an ad-hoc substitute for the unknown pre-post correlation might distort the resulting meta-analytic estimates. Additionally, the proposed multivariate approach could improve current practice by modeling the RCT effect as dependent effects in line with current state-of-the-art approaches to model dependencies in meta-analytic research.[11,42] Therefore, we present a comprehensive Monte Carlo simulation that evaluated the precision of different meta-analytic methods for pooling RCT effect sizes under various realistic conditions. Based on these results, we provide recommendations for future meta-analytic practice.

5 | METHOD

5.1 | Meta-analytic models for RCT effects

The simulation compared three univariate random-effects meta-analyses of RCT effect sizes and two multivariate random-effects meta-analyses of independent group effect sizes. As a point of comparison, we also included a univariate meta-analysis of standardized differences in posttest means that ignored any pretest information. All models used a REML estimator with a maximum number of 1000 iterations for the optimizer to converge. Although REML results in slightly negatively biased heterogeneity estimates in univariate meta-analyses, it is less biased than maximum likelihood estimation[43] and also compares favorably to alternative estimators as reviewed by Veroniki and colleagues.[20] More importantly, REML is applicable for univariate as well multivariate meta-analyses.

5.1.1 | Univariate meta-analyses of RCT effect sizes

The reference approach (UMA-S with S for sample) consisted of inverse-variance weighted random-effects meta-analyses for which all primary studies reported the required sample statistics to calculate the effect size in (2) and its sampling variance in (5). In (5), the sample effect size d was used for Δ and the sample pre-post correlation r was used for ρ. The second approach (UMA-P with P for population) also assumed that all primary studies reported the required sample statistics, but since (5) requires the population value for the pre-post correlation, a two-step approach was adopted. First, a random-effects meta-analysis (REML estimation) pooled the inverse variance-weighted Fisher's Z transformed pre-post correlations that were reported in the primary studies to derive a pooled pre-post correlation $\widehat{\rho}$. Then, the pooled pre-post correlation was used in (5) for the calculation of the sampling variances of the RCT effect sizes. It was assumed that the pooled pre-post correlation would more precisely represent the population correlation ρ required in (5), particularly in small samples.[29] For the third approach (UMA-I with I for imputation), we simulated meta-analyses for which neither primary study reported the necessary pre-post correlations. Therefore, the sampling variances in (3) were calculated by imputing a constant value of either 0.5 or 0.8 for the unknown correlation ρ in the population (5). For all univariate meta-analyses, standard errors and 95% confidence intervals were adjusted following Knapp and Hartung[44] for better control of type I error rates in small samples. Prediction intervals (PI) were calculated as

$$PI = \widehat{\Delta} \pm t_{(1,k-2)} \cdot \sqrt{SE\left(\widehat{\Delta}\right)^2 + \widehat{\tau}^2} \qquad (10)$$

where $SE\left(\widehat{\Delta}\right)$ is the standard error of the estimated pooled effect $\widehat{\Delta}$, $\widehat{\tau}^2$ is the estimated between-sample variance, $t_{(1,k-2)}$ is the 97.5 percentile of the t-distribution, and k gives the number of samples.[45]

5.1.2 | Multivariate meta-analyses of pre- and posttest effect sizes

Two independent group effect sizes (d_{post} and d_{pred}) with their sampling variances were calculated for each simulated sample. Then, the pooled RCT effect was estimated

using the regression framework in (8). The first approach adopted a robust meta-analytic model (i.e., RVE) using (8) as a working model and assuming correlated errors of either 0.5 or 0.8. Then, cluster-robust variances with a bias-reduced linearization correction were calculated for the regression parameters to account for heteroscedasticity and unmodeled errors.[37,46] Moreover, confidence intervals incorporated the Satterthwaite correction for the degrees of freedom which has been shown to lead to more appropriate coverage rates of confidence intervals.[37] The second approach also pooled two independent group effect sizes but acknowledged the dependencies between effects by estimating a three-level meta-analytic model (TLM) as given in (9). Thus, the random variance terms modeled effect sizes nested within studies.[12,13] For all multivariate meta-analyses, prediction intervals were calculated following (10) replacing $\widehat{\Delta}$ with $\widehat{\beta}_1$ from (8) or (9) and $\widehat{\tau}^2$ with the total between-study heterogeneity.

5.1.3 | Univariate meta-analyses of posttest effect sizes

As the most basic strategy for analyzing treatment effects, inverse-variance weighted random-effects meta-analyses pooled the posttest effect sizes without considering the pretest (UMA-B with B for basic). The effect size for the standardized mean difference between independent groups was calculated as $\widehat{\delta}_{post} = c(df) \cdot \frac{(M_{T,post} - M_{C,post})}{SD_{post}}$ with SD_{post} given by (3) using the posttest standard deviations.[17,18] The sampling variance followed (5) when setting ρ to 0.5.[17] Again, standard errors and confidence intervals were adjusted following Knapp and Hartung.[44] Prediction intervals were calculated as in (10).

5.2 | Experimental design

The simulation aimed to mimic typical conditions of meta-analyses of pre-post intervention studies with continuous outcomes that are often encountered in evaluation studies across many disciplines such as clinical (e.g., psychotherapy) and educational research (e.g., teaching) or personnel psychology (e.g., employee training). The present study manipulated seven design factors to evaluate their impact on the simulation results (see Table 1). These included the number of effect sizes in a meta-analysis (K), the average samples size per effect size (n), the true change in the treatment group (Δ), the true pre-post correlation (ρ), the true posttest variance in the treatment group (Φ), the between-study heterogeneity (τ^2), and the presence of attrition bias. This resulted in a 5 (effect sizes) × 3 (sample sizes) × 3 (true changes) × 3 (true correlations) × 2 (true posttest variances) × 2 (between-sample heterogeneities) × 2 (attrition biases) fully crossed simulation design.

TABLE 1 Experimental conditions and constant settings for simulation.

Experimental condition	Values
Number of effect sizes per meta-analysis (K)	3, 5, 10, 20, 40
Average sample size in meta-analysis (n)	40, 80, 120
True change in the treatment group (Δ)	0.20, 0.40, 0.80
True pre-post correlation in treatment and control groups (ρ)	0.20, 0.50, 0.80
True posttest variance in the treatment group (Φ)	1.0, 1.5
Between-sample heterogeneity (τ_Δ)	0.10, 0.30
Attrition bias	0.00, 0.05
Constant settings	**Value**
Between-sample heterogeneity of pre-post correlation (τ_ρ)	$\tau_\rho \sim min\{\text{half-}N(0, 0.25), 0.5\}$
True change in control group	0.00
True pretest variances in treatment and control groups	1.00
True posttest variance in control group	1.00

5.2.1 | Number of effect sizes per meta-analysis

Meta-analyses in psychology and education often combine between 10 and 200 effect sizes.[47,48] Meta-analyses on training studies that often adopt RCT designs are usually located at the lower end of this distribution, regardless of the discipline. For example, Collins and Holton[49] reported meta-analytic effects on the effectiveness of managerial leadership development programs that included between 6 and 23 samples. Similarly, various meta-analyses on behavior modeling training effects for different outcomes were based on 14 to 66 effect sizes, with most of them including less than 40.[26] In meta-analyses of clinical trials, the number of pooled effect sizes is even substantially smaller. In psychology, meta-analyses on the effectiveness of clinical psychology treatments include a median of 18 studies.[50] For medical trials, a review of the *Cochrane Database of Systematic Reviews* found that of nearly 3000 meta-analyses on mental health, 90% pooled results from up to 10 studies, while half of them included no more than three studies.[28] Therefore, the number of effect sizes per meta-analysis was set to either 3, 5, 10, 20, or 40.

5.2.2 | Average sample size in primary studies

Sample sizes of primary studies in meta-analysis differ largely depending on the setting (e.g., educational, clinical, personnel) and the specificity of the group (e.g., students with learning disabilities, patients with Parkinson's disease, team leaders). For example, Taylor and colleagues[26] reported a mean sample size (including control and treatment group) in meta-analyses of training studies of 37 ($Min = 5$, $Max = 271$), which was similar to meta-analyses in clinical psychology.[50] In contrast, the above-mentioned review of the Cochrane database[28] found a median sample size of RCTs on mental health of $Mdn = 63$ with the 25th and 75th percentiles at 36 and 165. Therefore, the average sample size per effect size (including control and treatment group) was set to either 40, 80, or 120. However, sample sizes in psychological meta-analyses are often positively skewed.[51,52] Therefore, we did not simulate constant sample sizes for a given meta-analysis, but used three vectors, [22, 26, 28, 30, 94], [62, 66, 68, 70, 134], and [102, 106, 108, 110, 174] that each exhibited a Pearson skewness of 1.464 but different means (which is analogous to the approach of Sánchez-Meca & Marín-Martínez[51]). These vectors were replicated $k/5$ times in a given meta-analysis to meet the total number of simulated samples. The sample sizes in the treatment and control groups were equal.

5.2.3 | True change in the treatment group

Lipsey and Wilson[48] found a median standardized difference (Cohen's d) in meta-analyses of psychological treatment effects of about 0.47. However, meta-analyses of training studies also reported, depending on the observed outcome, pooled effects that reached up to 1.00.[26] In contrast, clinical studies often observe more modest effect sizes. A review of more than 100,000 clinical trials conducted between 1975 and 2014 showed that—independent of the year of study—the average effect is about 0.20.[53] Educational studies with randomized designs even produce average effects of only about 0.10 to 0.16.[54,55] Therefore, the standardized mean change in the treatment group was set to either 0.20, 0.40, or 0.80 representing small, medium, and large effect sizes. In the control group, a standardized mean change of 0.00 was assumed.

5.2.4 | True pre-post correlation

Pooled pre-post correlations in training studies typically fall between 0.43 and 0.82.[26] In various meta-analyses of psychiatric RCTs the median of the pooled pre-post correlations was 0.36 with the 25th and 75th percentile amounting to 0.22 and 0.58. Negative correlations were uncommon.[24] Therefore, we used pre-post correlations of either 0.20, 0.50, or 0.80 that were identical in the treatment and control groups.

5.2.5 | True posttest variance in treatment group

The effect size for RCT designs assumes homogenous variances at pre- and posttest as well for treatment and control conditions.[16] However, if participants are differently affected by the treatment, some of them will improve more strongly while others will improve less. Consequently, the posttest scores will exhibit a larger variance as compared to the control group or the pretest. For example, in a meta-analysis of training studies, the posttest standard deviations increased by about 7.6% in the treatment group.[56] Similarly, a review of meta-analysis of clinical trials reported that, on average, empirical pre-post correlations for treatment groups were about $\Delta r = 0.20$ smaller than the respective pre-post correlations in the control groups, thus, reflecting larger posttest variances in the treatment groups.[24] On the other hand, often meta-analyses do not identify pronounced treatment heterogeneity, thus, making the assumption of homogeneous variances plausible for many applications.[57,58] To study potential effects of heterogeneous variances, we set the posttest variance in the treatment group to either 1.0 or 1.5 times the population variance. Although a variance increase by 50% seems unrealistic in most cases, it was chosen as a worst-case scenario (see also Morris[16] for a similar condition).

5.2.6 | Between-study variances

Van Erp and colleagues[59] reviewed heterogeneity estimates in over 700 psychological meta-analyses and found a median between-study heterogeneity of $\tau_\Delta = 0.20$ ($IQR = [0.10,\ 0.33]$) for meta-analyses of standardized mean differences. A similar review by Linden and Hönekopp[60] identified a slightly larger mean between-study heterogeneity across 150 meta-analyses from cognitive, organizational, and social psychology of $\tau_\Delta = 0.30$, whereas multiple close replications were less variable with a mean τ_Δ of 0.09. Therefore, we used a between-study heterogeneity in (6) of either 0.10 or 0.30, thus, reflecting small and large heterogeneity, respectively.

5.2.7 | Attrition bias

Longitudinal studies often suffer from sample attrition because not all participants randomized to the control and

treatment group at the pretest also participate in the posttest measurement. In the past, average attrition rates between 13% and 19% have been reported for medical trials and educational interventions, respectively.[61,62] Differential attrition for treatment and control groups was typically small.[61,63] Importantly, attrition is primarily a concern if it introduces bias because the likelihood of nonparticipation is associated with pre- and posttest scores. Bias is sometimes considered problematic if it exceeds a threshold of $|d| = 0.05$.[64] However, a recent examination of attrition bias in 10 educational RCTs found only a mean absolute bias of 0.026.[65] Similarly, the average bias in medical RCTs was about 0.02, albeit attrition increased the between-study variance.[66] Thus, attrition bias might not be a widespread threat to the validity of RCTs. Nevertheless, we considered a situation where RCTs exhibited an average attrition bias of $d = 0.05$ and compared the respective results to a condition without bias.

5.3 | Data simulation and model estimation

For each experimental condition outlined above, the pooled treatment effect was calculated using different random-effects meta-analyses from randomly generated samples. The entire simulation procedure for a given condition followed seven steps:

1. For a given sample k included in a meta-analysis, the true change in the treatment group Δ_k was randomly drawn from a normal distribution $N(\Delta, \tau^2_\Delta)$ with Δ and τ^2_Δ representing the true change and between-sample heterogeneity depending on the experimental condition.
2. For a given sample k included in a meta-analysis, the true pre-post correlation ρ_k was randomly drawn from a normal distribution as $\tanh\left(N\left(\tanh^{-1}(\rho_k), \tau^2_\rho\right)\right)$ with $\tanh^{-1}(x)$ representing the inverse hyperbolic tangent function, thus, giving the Fisher's Z transformed true pre-post correlation ρ depending on the experimental condition, and τ^2_ρ giving the between-sample heterogeneity. For a given meta-analysis, τ_ρ was derived by a random draw from a half-normal distribution $min\{\text{half-}N(0, 0.25), 0.5\}$. This closely reproduced the empirical distribution of between-sample heterogeneities identified in over 700 psychological meta-analyses of correlation coefficients that gave a median of $\tau_\rho = 0.16$ (IQR = [0.08, 0.22]).[59]
3. For the treatment group in sample k, $(n/2)/0.8$ data points representing the pre- and posttest scores as well as a standard normally distributed attrition indicator were randomly drawn from a multivariate normal distribution $N\left(\begin{bmatrix}0\\ \Delta_k\\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho_k & 0.15\\ \rho_k & \Phi & 0.15\\ 0.15 & 0.15 & 1\end{bmatrix}\right)$, with n and Φ representing the total sample size and posttest variance depending on the experimental condition. The covariances for the attrition indicator were identified by trial and error to produce an average attrition bias of about 0.05 for an attrition rate of 20%. In the condition without attrition bias, the first $n/2$ simulated rows were retained, whereas in the condition with attrition bias, the $n/2$ rows with the lowest values on the attrition indicator were retained. In this way, attrition bias did not affect the manipulated sample size.
4. For the control group in sample k, $n/2$ data points representing the pre- and posttest scores were randomly drawn from a multivariate normal distribution $N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho_k\\ \rho_k & 1\end{bmatrix}\right)$ with n representing the total sample size depending on the experimental condition.
5. In each sample, one RCT effect size and two independent group effect sizes with their sampling variances were calculated according to (2, 5).
6. Steps 1 to 5 were repeated to generate K samples for a given meta-analysis according to the experimental condition.
7. The different meta-analytic models were applied to the simulated samples to derive the pooled effect $\widehat{\Delta}$, the heterogeneity estimate τ^2, a 95% confidence interval for $\widehat{\Delta}$, and a 95% prediction interval for $\widehat{\Delta}$.

These steps were replicated 1000 times for each experimental condition. By default, multivariate meta-analyses used a bound constraint quasi-Newton optimizer (*nlminb*),[67] but in case of a convergence failure resorted to the Nelder and Mead[68] method. Replications for which a meta-analytic model still failed to converge were discarded and replaced with a valid case. All analyses were conducted in *R* version 4.2.2 with the packages *metafor* version 3.8.1 and *clubSandwich* version 0.5.8.[69,70]

5.4 | Performance criteria

The accuracy of an estimator $\widehat{\theta}$ was compared using the average parameter bias and root mean squared error (RMSE) which were evaluated for the mean RCT effect $\widehat{\Delta}$ and the heterogeneity estimate τ^2:

$$\text{Bias}\left(\widehat{\theta}\right) = E\left(\widehat{\theta} - \theta\right) \quad (11)$$

$$\mathrm{RMSE}\left(\widehat{\theta}\right) = \sqrt{E\left[\left(\widehat{\theta} - \theta\right)^2\right]} \tag{12}$$

For both criteria, values close to 0 indicate preferable estimators. However, because RMSE can be simplified to $\mathrm{RMSE}\left(\widehat{\theta}\right) = \sqrt{\left[\mathrm{Bias}\left(\widehat{\theta}\right)^2 + Var\left(\widehat{\theta}\right)\right]}$, a more biased estimator might be more efficient if it yields a considerably smaller variance. Moreover, the coverage rates of the 95% confidence intervals were calculated as the percentage of replications for which the true effect Δ fell within the confidence interval. Similarly, the coverage rates of the 95% prediction intervals were compared for the different estimators to study the precision that treatment effects in hypothetical future samples could be predicted. These coverage rates were calculated by randomly drawing a value from $N(\Delta, \tau^2_\Delta)$ and determining the percentage of replications for which this true effect fell within the prediction interval. Accurate confidence and prediction interval estimators should exhibit a nominal probability of 95%. Finally, we also calculated the widths of the 95% confidence and prediction intervals.

The Monte Carlo error (MCE) for each performance criterion was estimated using the jackknife method.[71] Let R represent the number of replications with $\mathbf{X} = \{X_1, X_2, ..., X_R\}$ giving the estimated replicates from which the performance criterion $\theta(\mathbf{X})$ (e.g., bias, RMSE, coverage rate) is calculated. If $\mathbf{X_r}$ with $r \in \{1, ..., R\}$ represents the subset of $\mathbf{X}$ without the rth replicate, then the MCE for θ is given as

$$\mathrm{MCE}(\theta) = \sqrt{\frac{R-1}{R} \cdot \sum_{r=1}^{R} \left(\theta(\mathbf{X}_{-r}) - \frac{1}{R} \cdot \sum_{s=1}^{R} \theta(\mathbf{X}_{-s})\right)^2}. \tag{13}$$

The MCE allows quantifying the precision for each performance criterion to compare differences across different simulation conditions. However, because of the large number of replications used in our simulation, the obtained MCEs were rather small. Thus, we refrain from reporting confidence intervals but provide the median and maximum MCE for each performance criterion.

6 | RESULTS

The simulation results are summarized separately for the different performance criteria. Given the large number of experimental conditions that resulted in 1080 unique cells, the results presented in the following tables and figures refer to the condition for a medium treatment effect ($\Delta = 0.4$) with homogenous posttest variances ($\Phi = 1.0$). Moreover, we will focus on the conditions without attrition bias for the different meta-analyses of pre-post effects; specific results for UMA-B or attrition bias will be selectively pointed out in the text (full results are available in the online material). Moreover, factorial analyses of variance (ANOVA) evaluated the source of the variability in the performance criteria to determine which combinations of experimental conditions produced stronger effects (in terms of η^2) and warranted detailed scrutiny (see Table 2). Again, these analyses were limited to the different estimators of pre-post effects but did not include the meta-analyses of posttest effect sizes.

6.1 | Convergence rates

For all estimators and experimental conditions, the estimated models converged successfully after 1000 iterations. However, for about 1.7% of the TLMs the default optimizer (*nlminb*)[67] failed to converge requiring the use of an alternative optimization algorithm.[68] RVEs did not exhibit similar convergence problems.

6.2 | Average bias and root mean squared error of fixed-effect estimators

The Monte Carlo errors for the average bias in Δ and RMSE were negligible in all conditions ($Mdn = 0.003/0.002$, $Max = 0.012/0.009$), thus, allowing for valid comparisons of the respective point estimates. Factorial analyses of variance for the simulation conditions showed that the main effect of the meta-analytic method explained about 16.7% of the variance in the average bias (see Table 2). Moreover, this effect was qualified by small two-way interactions with the average sample size ($\eta^2 = 5.3\%$) and the true change in the treatment group ($\eta^2 = 3.5\%$). Figure 1 summarizes the average bias by average sample size, number of studies, and true pre-post correlation for the conditions with a small and large between-sample heterogeneity. These results show that all estimators were slightly negatively biased at smaller average sample sizes. Regarding the meta-analytic method, the largest bias was observed for UMA-S which used sample-specific pre-post correlations for the calculation of the RCT sampling variances. In contrast, UMA-P which pooled the pre-post correlations before calculating the sampling variances of the RCT effects exhibited a smaller bias and performed comparably to the different

TABLE 2 Effect sizes for simulation conditions.

Condition	Bias in Δ	RMSE in Δ	Bias in τ_Δ	RMSE in τ_Δ	Coverage rate for CI	Coverage rate for PI	Width of CI	Width of PI
Method	16.7	0.0	34.4	5.1	38.1	15.4	0.1	0.8
K	2.4	67.7	4.6	31.4	6.5	22.0	79.8	82.9
n	30.3	9.5	7.3	15.4	0.7	0.4	5.1	1.2
ρ	0.1	5.6	19.8	10.3	0.0	7.1	2.3	1.5
Δ	18.5	0.1	0.0	0.0	1.6	0.0	0.0	0.0
Φ	0.0	0.4	2.2	0.9	0.0	1.4	0.2	0.2
τ_Δ	0.5	9.2	1.4	12.0	2.0	0.1	3.4	4.3
Bias	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Method × K	0.4	0.0	0.4	1.4	5.9	14.8	0.4	0.3
Method × n	5.3	0.0	6.6	2.4	4.2	0.3	0.0	0.1
Method × ρ	1.2	0.0	6.7	2.5	5.2	4.7	0.2	0.2
Method × Δ	3.5	0.0	0.1	0.0	1.0	0.1	0.0	0.0
Method × Φ	0.1	0.0	0.3	0.1	0.3	0.2	0.0	0.0
Method × τ_Δ	0.3	0.0	2.4	0.3	10.8	3.7	0.2	0.1
Method × Bias	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: Presented are values of η^2 for main effects and two-way interactions of the method factor based on analyses of variance including all possible higher-order interactions up to the order of 3. Univariate meta-analyses of posttest effects were not included. Method = Meta-analytic method for RCT effect, K = Number of samples, n = Average sample size, ρ = True pre-post correlation, Δ = True change in treatment group, τ_Δ = True between-sample heterogeneity, Φ = True posttest variance in treatment group, Bias = Presence of attrition bias, RMSE = Root mean squared error, CI = 95% confidence interval, PI = 95% prediction interval.

approaches that did not make use of the sample pre-post correlations (UMA-I, RVE, TLM). The negative bias was more pronounced for smaller average sample sizes or larger true effects (see Figure S1 in the supplement material), whereas the other factors had no substantial impact. In these situations, $\text{RVE}_{(r=0.8)}$ exhibited a slightly smaller bias as compared to the other multivariate estimators.

UMA-B that ignored the pretest information exhibited accuracies that were comparable to the meta-analyses of pre-post effect sizes (see Figure 1), at least as long as the posttest statistics were not systematically distorted. Attrition bias led to noticeable larger biases that reached up to −0.08 for larger true effects. Moreover, posttest variance heterogeneity amplified this effect and resulted in biases up to −0.15 to −0.11 for the conditions with and without attrition bias, respectively (see Figure 2). In contrast, meta-analyses using the pretest information were not affected by attrition bias.

The RMSE of the different estimators was not affected by the meta-analytic method (see Table 2). Although it was strongly affected by the number of included primary studies and grew larger for meta-analyses with a smaller number of samples or larger between-sample heterogeneity (see Figure S2 in the supplement material), the meta-analytic method did not produce different effects. This suggests that despite the larger bias of UMA-S, the estimator seems to exhibit a smaller variance. As a consequence, it performed rather comparably in terms of efficiency in relation to the other estimators.

6.3 | Average bias and root mean squared error of random-effect estimators

The average bias for τ^2 exhibited negligible Monte Carlo error in all conditions ($Mdn < 0.002$, $Max = 0.014$). However, bias was affected by the chosen meta-analytic method ($\eta^2 = 34.4\%$) including their two-way interactions with the true pre-post correlation ($\eta^2 = 6.7\%$) and the average sample size ($\eta^2 = 6.6\%$). Figure 3 highlights that these effects were primarily driven by UMA-I and RVE. Imputing a constant pre-post correlation of 0.8 led to an overestimation of the between-study heterogeneity in situations where the true correlation was substantially smaller, particularly at small sample sizes when few primary studies were available. In contrast, for $\text{UMA-I}_{(r=0.5)}$ a mismatch between the imputed and true correlation was less severe. A highly similar pattern was observed for RVE that resulted in a positive bias when using a correlation that was larger than the true pre-post correlation. In contrast,

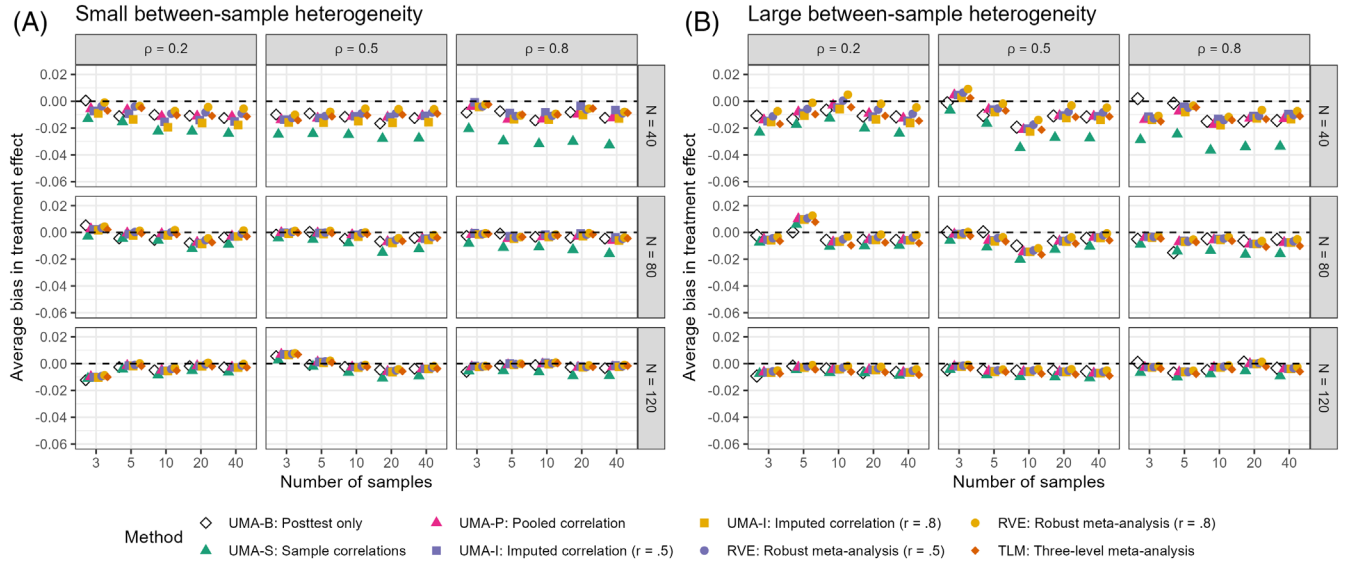

FIGURE 1 Average bias of fixed-effect estimators for a medium treatment effect, homogenous posttest variances, and no attrition bias. Detailed results are given in Tables S1 and S2 of the supplement material. [Colour figure can be viewed at wileyonlinelibrary.com]

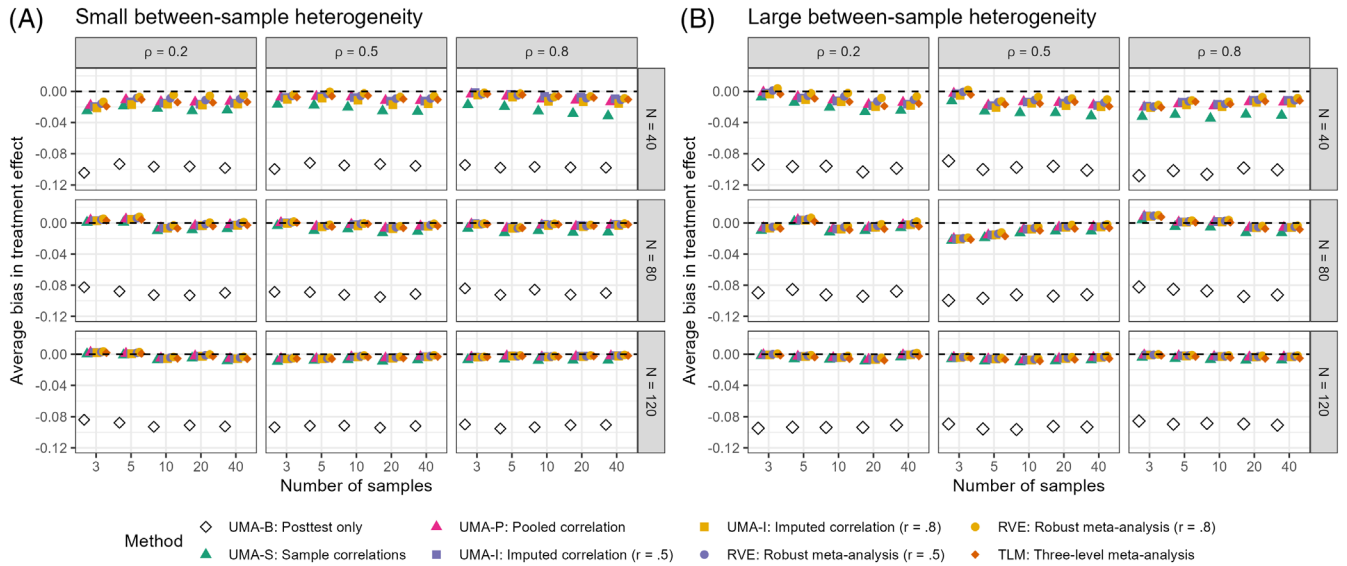

FIGURE 2 Average bias of fixed-effect estimators for a medium treatment effect, heterogeneous posttest variances, and attrition bias. [Colour figure can be viewed at wileyonlinelibrary.com]

TLM was less biased, except in conditions with large between-sample heterogeneity. In these instances, TLM resulted in slightly negative biases, particularly when the true pre-post correlation was large. Similarly, UMA-B was largely unbiased across most conditions. Posttest heterogeneity or attrition bias did not affect UMA-B or any of the other estimators (see Figure S3).

The results for the RMSE of the examined estimators (MCE: $Mdn = 0.002$, $Max = 0.009$) mirrored those for the bias. Again, UMA-I$_{(r = 0.8)}$ and RVE$_{(r = 0.8)}$ showed larger RMSE at small sample sizes and small pre-post correlations, while UMA-I$_{(r = 0.5)}$ and RVE$_{(r = 0.5)}$ were more efficient across all conditions (see Figure S4). In contrast, TLM seemed as efficient as the univariate meta-analyses. Again, posttest variance heterogeneity or attrition bias did not affect these results.

6.4 | Coverage rates of confidence intervals

The coverage rates of the 95% confidence intervals (MCE: $Mdn = 0.690$, $Max = 1.436$) were substantially affected by the meta-analytic method (see Table 2). The respective

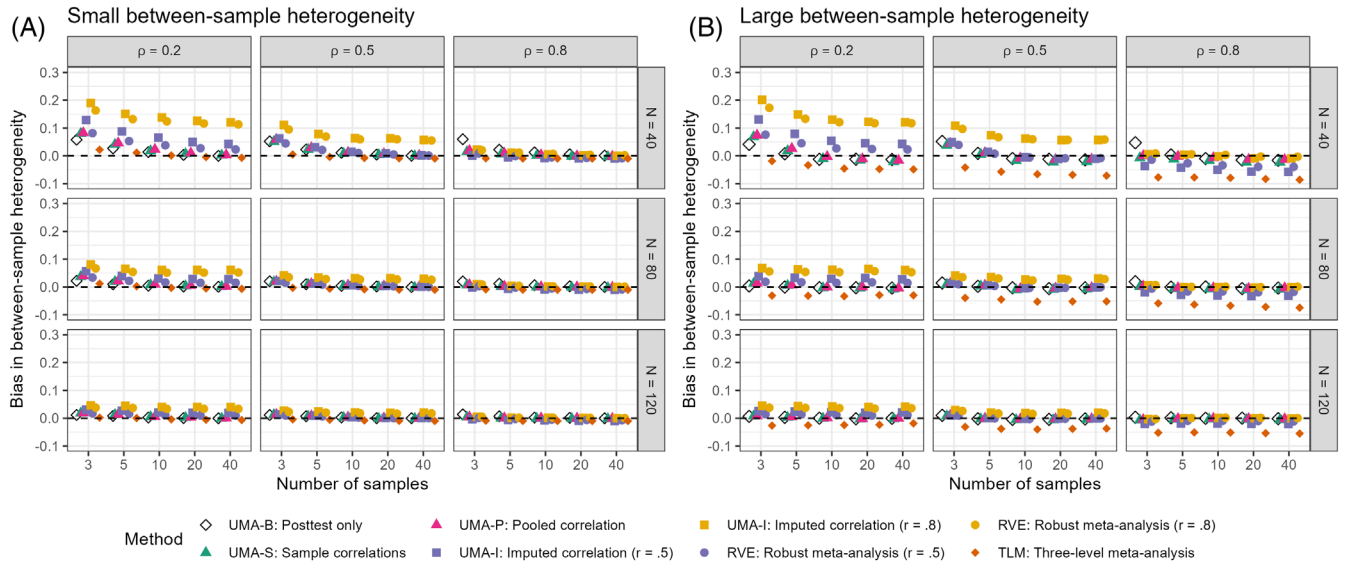

FIGURE 3 Average bias of random-effect estimators for a medium treatment effect, homogenous posttest variances, and no attrition bias. Detailed results are given in Tables S3 and S4 of the supplement material. [Colour figure can be viewed at wileyonlinelibrary.com]

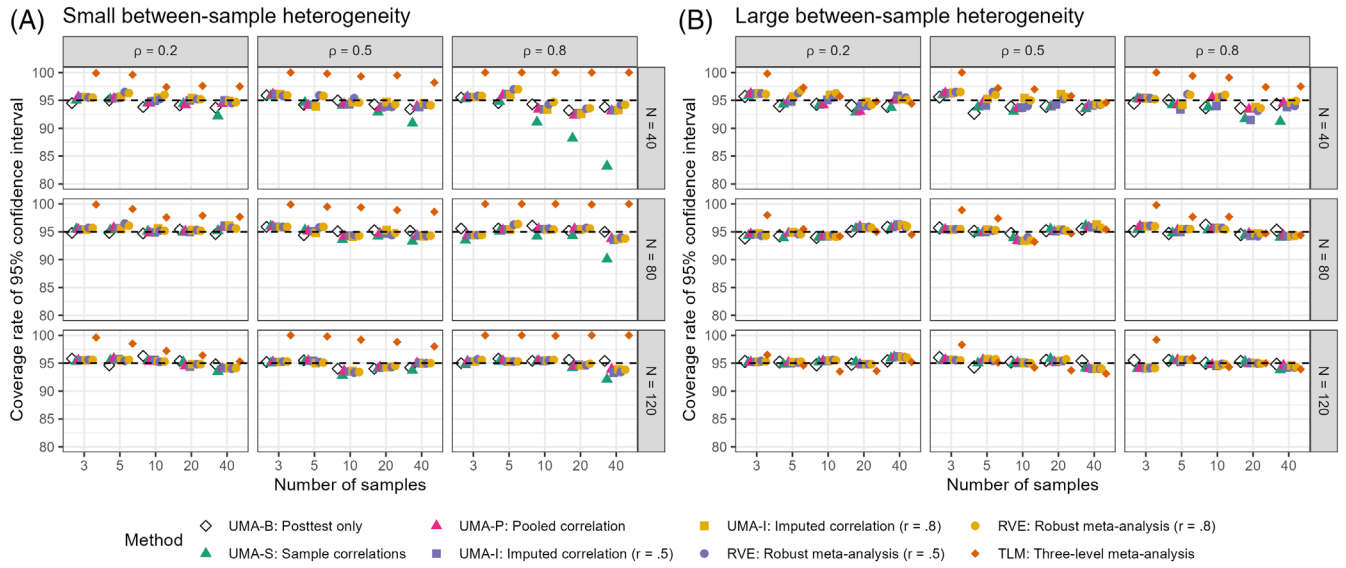

FIGURE 4 Coverage rates of 95% confidence intervals for a medium treatment effect, homogenous posttest variances, and no attrition bias. Detailed results are given in Tables S5 and S6 of the supplement material. [Colour figure can be viewed at wileyonlinelibrary.com]

main effect ($\eta^2 = 38.1\%$) was additionally qualified by two-way interactions with the between-sample heterogeneity ($\eta^2 = 10.8\%$) and, to a lesser degree, also the pre-post correlation ($\eta^2 = 5.2\%$), number of samples ($\eta^2 = 5.9\%$), and average sample size ($\eta^2 = 4.2\%$). The results in Figure 4 indicate that TLMs exhibited over-coverage, particularly when the between-sample heterogeneity was small or the true pre-post correlations were large. In contrast, the other estimators achieved coverage rates close to the nominal 95%. Only at small sample sizes, meta-analyses using the known pre-post correlations (UMA-S) showed undercoverage in a few conditions. Again, UMA-B which ignored the pretest information exhibited substantially lower coverage rates when attrition bias was present or posttest variances were larger as compared to the pretest (see Figure S5). In the most extreme cases, for example, for a large true effect, the respective coverage rate fell as low as 1.3%.

6.5 | Coverage rates of prediction intervals

The coverage rates of the 95% prediction intervals (MCE: $Mdn = 0.774$, $Max = 1.582$) were substantially affected by the meta-analytic method ($\eta^2 = 15.4\%$)

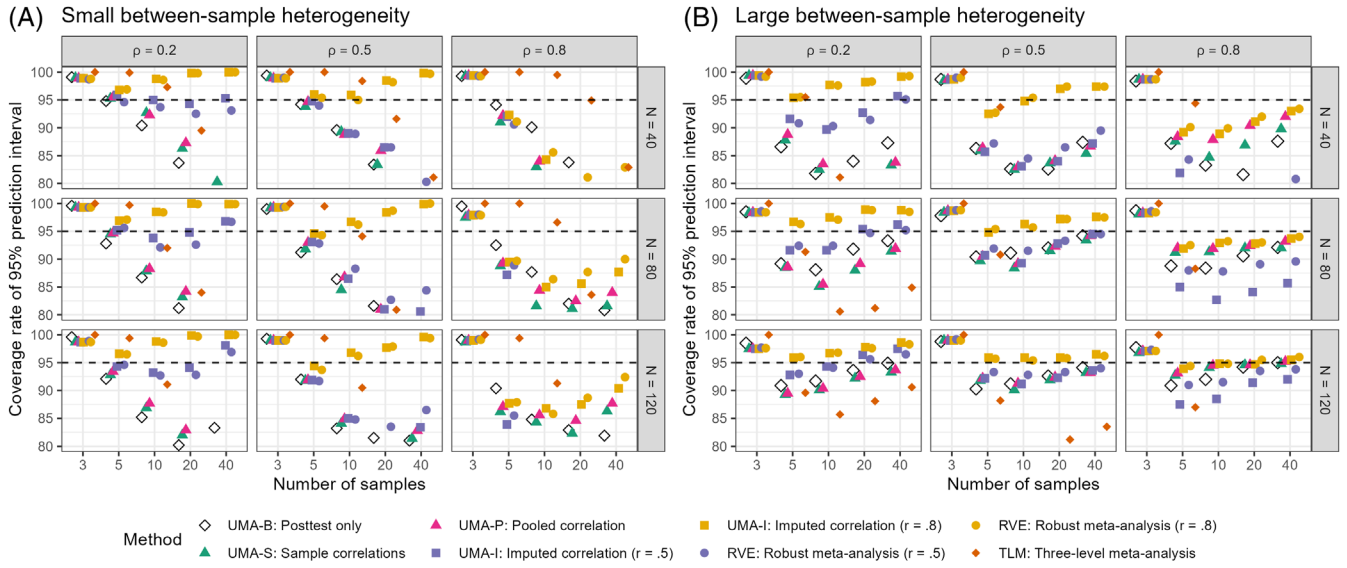

FIGURE 5 Coverage rates of 95% prediction intervals for a medium treatment effect, homogenous posttest variances, and no attrition bias. Results for values falling below 0.80 are not presented. Full are results are given in Tables S7 and S8 of the supplement material. [Colour figure can be viewed at wileyonlinelibrary.com]

which was qualified by two-way interactions with the number of samples ($\eta^2 = 14.8\%$). Across all conditions, the best coverage rates were observed for UMA-$I_{(r = 0.8)}$ and $RVE_{(r = 0.8)}$ with median coverage rates falling between 92% and 97% ($Min = 78\%$; see Figure 5). However, at small between-sample heterogeneity and large pre-post correlations these estimators resulted in substantial undercoverage, while they were close to the nominal level in the remaining conditions. Using $r = 0.5$ for the unknown pre-post correlation led to more heterogenous results with coverage rates falling as low as 39% and 49% for UMA-$I_{(r = 0.5)}$ and $RVE_{(r = 0.5)}$, respectively. TLM exhibited severe undercoverage in most conditions with a median coverage rate of 90%. Particularly, at large between-sample heterogeneity and large pre-post correlations coverage rates for TLM were rather low ($Min = 42\%$). Univariate meta-analyses using the known pre-post correlations (UMA-S or UMA-P) showed rather severe undercoverage in most conditions, falling as low as 73% ($Mdn = 90\%$). UMA-B exhibited coverage rates comparable to the other univariate estimators. Attrition bias or posttest variance heterogeneity amplified the undercoverage for UMA-B but had negligible effects on the other estimators (see Figure S6).

6.6 | Widths of 95% confidence and prediction intervals

The widths of the 95% confidence intervals (MCE: $Mdn = 0.003$, $Max = 0.049$) and the 95% prediction intervals (MCE: $Mdn = 0.014$, $Max = 0.304$) were hardly affected by the meta-analytic method (see Table 2). The estimators and their interactions explained between 0.0% and 0.8% of the variance in the performance indicator. Thus, the examined meta-analytic estimators did not substantially affect the widths of these intervals.

7 | ILLUSTRATIVE DATA EXAMPLE

To demonstrate the effect of the different approaches for the calculation of pooled RCT effects, let us consider a reanalysis of an existing RCT meta-analysis. This example aims to demonstrate that the choice of the meta-analytic method can matter and yield non-negligible variations in pooled RCT effects depending on the modeling approach. Carl and colleagues[72] evaluated the efficacy of virtual reality exposure therapy for the treatment of various anxiety-related disorders. The 14 RCTs on specific phobias and social anxieties included in the reanalysis provided mean pre- and posttest scores for a treatment group and an untreated control group (waitlist) with the respective standard deviations. As is common in clinical research reports, none of the primary studies provided the pre-post correlation. Therefore, we estimated the pooled effects with two univariate meta-analyses of RCT effect sizes that imputed a constant value of either 0.5 or 0.8 for the missing pre-post correlation. In addition, three multivariate meta-analyses of independent group effect sizes were conducted that either used RVE with working

TABLE 3 Reanalysis of Carl et al.[72]

		95% Confidence Interval				95% Prediction Interval		
	Δ	LB	UB	Width	τ	LB	UB	Width
Univariate meta-analysis of posttest effects	−0.96	−1.33	−0.59	0.74	0.53	−2.17	0.26	2.43
Univariate meta-analyses of RCT effects								
With imputed pre-post correlations ($r = 0.5$)	−1.06	−1.36	−0.77	0.59	0.40	−1.98	−0.15	1.83
With imputed pre-post correlations ($r = 0.8$)	−1.02	−1.33	−0.72	0.60	0.45	−2.05	0.00	2.06
Multivariate meta-analyses								
With robust standard errors ($r = 0.5$)	−1.06	−1.36	−0.76	0.60	0.47	−2.12	0.00	2.12
With robust standard errors ($r = 0.8$)	−1.06	−1.37	−0.76	0.61	0.50	−2.19	0.06	2.26
With an additional random effect	−1.04	−1.27	−0.81	0.46	0.23	−1.60	−0.48	1.12

Note: Based on 14 independent effect sizes with a median sample size of 32. Δ = Pooled effect; τ^2 = Between-sample heterogeneity; LB = lower bound; UB = upper bound; width interval width as UB − LB.

models assuming correlations of either 0.5 or 0.8 between the dependent effects or a TLM that accounted for dependencies with an additional random effect. As a point of comparison, we also report the results of a meta-analysis of posttest effect sizes that ignore any pretest information (Table 3).

As summarized in Table 3, the univariate meta-analyses of RCT effect sizes resulted in pooled point estimates Δ between −1.06 and −1.02, depending on the size of the imputed pre-post correlation. Similarly, the multivariate meta-analyses exhibited effect estimates around −1.06 to −1.04. In contrast, the meta-analysis of posttest effect sizes identified a slightly smaller effect of −0.96, thus, suggesting the presence of potential selection or time-selection interaction biases. The precision of the RCT effects was similar for all approaches except TLM and resulted in 95% confidence intervals of comparable widths. In contrast, the univariate meta-analysis of posttest effects exhibited a substantially larger interval.

The between-sample heterogeneity was more strongly affected by the meta-analytic estimator. The univariate meta-analyses showed, on average, slightly smaller between-sample heterogeneities as compared to the multivariate methods. Generally, the heterogeneity estimates slightly increased when imputing larger pre-post correlations or using larger correlations in the RVEs. As a result, the prediction intervals varied to some degree between the examined approaches leading to different conclusions about hypothetical effects predicted for future studies. For some estimators, the prediction intervals included 0, whereas for others they did not. In line with the simulation results, the TLM estimated a substantially smaller random effect and, consequently, a narrower prediction interval. Thus, the choice of the meta-analytic method affected the fixed effect estimate only modestly, but more so the heterogeneity estimates.

8 | DISCUSSION

In many disciplines such as clinical, psychological, and educational research, treatment or intervention effects are of primary interest to evaluate, for example, the effectiveness of novel therapies or training programs.[72,73] Because individual studies might be affected by a multitude of factors, meta-analyses try to consolidate the available evidence of multiple studies on a common topic by estimating whether an intervention yields robust effects in different settings to better understand the conditions under which an intervention might be more or less effective. Meta-analyses rely on sample statistics to calculate effect sizes in each sample. Unfortunately, relevant information for these calculations is frequently unavailable due to poor reporting practices in primary studies. Particularly the correlation between pre- and posttest scores which is required for RCT meta-analyses is often missing. As an ad-hoc solution, applied researchers frequently impute a constant value for the missing correlation without knowing how this might affect the pooled estimates. Therefore, the present study evaluated different meta-analytic estimators for the RCT effect. In addition to univariate meta-analyses of RCT effect sizes with known or imputed pre-post correlations, we also proposed two new multivariate meta-regression approaches which capitalize on recent advancements for the analysis of dependent effects in meta-analyses.[11,13] A comprehensive simulation study evaluated the different analytic approaches under different realistic conditions that are typically encountered in applied research. These analyses provided five major results:

First, traditional univariate meta-analyses of RCT effect sizes resulted in more biased point estimates as compared to the other estimators and tended to underestimate the true effect, particularly when sample sizes

were small and the true pre-post correlation was large. In contrast, a substantial improvement was observed when pooling the pre-post correlations and using the pooled estimates for the calculation of the sampling variances of the RCT effect sizes. Respective univariate meta-analyses were largely unbiased in most conditions. However, the confidence intervals for both estimators held the nominal error level in most of the examined conditions and, thus, did not indicate substantially different interval estimates.

Second, univariate meta-analyses of RCT effect sizes with imputed pre-post correlations also yielded largely unbiased estimates of the true effect and appropriate coverage rates of respective confidence intervals. However, estimates of the between-study heterogeneity were substantially biased when there was a mismatch between the imputed and the true pre-post correlation. This bias was larger for small sample sizes and when imputing a correlation that was too large as compared to imputing a too small of a correlation.

Third, multivariate meta-analyses exhibited largely unbiased estimates of the true effect in most conditions. Albeit, at smaller average sample sizes or larger true effects $RVE_{(r = 0.8)}$ was slightly less biased. However, TLM often resulted in overcoverage of the confidence intervals, while RVE held the nominal error rates in most conditions. Importantly, little differences were observed for RVE that assumed a correlation of 0.5 or 0.8 in the working model. However, the random effect estimates were slightly overestimated at small sample sizes when few primary studies were available, slightly stronger so for RVEs assuming larger correlations in the working model. In contrast, TLM resulted in more biased random effects across most conditions.

Fourth, all examined estimators had difficulties holding the nominal error rates for the prediction intervals, thus, mirroring previous results for simpler meta-analytic designs that also revealed far too low coverage rates for prediction intervals in most studied conditions.[74] Although prediction intervals reached close to nominal levels for $RVE_{(r = 0.8)}$ and $UMA\text{-}I_{(r = 0.8)}$ in many conditions, they tended to exhibit undercoverage at small between-sample heterogeneity with large pre-post correlations. In contrast, TLM showed substantial undercoverage across most conditions, particularly in the presence of large between-sample heterogeneity. Also, the univariate meta-analyses of individual or pooled correlations did not hold the coverage probabilities. Thus, the generalization of effects is seriously hampered because the estimation of reasonably expected effects in future RCT studies is subject to substantial imprecision.

Finally, a rather robust finding pertained to the effects of posttest variance heterogeneity and the presence of attrition bias. Neither attrition bias nor a rather large variance heterogeneity in the treatment group at the posttest affected the point estimates of fixed and random effects or the respective interval estimates, as long as the meta-analytic estimator incorporated the pretest information.[3] In contrast, meta-analyses of posttest effect sizes that ignored the pretest information were affected by both sources of error. Consequently, this estimator yielded substantially biased point estimates and also distorted confidence intervals.

Revisiting the introductory example on the efficacy of virtual reality exposure therapy for the treatment of anxiety-related disorders,[72] the simulation results might inform about the trustworthiness of the results juxtaposited in Table 1. Given that the empirical meta-analysis approximates the simulation condition with a small average sample size, a medium number of samples, a large true effect, and large between-sample heterogeneity, the larger fixed effect reported by $RVE_{(r = 0.8)}$ seems more plausible than the smaller effects. Moreover, the heterogeneity estimate identified by TLM seems less trustworthy because in contrast to the other estimators it systematically underestimates the between-study variance. For RVE or UMA-I, the respective prediction interval might be too small or too wide depending on the unknown pre-post correlation. However, given the poor coverage rates of the prediction intervals for all estimators, these generally need to be interpreted cautiously.

8.1 | Recommendations for meta-analytic practice

Even though the true RCT effect and the true pre-post correlation are unknown in practice, a consideration of the simulation results under the examined conditions led us to put forward the following recommendations. If most of the primary studies report sample-specific pre-post correlations (and there are no obvious systematic omissions), a univariate meta-analysis of RCT effect sizes following the two-step approach is recommended. Thus, first, a meta-analysis of the available pre-post correlations is conducted and, then, the pooled correlation is used for the calculation of the sampling variances of the RCT effect sizes. This approach yielded largely unbiased point estimates of the fixed and held the nominal coverage rate for the 95% confidence interval. If only a few or no pre-post correlations are available, we recommend using multivariate meta-analyses of independent group effect sizes with RVE that adopts a large correlation (e.g., $r = 0.8$) in the working model. In our simulation, this approach yielded largely comparable results to the univariate approach in most conditions and also yields more precise estimates of the between-sample variance.

Alternatively, a univariate meta-analysis of RCT effects with imputed pre-post correlations of $r = 0.8$ might be used which fared comparably to $RVE_{(r = 0.8)}$. Although none of the estimators resulted in trustworthy prediction intervals, $RVE_{(r = 0.8)}$ or $UMA\text{-}I_{(r = 0.8)}$ might be preferred because it resulted in more consistent coverage rates that were less affected by the unknown pre-post correlations. In contrast, univariate meta-analytic approaches with known pre-post correlations or assuming small pre-post correlations such as $UMA\text{-}I_{(r = 0.5)}$ or $RVE_{(r = 0.5)}$ yielded slightly worse coverage rates and, thus, are not recommended. However, generally, the differences between the studied estimators were rather small in most conditions. Therefore, the choice of estimator has likely only minor implications for meta-analytic results in practice. Finally, we want to caution against the imprudent use of meta-analyses using posttest effect sizes that ignore pretest information. These can result in substantially biased estimates of fixed effects unless negligible attrition bias and variance homogeneity can be guaranteed.

8.2 | Limitations and future directions

Although the aim of the study was the formulation of clear-cut recommendations for the analysis of RCT effect sizes based on a comprehensive simulation study, some weaknesses limit the generalizability of the presented findings and open avenues for future research. First of all, our results only pertain to meta-analyses of standardized mean differences for metric outcomes. Although continuous variables dominate psychological research, particularly clinical studies often also employ dichotomous (or less frequently, multinomial) outcomes that classify individuals into different groups such as improved versus not improved, recidivistic versus not recidivistic, or symptomatic versus asymptomatic. These results might be similarly synthesized across multiple samples, but require different effect sizes (e.g., odds ratios, risk ratios). However, the choice of the effect size might alter recommendations for a meta-analytic estimator.[20] Until these are available, applied researchers are encouraged to conduct sensitivity analyses with different estimators to compare the robustness of the meta-analytic results.

Second, in line with prevalent practice, the compared estimators assumed normally distributed effects which might not be tenable in some situations, particularly when the number of studies is small. Although point and interval estimates of meta-analyses including this normality assumption are quite robust, even when the true effects are severely skewed,[50,75] alternative parametric or mixture distributions might improve the accuracy of the heterogeneity estimates and prediction intervals (see Higgins and colleagues[76] for a review). Therefore, future research could evaluate the precision of RCT meta-analyses for different distributional assumptions of the between-study effect.

Third, as has been previously shown in the context of univariate meta-analysis and RVE, small-sample corrections are important to estimate precise standard errors and confidence intervals.[37,77] However, for TLM respective adjustments such as the Kenward-Rogers[78] correction have not yet been thoroughly evaluated and, thus, are hardly used.[42] To overcome the problematic coverage rates of TLM that were observed in the present study, we strongly encourage further research on the development of small-sample adjustments for these settings.

Fourth, the precise estimation of between-study heterogeneity is an unresolved challenge in meta-analytic research. Simulation studies showed that in many scenarios the coverage rates of prediction intervals are far too low, particularly for heterogeneous study sample sizes.[74] Therefore, future research is encouraged to improve prediction intervals for RCT meta-analyses, for example, using a bootstrap approach.[79]

Fifth, although the simulation studies tried to cover a broad range of realistic conditions, empirical data typically is noisier and might not fully match the simulated conditions. For example, we did not specifically evaluate how outliers (i.e., extreme effect sizes), sample attrition, or publication bias might have affected the meta-analytic results. A fruitful extension could also focus on differences between the proposed estimators for the identification of moderating effects.

Finally, our simulation was limited to estimators commonly implemented in standard software that is used by applied researchers. We readily acknowledge that alternative approaches can also account for missing correlations in multivariate meta-analyses. For example, Hong and colleagues[80] proposed a multivariate RVE model that specifies an overall marginal correlation between dependent outcomes, thus, not requiring within-study correlations. However, the reported simulations indicated an unacceptable precision of this approach for meta-analyses with few primary studies (i.e., less than 50) that dominate RCT research. Alternatively, Bayesian methods could be adapted by assuming a distribution for the missing pre-post correlations rather than imputing a constant value.[81]

9 | CONCLUSION

Meta-analyses of treatment or intervention effects in RCTs often struggle with missing information to calculate effect sizes and their sampling variances. In practice, often

ad-hoc solutions are adopted such as imputing a constant value for missing pre-post correlations without knowing the consequences for the meta-analytic results. The presented simulation study suggested that imputing a constant correlation of 0.8 might work well for estimating the pooled effect, but slightly distorts the between-study heterogeneity. Alternatively, we recommend a multivariate meta-regression approach with RVE that estimates the difference in independent group effect sizes without relying on known pre-post correlations.

AUTHOR CONTRIBUTIONS

Timo Gnambs: Conceptualization; formal analysis; investigation; methodology; software; writing – original draft. **Ulrich Schroeders:** Investigation; methodology; writing – review and editing.

ACKNOWLEDGMENTS

We thank two anonymous reviewers and the editor for providing valuable feedback and recommendations to improve the manuscript during the review process. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

We have no conflicts of interest to disclose. The study was not preregistered. https://osf.io/an2dg/

DATA AVAILABILITY STATEMENT

All data, analysis code, and research material generated to produce the reported results are provided at https://osf.io/an2dg/.

ORCID

Timo Gnambs https://orcid.org/0000-0002-6984-1276
Ulrich Schroeders https://orcid.org/0000-0002-5225-1122

ENDNOTES

[1] An anonymous reviewer emphasized that the model can be equivalently defined with a random slope specification by replacing u_{kt} in (8) with $u_{kt,0} + u_{kt,1} \cdot t$. Then, the random slope variance τ_1^2 represents the total between-study heterogeneity for the RCT effect size, that is, τ^2 in (6).

[2] The total between-study heterogeneity for the RCT effect size in the TLM corresponds to the variance of the difference in pre-post effect sizes such that $Var(\Delta) = Var(\delta_{k1} - \delta_{k0}) = Var([\beta_0 + \beta_1 + u_{k1} + u_k] - [\beta_0 + u_{k0} + u_k]) = Var(\beta_1 + u_{k1} - u_{k0}) = Var(u_{k1} - u_{k0}) = Var(u_{k1}) + Var(u_{k0})$. Because the TLM assumes similar between-study heterogeneities for all effect sizes, that is, $Var(u_{k0}) = Var(u_{k1}) = \tau_w^2$, the variance of the RCT effect size reduces to $2 \cdot \tau_w^2$.

[3] Supplement J also demonstrates that pretest imbalance, that is, random between-study variance in pretest scores does not affect the studied meta-analytic estimators differently. Thus, the reported results are expected to generalize to conditions with baseline imbalance.

REFERENCES

1. Jones DS, Podolsky SH. The history and fate of the gold standard. *Lancet*. 2015;285:1502-1503. doi:10.1016/S0140-6736(15)60742-5
2. Shadish WR, Cook D, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton; 2002.
3. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods*. 2002;7:105-125. doi:10.1037/1082-989X.7.1.105
4. Huang D, Yu H, Wang T, Yang H, Yao R, Liang Z. Efficacy and safety of umifenovir for coronavirus disease 2019 (COVID-19): a systematic review and meta-analysis. *J Med Virol*. 2021;93:481-490. doi:10.1002/jmv.26256
5. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:340. doi:10.1136/bmj.c221
6. Hardwicke TE, Thibault RT, Kosie JE, Wallach JD, Kidwell MC, Ioannidis JP. Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspect Psychol Sci*. 2022;17(1):239-251. doi:10.1177/1745691620979806
7. Nutu D, Gentili C, Naudet F, Cristea IA. Open science practices in clinical psychology journals: an audit study. *J Abnorm Psychol*. 2019;128(6):510-516. doi:10.1037/abn0000414
8. Grund S, Lüdtke O, Robitzsch A. Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychological Methods*. 2022. Advance online publication. doi:10.1037/met0000526
9. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. *Am Psychol*. 2018;73(1):3-25. doi:10.1037/amp0000191
10. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods*. 2010;1(1):39-65. doi:10.1002/jrsm.5
11. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: expanding the range of working models. *Prev Sci*. 2022;34:435-438. doi:10.1007/s11121-021-01246-3
12. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. *Res Synth Methods*. 2011;2(1):61-76. doi:10.1002/jrsm.35
13. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. *Behav Res Methods*. 2013;45(2):576-594. doi:10.3758/s13428-012-0261-6
14. Becker BJ. Synthesizing standardized mean-change measures. *Br J Math Stat Psychol*. 1988;41(2):257-278. doi:10.1111/j.2044-8317.1988.tb00901.x
15. Hedges LV, Tipton E, Zejnullahi R, Didaz KG. Effect sizes in ANCOVA and difference-in-differences designs. *Br Math J Stat Psychol*. 2023;76(2):259-282. doi:10.1111/bmsp.12296
16. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. *Org Res Methods*. 2008;11(2):364-386. doi:10.1177/1094428106291059

17. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat*. 1981;6(2):107-128. doi:10.3102/10769986006002107
18. Hedges LV. Estimation of effect size from a series of independent experiments. *Psychol Bull*. 1982;92:490-499. doi:10.1037/0033-2909.92.2.490
19. Rubio-Aparicio M, Marín-Martínez F, Sánchez-Meca J, López-López JA. A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behav Res Methods*. 2018;50(5):2057-2073. doi:10.3758/s13428-017-0973-8
20. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79. doi:10.1002/jrsm.1164
21. Raudenbush SW. Random effects models. In: Cooper H, Hedges LV, eds. *The Handbook of Research Synthesis*. Russell Sage; 1994:301-321.
22. Klimek P, Wei B, Blashill AJ. Exploring moderators of mirror exposure on pre-to post changes in body image outcomes: systematic review and meta-analysis. *Eat Disord*. 2022;30(1):77-98. doi:10.1080/10640266.2020.1791665
23. Rosenthal R. *Meta-Analytic Procedures for Social Science Research*. Sage. 1991.
24. Balk EM, Earley A, Patel K, Trikalinos TA, Dahabreh IJ. Empirical Assessment of Within-Arm Correlation Imputation in Trials of Continuous Outcomes (Methods Research Report). AHRQ Publication No. 12(13)-EHC141-EF. Agency for Healthcare Research and Quality. 2012 http://www.effectivehealthcare.ahrq.gov/reports/final.cfm
25. Cuijpers P, Weitz E, Cristea IA, Twisk J. Pre-post effect sizes should be avoided in meta-analyses. *Epidemiol Psychiatr Sci*. 2017;26(4):364-368. doi:10.1017/S2045796016000809
26. Taylor PJ, Russ-Eft DF, Chan DW. A meta-analytic review of behavior modeling training. *J Appl Psychol*. 2005;90(4):692-709. doi:10.1037/0021-9010.90.4.692
27. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *J R Stat Soc: Ser A (Stat Soc)*. 2009;172(4):789-811. doi:10.1111/j.1467-985X.2008.00593.x
28. Hulshof TA, Zuidema SU, van Meer PJ, Gispen-de Wied CC, Luijendijk HJ. Baseline imbalances and clinical outcomes of atypical antipsychotics in dementia: a meta-epidemiological study of randomized trials. *Int J Methods Psychiatr Res*. 2019;28(1):e1757. doi:10.1002/mpr.1757
29. Schönbrodt FD, Perugini M. At what sample size do correlations stabilize? *J Res Pers*. 2013;47(5):609-612. doi:10.1016/j.jrp.2013.05.009
30. Feingold A. A regression framework for effect size assessments in longitudinal modeling of group differences. *Rev Gen Psychol*. 2013;17(1):111-121. doi:10.1037/a0030048
31. Lin Y, Zhu M, Su Z. The pursuit of balance: an overview of covariate-adaptive randomization techniques in clinical trials. *Contemp Clin Trials*. 2015;45:21-25. doi:10.1016/j.cct.2015.07.011
32. Park EG, Hahn S. An approach to exploring patterns of imbalance and potential missingness in reports of the randomized baseline values for primary outcomes measurable at baseline in randomized controlled trials for meta-analyses. *BMC Med Res Methodol*. 2022;22:22. doi:10.1186/s12874-022-01620-x
33. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11(1):1-11. doi:10.1186/1471-2288-11-160
34. Raue A, Kreutz C, Maiwald T, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*. 2009;25(15):1923-1929. doi:10.1093/bioinformatics/btp358
35. Wieland FG, Hauber AL, Rosenblatt M, Tönsing C, Timmer J. On structural and practical identifiability. *Curr Opin Syst Biol*. 2021;25:60-69. doi:10.1016/j.coisb.2021.03.005
36. Viechtbauer W. Hypothesis tests for population heterogeneity in meta-analysis. *Br J Math Stat Psychol*. 2007;60(1):29-60. doi:10.1348/000711005X64042
37. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods*. 2015;20(3):375-393. doi:10.1037/met0000011
38. Tipton E, Pustejovsky JE. Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *J Educ Behav Stat*. 2015;40:604-634. doi:10.3102/1076998615606099
39. Cheung MWL. *Meta-Analysis: A Structural Equation Modeling Approach*. Wiley; 2015.
40. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Meta-analysis of multiple outcomes: a multilevel approach. *Behav Res Methods*. 2015;47(4):1274-1294. doi:10.3758/s13428-014-0527-2
41. Fernández-Castilla B, Jamshidi L, Declercq L, Beretvas SN, Onghena P, Van den Noortgate W. The application of meta-analytic (multi-level) models with multiple random effects: a systematic review. *Behav Res Methods*. 2020;52(5):2031-2052. doi:10.3758/s13428-020-01373-9
42. Park S, Beretvas SN. Synthesizing effects for multiple outcomes per study using robust variance estimation versus the three-level model. *Behav Res Meth*. 2019;51(1):152-171. doi:10.3758/s13428-018-1156-y
43. Novianti PW, Roes KC, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemp Clin Trials*. 2014;37(1):129-138. doi:10.1016/j.cct.2013.11.012
44. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22(17):2693-2710. doi:10.1002/sim.1482
45. Brannick MT, French KA, Rothstein HR, Kiselica AM, Apostoloski N. Capturing the underlying distribution in meta-analysis: credibility and tolerance intervals. *Res Synth Methods*. 2021;12(3):264-290. doi:10.1002/jrsm.1479
46. Pustejovsky JE, Tipton E. Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *J Bus Econ Stat*. 2018;36(4):672-683. doi:10.1080/07350015.2016.1247004
47. Ahn S, Ames AJ, Myers ND. A review of meta-analyses in education: methodological strengths and weaknesses. *Rev Educ Res*. 2012;82(4):436-476. doi:10.3102/0034654312458162
48. Tipton E, Pustejovsky JE, Ahmadi H. Current practices in meta-regression in psychology, education, and medicine. *Res Syn Methods*. 2019;10(2):180-194. doi:10.1002/jrsm.1339
49. Collins DB, Holton EF III. The effectiveness of managerial leadership development programs: a meta-analysis of studies

from 1982 to 2001. *Hum Resour Dev Q*. 2004;15(2):217-248. doi:10.1002/hrdq.1099

50. Rubio-Aparicio M, López-López JA, Sánchez-Meca J, Marín-Martínez F, Viechtbauer W, Van der Noortgate W. Estimating an overall effect size in random-effects meta-analysis when the distribution of random effects departs from normal. *Res Syn Methods*. 2018;9(3):489-503. doi:10.1002/jrsm.1312
51. Sánchez-Meca J, Marín-Martínez F. Weighting by inverse variance or by sample size in meta-analysis: a simulation study. *Educ Psychol Meas*. 1998;58:211-220. doi:10.1177/0013164498058002005
52. Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods*. 2008;13:31-48. doi:10.1037/1082-989X.13.1.31
53. Lamberink HJ, Otte WM, Sinke MR, et al. Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. *J Clin Epidemiol*. 2018;102:123-128. doi:10.1016/j.jclinepi.2018.06.014
54. Evans DK, Yuan F. How big are effect sizes in international education studies? *Educ Eval Policy Anal*. 2022;44:532-540. doi:10.3102/01623737221079646
55. Kraft MA. Interpreting effect sizes of education interventions. *Educ Res*. 2020;49:241-253. doi:10.3102/0013189X20912798
56. Carlson KD, Schmidt FL. Impact of experimental design on effect size: findings from the research literature on training. *J Appl Psychol*. 1999;84(6):851-862. doi:10.1037/0021-9010.84.6.851
57. Plöderl M, Hengartner MP. What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis. *BMJ Open*. 2019;9(12):e034816. doi:10.1136/bmjopen-2019-034816
58. Volkmann C, Volkmann A, Müller CA. On the treatment effect heterogeneity of antidepressants in major depression: a Bayesian meta-analysis and simulation study. *PloS One*. 2020;15(11):e0241497. doi:10.1371/journal.pone.0241497
59. Van Erp S, Verhagen J, Grasman RP, Wagenmakers EJ. Estimates of between-study heterogeneity for 705 meta-analyses reported in psychological bulletin from 1990–2013. *J Open Psychol Data*. 2017;5(1). doi:10.5334/jopd.33
60. Linden AH, Hönekopp J. Heterogeneity of research results: a new perspective from which to assess and promote progress in psychological science. *Perspect Psychol Sci*. 2021;16(2):358-376. doi:10.1177/1745691620964193
61. Crutzen R, Viechtbauer W, Kotz D, Spigt M. No differential attrition was found in randomized controlled trials published in general medical journals: a meta-analysis. *J Clin Epidemiol*. 2013;66(9):948-954. doi:10.1016/j.jclinepi.2013.03.019
62. Demack S, Maxwell B, Coldwell M, et al. *Review of EEF Reports*. Education Endowment Foundation; 2021.
63. Crutzen R, Viechtbauer W, Spigt M, Kotz D. Differential attrition in health behaviour change trials: a systematic review and meta-analysis. *Psychol Health*. 2015;30(1):122-134. doi:10.1080/08870446.2014.953526
64. What Works Clearinghouse. What Works Clearinghouse Standards Handbook (Version 4.1). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. 2020 https://ies.ed.gov/ncee/wwc/handbooks
65. Weidmann B, Miratrix L. Missing, presumed different: quantifying the risk of attrition bias in education evaluations. *J R Stat Soc Ser A Stat Soc*. 2021;184(2):732-760. doi:10.1111/rssa.12677
66. Hewitt CE, Kumaravel B, Dumville JC, Torgerson DJ. Assessing the impact of attrition in randomized controlled trials. *J Clin Epidemiol*. 2010;63(11):1264-1270. doi:10.1016/j.jclinepi.2010.01.010
67. Gay DM. *Usage Summary for Selected Optimization Routines (Computing Science Technical Report 153)*. AT&T Bell Laboratories; 1990.
68. Nelder JA, Mead R. A simplex algorithm for function minimization. *Comp J*. 1965;7:308-313. doi:10.1093/comjnl/7.4.308
69. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48. doi:10.18637/jss.v036.i03
70. Pustejovsky JE. clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections (R package version 0.5.7). 2021 https://CRAN.R-project.org/package=clubSandwich
71. Koehler E, Brown E, Haneuse SJP. On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am Stat*. 2009;63(2):155-162. doi:10.1198/tast.2009.0030
72. Carl E, Stein AT, Levihn-Coon A, et al. Virtual reality exposure therapy for anxiety and related disorders: a meta-analysis of randomized controlled trials. *J Anxiety Disord*. 2019;61:27-36. doi:10.1016/j.janxdis.2018.08.003
73. Benavides-Varela S, Callegher CZ, Fagiolini B, Leo I, Altoe G, Lucangeli D. Effectiveness of digital-based interventions for children with mathematical learning difficulties: a meta-analysis. *Comput Educ*. 2020;157:103953. doi:10.1016/j.compedu.2020.103953
74. Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. *Stat Med*. 2017;36(2):301-317. doi:10.1002/sim.7140
75. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Stat Methods Med Res*. 2012;21(4):409-426. doi:10.1177/0962280210392008
76. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc: Ser A*. 2009;172(1):137-159. doi:10.1111/j.1467-985X.2008.00552.x
77. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014;14(1):1-12. doi:10.1186/1471-2288-14-25
78. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997;53:983-997. doi:10.2307/2533558
79. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. *Stat Methods Med Res*. 2019;28(6):1689-1702. doi:10.1177/0962280218773520
80. Hong CD, Riley R, Chen Y. An improved method for bivariate meta-analysis when within-study correlations are unknown. *Res Synth Methods*. 2018;9(1):73-88. doi:10.1002/jrsm.1274

81. Wei Y, Higgins JP. Bayesian multivariate meta-analysis with multiple outcomes. *Stat Med*. 2013;32(17):2911-2934. doi:10.1002/sim.5745

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gnambs T, Schroeders U. Accuracy and precision of fixed and random effects in meta-analyses of randomized control trials for continuous outcomes. *Res Syn Meth*. 2023;1-21. doi:10.1002/jrsm.1673

APPENDIX A

A.1 | Robust variance estimation

Following Pustejovsky and Tipton[11], the mathematical details of robust variance estimation using weighted least squares with fully inverse-variance weights are briefly outlined. If $\mathbf{D}_k$ represents a vector of two effect sizes (i.e., at pre- and posttest) in sample $k \in \{1, ..., K\}$, $\mathbf{X}_k$ the 2×2 design matrix of covariates including the intercept and the time variable t ($0 =$ pretest, $1 =$ posttest), $\mathbf{u}_k$ the vector of two random effects, and $\mathbf{e}_k$ the vector of two sampling errors, then the random-effect meta-analytic model can be written as

$$\mathbf{D}_k = \mathbf{X}_k \cdot \boldsymbol{\beta} + \mathbf{u}_k + \mathbf{e}_k.$$

Let $\boldsymbol{\Phi}_k$ represent the 2×2 variance-covariance matrix giving the true dependency structure of the effect sizes in study k. Then, the weighted least squares estimate and the sampling variance of $\boldsymbol{\beta}$ are given by

$$\widehat{\boldsymbol{\beta}} = \mathbf{M} \cdot \left(\sum_{k=1}^{K} \mathbf{X}_k' \mathbf{W}_k \mathbf{D}_k \right),$$

where $\mathbf{M} = \left(\sum_{k=1}^{K} \mathbf{X}_k' \mathbf{W}_k \mathbf{X}_k \right)^{-1}$

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{M} \cdot \left(\sum_{k=1}^{K} \mathbf{X}_k' \mathbf{W}_k \boldsymbol{\Phi}_k \mathbf{W}_k \mathbf{X}_k \right) \cdot \mathbf{M}$$

with $\mathbf{W}_k$ denoting the 2×2 matrix of weights for study k. If $\boldsymbol{\Phi}_k$ were known, then the optimal weight matrix would be given by $\mathbf{W}_k = \boldsymbol{\phi}_k^{-1}$. Otherwise, RVE adopts a "working model" resulting in a general set of weights and approximates the study-specific variance–covariance matrix using the observed residuals $\widehat{\mathbf{e}}_k = \mathbf{D}_k - \mathbf{X}_k \widehat{\boldsymbol{\beta}}$. Then, the robust estimator for the sampling variance of $\widehat{\boldsymbol{\beta}}$ is

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{M} \cdot \left(\sum_{k=1}^{K} \mathbf{X}_k' \mathbf{W}_k \mathbf{A}_k \widehat{\mathbf{e}}_k \widehat{\mathbf{e}}_k' \mathbf{A}_k \mathbf{W}_k \mathbf{X}_k \right) \cdot \mathbf{M}$$

with $\mathbf{A}_k$ representing adjustments for small-sample bias.[37,38]