Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models

James E Pustejovsky[1] · Elizabeth Tipton[2]

Abstract

In prevention science and related fields, large meta-analyses are common, and these analyses often involve dependent effect size estimates. Robust variance estimation (RVE) methods provide a way to include all dependent effect sizes in a single meta-regression model, even when the exact form of the dependence is unknown. RVE uses a working model of the dependence structure, but the two currently available working models are limited to each describing a single type of dependence. Drawing on flexible tools from multilevel and multivariate meta-analysis, this paper describes an expanded range of working models, along with accompanying estimation methods, which offer potential benefits in terms of better capturing the types of data structures that occur in practice and, under some circumstances, improving the efficiency of meta-regression estimates. We describe how the methods can be implemented using existing software (the "metafor" and "clubSandwich" packages for R), illustrate the proposed approach in a meta-analysis of randomized trials on the effects of brief alcohol interventions for adolescents and young adults, and report findings from a simulation study evaluating the performance of the new methods.

Keywords Meta-analysis · Meta-regression · Robust standard errors · Dependent effect sizes

In the prevention sciences as in other areas of social inquiry, systematic reviews and meta-analyses are increasingly common modes of research. The need for systematic reviews has grown particularly over the past 10 years, in response to increasing numbers of randomized trials and quasi-experiments evaluating interventions targeting youth (e.g., Chhin et al., 2018). Over the past decade, *Prevention Science* alone has published thirty such reviews, including reviews on substance abuse prevention programs (Hennessy & Tanner-Smith, 2015), mental health interventions (Conley et al., 2015), and parenting interventions (Baudry et al., 2017).

The meta-analyses produced within prevention science and many other areas of social science tend to be large in scope, often including results from more than 50 studies (Tipton et al., 2019b). In reviews this large, the goal of meta-analytic synthesis is not simply to provide an estimate of the average effect but also to characterize heterogeneity and investigate sources of variation in effects. Meta-analyses routinely examine whether effect sizes vary in relation to methodological characteristics (e.g., study design), sample characteristics (e.g., average age of participants), and features of the intervention (e.g., hours of intervention, delivery mode). The primary statistical tool for investigating such questions is the meta-regression model—essentially the meta-analytic version of multiple regression.

When the goal of a research synthesis is to explore heterogeneity, it is often desirable to combine all relevant sources of evidence into a single meta-regression model. This allows tests of hypotheses regarding focal variables, such as intervention components, while holding constant methodological characteristics that might otherwise confound the analyses (Lipsey, 2003; Tipton et al., 2019a). However, including all collected effect size estimates in a single model creates complications due to the statistical dependence among the effect size estimates generated from the same study. Statistical dependence can arise from shared features of labs or research groups or as a result of multiple measurements being collected on the same individuals (e.g., outcomes at 1, 6, and 12 months). The structure of such dependence needs to be taken into account if conclusions based on

✉ James E Pustejovsky
pustejovsky@wisc.edu

[1] Educational Psychology Department, University of Wisconsin - Madison, 1082C Educational Sciences, 1025 West Johnson St, Madison, WI 53706-1706, USA

[2] Statistics Department, Northwestern University, Evanston, USA

meta-regression models are to remain valid, but doing so presents analytic challenges because the information needed to quantify the degree of dependence is rarely reported in primary studies.

Methods for handling dependent effect sizes have been available in meta-analysis since the beginning of the field (see Hedges & Olkin, 1982). Early methods, sometimes called multivariate meta-analysis, required knowledge of the exact dependence structure of the effect sizes, thus rendering them useful only in exceptional cases. Instead, analysts have typically sought to avoid dependence by dividing a larger meta-analysis into several smaller analyses, each focused on a particular outcome or particular subgroup of effects (e.g., effects measured immediately after intervention; effects based on researcher-developed outcome assessments). This approach can produce statistically valid results for each subgroup, but running separate analyses on each subgroup makes it difficult to quantify differences across levels of a moderator or to adjust for multiple moderators.

Over the past decade, focus has turned increasingly to a new method known as robust variance estimation (RVE), which provides a way to include dependent effect sizes in meta-regression, even when the nature of the dependence structure is unknown (Hedges et al., 2010; Tipton, 2013, 2015; Tipton & Pustejovsky, 2015). RVE is now widely used in meta-analyses, and software packages implementing the methods are available in both R (Fisher et al., 2017; Pustejovsky, 2020) and Stata (Hedberg, 2014; Tyszler et al., 2017). RVE methods do not require knowledge of the exact dependence structure between effect size estimates. Instead, RVE involves use of a *working model* for dependence, which approximates the dependence structure but does not need to be entirely correct. Even when the working model is misspecified, meta-regression coefficient estimates will be unbiased, and their standard errors (along with hypothesis tests and confidence intervals) will provide valid quantification of uncertainty. The primary benefit of using a more accurate working model comes from increased efficiency. Specifically, using a working model that more accurately captures the dependence structure will lead to meta-analysis and meta-regression coefficient estimates that are more precise and accurate (Tipton, 2015; Tipton & Pustejovsky, 2015).

To date, two different working models have been proposed for use with RVE methods. The first of these, called the *hierarchical effects* model, assumes that dependence arises solely through common features of a lab or research group, while within a lab or group, each effect size is estimated on an independent sample. The second of these, called the *correlated effects* model, assumes that dependence arises because effect sizes are estimated based on the same sample (e.g., multiple measures of a common outcome construct or one outcome assessed over multiple time points). The choice of one of these working models then forms the basis for a set of "approximately efficient" weights used in the meta-regression model (Hedges et al., 2010).

In practice, however, meta-analyses frequently involve *both* hierarchical and correlated effects structures. Existing guidance on use of RVE encourages analysts to select a hierarchical effects or correlated effects model based upon the structure that is most common in their analyses, noting that for the purposes of weighting and inference about the meta-regression coefficients, using the exactly correct dependence structure is not necessary (Tanner-Smith & Tipton, 2014; Tanner-Smith et al., 2016). When taking this approach, the model's variance components (e.g., the between study variance, τ^2) are treated as incidental to the analysis—only there for efficiency improvements—rather than as a focal parameter used for description or inference. This is where RVE diverges markedly from standard meta-analytic methods, where variance component estimates are considered an important component of synthesis results. Thus, analysts are left to choose between using RVE—which guards against model misspecification yet does not emphasize descriptions of heterogeneity—and multivariate meta-analysis—which can provide heterogeneity estimates but does not guard against misspecification.

We think that this forced choice is artificial and unnecessary. In this paper, we propose a hybrid approach that melds RVE with existing approaches to multivariate meta-analysis instead of treating them as separate analytic strategies. We argue that the flexible variance structures available with multivariate models offer benefits in terms of better capturing the types of dependence that occur in practice, including dependence that has *both* hierarchical effects and correlated effects. At the same time, treating the multivariate variance structures as working models within the RVE framework provides a safeguard against model misspecification. We show that this melding of methods can be implemented with existing software (the metafor and clubSandwich packages for R), and we study the properties of the approach using Monte Carlo simulations. We illustrate the proposed approach with an example based upon a meta-analysis of randomized trials examining the effects of brief-alcohol interventions for adolescents and young adults (Tanner-Smith & Lipsey, 2015).

Meta-Regression with Robust Variance Estimation

We begin by providing a general review of meta-analysis with dependent effect sizes using robust variance estimation (RVE). Consider a collection of J studies to be included in a meta-analysis, where each study contributes n_j effect size estimates, for $j = 1, .., J$. Let T_{ij} denote effect size estimate i from study j, with corresponding standard error s_{ij},

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$. We assume that T_{ij} is an unbiased estimate of an effect size parameter θ_{ij} and that s_{ij} is fixed and known. Letting $\mathbf{x}_{ij}$ denote a row vector of p covariates (possibly including an intercept term) and $\boldsymbol{\beta}$ denote a vector of p regression coefficients, we can relate the observed effect size estimates to these covariates using the meta-regression model

$$T_{ij} = \mathbf{x}_{ij}\beta + u_{ij} + e_{ij} \tag{1}$$

where $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, with $E(e_{ij}) = 0$ and $\mathrm{Var}(e_{ij}) = s_{ij}^2$. The error term u_{ij} describes variation in the effect size parameters above and beyond the variation explained by the covariates. Throughout, we assume that the effect size estimates from different studies are uncorrelated, so $\mathrm{cor}(e_{hj}, e_{ik}) = 0$ when $j \neq k$. To capture potential dependence, however, we will allow that effect size estimates from the same study may be correlated, although typically, the analyst will not know the exact degree of correlation between effect sizes from the same study. We will consider several different assumptions about the structure of u_{ij} within studies.

Robust Variance Estimation

In the RVE framework, meta-regression coefficients are estimated using weighted least squares (WLS).[1] WLS involves regressing the effect size estimates (T_{ij}) on the predictors ($\mathbf{x}_{ij}$). Unlike with ordinary regression, WLS incorporates a set of weighting matrices to improve the efficiency of the resulting coefficient estimates. For each study $j = 1, \ldots, J$, let $\mathbf{W}_j$ denote an $n_j \times n_j$ matrix of weights and let $\boldsymbol{\Phi_j}$ be an $n_j \times n_j$ variance–covariance matrix that describes the true dependence structure of the effect sizes from study j. The entries in true variance–covariance matrix for study j describe the covariance between pairs of effect sizes within the study, so that row h and column i of the matrix has entry $\phi_{hij} = \mathrm{Cov}(u_{hj} + e_{hj}, u_{ij} + e_{ij})$ for $h, i = 1, \ldots, n_j$.

If we knew the exact dependence structure of the effect sizes, then we could choose weights that are exactly inverse of the true variance–covariance for each study (i.e., $\mathbf{W_j} = \boldsymbol{\Phi}_j^{-1}$), resulting in an estimator of $\boldsymbol{\beta}$ that is fully efficient, meaning that it has the smallest possible sampling variance. Furthermore, if we knew the true dependence structure, we could exactly calculate standard errors, test statistics, and confidence intervals for the WLS estimator using standard formulas (Hedges et al., 2010). In contrast, if we are not certain of the dependence structure (i.e., the form of the $\boldsymbol{\Phi}_j$'s), then we cannot define exactly inverse variance

weights, nor can we estimate the variance of the WLS estimator using standard methods.

The RVE approach separates the choice of a set of weight matrices from the method of estimating standard errors, so that standard errors can be obtained without having to assume a known dependence structure. Rather, RVE standard errors are calculated by using products of the regression residuals to roughly approximate the variance–covariance structure of the errors (i.e., to roughly estimate $\boldsymbol{\Phi}_j$ for each study). Even though each product of residuals provides only a very crude estimate, the RVE standard errors are nonetheless valid because they involve an *average* of the residual products, which will be accurate if calculated across a sufficiently large number of studies. Furthermore, adjustment methods have been developed to reduce small-sample bias of the RVE standard errors (Tipton, 2015; Tipton & Pustejovsky, 2015).

Although RVE standard errors are statistically valid under *any* set of weights, the choice of weights does impact the precision of the meta-regression coefficient estimator ($\hat{\boldsymbol{\beta}}$). The most precise estimator results when the weights are exactly inverse of the true variance–covariance matrix. However, because the true covariance matrix $\boldsymbol{\Phi}_j$ is unknown, we must in practice use a working model—meaning a good guess or rough approximation—for purposes of developing weight matrices. If the working model is correct, the resulting weights are exactly inverse variance, and the meta-regression estimator is fully efficient. If the working model is only close to correct, the resulting weights are still *approximately* inverse variance, and the meta-regression estimator is still close to efficient. In contrast, if the working model is quite discrepant from the true covariance structure, the meta-regression estimator may be much less efficient, although it is still unbiased and inferences based on RVE remain statistically valid. Thus, the choice of working model and associated weight matrices offer a means to improve the precision of meta-regression coefficient estimates.

Currently Available Working Models

When they introduced RVE, Hedges et al. (2010) included two working models, "hierarchical" and "correlated" effects. For each working model, they also proposed approximately efficient weights. We now review these models.

Model 1: Hierarchical Effects (HE)

The first of the original working models is the hierarchical effects (HE) model, which has the form

$$T_{ij} = \mathbf{x}_{ij}\beta + u_j + v_{ij} + e_{ij}, \tag{2}$$

---

[1] This section describes RVE at a heuristic level. Section S1 of the supplementary materials includes exact mathematical details using matrix notation.

where $\mathrm{Var}(u_j) = \tau^2$, $\mathrm{Var}(v_{ij}) = \omega^2$, $\mathrm{Var}(e_{ij}) = s_{ij}^2$, and $\mathrm{Cov}(e_{hj}, e_{ij}) = 0$. Here, τ^2 is the between-study variation in study-average true effect sizes, ω^2 is the within-study variation in true effect sizes, and s_{ij} is the known standard error from estimation.[2] Note that the HE model assumes that conditional on being in the same study, the effect size *estimates* are independent, so that the only source of dependence between effect sizes in the same studies is with regard to the true effect sizes, not estimation error.

Model 2: Correlated Effects (CE)

The second working model is the correlated effects (CE) model, which has the form

$$T_{ij} = \mathbf{x}_{ij}\beta + u_j + e_{ij} \tag{3}$$

where now $\mathrm{Var}(u_j) = \tau^2$, $\mathrm{Var}(e_{ij}) = s_j^2$, and $\mathrm{Cov}(e_{hj}, e_{ij}) = \rho s_j^2$. Here, ρ is the correlation between two effect size *estimates* in the same study (for $h \neq i$), and $s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j} s_{ij}^2$ is the average (known) sampling variance in the study.[3]

The CE model imposes several simplifying assumptions about the dependence structure. First, it is assumed that there is no within-study variation in the true effect size parameters beyond what is explained by the covariates. Second, it is assumed that sampling variances are roughly equivalent, so that $s_{ij}^2 \approx s_j^2$. Third, it is assumed that the correlation between effect size estimates is the *same* for all pairs of effect size estimates in study j (and in every study). Thus, this working model has only a single variance component, τ^2, representing between-study variation in true effect sizes.

Weights and Estimation

The correlated effects and hierarchical effects working models involve, respectively, one or two unknown variance component parameters that must be estimated. Existing software for RVE estimates these variance component parameters using special method-of-moment estimators proposed by Hedges et al. (2010). Estimates of the variance components are used to calculate weighting matrices, which are then used for weighted least squares estimation of the regression coefficients $\boldsymbol{\beta}$. For the hierarchical effects model, the weight matrices are given by

$$\mathbf{W}_j = diag\left(w_{1j}, \ldots, w_{n_j j}\right) \tag{4}$$

where $w_{ij} = 1/\left(\hat{\tau}^2 + \hat{\omega}^2 + s_{ij}^2\right)$ and $\hat{\tau}^2$ and $\hat{\omega}^2$ are the method-of-moments estimators of τ^2 and ω^2 (Hedges et al., 2010). For the correlated effects model, the weight matrices are also diagonal, with every diagonal entry equal to

$$w_{ij} = \frac{1}{n_j\left(\hat{\tau}^2 + s_j^2\right)} \tag{5}$$

where $\hat{\tau}^2$ is the method-of-moments estimator of τ^2. This estimator requires specification of an assumed correlation ρ; the software default is set as $\rho = 0.80$ (Fisher et al., 2017).

Crucially, these weight matrices are not fully inverse-variance weights, even when the working models are correct. Rather, Hedges and colleagues describe them as "approximately efficient" weights, which are close to the optimally efficient inverse-variance weights but are easier to calculate because they do not involve inverting $n_j \times n_j$ matrices. In contrast, we follow a different strategy and use fully inverse-variance weighting when implementing new working models for RVE. As we demonstrate in the following sections, this change in strategy can lead to non-trivial efficiency improvements under certain conditions while still being feasible to implement with existing software tools.

A New Class of Working Models for RVE

Given the two available working models, it is not hard to see that there are situations in which a fusion of the models would be desirable. Indeed, in our consultations with researchers using RVE, this has been a common request. In this section, we propose two new working models, as well as describing an approach for developing further extensions and refinements. We start with a model that combines the features of the HE and CE models because we anticipate that this working model may be the most broadly applicable.

Model 3: Correlated and Hierarchical Effects (CHE)

In our experience, meta-analytic data rarely have a purely correlated or purely hierarchical effects structure. Rather, it is far more common for both types of structure to be present. A model that combines both dependence structures is given by

$$T_{ij} = \mathbf{x}_{ij}\beta + u_j + v_{ij} + e_{ij} \tag{6}$$

where $\mathrm{Var}(u_j) = \tau^2$, $\mathrm{Var}(v_{ij}) = \omega^2$, $\mathrm{Var}(e_{ij}) = s_j^2$, and $\mathrm{Cov}(e_{hj}, e_{ij}) = \rho s_j^2$. Like the original CE model, the correlated and hierarchical effects (CHE) model makes the simplifying assumption that there is a single, known correlation ρ between pairs of effect sizes from the same study, which is

---

[2] These assumptions define a certain structure for the variance–covariance matrices, $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_J$. See Section S2.1 of the supplementary materials for further details.

[3] Section S2.2 of the supplementary materials provides further details on the structure of the variance–covariance matrices, $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_J$ under this model.

the same across all studies. We shall refer to this as the "constant sampling correlation" assumption.[4]

This CHE model combines the desirable features of the HE and CE working models. Like the HE working model, it allows for both between-study heterogeneity (quantified by the between-study SD τ) and within-study heterogeneity (quantified by the within-study SD ω) in true effect sizes. Like the CE model, it also allows there to be some correlation between the effect size estimates within each study. Combining these features into one working model may be particularly attractive in meta-analyses that include studies with a broad variety of outcomes, follow-up times, or other operational variations—exactly the circumstances where meta-regression and RVE methods are most useful.

We expect that the CHE model will be a first choice as a working model in many applications, especially when little information is available about correlations between included effect size estimates. However, other working models are possible and may be of interest when effect sizes from included studies can be classified into distinct dimensions or categories.

Model 4: Subgroup Correlated Effects (SCE)

When a meta-analytic database includes multiple dimensions or categories of effect sizes, meta-analysts often conduct sub-group analysis, estimating separate results within each category of effect sizes. For example, in a synthesis of adolescent mental health interventions, Skeen et al. (2019) conducted separate meta-analyses of effects on positive mental health, depression and anxiety symptoms, aggressive behavior, and substance use. Their approach resulted in five separate meta-regressions, with separate pooled effect size estimates, between-study heterogeneity estimates, and hypothesis tests.

Using separate sub-group analyses may be appealing because of its feasibility and conceptual simplicity. In particular, the results for each sub-group are based *only* on the effect size estimates from that sub-group, rather than coming from a model for the full data including all effect size estimates. Running separate analyses also allows the between-study variance to be different for each sub-group, rather than assuming that it is common across categories. However, running separate meta-regressions for each of several sub-groups can become unwieldy because each analysis has a reduced sample size, potentially making it infeasible to estimate a meta-regression with many predictors. More fundamentally, running separate analysis does not provide any way to conduct statistical tests or calculate confidence

intervals for comparisons between average effects from different sub-groups. This is because, if the sub-groups include overlapping sets of studies, the pooled effect sizes from each sub-group are not independent.

Here, we propose an alternative working model, the sub-group correlated effects (SCE) working model, that overcomes these problems and allows for comparisons across sub-groups, while preserving the conceptual clarity of sub-group analysis. The SCE working model embeds the assumptions of separate meta-regression analyses into a working model for the full data, including effects across all categories. By using a working model that treats the effect size estimates from different sub-groups as independent, we can obtain meta-analytic estimates that are identical to the results of separate sub-group analyses, but that can be statistically compared using RVE because they are all part of one model.

Suppose that every effect size estimate can be classified into one of C categories, and define the indicators $d_{ij}^1, \ldots, d_{ij}^C$, where $d_{ij}^c = 1$ if effect size i in study j falls into category c (e.g., a depression outcome), with $d_{ij}^c = 0$ otherwise. If they were running a sub-group analysis of the effects in category c, the analyst would estimate the model

$$T_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}_c + u_{cj} + e_{ij}$$

using only on the subset of effect sizes that has $d_{ij}^c = 1$. The set of C such models can be expressed in one meta-regression model by interacting the covariates ($\mathbf{x}_{ij}$) with indicators for each category (d_{ij}^c) and similarly interacting the random-effects term (u_{cj}) with indicators for each category. This yields the model

$$\begin{aligned} T_{ij} = &d_{ij}^1\mathbf{x}_{ij}\boldsymbol{\beta}_1 + d_{ij}^2\mathbf{x}_{ij}\boldsymbol{\beta}_2 + \cdots + d_{ij}^C\mathbf{x}_{ij}\boldsymbol{\beta}_C + d_{ij}^1 u_{1j} \\ &+ d_{ij}^2 u_{2j} + \cdots + d_{ij}^C u_{Cj} + e_{ij}, \end{aligned}$$

which can be expressed more compactly as

$$T_{ij} = \sum_{c=1}^{C} d_{ij}^c \left(\mathbf{x}_{ij}\boldsymbol{\beta}_c + u_{cj} \right) + e_{ij} \tag{7}$$

Note that this model includes separate terms for the regression coefficients from each of the C categories. To ensure that the coefficient estimates from each category are based only on the effect size estimates from that category, the assumptions regarding the random effects terms and sampling errors need to be slightly modified. First, rather than using the constant sampling correlation assumption, the analyst can use what we shall call a "constant sampling correlation within subgroups" assumption. Here, we assume that effect size estimates from the same study are correlated if they fall within the same sub-group, but are *uncorrelated* if they fall in different sub-groups. This assumption can be expressed mathematically as

---

[4] Section S2.3 of the supplementary materials gives form of the variance–covariance matrices implied by the CHE model.

$$Cov(e_{hj}, e_{ij}) = \rho s_j^2 \sum_{c=1}^{C} d_{hj}^c d_{ij}^c \qquad (8)$$

Second, the sub-group analysis approach requires specifying random effects for each category that are mutually independent, so that $Var(u_{cj}) = \tau_c^2$ and $Cov(u_{bj}, u_{cj}) = 0$.[5]

In contrast to the original CE model, this SCE working model produces several estimates of between-study heterogeneity—one for each category of effects. Furthermore, using the subgroup working model produces estimates of the meta-regression coefficients $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_C$ and variance components $\tau_1^2, \dots, \tau_C^2$ that are *identical* to results based on estimating separate models for each sub-group of effect sizes. The main benefit of this working model, then, is that using a single model allows statistical comparisons of the meta-regression coefficients across categories. For instance, in the meta-analysis of mental health interventions (Skeen et al., 2019), one could use RVE methods to test the hypothesis that average effect sizes are equal across the dimensions of positive mental health, symptomatology, aggressive behavior, and substance use, or that the *differences* in program impacts between in-person interventions and digital interventions are consistent across these dimensions. Similarly, one could calculate a robust confidence interval for a difference between two categories (e.g., the difference in average effect on symptomatology and the average effect on substance use).

Additional Variants and Extensions

We expect that the CHE and SCE working models may be more widely applicable than the original CE or HE working models. However, we would also emphasize that these models will not address all of the problems that meta-analysts face when applying RVE. We view these two working models as leading examples of a larger class of working models that can be expressed using multivariate or multilevel models. Other working models, as well as refinements to the models that we have described, are possible and simply require expressing other assumptions regarding the variance–covariance structure of the effect sizes. Here, we briefly describe three potential refinements that may be useful in practice.

Add a Level (+)

Thus far, the working models we have described all assume a two-level meta-analytic structure, wherein effect sizes are nested within independent studies. However, some meta-analytic databases include one or more higher levels of

nesting that could also be modeled. For example, it may be that multiple effect sizes are nested in samples, with multiple samples nested within larger studies (e.g., multi-site trials that report site-specific results) or within independent research labs. In any of these cases, including this additional level amounts to including an additional random effect and a corresponding variance component. For notational purposes, we refer to such models with a " + " sign to indicate the additional level, as in "CHE + ".[6]

Non-Constant Correlation Using Auxiliary Data

The working models described thus far are premised on the assumption that the analyst will typically have information on the variances of effect size estimates but little or no information about their correlations or covariances. In some meta-analytic databases, however, the analyst may be able to directly compute the covariances between effect sizes for some pairs of effect sizes within studies.[7] Further, some primary studies might also report a correlation matrix across outcome measures (or across repeated measures of an outcome), which can be used to compute the covariances between effect size estimates for those outcomes.

In practice, it is unlikely that sampling covariances could be directly estimated for *all* included studies and *all* pairs of effect size estimates. Still, the analyst may be able to specify the covariances by using a combination of known values (i.e., for studies that report correlations between outcomes) and potentially arbitrary guesses about covariances between effect sizes where no information is available. We shall call this approach the "partially empirical correlations" assumption or "PEC" assumption. Using it in place of the constant sampling correlation assumption, Model 3 would then be called "PECHE."

Auto-Correlated Errors

Some meta-analyses may involve very narrowly defined outcomes, but where there is interest in understanding change over time. In these cases, effect size dependence arises primarily from outcomes measured repeatedly over time, and the constant sampling correlation assumption might be considered

---

[5] Section S2.4 of the supplementary materials give the form of the variance–covariance matrices in the SCE model.

[6] More details about the specification of CHE + are given in Section S3.1 of the supplementary materials.

[7] For example, a meta-analysis might include some experiments with multiple treatment conditions, where the effect size estimates represent treatment effects compared to a common control condition. The formula for the covariance between pairs of effect sizes estimated from a common control group can be calculated from information that is usually reported in primary studies (Gleser & Olkin, 2009; Wei & Higgins, 2013), so that there is no need to make arbitrary assumptions about the correlation between sampling errors.

somewhat unrealistic. Instead, it may be more plausible to assume that the correlation between a pair of effect size estimates generated from a common sample depends on the time interval between assessments (cf. Trikalinos & Olkin, 2012). For example, we might expect the correlation between effect sizes at the end of an intervention, and 1 month later to be larger than the correlation with effect sizes 6 or 12 months later. In the supplementary materials (Sect. 3.2), we suggest an alternative to the constant sampling correlation assumption, based on a simple first-order auto-regressive working model. Similar to the constant sampling correlation assumption, the AC assumption involves a single parameter, the choice of which might be arbitrary (and paired with sensitivity analysis) or informed by empirical data provided by one or more of the primary studies. We call this the "sampling auto-correlation" or "AC" assumption. Using it in place of the constant sampling correlation assumption, Model 3 would then be called "ACHE."

Choosing and Implementing a Working Model

The new working models and refinements that we have described are more complicated than the original RVE working models, and readers may worry about how feasible it is to implement them in practice. As we illustrate here and in the example, it is possible to implement these new working models using readily available, existing software. The implementation process involves three steps, which we detail in this section: (1) identify a working model and flesh it out; (2) assume the working model is true and estimate the meta-regression coefficients based upon it; and (3) guard against misspecification of the working model by calculating standard errors and hypothesis tests using RVE.

Identify an Appropriate Working Model

The benefit of the approach we have described is that there are now a wider variety of possible working models available to analysts, thus allowing selection of a model that more closely aligns with the dependence structure of the effect size data. This flexibility does come with some potential risks, as it requires analysts to make a larger number of decisions and thus expands the possibilities for analytic flexibility or "researcher degrees of freedom"—even the meta-analytic equivalent of p-hacking.

To guard against such concerns, we recommend that the working model be chosen based upon a broad understanding of the data structure and data-generating process—*not* by comparing how meta-regression results change (or do not) over different working models or assumptions. In Fig. 1, we provide a decision-tree that analysts can use to select a

working model. This involves three decisions: (1) choosing an assumption for the within-study correlation between effect size estimates, (2) identifying a structure for the random effects, and (3) determining whether to include any additional levels. Ideally, this decision-tree would be used to select a working model during the planning stage of the meta-analysis (before analysis begins), and the intended working model would be pre-registered. Even if it is not possible to identify a specific working model during the planning stage, researchers might still pre-register the commitment to follow the decision tree (or a suitably modified version of the decision tree), once the structure of the data has been determined.

Estimate the Meta-Regression while Treating the Working Model as True

After a working model is selected, it can be implemented in R using a combination of the clubSandwich (Pustejovsky, 2020) and metafor packages (Viechtbauer, 2010). The first step in the analysis is to specify the correlation structure of the effect size estimates within studies. For the constant sampling correlation (C) or sampling auto-correlation models (AC), as well as the corresponding sub-group versions, this can be accomplished using the function impute_covariance_matrix() from the clubSandwich package. The partially empirical sampling correlation model (PE) can be implemented using the pattern_covariance_matrix() from clubSandwich.

The second step is to estimate the random effects part of the working model and to use the results to estimate the meta-regression coefficients. This can be accomplished using the function rma.mv() from the metafor package. To use this function, the analyst must specify (a) the form of the meta-regression model, using R's regression equation syntax, (b) the chosen sampling correlation working model, and (c) the form of the random effects working model, using syntax and arguments described in the metafor package. Given these inputs, the function automatically generates restricted maximum likelihood estimates of the random effects variance components, calculates inverse-variance weight matrices, and estimates the meta-regression coefficients.

Guard Against Misspecification by Using RVE

The third step is to calculate RVE standard errors, hypothesis tests, or confidence intervals for the meta-regression coefficient estimates. This can be accomplished using the functions coef_test() or conf_int() from the clubSandwich package. These functions take as input the results of the rma.mv()model, calculate RVE standard errors, hypothesis tests for each coefficient, and confidence intervals for each coefficient. The function Wald_test() from clubSandwich

Fig. 1 Decision-tree for selecting working model based upon data generating model. ES effect size, RE random effects, HE hierarchical effects, CE correlated effects, CHE correlated and hierarchical effects, SCE subgroup correlated effects
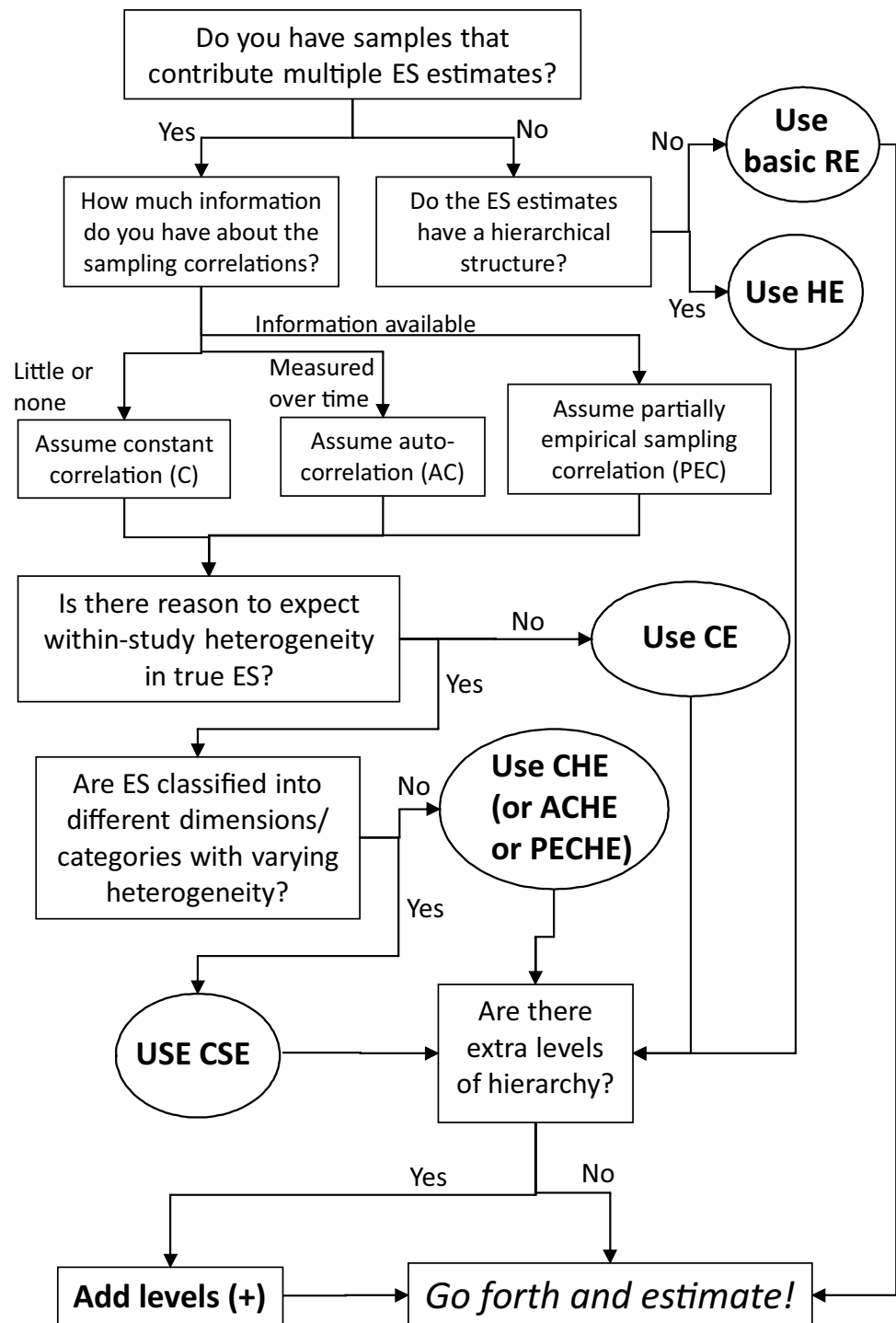
can also be used to carry out tests of multiple-contrast hypotheses, such as tests for equality of levels across multiple categories of effect sizes. In the accompanying supplementary materials, we provide annotated R syntax demonstrating how to use all of these functions to carry out the empirical analysis reported in the next section.

Comparison to the Standard RVE Implementation

Our proposed approach differs in two subtle ways from the current approach as instantiated in the robumeta package (Fisher et al., 2017) in R. First, instead of estimating variance components using method-of-moments estimators, our

approach uses restricted maximum likelihood (REML) estimation, which obtains parameter estimates by finding the parameter values that maximize the restricted log likelihood of a specified working model. The primary advantage of REML methods is that they can be applied to a very broad set of models, including models with multiple levels of nesting, multivariate models, and combinations thereof, even beyond the range of working models that we have described. In contrast, method-of-moments estimation methods are not as extensible because they have to be developed anew for every working model. Furthermore, for univariate models with only a single variance component, REML estimation has been generally recommended over moment estimation methods (Veroniki et al., 2016).

Second, instead of using diagonal weight matrices that are only approximately efficient, our approach uses weighting matrices that are exactly inverse of the variances defined by the working model, and thus fully efficient when the working model is correct. There are two main reasons for taking this approach. The first is simply computational convenience. Available software automatically carries out the matrix inversion and weighted least squares regression estimation, and we see no strong reason to modify the inverse-variance weights—especially if doing so sacrifices efficiency. The second reason is that the originally proposed weighting schemes are sometimes less efficient than one might expect, particularly when the model includes predictors that vary within studies. As we demonstrate in the next two sections, using the approximately efficient weight matrices can, under some circumstances, lead to less precise estimates of meta-regression coefficients. Thus, using fully inverse-variance weights may avoid a subtle drawback of the existing weighting approach.

Application: Brief Alcohol Interventions

Tanner-Smith and Lipsey (2015; henceforth TSL15) conducted a comprehensive synthesis examining the effects of brief alcohol interventions for adolescent and young adults. Here, we use a subset of the studies from TSL15 to demonstrate use of the new working models and estimation method that we have proposed. Our re-analysis illustrates how the new working models better capture the structure of the effect sizes included in the synthesis, providing descriptive information about both between- and within-study heterogeneity. The re-analysis also indicates how the new working models might, under certain circumstances, yield more precise estimates of meta-regression coefficients compared to existing RVE approaches—a possibility that we explore further in Monte Carlo simulations. Importantly, these analyses are presented for illustrative purposes only; readers should refrain from drawing substantive conclusions or policy implications based upon them.

The full TSL15 synthesis included quasi-experimental designs, cluster-randomized trials, and individually randomized trials; eligible outcomes included measures of alcohol consumption and alcohol-related problems. Effect sizes were operationalized as standardized mean differences, coded so that positive values indicated better outcomes (e.g., lower alcohol consumption). For illustration, we focus on a subset of the included studies that used individually randomized designs and reported a continuous measure of alcohol consumption. After excluding effect sizes that were missing information on control variables (described further below), this subset consists of 117 studies and a total of 1198 effect size estimates; individual studies contributed between 1 and 108 effect sizes (median = 6, IQR = 3–12).

The dependence structure of TSL15′s database is quite complex. Many studies reported effects for multiple measures of alcohol consumption (e.g., frequency of use, quantity of use, peak consumption), effects across multiple follow-up times, or both. Many studies also examined multiple interventions, multiple variants of an intervention, or compared an intervention to multiple non-intervention conditions. Thus, correlated sampling errors are a major feature of the included effect sizes.

Analysis and Working Models

TSL15 investigated a range of moderator variables related to intervention, participant, and measurement characteristics. For illustrative purposes, we focus on two types of moderators, each of which are common in practice. First, we examine differences by type of dependent variable, classified into six categories; this moderator varies substantially *within* studies (i.e., an effect size level moderator).[8] Second, we examine differences by type of control condition, classified into four categories; this moderator varies substantially *between* studies but not within studies (i.e., a study level moderator).[9] As the results illustrate, this distinction affects how the new models perform relative to current RVE working models.

We estimated separate meta-regression models for each moderator, using a no-intercept specification so that coefficients represent average effect sizes for the corresponding category. Each model also included predictors for follow-up time, the proportion of the sample that was in college, the proportion of the sample that was male, and the overall level of attrition. These predictors were centered so that intercepts correspond to average effects 12 weeks after treatment for a college-age sample that is 50% male and with attrition of 16% (the median attrition level). With

---

[8] See Supplementary Table S1 for further details.

[9] See Supplementary Table S2 for further details.

each model, we report a robust Wald test for equality of effects across levels of the focal moderator. Because the dependence structure of TSL15 was predominantly one of correlated sampling errors, the original analysis used the correlated effects (CE) working model as proposed by Hedges et al. (2010), with moment estimation for the between-study variance τ^2. We therefore used this working model as a benchmark, then explored alternative working models for the random-effects structure.

To determine our primary working model, we followed the decision tree depicted in Fig. 1. First, because we had very limited information about the sampling correlations among effect sizes, we assumed a constant sampling correlation of $\rho = .6$. The supplementary materials include sensitivity analyses for varying values of ρ. Second, given the wide variation in outcome measures, follow-up times, treatment conditions, and control conditions, there was strong reason to expect within-study heterogeneity in effects. Because the original analysis did not distinguish categories or dimensions of effect sizes with varying degrees of heterogeneity, our primary analysis therefore used the CHE working model. Third, the database did not include information about hierarchical groupings of studies, and so we did not include additional levels of random effects.

In addition to analyses based on CHE, we also present results based on the SCE working model. We included this additional approach to illustrate similarities and differences with the CE and CHE models—*not* because we think that all of the models should be applied or reported in practice. Results based on SCE also illustrate how this working model enhances a traditional sub-group analysis, which some meta-analysts might have preferred for handling dependence in these data. As with the primary CHE model, we used the constant sampling correlation assumption with $\rho = .6$.

Results

Effect Size Level Moderator In Table 1, we report estimated average effect sizes by type of dependent variable, along with variance component estimates for each working model. Column A contains results based on the correlated effects (CE) working model, which includes only a single variance component. Estimated effects range from 0.087 SD (SE = 0.044) for frequency of use measures to 0.167 SD for frequency of heavy use and blood alcohol concentration (BAC) measures (SE = 0.039 and 0.034, respectively). Average effects are statistically distinguishable from zero for frequency of heavy use, quantity of use, peak, consumption, and BAC. Based on a robust Wald test, we cannot rule out the hypothesis that average effects are equal across dependent variables ($p = 0.463$).

Point estimates for the average effects based on CHE (Column B) are quite consistent with results from the basic CE model. The results suggest two potential advantages over the original CE model. First, the CHE model provides both between- and within-study variance components, yielding a more useful description of the structure of heterogeneity in these effect sizes. The between-study SD (τ) estimate from this working model is much smaller than the moment estimate from the correlated effects model, but the within-study SD (ω) estimate indicates that there is substantial heterogeneity across effect sizes within studies. This is helpful diagnostic information, suggesting that it may be useful to identify further moderators that vary at the within-study level in order to explain this heterogeneity. Second, the estimates from the CHE working model are noticeably more precise than those from the original CE working model, with standard errors that are 14 to 43% smaller. These two working models differ in several respects, including the number of variance components, how the variance component(s)

Table 1 Comparison of working models for a within-study moderator (type of dependent variable)

Coefficient	Studies	Effect sizes	(A) Correlated effects Est. [SE]	(B) Correlated hierarchical effects Est. [SE]	(C) Sub-group correlated effects Est. [SE]	tau
Frequency of use	49	180	0.087 [0.044]	0.118 [0.027]	0.104 [0.035]	0.123
Frequency of heavy use	55	214	0.167 [0.039]	0.170 [0.026]	0.149 [0.031]	0.155
Quantity of use	103	470	0.169 [0.027]	0.151 [0.023]	0.148 [0.023]	0.101
Peak consumption	27	122	0.119 [0.044]	0.154 [0.033]	0.153 [0.047]	0.135
Blood alcohol concentration	31	156	0.167 [0.034]	0.181 [0.029]	0.186 [0.030]	0.007
Combined measures	21	56	0.114 [0.063]	0.160 [0.036]	0.159 [0.060]	0.216
$\hat{\tau}$			0.182	0.123		
$\hat{\omega}$				0.116		
Wald test p value			0.463	0.232	0.602	

Estimates for all working models are adjusted for follow-up time, the proportion of the sample that was in college, the proportion of the sample that was male, and the overall level of attrition. All working models assume a sampling correlation of $\rho = .6$

Table 2 Comparison of working models for a between-study moderator (type of control group)

Coefficient	Studies	Effect sizes	(A) Correlated effects	(B) Correlated hierarchical effects	(C) Sub-group correlated effects	
			Est. [SE]	Est. [SE]	Est. [SE]	tau
Straw man	27	340	0.099 [0.050]	0.119 [0.046]	0.112 [0.044]	0.114
Attention/sham	13	137	0.135 [0.074]	0.166 [0.072]	0.150 [0.081]	0.208
Treatment as usual	8	36	− 0.038 [0.063]	0.040 [0.083]	− 0.048 [0.053]	0.000
No treatment	72	685	0.177 [0.026]	0.168 [0.026]	0.178 [0.028]	0.136
$\hat{\tau}$			0.183	0.126		
$\hat{\omega}$				0.116		
Wald test p value			0.076	0.553	0.027	

Estimates for all working models are adjusted for follow-up time, the proportion of the sample that was in college, the proportion of the sample that was male, and the overall level of attrition. All working models assume a sampling correlation of $\rho = .6$

are estimated, and use of semi-efficient versus inverse-variance weighting. Of these differences, the major driver of increased precision is use of the fully efficient inverse-variance weight matrices.[10] Because the levels of the focal moderator vary at the effect size level, the weighting scheme used in the original correlated effects model sacrifices a substantial amount of precision.

Results based on the SCE working model are generally consistent with those of the CHE working model, although the standard errors from the SCE working model are larger and more similar to those from the CE model. A potential advantage of the SCE working model is that it allows the degree of heterogeneity to vary across type of dependent variable.

Study Level Moderator In Table 2, we report estimated average effect sizes by type of control group, along with variance component estimates based on the CE, CHE, and SCE working models. Estimated average effect sizes are generally consistent across working models, with the exception of the average effect for treatment-as-usual control groups. Only eight studies have effect sizes from this category, and two of those studies also included a no-treatment control group (see supplementary Table S2). For three of the four types of control groups, average effect estimates from the CHE working model are more precise than those from the original CE model, although the precision gains are not as strong as for the analysis by type of dependent variable (which had considerable variation within studies). Similar to the previous analysis, the CHE working model has the advantage of providing an estimate of within-study heterogeneity. Estimates based on the SCE working model are very similar to those from the CE working model, without any clear pattern of differences in precision.

Sensitivity Analysis All of the estimates in Tables 1 and 2 are based on a constant sampling correlation model with $\rho = .6$. In Supplementary Figures S2 through S5, we report sensitivity analysis for all model parameters, varying the assumed correlation between $\rho = .0$ and $\rho = .95$. For both the study level and effect size level moderators, the average effects and variance component estimates from the original CE model are nearly identical across the entire range of ρ values. In comparison, meta-regression coefficient estimates based on the CHE model were somewhat more sensitive, but still generally quite stable. Coefficient estimates based on the SCE working models became a bit sensitive at values of ρ above 0.8, but were otherwise stable. For both moderator analyses, however, the variance component estimates from the CHE and SCE working models were substantially more sensitive to the assumed value of ρ, particularly for values above 0.8. This sensitivity represents an important limitation to the conclusions that one can draw regarding the dependence structure based on the working models.

Simulation Study

In the previous section, we noted that when the focal moderator had substantial variation at the effect size level, estimates based on the new working models tended to be more precise than those based on the original CE working model, as indicated by differences in the robust standard errors based on each working model. However, one must be cautious in interpreting differences in the standard errors as indicative of systematic differences in the performance of the working models because the standard errors are themselves random quantities, affected by sampling variability, and because all of the analysis was based on just a single dataset.

In order to investigate whether the patterns noted in the empirical application hold more generally, we conducted a

---

[10] Section S4.2 of the supplementary materials provides further details.

set of Monte Carlo simulations. These simulations allowed us to assess the performance of the new CHE and SCE working models relative to the CE working model (the current default) in a more systematic way, by repeatedly generating artificial data under conditions with known parameters.

Simulation Design

We designed the simulations to generate data that had a structure similar to that of the TSL15 data (but using smaller sample sizes of 30 or 60 primary studies per meta-analysis), so that the simulation results would inform our interpretation of the re-analysis that we have presented.[11] We assessed several aspects of the performance of each working model, including the precision of meta-regression coefficient estimators (as measured by root mean-squared error or RMSE), confidence interval coverage rates, and the accuracy of the variance component estimators (as measured by RMSE). We examined performance under several different conditions, including conditions of mild mis-specification, where the general form of the CE or CHE working model was correct but the assumed sampling correlation was incorrect, as well as stronger mis-specification, where the structure of the CE, CHE, or both the CE and CHE working models was not close to the true data-generating structure.

Simulation Results

Consistent with the patterns noted in the empirical application, we found that using the new working models could make a substantial difference for models that included effect size level predictors but made little difference for models with only study level predictors. For models with effect size level predictors, using the CHE working model rather than CE led to systematic improvements in the precision of coefficient estimates. Gains in precision were substantial (10–50% reductions in RMSE) when the CHE working model was correct or mildly mis-specified and more moderate (10–44% reductions) when CHE was more strongly mis-specified. Notably, these gains were also present even when the CE model was correct or only mildly mis-specified. Using the SCE working model led to smaller improvements (2–13% reductions in RMSE compared to using CE) that were consistent across data-generating conditions. For meta-regression models of study level predictors, using the CHE working model led to very small improvements in precision, and using the SCE model led to small improvements or, for

some predictors, modest reductions in precision compared to using the CE.

Under the various data-generating conditions we examined, confidence intervals generated based on the new CHE and SCE working models were appropriately calibrated (as were those based on the original CE working model), with coverage rates close to or above the nominal 95% rate. Confidence intervals had near-nominal coverage rates even when the working model was quite discrepant from the true data-generating model.

In the empirical application, we also noted that the REML variance component estimators appeared to be more sensitive to modeling assumptions than the moment estimator from the CE working model. We observed a similar pattern in the simulations but also found that this increased sensitivity was not necessarily indicative of reduced accuracy. Comparing the REML estimator of between-study variance from the CHE model to the moment estimator from the CE model, we found that the estimators had similar accuracy when the CE working model was correct or mildly mis-specified. Further, the CHE REML estimator had consistently *better* accuracy than the CE moment estimator when the true data-generating model included within-study heterogeneity. Thus, REML estimation of the CHE working model represents an improvement in that it can describe both between-study and within-study heterogeneity, albeit with some sensitivity to the assumed value of the sampling correlation.

Discussion

In this paper, we have described and demonstrated an expanded set of working models for use with RVE meta-regression, along with an accompanying estimation strategy based on REML and fully inverse-variance weighting. By melding RVE with the powerful and extensible methods for multivariate meta-analysis, these new working models expand the range of options available for meta-analytic application, which were previously limited to two working models.

The new models and new estimation strategy have several potential advantages relative to existing tools for meta-analysis with RVE. First and foremost, as illustrated in our re-analysis of TSL15 and supported in simulations, the new models can sometimes provide more precise estimates of average effect sizes than the original working models—particularly in meta-regressions involving moderators that vary at the effect size level (within study). Under such conditions, the improved precision is due to the use of working model assumptions that better approximate the true dependence structure and to the shift to fully inverse-variance weighting. Second, the availability of working models with both correlated sampling errors and hierarchical random effects

---

[11] Section S5 of the supplementary materials include a full description of the simulation design and results, including graphical depictions of the main findings. Supplementary materials also include the full computer code and complete numerical results.

provides richer (albeit still imperfect) descriptive information about effect heterogeneity, allowing the analyst to better tailor their analytic approach to the complexities and nuances of their data. Finally, the new approach is feasible to implement with existing software packages in R (metafor and clubSandwich) that are already widely used.

An important part of meta-regression analysis with RVE is to consider how the working model assumptions, such as the assumed sampling correlation (ρ), influence one's conclusions. We have illustrated one approach to sensitivity analysis in the empirical application, where we found that meta-regression coefficient estimates were largely insensitive to assuming different values of ρ between 0.0 and 0.8 (although they were not as insensitive as results based on the original CE working model). However, the REML variance component estimates from the CHE and other new working models were substantially more sensitive to the choice of ρ, whereas the moment estimator used in the original CE working model was almost entirely invariant.

Some might argue that the sensitivity of the REML estimators of variance components points toward a limitation of using the approach with more complex working models. Clearly, the sensitivity of the variance component estimates does mean that analysts must be cautious in drawing any substantive conclusions based upon their magnitude. For example, in the analyses of TSL15, we would refrain from drawing any conclusion about the relative magnitude of within- versus between-study variation because the ratio depends very strongly on ρ. Still, based upon the overall pattern of results, it seems reasonable to infer that one should be concerned with heterogeneity on both levels, rather than limiting consideration to between-study variation alone. Further, our simulation results indicated that the REML estimator of between-study variance under the CHE working model had similar or better accuracy than the moment estimator used in the original CE model, suggesting that the sensitivity of REML is not necessarily indicative of poorer performance. On balance, we would argue that it is better to apply a working model that captures the structure of one's data—even if the variance component estimates are sensitive to assumptions—than to use one that imposes stronger and less plausible assumptions (i.e., that there is no within-study heterogeneity).

In addition to the sensitivity of REML estimation, it is important to note two other limitations of the new working models and estimation methods that we have proposed. First, we have assessed their performance under conditions similar to the TSL15 data, but performance in other scenarios warrants further investigation—especially in scenarios where the data have a highly imbalanced structure (e.g., one or a small number of studies that contribute many more effect sizes than the other studies) or where the data include moderators with less extensive variation at the effect size level. Second, the broader variety of working models that we have proposed does create more room for analytic flexibility, and we have cautioned that the new working models should *not* be used for specification searches to reach desired results. We have briefly sketched how our approach might be built into a pre-registered analytic protocol, but it remains to see how feasible or effective this will be when planning future research syntheses.

Compared to the original development of RVE, the new working models and estimation methods that we have proposed place more emphasis on the match between the data structure and the working model. Along with this shift in emphasis, we would encourage meta-analysts to shift their understanding of RVE as a method. Currently, most researchers seem to understand RVE as a distinct, self-contained, and automatic method for meta-analysis of dependent effect sizes. Rather than viewing it this way, we suggest that it would be better to view RVE as one component tool—a technique for guarding against model mis-specification—that can be used in combination with other available strategies for modeling dependent effect sizes. We hope that this shift in emphasis might lead meta-analysts to develop and report working models that better fit the complex, often multi-level structures found in large-scale research syntheses.

Data Availability Supplementary analysis and replication materials for this study are available on the Open Science Framework at https://osf.io/x8yre/.

Declarations

Ethics Approval Not applicable. This study includes a re-analysis of data from a previously published systematic review, based on published summary data.

Conflict of Interest The authors declare they have no financial interests. James Pustejovsky is the author of the clubSandwich package for R. Elizabeth Tipton is an author of the robumeta packages for R and Stata. Neither author receives compensation from use of these software packages.

References

Baudry, C., Tarabulsy, G. M., Atkinson, L., Pearson, J., & St-Pierre, A. (2017). Intervention with adolescent mother–child dyads and cognitive development in early childhood: a meta-analysis. *Prevention Science, 18*, 116–130. https://doi.org/10.1007/s11121-016-0731-7

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a culture of replication: an examination of education and special education research grants funded by the Institute of Education Sciences. *Educational Researcher, 47*, 594–605. https://doi.org/10.3102/0013189X18788047

Conley, C. S., Durlak, J. A., & Kirsch, A. C. (2015). A meta-analysis of universal mental health prevention programs for higher education

students. *Prevention Science, 16*, 487–507. https://doi.org/10.1007/s11121-015-0543-1

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *robumeta: Robust Variance Meta-Regression* (2.0) [Computer software]. https://github.com/zackfisher/robumeta

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed., pp. 357–376). Russell Sage Foundation.

Hedberg, E. C. (2014). *ROBUMETA: Stata module to perform robust variance estimation in meta-regression with dependent effect size estimates*. Statistical Software Components S457219, Boston College Department of Economics.

Hedges, L. V., & Olkin, I. (1982). *Statistical methods for meta-analysis*. Academic press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65. https://doi.org/10.1002/jrsm.5

Hennessy, E. A., & Tanner-Smith, E. E. (2015). Effectiveness of brief school-based interventions for adolescents: A meta-analysis of alcohol use prevention programs. *Prevention Science, 16*, 463–474. https://doi.org/10.1007/s11121-014-0512-0

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The ANNALS of the American Academy of Political and Social Science, 587*, 69–81. https://doi.org/10.1177/0002716202250791

Pustejovsky, J. E. (2020). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections* (0.4.2) [R package]. https://github.com/jepusto/clubSandwich

Skeen, S., Laurenzi, C. A., Gordon, S. L., Toit, S. du, Tomlinson, M., Dua, T., Fleischmann, A., Kohl, K., Ross, D., Servili, C., Brand, A. S., Dowdall, N., Lund, C., Westhuizen, C. van der, Carvajal-Aguirre, L., Carvalho, C. E. de, & Melendez-Torres, G. J. (2019). Adolescent mental health program components and behavior risk reduction: A meta-analysis. *Pediatrics*, *144*. https://doi.org/10.1542/peds.2018-3488

Tanner-Smith, E. E., & Lipsey, M. W. (2015). Brief Alcohol Interventions for Adolescents and Young Adults: A Systematic Review and Meta-Analysis. *Journal of Substance Abuse Treatment, 51*, 1–18. https://doi.org/10.1016/j.jsat.2014.09.001

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5*, 13–30. https://doi.org/10.1002/jrsm.1091

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology, 2*, 85–112. https://doi.org/10.1007/s40865-016-0026-5

Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods, 4*, 169–187. https://doi.org/10.1002/jrsm.1070

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*, 375–393. https://doi.org/10.1037/met0000011

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*, 604–634. https://doi.org/10.3102/1076998615606099

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019a). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods, 10*, 161–179. https://doi.org/10.1002/jrsm.1338

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

Trikalinos, T. A., & Olkin, I. (2012). Meta-analysis of effect sizes reported at multiple time points: A multivariate approach. *Clinical Trials, 9*, 610–620. https://doi.org/10.1177/1740774512453218

Tyszler, M., Pustejovsky, J. E., & Tipton, E. (2017). REG_SANDWICH: Stata module to compute cluster-robust (sandwich) variance estimators with small-sample corrections for linear regression. In *Statistical Software Components*. Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458352.html

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*, 55–79. https://doi.org/10.1002/jrsm.1164

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

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