Power Approximations for Overall Average Effects in Meta-Analysis with Dependent Effect Sizes

CORRESPONDING AUTHOR

Mikkel Holding Vembye
Doctoral Student
Department of Educational Sociology
Danish School of Education, Aarhus University
Mail: mihv@edu.au.dk
Phone: (+45) 3131 9209

CO-AUTHORS

James Eric Pustejovsky
University of Wisconsin-Madison

Therese Deocampo Pigott
Georgia State University

June 1, 2022

BIOGRAPHIES

MIKKEL H. VEMBYE is a PhD candidate in the Department of Educational Sociology at the Danish School of Education, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C.; mihv@edu.au.dk. His interests include methods for conducting state-of-the-art systematic reviews and meta-analyses, especially methods for meta-analysis of dependent effect sizes.

JAMES E. PUSTEJOVSKY is Associate Professor in the Educational Psychology Department, University of Wisconsin-Madison, 1025 W. Johnson Street, Madison, WI 53706-1796; pustejovsky@wisc.edu. His research interests include statistical methods for research synthesis and meta-analysis, including methods for investigating problems of selective reporting, methods for handling complex dependence structures, and methods for synthesizing findings from single-case experimental designs.

TERRI D. PIGOTT is Professor at Georgia State University with a joint appointment in the School of Public Health and the College of Education & Human Development, P.O. Box 3995, Atlanta, GA 30302-3995; tpigott@gsu.edu. Her research interests include statistical methods for meta-analysis, including statistical power and missing data methods.

Abstract

Meta-analytic models for dependent effect sizes have grown increasingly sophisticated over the last few decades, which has created challenges for a priori power calculations. We introduce power approximations for tests of average effect sizes based upon several common approaches for handling dependent effect sizes. In a Monte Carlo simulation, we show that the new power formulas can accurately approximate the true power of meta-analytic models for dependent effect sizes. Lastly, we investigate the Type I error rate and power for several common models, finding that tests using robust variance estimation provide better Type I error calibration than tests with model-based variance estimation. We consider implications for practice with respect to selecting a working model and an inferential approach.

KEYWORDS: *power, meta-analysis, dependent effect sizes, robust variance estimation*

**Power Approximations for Overall Average Effects
in Meta-Analysis with Dependent Effect Sizes**

Meta-analyses in the social and behavioral sciences typically include studies that report on multiple outcomes measured on the same sample. Recent research in meta-analysis (Pustejovsky & Tipton, 2021; van den Noortgate et al., 2013) provides models that better reflect the complex error structure of such effect size data, recognizing the dependence among effect sizes within studies and accounting for the multilevel nature of the data. As these models come into wider use, it is important to understand their performance given the complex structure of many meta-analysis data sets. One critical aspect of performance is the statistical power of the model to detect a non-null average effect size.

Power analysis in meta-analysis can provide insight about the potential utility of a planned systematic review. Conducting an *a priori* power analysis helps researchers determine whether the existing evidence base is large enough to detect an effect size of substantive importance. Similarly, funders often request power analyses as part of grant proposals to establish whether a literature is mature enough to support a proposed research synthesis project. An *a priori* power analysis can also guide decisions about potential meta-analytic models. Meta-analysts are employing more complex models that reflect the multilevel and correlated nature of effect size data, and these models have greater data requirements than traditional models for independent effect sizes. As illustrated later in this paper, estimates of statistical power to detect a non-null average effect size may differ depending on both the nature of effect size data and the model used to approximate the distribution of effect sizes.

Available methods for calculating *a priori* power of the statistical tests used in meta-analysis are limited to models for independent effect sizes, that is, where each study contributes one independent effect size estimate to the meta-analysis (Hedges & Pigott, 2001, 2004; Jackson

& Turner, 2017; Valentine et al., 2010). However, the assumption of independent effect sizes tends to hold only for narrowly-focused and smaller-scale meta-analyses (Ahn et al., 2012; Tipton et al., 2019). As researchers adopt meta-analysis models that reflect the multivariate and multilevel nature of effect size data, information is needed about the power of these newer models, given the distinct assumptions and data structures on which they are based. In this paper, we develop new power approximations and examine the power of the test of *the mean effect size* under different strategies for modeling dependent effect sizes nested within studies. Below we review current models for dependent effect sizes nested within studies and then discuss the aims of this research.

Models for Dependent Effect Sizes

Research syntheses in the social and behavioral sciences often include multiple effect sizes from a single primary study, leading to dependent effect sizes. Dependency can occur for a variety of reasons, for example, by studies measuring multiple relevant outcomes (e.g. math and science scores, respectively) on the same sample of individuals, or by studies reporting effect sizes across multiple independent samples (e.g. results for primary and secondary school students, respectively). In the past, researchers often handled effect size dependency through ad hoc modifications of the data. For instance, researchers might calculate a synthetic effect size for each study, averaging across different outcomes and/or time points (Tipton et al., 2019), or choose a single effect size from each study for analysis. These strategies then allowed the use of univariate meta-analysis methods.

Multivariate effect size models that reflect effect size dependencies were first introduced by Hedges and Olkin (1985) and further developed by Raudenbush, Becker, and Kalaian (1988). These methods did not see widespread use in meta-analysis because they required the correlation

matrix among effect size estimates—information not usually available from primary studies. A key advance in the modeling of effect size dependencies occurred when Hedges, Tipton, and Johnson (2010) introduced the use of robust variance estimation (RVE), a technique that allows for the estimation of meta-analysis models even when the exact correlation matrix among effect sizes is unknown. More recent research (Tipton, 2015; Tipton & Pustejovsky, 2015) extended this approach, providing small-sample corrections for standard errors and hypothesis tests.

A key difference between RVE and previous approaches is that inferences under earlier multivariate models were *model-based*, meaning that they required the distributional assumptions of the model to be correctly specified for hypothesis tests and confidence intervals to work properly. In contrast, RVE makes use of a *working model* for dependence among the effect sizes, which is an approximation to the dependence structure that need not be entirely correct. Initially, Hedges and colleagues (2010) introduced two working models, called the correlated effects (CE) model and the hierarchical effects (HE) model, to approximate different aspects of dependence. They showed that even when the working model is mis-specified, it can still provide reasonably precise estimates of the mean effect size or meta-regression coefficients. Furthermore—and in contrast to model-based methods—RVE methods produce properly calibrated hypothesis tests and confidence intervals, even if the working model is mis-specified.

A limitation of the CE and HE working models is that each describes only a single type of dependence, yet in practice it is common to encounter data with multiple forms of dependence. A recently proposed strategy—coined by Pustejovsky and Tipton (2021) as the correlated-hierarchical effects (CHE) working model—recognizes both the correlated nature of effect size estimates and the multilevel structure of effect sizes nested within studies. Compared to the previously proposed CE and HE working models, the CHE working model provides

researchers with the option of more closely approximating the actual structure of meta-analytic data while also guarding against mis-specification using robust variance estimation techniques.

An alternative strategy, suggested by Van den Noortgate and colleagues (2013), is to use a multi-level meta-analysis (MLMA) model along with conventional, model-based inference methods. Van den Noortgate and colleagues (2013, 2014) demonstrated that model-based inferences from the MLMA work well in the presence of dependent effect sizes, even though aspects of the model may be mis-specified. They argued that the MLMA is therefore robust and, just as with the RVE approach, can be applied without knowledge of the dependence structure of the data. More recently, Moeyaert and colleagues (2017) conducted head-to-head comparisons of RVE (with a correlated effects working model) and MLMA. Their findings indicate that both methods perform similarly when the data include a large number of studies, but that RVE provided more accurate uncertainty assessments when the number of studies was limited. Further, Fernandez-Castilla and colleagues (2020) suggested that MLMA could be treated as a working model and combined with RVE to gain robustness against model-mis-specification.

Aims

In this article, we investigate the power of current methods for handling dependent effect sizes in meta-analysis. We pursue three aims: 1) to develop approximations for the power of hypothesis tests based upon models that reflect the multivariate and multilevel nature of effect size data, 2) to validate these approximations using simulations, and 3) to provide guidance to researchers applying these methods with respect to Type I error and power of different working models and tests. To illustrate the approximations and to provide context for the simulation conditions, we use a recent meta-analysis conducted by Dietrichson, Bøg, Filges, and Jørgensen (2017, henceforth DBFJ17) that investigated interventions for increasing the academic

achievement (i.e. mathematics and reading performance) of students with low socioeconomic status.

We develop new approximations for the power of several different hypothesis tests in meta-analysis of dependent effect sizes. For developing prospective power calculations, it is necessary to posit a true data-generating process. We take as a starting point the CHE model because it nests many other simpler specifications of interest. Under the CHE, we provide power approximations for 1) a model-based test based on a correctly specified working model (CHE-model); 2) a robust test based on a correctly specified model (CHE-RVE); 3) a robust test based on a simpler correlated effects (CE) working model, which may not be correctly specified; 4) a model-based test based on a potentially mis-specified MLMA model (MLMA-model); and 5) a robust test that uses the MLMA as a working model (MLMA-RVE). We then provide examples of how to use these approximations to determine power for testing an overall average effect. Next, we test and validate the performance of the new power approximations via Monte Carlo simulation by comparing the true simulated and approximated power across various model conditions. Before describing the new power approximations, we review extant methods for statistical power in univariate meta-analysis.

Power Approximation for Univariate Meta-Analysis

Current methods for *a priori* power calculations are limited to models that include a single, independent effect size estimate from each study. Consider such a meta-analysis, based on data from J studies, where the primary aim is to test the null hypothesis that the overall average effect size μ is equal to a specific value d . Let σ_j denote the standard error of the effect size estimate from study j , for $j = 1, \dots, J$. Under a univariate random-effects model (RE), the null hypothesis $H_0: \mu = d$ would typically be tested using the Wald statistic

$$t^U = \frac{\hat{\mu} - d}{\sqrt{\hat{V}}} \quad (1)$$

where $\hat{\mu}$ is the random effects estimate of the overall average effect size and $\hat{V}$ is its estimated sampling variance (Hedges & Pigott, 2001). When the null hypothesis holds, the test statistic t^U approximately follows a central Student-t distribution with $J - 1$ degrees of freedom; when the null does not hold, it approximately follows a non-central Student-t distribution with non-centrality parameter $\lambda = (\mu - d)/\sqrt{V}$ and $J - 1$ degrees of freedom, where V is the expected sampling variance (Hartung & Knapp, 2001). Power is therefore given by

$$F_t(-c_{\alpha/2, J-1} | J - 1, \lambda) + 1 - F_t(c_{\alpha/2, J-1} | J - 1, \lambda) \quad (2)$$

where $F_t(x | \nu, \lambda)$ is the cumulative distribution function of a non-central Student-t distribution, and $c_{\alpha, \zeta}$ is the upper α -level critical value for the central Student-t distribution with ζ degrees of freedom, so $F_t(c_{\alpha/2, \zeta} | \zeta, 0) = 1 - \alpha/2$.

The usual way of approximating *a priori* power under a univariate model is first to determine the minimum effect size of practical significance, and second to estimate the variance of the weighted overall mean effect size based on a) the average sampling variance of an effect size estimate in a “typical” study; b) the true between-study variance, τ^2 ; and c) the anticipated number of studies in the meta-analysis, J . The variance of the weighted mean effect size is approximately $V = 1/W^{RE}$, where $W^{RE} = \sum_{j=1}^J w_j^*$ and $w_j^* = (\tau^2 + \sigma_j^2)^{-1}$ are the study-specific inverse variance weights under the random effects model. If complete balance of sample sizes is assumed, so that $\sigma_1 = \sigma_2 = \dots = \sigma_J = \sigma$, then V simplifies to $(\tau^2 + \sigma^2)/J$.

In meta-analyses of standardized mean difference effect sizes comparing two groups, the effect size estimate’s sampling variance is closely related to the overall sample size (Valentine et al., 2010). Assuming the groups are of equal size,

$$\sigma^2 \approx \left(\frac{4}{N} + \frac{\mu^2}{2(N-2)} \right), \quad (3)$$

where N is the average effective sample size. Thus, if we know the average effective sample size of studies in a given area, we can approximate the average sampling variance. To arrive at a value for the between-study variance τ^2 , Pigott (2012) suggested that $\tau^2 = (1/3)\sigma^2$ could be considered a low degree of heterogeneity, $\tau^2 = \sigma^2$ could be considered a moderate degree of heterogeneity, and $\tau^2 = 3\sigma^2$ could be considered a large degree of heterogeneity.

Suppose we aim to estimate the power of the test for $H_0: \mu = 0$, with the usual level of $\alpha = .05$, in a meta-analysis of standardized mean difference effect sizes. With a low degree of heterogeneity, Pigott's guidelines would suggest a sampling variance of approximately $V = 4\sigma^2/(3J)$. Suppose that we expect to identify at least 12 studies and that the average effective sample size is $N = 100$. Therefore, $\sigma^2 \approx 4/100$ and the expected sampling variance is at most $V = 16/3600$. Using this value in Equation (2), we find power of 0.278 for an average effect of $\mu = 0.1$, power of 0.780 for $\mu = 0.2$, and power of 0.983 for $\mu = 0.3$.

These power approximations do not apply directly in meta-analyses involving dependent effect sizes. However, one could try applying them by calculating power assuming that there is just one effect size estimate per study—as would be the case if the meta-analyst calculated a single, synthetic effect size per study. Following this approach, we would anticipate power of 0.278 to detect an average effect of $\mu = 0.1$ in a meta-analysis of 12 studies with an average sample size of $N = 100$, regardless of whether each study included a single or multiple effect size estimates. The performance of this approximation in terms of predicting the true power of models with synthetic effect sizes is not known. Furthermore, because this approximation underdetermines key quantities needed for calculating power under models for dependent effect sizes, we now turn to the development of new approximations.

Power Approximations for Meta-Analysis of Dependent Effect Sizes

We now describe approximations for the power of tests for an overall average effect in a meta-analysis of dependent effect sizes. We assume that the data-generating process conforms to the correlated-and-hierarchical effects (CHE) model as described by Pustejovsky and Tipton (2021). Under this data-generating process, we consider several different testing procedures, including both model-based tests and robust tests based on three distinct working models. Unlike the univariate approximations described in the previous section, we allow for sampling variances and other features to differ from study to study, so that we can examine the implications of assuming that study features are homogeneous.

Consider a collection of J studies to be included in a meta-analysis, where each study contributes k_j effect size estimates, for $j = 1, \dots, J$. Let T_{ij} denote effect size estimate i from study j , with corresponding standard error σ_{ij} , for $i = 1, \dots, k_j$ and $j = 1, \dots, J$. As usual in meta-analysis, we shall assume that each T_{ij} is an unbiased estimator of an effect size parameter θ_{ij} and that σ_{ij} is fixed and known. These assumptions can be expressed by the model

$$T_{ij} = \theta_{ij} + e_{ij}, \quad (4)$$

where $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, with $E(e_{ij}) = 0$ and $\text{Var}(e_{ij}) = \sigma_{ij}^2$. We assume that the effect size estimates from different studies are uncorrelated, so $\text{cor}(e_{hj}, e_{il}) = 0$ when $j \neq l$, but that effect size estimates from the same study may be correlated. For simplicity, we also assume that the sampling variances are constant within each study, so $\sigma_{1j}^2 = \sigma_{2j}^2 = \dots = \sigma_{k_j j}^2 = \sigma_j^2$, and that the correlations between sampling errors within a given study are all equal to a known constant, $\text{cor}(e_{hj}, e_{ij}) = \rho$.

Following the CHE model, we assume that the effect size parameters represent a sample from an underlying population of effects that has a hierarchical structure, according to

$$\theta_{ij} = \mu + u_j + v_{ij}, \quad (5)$$

where the study-level error term u_j has mean zero and variance τ^2 and the effect size-level error term v_{ij} has mean zero and variance ω^2 . The main parameters of the data-generating model are then the overall average effect size μ ; the between-study heterogeneity τ^2 ; the within-study heterogeneity ω^2 ; and the sampling correlation ρ . Under this model, we consider tests of the null hypothesis $H_0: \mu = d$ versus a two-sided alternative, with specified Type-I error level α .

Estimation of CHE

If one treats the CHE model as correctly specified, then estimation of the overall average effect size μ entails first estimating the variance components and then using the estimated variance components to take an inverse-variance weighted average of the effect size estimates. Let $\hat{\tau}^2$ and $\hat{\omega}^2$ denote full or restricted maximum likelihood estimators of the variance components, which are calculated given the true sampling correlation ρ . Given values of these estimators, the overall average effect size estimate is a weighted average of the study-specific average effect size estimates, with weights given by

$$w_j = \frac{k_j}{k_j \hat{\tau}^2 + k_j \rho \sigma_j^2 + \hat{\omega}^2 + (1 - \rho) \sigma_j^2}. \quad (6)$$

The overall weighted average is then

$$\hat{\mu} = \frac{1}{W} \sum_{j=1}^J w_j \bar{T}_j, \quad (7)$$

where $\bar{T}_j = \frac{1}{k_j} \sum_{i=1}^{k_j} T_{ij}$ and $W = \sum_{j=1}^J w_j$. If the CHE model is correctly specified, then

$$\text{Var}(\hat{\mu}) \approx \frac{1}{W}. \quad (8)$$

The approximation here arises because W is calculated using estimated variance components rather than known parameter values. Note that the weights given in Equation (6) are inverse-variance and therefore minimize the variance of the weighted average in Equation (7).

Model-Based Hypothesis Test

One way to test the null hypothesis $H_0: \mu = d$ is via a conventional Wald test. The model-based Wald test statistic is

$$t^M = \frac{\hat{\mu} - d}{\sqrt{V^M}}, \quad (9)$$

where $V^M = 1/W$. Consider the scenario in which the CHE model is correctly specified and the number of independent studies is large. If the null hypothesis holds, then t^M follows a standard normal distribution. If the null hypothesis does not hold, then t^M approximately follows a normal distribution with mean

$$\lambda = \sqrt{W}(\mu - d). \quad (10)$$

and unit variance. However, such large-sample approximations do not necessarily provide an adequate guide for sample sizes encountered in practice (i.e., fewer than 40 primary studies) because of the uncertainty in the variance component estimates used to calculate W . It is thus desirable to develop an approximation that works even with a smaller number of studies.

In practice, researchers might use a Student-t distribution with $J - 1$ degrees of freedom as a reference distribution in the model-based tests. This is a fairly rough approximation to the sampling distribution of the model-based test. Alternatives would be to use a Satterthwaite approximation (Satterthwaite, 1941; Giesbrecht & Burns, 1985) for the degrees of freedom or Kenward and Roger (2009) approximation for the sampling variance estimator and degrees of

freedom. We consider the Satterthwaite approximation because it is simpler and more tractable and because it has been found to perform well in the context of linear mixed models (e.g., Luke, 2017). Section S2 of the supplementary materials provides further technical details.

We propose to approximate the power of the model-based Wald test by assuming that t^M follows a non-central Student-t distribution with non-centrality parameter λ and degrees of freedom ζ , with the latter quantity determined using Satterthwaite approximation. As previously, let $F_t(x|\zeta, \lambda)$ be the cumulative distribution function of the Student-t and let $c_{\alpha, \zeta}$ be the upper α -level critical value from a central Student-t distribution. The power of the model-based Wald test against a two-sided alternative can then be approximated by

$$F_t(-c_{\alpha/2, \zeta}|\zeta, \lambda) + 1 - F_t(c_{\alpha/2, \zeta}|\zeta, \lambda). \quad (11)$$

Under the CHE model, the Satterthwaite degrees of freedom are given by

$$\zeta = \frac{st - u^2}{sy^2 + tx^2 - 2uxy}, \quad (12)$$

where

$$x = \frac{1}{W} \sum_{j=1}^J w_j^2, \quad y = \frac{1}{W} \sum_{j=1}^J \frac{w_j^2}{k_j}, \quad s = x^2 + Wx - \frac{2}{W} \sum_{j=1}^J w_j^3, \\ t = y^2 + \sum_{j=1}^J \frac{w_j^2}{k_j^2} + \sum_{j=1}^J \frac{k_j - 1}{(\hat{\omega}^2 + (1 - \rho)\sigma_j^2)^2} - \frac{2}{W} \sum_{j=1}^J \frac{w_j^3}{k_j^2}, \quad \text{and} \quad u = xy + Wy - \frac{2}{W} \sum_{j=1}^J \frac{w_j^3}{k_j}$$

(Supplementary Materials Section S2.1 provides a derivation). If all studies include the same number of effect sizes ($k_1 = k_2 = \dots = k_J = k$) and have equal standard errors ($\sigma_1 = \sigma_2 = \dots = \sigma_J = \sigma$), we describe the meta-analytic sample as “completely balanced.” With a completely balanced sample, the weights will be equal for any values of the variance components τ^2 and ω^2

and the degrees of freedom will simplify to $\zeta = J - 1$. In a sample that is not completely balanced, ζ will be less than $J - 1$.

We note that the Satterthwaite approximation is not commonly applied in practice, nor is it readily available in software. In principle, one could use the power approximation with other degrees of freedom, such as by substituting the critical value $-c_{\alpha/2, J-1}$. However, such a test would have distorted Type-I error rate to the extent that the Satterthwaite degrees of freedom deviate from $J - 1$. We explore the extent of such size distortions in the simulation study.

In order to implement this power approximation prospectively, one will need to calculate weights for each of the J included studies. We propose to make such calculations using assumed values for the variance component estimates $\hat{\tau}^2$ and $\hat{\omega}^2$ and sampling correlation ρ , as well as assumptions about the distribution of primary study sample sizes and effect sizes per study. We demonstrate these calculations and discuss these assumptions further at the end of this section.

Robust Hypothesis Test

Even when using Satterthwaite degrees of freedom, the model-based test will have close-to-correct Type I error only when the assumptions of the CHE working model hold. Given the typical lack of information about the sampling correlations between effect size estimates, meta-analysts may prefer to use tests based on robust variance estimation methods, which maintain close-to-correct size even if the CHE model is mis-specified. With the CHE working model, a robust estimator for the variance of $\hat{\mu}$ is given by

$$V^R = \frac{1}{W^2} \sum_{i=1}^J \frac{w_j^2 (\bar{T}_j - \hat{\mu})^2}{\left(1 - \frac{w_j}{W}\right)}. \quad (13)$$

This formula incorporates a small-sample correction (the CR2 correction), as proposed by Tipton (2015). Specifically, the denominator in the summand of Equation (13) is a small-sample

adjustment that makes V^R equivalent to the CR2 estimator. V^R is an exactly unbiased estimator of $\text{Var}(\hat{\mu})$ when the working model is correctly specified. However, even if the assumptions of the working model do not hold, V^R remains close to unbiased. We focus on the CR2 estimator because it has been found to outperform other robust variance estimators and is recommended for general use in the context of meta-analysis of dependent effect sizes (Tipton, 2015; Tipton and Pustejovsky, 2015).

A robust Wald test statistic based on V^R is

$$t^R = \frac{\hat{\mu} - d}{\sqrt{V^R}}. \quad (14)$$

Again consider the scenario in which the CHE model is correctly specified and the number of independent studies is large. If the null hypothesis holds, then t^R follows a standard normal distribution. If the null hypothesis does not hold, then t^R approximately follows a normal distribution with mean λ (as given in Equation 10) and unit variance. Therefore, with a sufficiently large number of studies, the robust test has power equivalent to that of the model-based test. However, Tipton (2015) and Tipton and Pustejovsky (2015) observed that there is no clear rule of thumb regarding a sufficient number of studies to trust the large-sample approximation because its adequacy depends on study features besides the total number of studies. Thus, large-sample approximations do not generally provide an adequate guide for sample sizes encountered in practice.

Tipton (2015) proposed approximating the distribution of t^R under the null hypothesis by a Student-t distribution with ξ degrees of freedom, where ξ is derived based on a Satterthwaite approximation under the assumption that the working model is correct. Here, we propose to use the same approximation when the null does not hold, so that t^R approximately follows a non-

central Student-t distribution with ξ degrees of freedom and non-centrality parameter λ . The power of the robust Wald test against a two-sided alternative can then be approximated by

$$F_t(-c_{\alpha/2, \xi} | \xi, \lambda) + 1 - F_t(c_{\alpha/2, \xi} | \xi, \lambda). \quad (15)$$

If the working model is correctly specified (and treating the variance components as known), then the degrees of freedom for the robust test are given by

$$\xi = \left[\sum_{j=1}^J \frac{w_j^2}{(W - w_j)^2} - \frac{2}{W} \sum_{j=1}^J \frac{w_j^3}{(W - w_j)^2} + \frac{1}{W^2} \left(\sum_{j=1}^J \frac{w_j^2}{W - w_j} \right)^2 \right]^{-1} \quad (16)$$

(see Supplementary Materials Section S2.2 for a derivation). In a completely balanced sample, the degrees of freedom simplify to $\xi = J - 1$. When the sample is not completely balanced, the degrees of freedom will be less than $J - 1$ to an extent that depends on the degree of imbalance. One implication is that, for a completely balanced meta-analytic sample, the robust test has power approximately equivalent to that of the model-based test. The tests might diverge in power, however, when the primary study features are imbalanced.

RVE with CE Working Model

The original implementation of RVE introduced working models that were simplifications of the CHE model, as well as using weights that were not exactly inverse-variance under those simplified working models. The default working model, called the correlated effects (CE) model, has only a single, between-study variance component, estimated using a method-of-moments formula. Let $\tilde{\tau}^2$ denote this method-of-moments estimator. If the true data-generating process follows the CHE model, then this estimator has expectation

$$E(\tilde{\tau}^2) = \tau^2 + \omega^2 \left(\frac{1 - \sum_{j=1}^J \frac{1}{k_j \sigma_j^4}}{1 - \sum_{j=1}^J \frac{1}{\sigma_j^4}} \right). \quad (17)$$

For purposes of power calculations, we will approximate the estimator $\bar{t}^2$ by its expectation.

The weights used with the CE model are given by

$$\ddot{w}_j = \frac{1}{(\bar{t}^2 + \sigma_j^2)}, \quad (18)$$

with overall average effect size estimator $\ddot{\mu} = \sum_{j=1}^J \ddot{w}_j \bar{T}_j / \ddot{W}$, where $\ddot{W} = \sum_{j=1}^J \ddot{w}_j$. If the CE model is applied when the true data-generating process follows the CHE model, then the variance of the overall average effect size estimator will be

$$\text{Var}(\ddot{\mu}) = \ddot{S} = \frac{1}{\ddot{W}^2} \sum_{j=1}^J \ddot{w}_j^2 \left(\tau^2 + \rho \sigma_j^2 + \frac{1}{k_j} [\omega^2 + (1 - \rho) \sigma_j^2] \right) = \frac{1}{\ddot{W}^2} \sum_{j=1}^J \frac{\ddot{w}_j^2}{w_j}, \quad (19)$$

which will generally be larger than $1/W$ because the CE model weights differ from the variance-minimizing weights given in Equation (6).

This approximation for the power of the robust test with the CE working model entails two simplifications. First, the robust variance estimator itself is not exactly unbiased because the working model is not correctly specified (although the estimator is still asymptotically consistent as the number of studies increases). Second, the Satterthwaite degrees of freedom approximation is derived under the assumption that the working model is correctly specified, which is not the case here. As a result, the approximation might not provide the correct Type-I error rate.

Ignoring both of these complications for the time being, we propose to approximate the power of the robust test based on the CE model using the same Student-t approximation as above, but with non-centrality parameter

$$\ddot{\lambda} = \frac{\mu - d}{\sqrt{\ddot{S}}} \quad (20)$$

and degrees of freedom $\ddot{\xi}$, calculated just as in Equation (16), but with $\ddot{w}_j$ in place of w_j . In the completely balanced case, $\ddot{S} = 1/W$, $\ddot{\lambda} = \lambda$, and $\ddot{\xi} = \xi = J - 1$, and so the test will have power

equal to the other tests. If the data are not completely balanced, then the power of the CE test might diverge from that of the robust test based on the CHE working model.

Multi-level Meta-Analysis

Van den Noortgate et al. (2013, 2014) proposed handling dependent effect sizes via a multi-level meta-analysis model (MLMA), which includes both between-study and within-study random effects but ignores the possible correlation of effect size estimates drawn from the same sample. This model is a special case of the CHE, under the assumption that the correlation between sampling errors is $\rho = 0$. When the true sampling correlation is non-zero, the model is mis-specified. However, Van den Noortgate et al. (2013, 2014) provided simulation evidence that model-based standard errors can still be accurate despite the model mis-specification.

A challenge in analyzing the power of the MLMA model is that the variance component estimates may be systematically biased when the true sampling correlation is non-zero. For purposes of power calculations, we approximate the variance component estimators using the values that minimize the Kullback-Liebler divergence between the MLMA and the true data-generating model (White, 1982). Let $\tilde{\tau}^2$ and $\tilde{\omega}^2$ denote the minimizing values of the between-study and within-study variance components, respectively. Section S3 of the supplementary materials provides further details about how these quantities are calculated.

The weights used with the MLMA model are then given by

$$\tilde{w}_j = \frac{k_j}{(k_j \tilde{\tau}^2 + \tilde{\omega}^2 + \sigma_j^2)}, \quad (21)$$

with overall average effect size estimator $\tilde{\mu} = \sum_{j=1}^J \tilde{w}_j \bar{T}_j / \tilde{W}$, where $\tilde{W} = \sum_{j=1}^J \tilde{w}_j$. The variance of the overall average effect size estimator is

$$\text{Var}(\tilde{\mu}) = \tilde{S} = \frac{1}{\tilde{W}^2} \sum_{j=1}^J \tilde{w}_j^2 \left(\tau^2 + \rho \sigma_j^2 + \frac{1}{k_j} [\omega^2 + (1 - \rho) \sigma_j^2] \right) = \frac{1}{\tilde{W}^2} \sum_{j=1}^J \frac{\tilde{w}_j^2}{w_j}. \quad (22)$$

For the MLMA, the model-based variance estimator is $1/\tilde{W}$, which may be a biased estimator for $\text{Var}(\tilde{\mu})$ due to mis-specification.

The MLMA model is commonly applied with model-based variance estimation and $J - 1$ degrees of freedom. However, for consistency with the other models that we have examined, we consider approximating the power of the test using Satterthwaite degrees of freedom. We calculate these using Equation (12), but substituting $\tilde{\omega}^2$ for $\hat{\omega}^2$ and $\tilde{w}_j$ for w_j . Let $\tilde{\zeta}$ denote the MLMA model-based degrees of freedom and let $\tilde{\lambda} = (\mu - d)/\tilde{S}$. We approximate the power of the model-based Wald test with Satterthwaite degrees of freedom as

$$F_t(-g \times c_{\alpha/2, \tilde{\zeta}} | \tilde{\zeta}, \tilde{\lambda}) + 1 - F_t(g \times c_{\alpha/2, \tilde{\zeta}} | \tilde{\zeta}, \tilde{\lambda}), \quad (23)$$

where $g = 1/\sqrt{\tilde{W}\tilde{S}}$. For the test with $J - 1$ degrees of freedom, we replace $c_{\alpha/2, \tilde{\zeta}}$ with $c_{\alpha/2, J-1}$.

Fernandez-Castilla et al. (2020) suggested combining MLMA with robust variance estimation. We approximate the power of the robust test based on the MLMA model by following the same approach as with the CE model. We denote the Satterthwaite degrees of freedom based on the MLMA working model as $\tilde{\xi}$, calculated by using $\tilde{w}_j$ in place of w_j in Equation (16). We then approximate the power of the robust test using Equation (15), with $\tilde{\lambda}$ in place of λ and $\tilde{\xi}$ in place of ξ .

Using the Power Approximations: A Computational Example

To put each of these power approximations into practice, we need to determine the non-centrality parameters and the degrees of freedom of each of the tests. These quantities are a function of a) the number of included studies, J ; b) the parameters of the data-generating model, τ , ω , and ρ ; and c) the sample characteristics, including the primary study sample sizes and the

TABLE 1. *Hypothetical Studies in a Meta-Analysis*

Study	N_j	σ_j^2	k_j	CHE weight (w_j)	CE weight ($\ddot{w}_j$)	MLMA weight ($\tilde{w}_j$)	RE weight (w_j^*)
A	28	1 / 7	1	6.02	6.23	5.77	6.40
B	32	1 / 8	3	10.00	7.01	13.87	7.23
C	40	1 / 10	2	10.71	8.50	12.43	8.82
D	48	1 / 12	3	13.85	9.91	17.17	10.34
E	56	1 / 14	4	16.54	11.23	20.70	11.80
F	64	1 / 16	2	15.34	12.48	16.21	13.19
G	80	1 / 20	2	17.91	14.79	18.03	15.79
H	96	1 / 24	1	15.38	16.87	13.87	18.18
I	128	1 / 32	2	23.94	20.46	21.70	22.43
J	180	1 / 45	3	31.76	25.10	26.41	28.12
K	192	1 / 48	5	35.93	26.00	28.88	29.27
L	256	1 / 64	2	33.28	30.08	26.13	34.53

number of effect size estimates in each primary study. We now demonstrate the mechanics of the power calculations using a hypothetical example (with quantities chosen for ease of calculation rather than verisimilitude).

Consider an on-going review in which the investigators have identified $J = 12$ studies and determined the (average) sampling variances and number of eligible outcomes available in each study. Table 1 lists these quantities. Recall that in our prior univariate power example, we assumed an average sampling variance of $\sigma^2 = 4/100$ and a low degree of heterogeneity, with $\tau = \sqrt{\sigma^2/3} = 1/\sqrt{75} = 0.115$. Let us also assume $\omega = .10$ and $\rho = .5$ and determine power under the CHE, CE, and MLMA models to detect an average effect size of $\mu = 0.1$.

Given the assumed values of the variance components, we can calculate weights under the CHE, CE, and MLMA models, as well as under the univariate random effects model (i.e., ignoring that studies include multiple, dependent effect size estimates). These weights are reported in the last four columns of Table 1. Given the CHE weights, we calculate $W = 230.66$ and $\lambda = 1.519$. For the model-based test, Equation (12) gives $x = 23.798$, $y = 9.5751$, $s =$

TABLE 2. *Power calculations based on hypothetical study characteristics*

Working model and test	Sampling Variance	Degrees of freedom	Power (%)
<i>CHE working model</i>			
Model-based, large-sample d.f.	0.00434	11	29.40
Model-based, Satterthwaite d.f.	0.00434	8.37	27.06
Robust, Satterthwaite d.f.	0.00434	8.71	27.27
<i>CE working model</i>			
Robust, Satterthwaite d.f.	0.00441	8.71	26.90
<i>MLMA working model</i>			
Model-based, large-sample d.f.	0.00444	11	27.87
Model-based, Satterthwaite d.f.	0.00444	9.54	26.66
Robust, Satterthwaite d.f.	0.00444	9.72	27.28
<i>Univariate random effects model</i>			
Model-based, Knapp-Hartung test	0.00485	11	25.90

4741.0, $t = 24455$, $u = 1944.5$, and Satterthwaite degrees of freedom $\zeta = 8.37$. From Equation (16), the robust test has degrees of freedom $\xi = 8.71$. Based on the CE weights and assumed model parameters, we calculate $\tilde{\tau}^2 = 0.0175$ (Equation 17), $\tilde{S} = 0.004409$ (Equation 19), $\tilde{\lambda} = 1.506$ (Equation 20), and $\tilde{\xi} = 8.71$. Based on the MLMA weights, we calculate $\tilde{\tau}^2 = 0.0303$, $\tilde{\omega}^2 = 0$, $\tilde{S} = 0.004442$, $\tilde{\lambda} = 1.500$, $g = 1.0088$, $\tilde{\zeta} = 9.54$, and $\tilde{\xi} = 9.71$.

Table 2 reports the variance estimates, degrees of freedom, and power based on each of these working models and approximations. In this particular example, the model-based CHE test, the robust CHE test, the robust CE test, the model-based MLMA test, and the robust MLMA test all have quite similar power. Using the effective sample sizes listed in Table 1, the univariate approximation described in the previous section gives power of 25.9%, slightly lower than the power of the more complex approximations.

Using the Power Approximations in Practice

Often, researchers will need to make prospective power calculations before completing the search and screening process of a systematic review. In this situation, the number of included studies and properties of those studies will not yet be known, and so the researcher will have to

make assumptions about the distribution of sampling variances and number of effect sizes per study. Assuming complete balance will generally yield optimistic power calculations. Alternative approaches would be to simulate σ_j^2 and k_j from stylized distributions with specified parameters or to sample σ_j^2 and k_j from an empirical distribution of study characteristics—perhaps based on pilot data or previous syntheses on similar research topics. With approaches that sample study characteristics, the power approximations given in Equations (11), (15), and (23) become random quantities, with distributions governed by the distribution of σ_j^2 and k_j . For prospective power analysis, we can calculate power as the expectation over this distribution, such as by drawing repeated samples of size J , calculating power, and then averaging over the samples.

We now demonstrate the power calculations as they might be used in practice, by developing power estimates based on the characteristics of primary studies included in the DBFJ17 meta-analysis. For purposes of illustration, we used the subsample of 77 studies (including 317 unique effect sizes) comprised of all studies with effective sample sizes of no more than 500 and no more than 20 effect sizes per study. Many of the included studies were cluster-randomized trials, for which sampling variances were computed using cluster-adjustment formulas from Hedges (2007). In the analytic sample of 77 studies, effective sample sizes ranged from 19 to 485, with a median of 87, a mean of 140, and a standard deviation of 125. The average sampling variance was $\sigma^2 = 0.068$. Included studies reported between 1 and 18 effect sizes, with a median of 3, a mean of 4.1, and a standard deviation of 3.5.

We calculate power to detect an average effect of $\mu = 0.1$, again assuming $\tau = 0.115$, $\omega = 0.10$, and $\rho = .5$ for sample sizes ranging from $J = 5$ to $J = 40$. Figure 1 displays the power of each model for which we have developed approximations. Each panel corresponds to a different method of determining the distribution of study characteristics. In the left panel, we

POWER APPROXIMATIONS FOR META-ANALYSIS OF DEPENDENT EFFECT SIZES

assume a completely balanced sample with $\sigma_j^2 = .068$ and $k_j = 4.1$, the average values of the studies in DBFJ17. Because the sample characteristics are perfectly balanced, the power of all three working models for dependent effect sizes coincide and can be calculated directly from the formulas, without re-sampling. In the middle panel of Figure 1, we determined the sample characteristics by drawing $4/\sigma_j^2$ from a gamma distribution with shape $\alpha = 1.33$ and $rate = .0095$ [which we obtained from fitting to the effective sample sizes from DBFJ17 by maximum likelihood using the `fitdistr` function from the MASS package (Venables & Ripley, 2002)] and by sampling $k_j \sim 1 + \text{Poisson}(3.1)$. The distribution of sampling variances therefore closely matches the empirical distribution from DBFJ17, while the distribution of number of effect sizes has a similar average but lower dispersion than the empirical distribution. In the right panel of Figure 1, we determine the sample characteristics by repeatedly sampling directly from the empirical distribution of sampling variances and number of effect sizes found in DBFJ17.

Across all three panels, the power of the aggregate-level approximation is notably lower because it does not account for the availability of multiple effect sizes per study. In each of the panels, the power of the model-based and robust tests under the correctly specified CHE working model are very similar. Because the CE working model uses weights that are not entirely efficient when the study characteristics are not balanced, the CE-RVE test has slightly lower power than the tests based on the CHE, but the difference is only noticeable when k_j and σ_j^2 are sampled from the empirical data. Similarly, the MLMA tests have lower power than the CHE tests because the MLMA tests use weights that are not entirely efficient.

POWER APPROXIMATIONS FOR META-ANALYSIS OF DEPENDENT EFFECT SIZES

FIGURE 1. Power for finding $\mu = 0.1$ with $\tau = 0.115$, $\omega = 0.1$ and $\rho = .5$ across three different methods for obtaining n_j and σ_j^2 . For the stylized and pilot sample (empirical) methods, the average power is estimated across 100 iterations. Dashed lines indicate power of 80 percent.

Comparing across panels, the power levels of each test are substantially higher when based on balanced study characteristics than when based on the stylized distributions or empirical distributions. For instance, with $J = 25$ studies, the CHE-RVE test has power of 0.79 when assuming balanced study characteristics, but power of only 0.70 when using the stylized distribution or 0.65 when using the empirical distribution. A very similar pattern holds for the other model-based and robust tests (see Supplementary Figure S1 for further details).

Simulation Study

We used Monte Carlo simulation to validate the new power approximations and investigate the performance of different working models and inferential approaches for testing overall average effects. We designed the simulations to address three specific aims. First, we examine the accuracy of the proposed power approximations by comparing predicted power levels to simulation-based estimates of power, which fully capture the uncertainties of estimating the working models from limited data. In these analyses, we are interested both in the overall accuracy of the approximations as well as the extent to which the assumed distribution of σ_j^2 and k_j matters for obtaining accurate power estimates. Second, we evaluate the empirical Type I

error rates of tests based on the different working models and inferential approaches for which we have provided power approximations. Third, we examine the relative power of tests that adequately control Type I error rates. Across all three aims, we seek to provide a basis for clearer recommendations about how to select a working model and an inferential approach in meta-analyses of dependent effect sizes.

Data Generation Process

The simulations focused on a data-generating process in which the true error structure followed the correlated-hierarchical effect (CHE) working model from Equations (4) and (5) because this model nests the simpler correlated effect model and MLMA model. The data generating procedures followed the same process as the simulations reported by Pustejovsky and Tipton (2021), except that we used the DBFJ17 data to inform the distribution of study characteristics. We imposed the same restrictions as in the example described in the previous section, after which the analytic sample was comprised of 77 studies, with an average effective sample size of 140 and an average of 4.1 effect sizes per study.

We simulated standardized mean difference (SMD) effect size estimates because this is one of the most common metrics encountered in meta-analyses in education (Ahn et al., 2012; Tipton et al., 2019). We generated effect size estimates by first simulating study-specific characteristics and effect size parameters. We simulated effective sample sizes N_j and the number of effect sizes k_j by sampling from the study characteristics of DBFJ17. We then simulated effect size parameters based on Equation (5), given values of the overall average effect size μ , between-study SD τ , and within-study SD ω . We assumed that the effect size estimates from a given study were equi-correlated with a common correlation ρ . We focus on this case in order to compare the approximations against the true simulated power when the CHE working

model is correctly specified. Although simplistic, assuming a constant correlation between sampling errors is concordant with the degree of detail that meta-analysts will usually be able to provide in practice—particularly when conducting power analysis in the early stages of a project. Given the study-specific parameters, we simulated unstandardized mean difference effect size estimates and sampling variances. Supplementary Materials Section S4 provides further implementation details.

Estimators

For each simulated dataset, we applied eight different tests that varied in terms of the working model, the variance estimator, and the method for calculating degrees of freedom (d.f.). Specifically, we calculated all five tests for which we have developed power approximations, including: the CHE working model with model-based variance and with robust variance estimation, the CE working model with robust variance estimation, the MLMA model with model-based variance and with robust variance estimation. For the robust tests, we used the CR2 variance estimator (Equation 13) and Satterthwaite d.f. with each of the working models because this combination of small-sample adjustments has been recommended for use in practice (Tipton, 2015). We also used the Satterthwaite d.f. for the model-based variance estimators. However, because this strategy is novel and not typically applied in practice, we also examined tests based on the CHE working model and the MLMA working model with model-based variance and the more conventional choice of $J - 1$ d.f. Finally, we also included a test based on the common approach of aggregating effect sizes to the study level. For the aggregated effect sizes, we used a univariate random effects model, with Knapp-Hartung adjusted standard error (Hartung & Knapp, 2001) and $J - 1$ d.f. We estimated all the above models using the `metafor`

(Viechtbauer, 2010), `robumeta` (Fisher & Tipton, 2015), and `clubSandwich` (Pustejovsky, 2020) packages in R.

Experimental Design

We examined the performance of the tests using a full factorial design with 768 unique conditions. Meta-analyses in the social and behavioral sciences include a wide range of studies; a recent review of meta-analyses published in *Psychological Bulletin* between 1990 and 2017 found that the number of included studies ranged from 12 to 1,753 with a median of 75; about 19% of meta-analyses included 40 or fewer primary studies and 39% included 60 or fewer (Polanin et al. 2020). We therefore varied the number of independent studies from $J = 10$ to 60 (see Table 3). These represent a small to moderate number of studies compared to sample sizes encountered in meta-analyses in education (Tipton et al., 2019). We used a maximum of 60 studies because power tended to reach ceiling levels beyond this range.

As shown in Table 3, we set the true average effect size to values of $\mu = 0$ (to investigate the Type I error rate), 0.05, 0.1, or 0.2 (to examine power). The latter values represent a small, moderate, and large effect sizes for educational interventions, as suggested by Kraft (2020). We chose $\tau = 0.05$, 0.2, or 0.4 to represent a small, medium, or large amount of between-study heterogeneity, respectively. These values cover a broad range of heterogeneity levels observed in social science syntheses, including both meta-analyses of direct replications and meta-analyses of broader literatures (Linden & Hönokopp, 2021; Olsson-Collentine et al., 2020). We used $\omega =$

TABLE 3. *Design factors for the simulation study*

Factor	Parameter values
Number of studies (J)	10, 20, 40, 60
Average effect size (μ)	0.00, 0.05, 0.10, 0.20
Between-study heterogeneity (τ)	0.05, 0.20, 0.40
Within-study heterogeneity (ω)	0.00, 0.05, 0.10, 0.20
Sampling correlation (ρ)	.0, .2, .5, .8

0.0, 0.05, 0.1, or 0.2 to represent a no, small, medium, or large amount of within-study heterogeneity, on the assumption that within-study heterogeneity will typically be smaller than between-study heterogeneity. Lastly, we let values of $\rho = 0, .2, .5, .8$ represent no, small, moderate, and large levels of correlation between effect size estimates from the same study, covering a wide range of plausible values. In conditions where $\rho = 0$, the MLMA model is correctly specified (as is the CHE), whereas in conditions where $\rho > 0$, the MLMA model is increasingly mis-specified.

Performance Assessment

The main performance criterion of interest was the rejection rate of each test, which we estimated by calculating the proportion of replications in which a test returned a p -value less than a specific α -level. For conditions where $\mu = 0$, the rejection rate corresponds to Type I error. For conditions with $\mu > 0$, the rejection rate is the power of the test. We calculated rejection rates of each test for $\alpha = .01, .05, .10$, although we mainly concentrate on the conventional level of $\alpha = .05$. For each simulation condition, we generated 4000 replications, which led to Monte Carlo standard errors for simulated rejection rates of no more than .008.

Assumptions about Study Characteristics

For each condition in the design and each hypothesis test, we compared the simulated rejection rates to the approximate power rates under three different sets of assumptions regarding

the distribution of study characteristics. First, we estimated power based on the empirical distribution of sampling variances and number of effects per study in the DBFJ17 data. These features varied across studies, leading a degree of imbalance similar to what researchers might encounter in other systematic reviews of educational intervention studies. This approach also matched the actual distribution of study characteristics used in the data-generating process. Second, we estimated power under stylized distributional assumptions, as described in the example presented in the previous section. These assumptions matched the means of the study characteristics in the DBFJ17 data, as well as the variance of the σ_j^2 distribution, but differed in the shape of the distributions. This approach allowed us to assess the accuracy of the approximations if one's assumptions were similar to, but not exactly aligned with, the true distribution. Third, we estimated power assuming perfect balance, where all included studies had $\sigma_j^2 = .068$ and $k_j = 4.1$ (the average values in the DBFJ17 data).

Replication Materials

R code for replicating the simulations and numerical results from all simulation conditions are available on the Open Science Framework at <https://osf.io/yhkq4/>.

Results

We describe results of the simulation study pertaining to each of the three aims.

Finding 1a: Power approximations are accurate when based on empirical study characteristics

Our first aim was to validate the proposed power approximations for meta-analysis models of dependent effect sizes. Figures 2A and 2B plots the *power difference* between the approximated and the simulated (true) power on the vertical axis versus the approximate power estimate on the horizontal axis. The figures is faceted by the type of working model and the methods of sampling k_j and σ^2 . Points above 0 indicate conditions where the approximation over-states the true power

level. Different colors indicate varying numbers of studies and different shapes correspond to different amounts of between-study heterogeneity. We represent these features because we found that J and τ^2 are the two largest and most consistent sources of variation in accuracy of the approximation (see Supplementary Section S5). The approximation formulas for the CHE, CE, and MLMA working models are quite accurate when the approximations are based on sampling from the empirical data. The approximations nearly perfectly reproduce the simulated power levels for the robust tests (CHE-RVE, CE-RVE, and MLMA-RVE) when sampling k_j and N_j from the DBFJ17 data (top row in Figures 2A and 2B). The approximations only slightly under-estimate power in a few cases when $\tau = 0.05$. For the CHE and MLMA model-based tests with Satterthwaite d.f., the power approximations were sometimes too optimistic (exceeding the simulated power level) when based on 10 studies and $\tau = 0.05$, whereas using the model-based tests with $J - 1$ degrees of freedom sometimes led to overly cautious power levels when based on 10 studies (see plots 1, 5, and 9 in Figure 2A and Plots 1, 4, and 7 in Figure 2B). This indicates that the approximations for the RVE-based tests are more accurate than those for the model-based tests when the analyst has empirical data available.

Finding 1b: Power approximations based on stylized distributions rarely overestimate the true power by more than five percentage points.

As depicted in the second row of Figures 2A and 2B, power approximations based on stylized distributions of k_j and N_j only slightly over-estimate true power, typically to within five percentage points. For the power approximations for CHE-RVE and MLMA-RVE models, the approximated and true power rarely diverge by more than five percentage points, and discrepancies only occur when $\tau = 0.05$. However, models with Satterthwaite d.f. can yield extreme overestimation (i.e., more than 20 percentage points, see Plot 6 in Figure 2A and Plot 5

POWER APPROXIMATIONS FOR META-ANALYSIS OF DEPENDENT EFFECT SIZES

in Figure 2B) when based on few studies and a low amount of between-study heterogeneity. The approximated power for the CE-RVE model mainly exceeds five percentage points discrepancy when based on a low amount of heterogeneity. Across all model approximations based on stylized distributions of k_j and N_j , a low amount of τ is the main reason for overestimation of true power.

FIGURE 2B. Power difference between approximated and true (simulated) power versus approximated power for the MLMMA working models, across different methods of sampling k_j and N_j . Solid lines indicate no discrepancy between approximated and simulated power. Dashed lines indicate five percentage points over- or underestimation of true power

Finding 1c: Power approximations assuming completely balanced study characteristics tend to substantially over-state true power.

The bottom rows of Figures 2A and 2B further indicate that the power approximations generally over-state the true power of all models by 5 to 20 percentage points when the approximations are based on the assumption of complete balance. This pattern is especially pronounced for the CE-RVE and the CHE and MLMA models with Satterthwaite d.f. In cases with few studies, these approximations sometimes overestimate true power by more than 30 percentage points (see Plot 10 in Figure 2A and Plot 8 in Figure 2B). These results suggest that power approximations premised upon the assumption of complete balanced study characteristics generally perform poorly across all models.

Finding 1d: Simple power approximations do not accurately predict true power levels

Researchers might also wonder about how the original, simpler power approximations for univariate meta-analysis (Hedges & Pigott, 2001) perform for anticipating power in meta-analyses involving dependent effect sizes. Figure S2 in the supplementary material illustrates the performance of the univariate approximation formula to predict the true power both for the RE model estimated using synthetic effect sizes and the more complex models using RVE. From these supplementary investigations, the original power approximation performs inadequately as a means for estimating the true power of all models handling dependency, including the RE model. Across conditions, the univariate approximations often over- or under-estimate the true simulated power by 20 percentage points or more. The patterns are most extreme when the original approximations are based on a low amount of heterogeneity and become smaller as a function of τ . Thus, we do not recommend using the original univariate power approximations for estimating power of the overall average effect size in the presence of dependent effect sizes.

Finding 2: Robust variance estimation guards against Type I error with all working models

Figure 3 and Supplementary Figure S3 display the distribution of simulated Type I error rates for the eight different tests under consideration. Tests using $J - 1$ d.f. yield Type I error rates that were substantially above nominal levels. This pattern is especially evident when the number of studies is small (10) to moderate (40). Even with $J = 60$ studies, the aggregated model fails to control the nominal Type I error when $\rho = 0$ or $\rho = .2$.

Tests based on model-based variance estimation and Satterthwaite d.f. (CHE-Model+Satt and MLMA-Model+Satt) appear conservative, sometimes yielding Type I error rates substantially below nominal when the number of studies is $J = 20$ or fewer. Under these scenarios, they also cover the widest range of rejection rates across the different simulation conditions (based on the interquartile range of the boxplots). Concretely, this indicates that the Type I error rate of this set of models fluctuates substantially when the number of studies is small. Although conservative, these models should be prioritized relative to the models with Type I error rates exceeding the nominal level.

Ideally, a hypothesis testing procedure should not only control the Type I error rate but should also come as close to the nominal level as possible. In this regard, it can be seen that all tests based on RVE with small-sample adjusted standard errors and Satterthwaite d.f. are close to or equal to the nominal rejection rate. Using small sample adjustments is particularly relevant for multilevel meta-analysis models because these methods usually use model-based tests with large-sample approximations, which can be inaccurate when the total number of studies is small. Indeed, results in Figure 3 demonstrate that the conventional MLMA test with $J - 1$ degrees of freedom requires a large number of studies ($J = 60$) to attain near-nominal Type I error—even when $\rho = 0$ so that the MLMA is correctly specified.

FIGURE 3. Type I error rate for $\alpha = .05$ of all estimated models by number of studies, J , and between outcomes within-study correlation, ρ . Solid lines indicate the .05 α -level and dashed lines indicate bounds for simulation error.

Finding 3: Only small power differences between RVE models

Figure 4 displays the power of the CE-RVE and MLMA-RVE models relative to the power of CHE-RVE, across varying numbers of studies, sizes of the within-study, between-outcome correlations, and average effect sizes; Supplementary Figure S4 depicts the same relationships by the amount of within-study heterogeneity. Points below 1 indicate a loss of power relative to the CHE-RVE model. Under the conditions examined, one would expect that tests based on the CHE will achieve the highest possible power because they use a working model that is consistent with the true data-generating process. In contrast, the CE-RVE tests are based on a mis-specified working model and also use weights that are not fully efficient. In light of this, it is interesting that the CE-RVE tests do not lose substantial power relative to CHE-RVE. Under most conditions, the relative power of CE-RVE tests was 80% or higher, and often closer to 95%. Similarly, the MLMA model is only correctly specified when $\rho = 0$. When correctly specified, it is equivalent to the CHE working model and thus the MLMA-RVE test has power identical to that of CHE-RVE. For $\rho > 0$, the MLMA working model is mis-specified yet it still retains most of the power of the CHE-RVE test, with relative power of 90% or more. These results suggest that all three working models for handling dependent effects are reliable for estimating the overall average effect size (even when the total number of studies is small) as long as these are guarded for misspecifications via RVE.

FIGURE 4. Relative power between the simulated power for the CHE-RVE model and the CE-RVE and MLMA-RVE models across the different values of between outcomes within-study correlation ρ , total number of studies, J , and average effect sizes, μ , respectively. Values less than 1 indicate loss of power relative to CHE-RVE.

Discussion and Conclusion

Methods for handling dependent effect sizes have grown increasingly complex, which has created challenges for conducting power calculations for meta-analysis of dependent effects. In this study, we developed new approximation formulas for several Wald-type tests based on the CHE, CE, and MLMA models, and we evaluated the performance of the approximations via Monte Carlo simulations assuming a correlated-and-hierarchical effects data-generating process. The new approximation formulas can closely match the true model power when the relevant primary study characteristics, including sampling variance, σ_j^2 (or sample size N_j), and the number of effect sizes per study, k_j , are sampled from pilot data with similar characteristics to the data used for the eventual meta-analysis.

We acknowledge that it will not always be possible for systematic reviewers to have access to reliable or relevant pilot data that can inform their power analysis. Therefore, we also tested the performance of power approximations when these are based either on a stylized distribution of k_j and σ_j^2 or on completely balanced study characteristics (i.e. all studies have equal sampling variance and the same number of effect sizes). We found that most of the power approximations over-estimate the true power to some extent, but approximations based on stylized distributions rarely overestimate the true power by more than five percentage points. This only happens for power approximations for models not using RVE when the between-study heterogeneity is low. Approximations based on the assumption of complete balance perform worse, yielding overly optimistic power estimates, compared to approximations that allowed imbalance across studies. Thus, we do not recommend researchers assume complete balance in practice. When no pilot data are available, we recommend reviewers use the approximations for working models using RVE based on stylized distributions of k_j and σ_j^2 (or N_j) because these approximations rarely

overestimate the true power by more than 5 percentage points. For the RVE models, we tentatively suggest that reviewers should anticipate a systematic power loss of 5 percentage points when conducting power analysis based on stylized distributions of study characteristics. We generally suggest avoiding the approximation formulas for model-based tests since these are rather sensitive to the assumed between-study heterogeneity and the number of studies, generally performing insufficiently when based on small values of J and τ .

From our simulation study, we also investigated Type I error rates and, for the models that adequately controlled the nominal Type I error rate, relative power. The simulation results provide further evidence that meta-analysts should routinely guard against model misspecification by using robust variance estimation. Consistent with findings from Fernández-Castilla et al. (2020), we find that combining the MLMA model with RVE adequately controls Type I error. If using model-based inference, meta-analysts should use the more conservative test based on Satterthwaite degrees of freedom, particularly when the total number of studies is small or moderate (i.e., 10-40). Our results support the previous recommendations from Tipton (2015) to routinely use both the CR2 small-sample adjustment and the Satterthwaite d.f. Compared to model-based variance approaches with $J - 1$ degrees of freedom, tests based on robust variance estimation better control the rejection rate and yield more adequate power estimates.

For tests of overall average effect sizes, using robust variance estimation has little cost in terms of power. In addition, the power differences between the CE, CHE, and MLMA models are minor when applying RVE. That said, and in line with Pustejovsky and Tipton (2021), we recommend using working models, such as the CHE, that capture the main features of the data structures that meta-analysts are likely to encounter in practice. Predicated upon our results, we generally recommend using the new power approximation for the models using RVE because

this approximation seems to perform most reliably across all techniques for obtaining k_j and σ_j^2 (or N_j).

Lastly, we find that the original univariate power approximation (Hedges & Pigott, 2001) performs insufficiently for purposes of estimating power of both the univariate model using synthetic effect sizes or the more complex family of models using RVE. Therefore, we recommend no longer using the simpler, univariate formulas to approximate power for models handling dependent effect sizes. Future research is needed to investigate how these univariate power formulas perform when the true data-generation process follows a univariate random effects structure.

The work in this article does have some clear limitations. Although we find that the approximations perform well when based on pilot data that approximates the real distribution of study features, it may be that available pilot data are not representative of the target population of studies (for instance, by imposing too much or too little imbalance in the data), which could distort the accuracy of the proposed approximations. Furthermore, our simulation results are limited by the selected data-generating model and parameters. The most prominent limitation here is that we have concentrated on the situation in which the CHE working model is consistent with the true data-generation process—a best-case scenario that implies that the CHE working model will have higher power than the CE or MLMA models. In future work, it might be useful to elaborate upon the power approximations by allowing for some degree of mis-specification of the working model, such as by assuming a correlation of $\rho = .6$ but allowing the true data-generating process to have a correlation of $.4 < \rho < .8$.

This study is limited in scope in that the simulations focused on the common case of standardized mean differences effect sizes. The power formulas can readily be applied to some

POWER APPROXIMATIONS FOR META-ANALYSIS OF DEPENDENT EFFECT SIZES

other effect size metrics such as Fisher's z -transformed correlation coefficient, but application to metrics such as log odds ratios or risk ratios requires making further assumptions. Future research needs to develop guidance about how to implement the power calculations under a range of scenarios encountered by working meta-analysts.

In this article, we have only focused on power of tests for the overall average effect size, which clearly limits the application of the proposed methods. For testing the overall average effect size, we found that the choice of working model (CHE or CE or MLMA) leads to only minor differences in power. However, this finding may not generalize to more complex models involving moderator variables. Rather, Pustejovsky and Tipton (2021) found that using CHE can lead to substantially more precise estimates than using CE for meta-regression models with predictor variables that vary within study. Thus, the choice of working model may be more consequential for models that involve potential moderator variables.

Developing power calculations for more intricate models such as meta-regressions with one or multiple predictors requires making assumptions about the distribution of covariates across studies and effect sizes, which may be challenging to specify *a priori*. However, if reviewers have access to detailed and relevant pilot data, power analysis for meta-regression can be conducted via Monte Carlo simulation. Although not trivial, future research could focus on making power simulation for meta-regression models more accessible to the applied meta-analyst.

We intend the results of this research to increase understanding of the performance of a range of meta-analysis models currently used in practice. The power approximations provided can guide researchers in obtaining *a priori* power estimates required for grant proposals and can inform choices about working models in a meta-analysis. Consistent with other methodologists'

recommendations regarding retrospective power analysis (Lakens, 2022; Zumbo & Hubley, 1998), we discourage use of these approximations to assess the *post hoc* power of a meta-analysis based on an observed effect size.

References

- Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, 82(4), 436–476.
- Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, 87(2), 243–282.
- Fernández-Castilla, B., Aloe, A. M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behavior Research Methods*, 53(2), 702–717.
<https://doi.org/10.3758/s13428-020-01459-4>
- Fisher, Z., & Tipton, E. (2015). robumeta: An R-package for robust variance estimation in meta-analysis. ArXiv Preprint ArXiv:1503.02220.
- Giesbrecht, F. G., & Burns, J. C. (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *Biometrics*, 477–486.
- Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, 20(12), 1771–1782.
<https://doi.org/10.1002/sim.791>
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. London: Academic Press.

POWER APPROXIMATIONS FOR META-ANALYSIS OF DEPENDENT EFFECT SIZES

- Hedges, L. V., & Pigott, T. D. (2001). The Power of Statistical Tests in Meta-Analysis. *Psychological Methods*, 6(3), 203.
- Hedges, L. V., & Pigott, T. D. (2004). The Power of Statistical Tests for Moderators in Meta-Analysis. *Psychological Methods*, 9(4), 426.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65.
<https://doi.org/10.1002/jrsm.5>
- Jackson, D., & Turner, R. (2017). Power Analysis for Random-Effects Meta-Analysis. *Research Synthesis Methods*, 8(3), 290–302.
- Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, 53(7), 2583–2595.
- Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, 8(1), 33267.
<https://doi.org/10.1525/collabra.33267>
- Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, 16(2), 358–376.
- Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49(4), 1494–1502. <https://doi.org/10.3758/s13428-016-0809-y>
- Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 0013189X20912798. <https://doi.org/10.3102/0013189X20912798>
- Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison

POWER APPROXIMATIONS FOR META-ANALYSIS OF DEPENDENT EFFECT SIZES

- between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, 20(6), 559–572.
<https://doi.org/10.1080/13645579.2016.1252189>
- Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, 146(10), 922.
- Pigott, T. (2012). *Advances in Meta-Analysis*. Springer Science & Business Media.
- Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, 15(4), 1026-1041
- Pustejovsky, J. E. (2020). ClubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.5.0 (0.5.2). cran.r-project.org.
- Pustejovsky, J. E., & Tipton, E. (2021). Meta-analysis with Robust Variance Estimation: Expanding the range of working models. *Prevention Science*, 1–14.
- Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling Multivariate Effect Sizes. *Psychological Bulletin*, 103(1), 111–120. <https://doi.org/10.1037/0033-2909.103.1.111>
- Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6(5), 309–316.
<https://doi.org/10.1007/BF02288586>
- Tipton, E. (2015). Small Sample Adjustments for Robust Variance Estimation With Meta-Regression. *Psychological Methods*, 20(3), 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(6), 604–634. <https://doi.org/10.3102/1076998615606099>

POWER APPROXIMATIONS FOR META-ANALYSIS OF DEPENDENT EFFECT SIZES

- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, 10(2), 180–194.
<https://doi.org/10.1002/jrsm.1339>
- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215–247. <https://doi.org/10.3102/1076998609346961>
- Van den Noortgate, W., López-López, J., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, 45(2), 576–594. <https://doi.org/10.3758/s13428-012-0261-6>
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2014). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, 47(4), 1274–1294. <https://doi.org/10.3758/s13428-014-0527-2>
- Venables, W. N. & Ripley B. D. (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York.
- Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1), 1. <https://doi.org/10.2307/1912526>
- Zumbo, B. D., & Hubley, A. M. (1998). A Note on Misconceptions Concerning Prospective and Retrospective Power. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(2), 385–388. <https://doi.org/10.1111/1467-9884.00139>