Received: 12 February 2024 | Revised: 24 June 2024 | Accepted: 5 August 2024

DOI: 10.1002/jrsm.1752

TUTORIAL

Research Synthesis Methods WILEY

Conducting power analysis for meta-analysis with dependent effect sizes: Common guidelines and an introduction to the POMADE R package

Mikkel Helding Vembye[1] | James Eric Pustejovsky[2] | Therese Deocampo Pigott[3]

[1]Department of Quantitative Methods, The Danish Center for Social Science Research, VIVE, Aarhus, Denmark

[2]University of Wisconsin-Madison, Madison, Wisconsin, USA

[3]Georgia State University, Atlanta, Georgia, USA

Correspondence
Mikkel Helding Vembye, Department of Quantitative Methods, The Danish Center for Social Science Research, VIVE, Copenhagen, Denmark.
Email: mihv@vive.dk

Abstract

Sample size and statistical power are important factors to consider when planning a research synthesis. Power analysis methods have been developed for fixed effect or random effects models, but until recently these methods were limited to simple data structures with a single, independent effect per study. Recent work has provided power approximation formulas for meta-analyses involving studies with multiple, dependent effect size estimates, which are common in syntheses of social science research. Prior work focused on developing and validating the approximations but did not address the practice challenges encountered in applying them for purposes of planning a synthesis involving dependent effect sizes. We aim to facilitate the application of these recent developments by providing practical guidance on how to conduct power analysis for planning a meta-analysis of dependent effect sizes and by introducing a new R package, *POMADE*, designed for this purpose. We present a comprehensive overview of resources for finding information about the study design features and model parameters needed to conduct power analysis, along with detailed worked examples using the POMADE package. For presenting power analysis findings, we emphasize graphical tools that can depict power under a range of plausible assumptions and introduce a novel plot, the traffic light power plot, for conveying the degree of certainty in one's assumptions.

KEYWORDS

dependent effect sizes, meta-analysis, power, robust variance estimation, traffic light power plot

Highlights

What is already known

- Power of meta-analysis models for handling statistically dependent effect sizes can be approximated but is challenging given the lack of common guidelines for estimating key power parameters.
- Power approximations for meta-analysis of dependent effect sizes perform reliably when based either on empirical or stylized assumptions about key design features.
- Power approximations generally overestimate the true power by more than 10% when assuming balanced data (i.e., equal numbers of effect sizes nested within studies).
- Power approximations involving robust variance estimation (RVE) are more accurate than other power approximation methods.

What is new

- General guidelines for the conduct of power analysis involving statistically dependent effect sizes using the correlated and hierarchical effect model (CHE) with RVE.
- Functions to find the minimum detectable effect size and the number of studies needed to find a certain effect with a prespecified amount of power in a meta-analysis context
- The *POMADE* R package for conducting **po**wer analysis of **m**eta-**a**nalysis of **d**ependent **e**ffects.
- Graphical tools for presenting a priori power analyses across a range of possible assumptions.

Potential impact for *Research Synthesis Methods* readers

- Makes power analysis for meta-analysis of dependent effect sizes easily accessible and expands its use in the context of systematic reviews involving meta-analysis.
- Expansion of open science and open data practices.

1 | INTRODUCTION

In meta-analyses on topics in the social and behavioral sciences, it is very common to include findings from primary studies that report multiple effect sizes, producing various types of dependency structures in the meta-analysis data. Often studies report multiple eligible results for the same sample of participants (e.g., across different time points or types of measurements), creating correlated sample errors, also known as a *correlated effects dependency structure*. Studies also often report multiple results across non-overlapping samples (e.g., for primary and secondary students, for each of several sites in a multi-site trial, or for several experiments conducted by the same group of researchers), creating a multi-level or *hierarchical effects dependency structure*. Although the results are drawn from non-overlapping samples, the fact the researchers apply the same estimation techniques, implementation strategies, measurement, etc., creates dependency among effect size parameters from different samples within the same study. Often, both dependency structures appear simultaneously in social science syntheses.

When faced with dependent effect sizes, Hedges and Olkin[1] and Raudenbush, Becker, and Kalaian[2] suggested the use of multivariate effect sizes models. However, these models were rarely used in practice[3] because they require knowledge of the true dependency structures among effect sizes (i.e., the full correlation matrix), and such information is rarely reported or retrievable from primary studies. A decade ago, methods[4–6] based on robust variance estimation (RVE) or multi-level modeling (MLMA) were developed to handle dependency among effect sizes when the true dependency structure is partly or fully unknown. A common form of the RVE, the correlated effects (CE) model, assumes that effect sizes are dependent because they are measured on the same samples. The multi-level meta-analysis (MLMA) model assumes a hierarchical structure to effect sizes, that is, that they are nested within studies but are measured on independent samples. Initial implementations of these models required a choice; either the researcher assumed that effect size estimates were all correlated or that effect size estimates were independent and nested within studies. However, when the meta-analysis data substantially diverge from the assumptions of either the CE model or MLMA model, the precision of the models is impacted.[7]

More recently, Pustejovsky and Tipton propose a new model,[7] known as the correlated-hierarchical effects (CHE) model, in which multi-level modeling and RVE are combined while simultaneously accounting for the correlated and hierarchical effects dependency structures (therefore also defined as the CHE-RVE model). CHE-RVE models more closely approximate the true dependency structures found in meta-analysis applications in the social sciences and increase the statistical power to find small effects when multiple dependency structures exist.

1.1 | A priori power considerations

In primary research, researchers often estimate a priori power for key statistical tests when planning for an appropriate sample size. Planning a systematic review is more challenging because researchers rarely know the number of eligible studies and other study characteristics prior to a literature search. Nevertheless, external funders often require systematic review authors to provide evidence that conducting a systematic review will be productive. The increased use of meta-analysis models for dependent effect sizes has also raised questions about the data requirements for applying these more complex models. Understanding the a priori power of meta-analysis models for dependent effect sizes provides insights about a planned systematic review, including its potential to support the use of models that better approximate the multilevel and correlated nature of effect size data.

Until recently, available power approximation techniques for meta-analysis[8–11] were restricted by the assumption of independence among effect sizes—that is, that studies provide only one effect size estimate each. Power approximations developed for models with independent effect sizes perform inadequately when used for approximating power for meta-analysis models with dependent effect sizes.[12] Furthermore, the assumption of independent effect sizes is rarely fulfilled in the social and behavioral sciences, where studies commonly report multiple effect sizes.

In a recent paper, Vembye, Pustejovsky, and Pigott[12] developed power approximation formulas for meta-analysis of dependent effect sizes across the CE, MLMA, and CHE models. Concurrently, Zhang and Konstantopoulos[13] developed power approximation formulas for MLMA models. However, these recent works focused only on the technical development of power formulas and evaluation of the accuracy of the proposed approximations, without providing guidance on how to apply the developed methods in practice. Thus, we believe there remains a need to consider the practical challenges that can be encountered by reviewers in obtaining the relevant quantities and developing reasonable assumptions as required to actually implement power calculations for meta-analyses of dependent effect sizes.

1.2 | Aims

In this article, we provide guidelines for conducting power calculations for meta-analyses of dependent effect sizes and introduce the *POMADE* R package[14] for this purpose. The paper has four major aims. First, we review recently developed power approximations and introduce novel extensions for approximating the number of studies required to detect a given effect size and for approximating the minimum detectable effect size given pre-specified levels of statistical significance and power. Second, we give an overview of resources where reviewers can find information regarding sample characteristics and parameters needed to actually conduct power calculations. Third, we provide worked examples of how to conduct power analyses in meta-analysis with dependent effects. Finally, we introduce new graphical tools (including the *traffic light power plot*) for presenting power analyses across a range of plausible scenarios of design and sample features as well as model parameters. Our overarching goal is to make power approximation formulas for meta-analysis of dependent effect sizes accessible for researchers planning to conduct meta-analyses.

In our exposition and examples, we focus on power analysis based on the CHE-RVE model for several reasons. First, the CHE-RVE model is more comprehensive than currently available alternatives, in that it allows both for correlated sampling errors (as in the CE model) and for heterogeneity both within and between studies (as in the MLMA model). These other models can be viewed as special cases of the CHE. For instance, the MLMA model guarding against misspecification via RVE[15] is a special case of the CHE-RVE model, assuming that the sample correlation among effect sizes is $\rho = 0$. However, it is important to note that in cases when either no within-study heterogeneity or no correlation between effect sizes are expected, the CE or the MLMA[15] models are the preferable models to be used. Power approximation functions for all of the common models for handling dependent effect sizes are available in the *POMADE* R package, and examples of how to use these methods will be incorporated in the accompanying vignette to the package. Power approximation formulas were also developed for the CHE and MLMA models, not guarding against misspecification via RVE,[16] but we do not recommend using these models because they do not control the nominal Type-I error rate when the number of studies is limited (i.e., less than 40 studies).

The remainder of the paper proceeds as follows. In Section 2, we review the statistical foundation of power, sample size (i.e., the number of studies needed), and minimum detectable effect approximation in meta-analysis of dependent effects. In Section 3, we present various strategies for how investigators can make qualified assumptions about the sample characteristics and parameters needed to conduct meta-analytical power analysis. We go through each factor one by one. In Section 4, we provide empirical examples of how to conduct and visualize various power, sample size, and minimum detectable effects analyses. In Sections 5 and 6, we reflect on the utility of power analysis in meta-analysis and on what it requires from the research community to make meta-analytical power analyses common practice.

2 | A PRIORI POWER APPROXIMATION FOR THE CHE-RVE MODEL

To illustrate the conduct of power analysis for meta-analysis of dependent effect sizes, we first describe the power approximation for a hypothesis test for an overall average effect based on standardized mean differences[17] in which the assumed data-generating process follows that of the correlated-and-hierarchical effects (CHE) model as described by Pustejovsky and Tipton.[7]

The CHE model can be applied for meta-analyzing a set of studies where some or all included studies contribute multiple, statistically dependent effect size estimates. Suppose that we have a collection of J studies to be included in a meta-analysis, where study j includes $k_j \geq 1$ effect size estimates, for $j = 1,...,J$. Let T_{ij} denote effect size estimate i from study j, with corresponding standard error σ_{ij}, for $i = 1,...,k_j$ and $j = 1,...,J$. For simplicity, we assume that the sampling variances are constant within each study, so $\sigma^2_{1j} = \sigma^2_{2j} = \cdots = \sigma^2_{k_j j} = \sigma^2_j$.

As usual in meta-analysis, the CHE model makes the assumptions 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 as

$$T_{ij} = \theta_{ij} + e_{ij}, \tag{1}$$

where $e_{ij} = T_{ij} - \theta_{ij}$ is the sampling error, which has expectation zero and variance $\text{Var}\left(e_{ij}\right) = \sigma^2_j$. Effect size estimates from different studies are assumed to be uncorrelated, so $\text{cor}\left(e_{hj}, e_{il}\right) = 0$ when $j \neq l$, but effect size estimates from the same study may be correlated. Because information about the sampling correlation between effect sizes is often not available from included studies, analysts will typically need to make a more-or-less arbitrary assumption about the degree of dependence. With the CHE model, the correlations between sampling errors within a given study are all assumed to be equal to a known constant, $\text{cor}\left(e_{hj}, e_{ij}\right) = \rho$, specified by the analyst. This feature of the CHE model captures the "correlated effects" structure of the data. In Section 3.7, we discuss in more detail how one can draw assumptions of the value of ρ.

The other component of the CHE model captures the "hierarchical effects" structure. Here, it is assumed that effect size parameters represent a sample from an underlying population of effects that has a hierarchical structure, according to

$$\theta_{ij} = \mu + u_j + v_{ij}, \tag{2}$$

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 CHE model are 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 power approximations for tests of the null hypothesis $H_0 : \mu = d$ versus a two-sided alternative, with specified Type-I error level α.

2.1 | Estimation of CHE

Estimation of the overall average effect size μ entails first estimating the variance components, τ^2 and ω^2, 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 an assumed 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

$$w_j = \frac{k_j}{k_j \hat{\tau}^2 + k_j \rho \sigma_j^2 + \hat{\omega}^2 + (1-\rho)\sigma_j^2} \tag{3}$$

The overall average effect size is estimated as

$$\hat{\mu} = \frac{1}{W} \sum\nolimits_{j=1}^{J} w_j \overline{T}_j, \tag{4}$$

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

$$Var(\hat{\mu}) \approx \frac{1}{W}. \tag{5}$$

Hypothesis tests or confidence intervals based on Equation (5) will perform properly if the assumptions of the CHE model are good approximations to the true data-generating process.

In light of the lack of information about the sampling correlations between effect size estimates, meta-analysts will often prefer to use tests based on RVE methods, which maintain close-to-correct Type I error calibration 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

$$V^R = \frac{1}{\mathrm{W}^2} \sum\nolimits_{i=1}^{J} \frac{w_j^2 \left(\overline{T}_j - \hat{\mu}\right)^2}{\left(1 - \frac{w_j}{W}\right)}. \tag{6}$$

When the working model is correctly specified and variance components are known, then V^R is an exactly unbiased estimator of $Var(\hat{\mu})$. However, even if the assumptions of the working model do not hold and if the variance components must be estimated, V^R remains close to unbiased.

A robust test of the hypothesis $H_0 : \mu = d$ is based on the robust Wald test statistic

$$t^R = \frac{\hat{\mu} - d}{\sqrt{V^R}}. \tag{7}$$

Tipton[18] 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. Specifically, the Satterthwaite degrees of freedom are calculated as

$$\xi = \left[\sum\nolimits_{j=1}^{J} \frac{w_j^2}{\left(W - w_j\right)^2} - \frac{2}{W} \sum\nolimits_{j=1}^{J} \frac{w_j^3}{\left(W - w_j\right)^2} + \frac{1}{W^2} \left(\sum\nolimits_{j=1}^{J} \frac{w_j^2}{W - w_j} \right)^2 \right]^{-1}. \tag{8}$$

The robust Wald test rejects the null hypothesis if $|t^R| > c_{\alpha/2,\xi}$, where $c_{\alpha/2,\xi}$ is the $\alpha/2$ critical value from a Student t distribution with ξ degrees of freedom.

2.2 | Power approximation

Vembye, Pustejovsky, and Pigott[12] proposed to approximate the power of the Wald robust test using a non-central Student-t distribution, with non-centrality parameter given by

$$\lambda = \sqrt{W}(\mu - d) \tag{9}$$

and degrees of freedom as given in Equation (8). The power of the robust Wald test against a two-sided alternative is then approximated as

$$F_t\left(-c_{\alpha/2,\xi} | \xi, \lambda\right) + 1 - F_t\left(c_{\alpha/2,\xi} | \xi, \lambda\right), \tag{10}$$

where $F_t(x|\xi,\lambda)$ is the cumulative distribution function of a non-central Student-t distribution and $c_{\alpha,\xi}$ is the upper α-level critical value for the central Student-t distribution with ξ degrees of freedom, so $F_t\left(c_{\alpha/2,\xi}|\xi,0\right) = 1 - \alpha/2$. This approximation assumes that the CHE model is correctly specified.

The power of the test based on CHE-RVE depends on several parameters: the true average effect size μ, the between-study variance τ^2, the within-study variance ω^2, and the assumed correlation between sampling errors ρ. In the next section, we discuss strategies for making assumptions regarding these parameters for purposes of prospective power analysis and sample size planning.

The power of the test also depends on several features of the set of studies to be included in the meta-analysis: the total number of studies (J), the magnitude of their sampling variances $\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2\right)$, and the number of effect sizes contributed by each included study $(k_1, k_2, \ldots, k_J)$. Prior to completing a systematic review, the sampling variances and number of effect sizes per study will not be known precisely. For prospective power analysis, Vembye, Pustejovsky, and Pigott[12] proposed treating these quantities as random variables that follow some distribution. The distribution might be based on empirical data from an initial scoping review or a previous meta-analysis on a similar topic, or it might be based on more stylized assumptions involving a parametric distribution. With this approach, the power of the test is calculated by taking the expected value of Equation (10) over the distribution of sampling variances and effect sizes per study. Practically, the expectation is approximated by drawing a random sample of J sets of study characteristics $\left(\sigma_j^2, k_j\right)$ from specified distributions, calculating λ and ξ based on the sample of study characteristics, and then calculating power with Equation (10). This process is repeated several times, with the expected power level calculated as the overall average power across repeated samples. In the *POMADE* package presented below, this process is by default repeated 100 times.

2.3 | Sample size planning and minimum detectable effects

The proposed methods provide a means of approximating the power of a test of the null hypothesis $H_0: \mu = d$ versus a two-sided alternative, given assumptions about the true overall average effect size, for a meta-analysis with a specified number of studies. Researchers in the planning stage of a meta-analysis might use the methods directly to answer the question "What is the power of this test?" However, they might also find it useful to frame the question somewhat differently. Two alternative framings are common: one that centers on a target sample size and one that centers on minimum meaningful effect sizes.

One alternative framing is to pose the question, "*How big a sample is needed to achieve a specified power level?*" To answer this question, one would first specify a desired power level P, such as the conventional level of $P = 0.8$, a minimum effect size of interest (μ), and a distribution of primary study sample sizes and effect sizes per study. Given these quantities, the number of included studies J affects power through the total weight W, which in turn determines the non-centrality parameter λ, and through the degrees of freedom ξ. Therefore, the target sample size is the smallest value of J that satisfies the equation

$$P = E\left[F_t\left(-c_{\alpha/2,\xi} | \xi, \sqrt{W}(\mu - d)\right) + 1 - F_t\left(c_{\alpha/2,\xi} | \xi, \sqrt{W}(\mu - d)\right)\right], \tag{11}$$

where the expectation is taken over the distribution of primary study sample sizes and effect sizes per study. The solution can be found through a direct grid search over a range of possible values for J. This feature is integrated into the find_J_* functions in the *POMADE* package.

Another alternative framing is to pose the question, "*How small an average effect size can be detected with a given sample size with a specified power level?*" To answer this question, we would again need to specify a desired power level P and a distribution of primary study sample sizes (or variance estimates) and effect sizes per study. We would also need to specify an anticipated sample size, J. Given these assumptions, we can find the average effect size μ that satisfies Equation (11). Just as with the previous question, the solution can be found through a direct grid search over a range of possible values for μ, and is integrated into the `MDES_*` functions in the *POMADE* package.

2.4 | Multi-level meta-analysis as a special case

MLMA models were originally developed for meta-analytic databases with a hierarchical dependence structure, where included studies report results based on multiple samples or experiments, but there is only one effect size estimate per sample.[19] A corresponding situation arises in a meta-analysis where studies each report only one effect size estimate, but studies can be grouped based on the lab or investigator who conducted them. In both situations, effect size estimates do not have correlated sampling errors, but dependence arises because the true effect size parameters may be correlated due to use of similar operational procedures. Because the CHE model allows for hierarchical dependence (through the inclusion of between-study and within-study random effects), the MLMA model can be understood as a special case of the CHE model, with zero correlation between sampling errors for effect size estimates from the same study (i.e., $\rho = 0$).

Zhang and Konstantopoulos[13] proposed power approximation formulas for MLMA models that are similar but not identical to the approximations described in this section. Their approximations are for the test of the overall average effect that uses model-based variance estimation rather than RVE. The *POMADE* package also implements power approximations for model-based variance estimation, using a t-distribution with degrees of freedom derived from a Satterthwaite approximation.[12] In contrast, Zhang and Konstantopoulos's approximation uses a normal distribution; the code implementing their approximations also assumes that the number of effect sizes per study is a fixed constant. Both of these differences imply that their approximations will tend to provide more optimistic estimates of power levels than those implemented in *POMADE*. We expect that the approximations based on the *t*-distribution with Satterthwaite degrees of freedom will be more accurate, especially when the total number of studies is limited or the number of effect sizes per study is imbalanced.

3 | SUGGESTIONS FOR HOW TO OBTAIN RELEVANT EMPIRICAL PARAMETERS AND QUANTITIES NEEDED FOR POWER APPROXIMATION

As we have highlighted in the prior sections, reviewers must make a range of assumptions to conduct reliable power analyses. This is, of course, a clear limitation of the methods. To mitigate this limitation, this section presents guidelines for how researchers could make plausible and empirically informed assumptions needed to execute reasonable power approximation. We discuss each parameter and quantity needed for the power approximation separately. In the sections below, we will use the conventional $\alpha = 0.05$ for all the presented power calculations. Researchers should, of course, change the α-level based on their research context.[20]

3.1 | Smallest effect size of practical concern, μ

The first thing reviewers will need to determine to conduct power analysis of meta-analysis is the smallest effect of practical concern, μ. The determination of the smallest effect size of practical importance exclusively hinges on the specific topic of the review literature. Although common practice in the social and behavioral sciences, we do not recommend using general effect size conventions for small, medium, and large effect sizes such as Cohen's[21] or Hattie's[22] standards. As others have argued, relying on such decontextualized standards amounts to "characterizing a child's height as small, medium, or large, not by reference to the distribution of values for children of similar age and gender, but by reference to a distribution for all vertebrate mammals."[23]

Therefore, the smallest effect size of practical importance should ideally be deduced from relevant content sources related to the given discipline(s) and topic(s) under review. Reviewers should consider a range of factors such as the cost, complexity, and scalability of the intervention. Furthermore, μ should be determined by comparing the intervention(s) to any structurally related and/or similarly resource-intensive interventions that have been reviewed in other syntheses.

In education, researchers interested in the effects of field experiments/interventions on student achievement could profitably apply Kraft's[24] empirical benchmarks for interpreting the smallest effect size of practical significance of educational interventions on standardized achievement outcomes. If reviewers are concerned with grade-specific effect sizes, they can also consult Lipsey and colleagues's[23] overview of effect sizes of annual achievement gains. In psychology, reviewers could consult Schäfer and Schwartz[25] to understand meaningful effect sizes across sub-disciplines.

3.2 | Expected number of studies, J

A major aim of conducting power analysis for meta-analysis is to gain knowledge about how many studies, J, are needed to find the smallest effect size of practical concern. The number of studies expected to be found will often be based on the reviewers' content-specific knowledge of the given review topic. However, reviewers should conduct power analyses across a range of assumptions about the expected number of studies to allow for the possibility that the literature search and author solicitation reveal further studies unknown to reviewers. If reviewers are uncertain about the anticipated number of included studies, they could consult previous syntheses and reviews on similar research topics and/or from similar disciplines.[3] In education, reviewers could consult Hattie[22] and Ahn et al.'s[26] overviews of meta-analyses across various topics. In medicine, reviewers could consult Davey et al.'s[27] overview of the typical numbers of studies within medical meta-analyses. Across education, psychology, and medicine, reviewers could look to Tipton, Pustejovsky, and Ahmadi[3] for an overview of the average number of studies included in meta-analyses in these disciplines. Another source for retrieving empirical meta-analytical data, including J, is the *metadat* R package,[28] in which a large number of datasets of previously conducted meta-analyses are stored.

3.3 | Number of effect sizes per study, k_j

Making assumptions about the number of effect sizes per study, k_j, in Equation (3) can be done in various ways. Ideally, reviewers should obtain this information from pilot data of previous reviews on related topics. In practice, however, this advice might be difficult to compile because many systematic reviews and meta-analyses fail to provide their data publicly and openly. If data are not publically available, reviewers could contact previous review authors to request access. However, this might be a complicated route since author responses are generally low.[29] If relevant data from previous systematic reviews is not available, the *metadat* R package[28] could again be used. Alternatively, reviewers could simulate k_j around the average k_j previously found in education, psychology, or medicine.[3,26] We have made this simulation function available in the *POMADE* package.

Researchers might be inclined to make the simplifying assumption that all studies in the synthesis will include the same number of effect sizes (i.e., a "balanced" design where $k_1 = k_2 = ... = k_J = k$). Except when this is true by design of the review, we recommend against using such an assumption because it rarely holds in practice and because, if the true k_j varies from study to study, then the power approximations will systematically overestimate the true power of the model.[12]

3.4 | Study sample sizes, N_j, or sampling variances, σ_j^2

To conduct reliable power approximations, reviewers must further put forward assumptions about the distribution of sampling variances, σ_j^2, in the included studies. Such information might be difficult to retrieve in practice, but we generally suggest that reviewers obtain this information either from pilot data of previously conducted reviews on similar research topics or from relevant meta-analytic datasets, such as from the *metadat* package. In medicine, reviewers could consult Davey et al.'s[27] overview of the typical study sample sizes within medical meta-analyses.

For a given effect size metric, the distribution of sampling variances can often be approximated from information about the distribution of sample sizes, N_j. For example, for the standardized mean difference effect size metric involving comparison of two groups of independent observations, the sampling variance of the effect size estimate is approximately

$$\sigma_j^2 \approx \left(\frac{4}{N_j} + \frac{\mu^2}{2(N_j - 2)}\right) \tag{12}$$

where μ denotes the anticipated overall average effect size.[17] For meta-analyses of correlation coefficients estimated from samples of independent observations, the sampling variance of the sample correlation coefficient is approximately

$$\sigma_j^2 \approx \frac{(1-\rho^2)^2}{N_j - 1} \tag{13}$$

where ρ denotes the anticipated overall average correlation.[30] In correlational meta-analysis, analysts often prefer to transform the effect size estimates into the metric of Fisher's z. For Fisher-z-transformed correlations, the sampling variance of the effect size estimate is

$$\sigma_j^2 \approx \frac{1}{N_j - 3} \tag{14}$$

to a very close approximation.[30]

Just as with k_j, we do not recommend the assumption of complete balance about N_j or σ_j^2 (i.e., assuming $N_1 = N_2 = ... = N_J = N$ or $\sigma_1^2 = \sigma_2^2 = ... = \sigma_J^2 = \sigma^2$), because it is rarely experienced in practice and, if the true N_j and σ_j^2 vary, the power approximations will overestimate the true power of the model.[12] The *POMADE* package also includes functions from which N_j can be simulated in cases where pilot data is inaccessible.

3.4.1 | Clustering

In education research as in many other settings, primary study samples often involve clusters of observations.[31] For instance, a primary study might use a cluster-randomized experimental design in which students are nested in classrooms and entire classrooms are assigned to different treatment conditions. Clustering has a major influence on the precision of effect size estimates, and needs to be taken into account to obtain accurate estimates of effect size variance.[32] Likewise, for power approximations to work properly, reviewers must account for clustering of observations in the primary study samples when calculating sampling variances of the effect size estimates.[33,34] Without accounting for clustering, the sampling variances may be much too small, leading to overly optimistic estimates of the true power of the given model.

If clustered studies are expected to be included in the review, it is pivotal that reviewers either apply *effective sample sizes*[10] (ESS) or sampling variances that account for variation from the individual and the cluster levels.[35] If reviewers have a vector of raw sample sizes, N_j, from clustered studies, these can be corrected for one level of clustering by roughly approximating the effective sample size for study j via

$$\text{ESS}_j = \frac{N_j}{\text{DE}} \tag{15}$$

where DE is the design effect of a two-stage sample given by

$$\mathrm{DE} = 1 + (n-1)\rho_{\mathrm{ICC}} \tag{16}$$

with n being the average cluster size and ρ_{ICC} the intraclass correlation coefficient (ICC) for the cluster level. Relevant compendiums of ICC in education can be found in Hedges and Hedberg,[36] in medicine from Gulliford, Ukoumunne, and Chinn[37] and Verma and Lee,[38] and in psychology from Murray and Blitstein.[39] The `effective_sample_sizes()` function from the *POMADE* package can be used to correct the raw sample size from cluster studies.

If reviewers have pilot data containing a vector of sampling variances not including cluster-level variation, these can roughly be adjusted for cluster bias by multiplying DE to each sample variance component. The `cluster_bias_adjustment()` function from the *POMADE* package can be used for this purpose. Ideally, reviewers should strive to obtain pilot data, including sampling variances estimated from multi-level models or cluster-robust standard errors or alternatively sampling variance components that have been cluster-bias corrected as in Tanner-Smith and Lipsey[40] and Dietrichson et al.[41]

3.5 | Between-study variance (study-level variance), τ^2

When making assumptions about a plausible value for the between-study variance, τ^2, reviewers could consult previous reviews of similar topics, just as with the other required assumptions. Linden and Hönekopp[42] reported a survey of heterogeneity levels observed across 150 meta-analyses in different areas of psychology, which may be useful for establishing benchmark levels of heterogeneity; however, their survey was limited in that they did not distinguish between-study versus within-study heterogeneity of effect sizes. In medicine, reviewers could consult Turner et al.[43] (c.f. Table 3 herein) in which they report typical values of the between-study heterogeneity in medical reviews across various types of outcomes and interventions.

If information from prior reviews is not available, reviewers could follow the guideline suggested by Pigott[44] in which $\tau^2 = \left(\frac{1}{3}\right)\overline{\sigma}^2$ is considered as a low degree of heterogeneity, $\tau^2 = \overline{\sigma}^2$ is considered as a moderate degree of heterogeneity, and $\tau^2 = 3\overline{\sigma}^2$ is considered as a large degree of heterogeneity, where $\overline{\sigma}^2$ is the average sample variance expected to be found in the given literature. Reviewers could consult Fraley and Vazire[45] to gain an overview of common study sample sizes in psychology journals. To make these calculations accessible to reviewers, we have made this procedure available via the `tau2_approximation()` function from the *POMADE* package. To recognize the uncertainty of the τ^2 estimation, we highly recommend that power approximations are conducted across a range of possible values of τ^2. To make more intuitive estimates of τ^2, it can be advantageous to think of the study-level heterogeneity in terms of between-study standard deviation (SD) units because these are on the same scale as the mean effect size, μ.

3.6 | Within-study variance (effect size level variance), ω

As with the τ^2 estimate, the true within-study variance, ω^2, could be obtained from result sections of previous reviews of similar research topics or estimated from relevant pilot data with dependent effect sizes. Similarly, we suggest that reviewers think of the effect-level heterogeneity in terms of within-study SD because it allows for a more intuitive interpretation of this variance component. It might also be helpful to think of ω relative to τ or vice versa. Say for example that reviewers expect one-third of the total true variance to come from within-study heterogeneity, then $\omega^2 = \tau^2 \times 0.5$. As with τ^2, we think it is good practice to conduct power analyses across a range of within-study SD estimates to acknowledge the uncertainty of one's assumptions, perhaps highlighting the most likely scenario. We elaborate more thoroughly on this procedure in Section 4.

3.7 | Assumed sample correlation, ρ

Finally, reviewers have to make assumptions about the expected sampling correlation among outcomes coming from the same study. This is indeed a tricky part of the power approximation of the CHE-RVE model. However, there are certain ways that reviewers can make reliable estimates of ρ. First, reviewers could search for literature in relevant disciplines for common sample correlations among the outcome measures relevant for the review. Second, if raw primary data containing multiple eligible outcomes measures are available to the reviewers, ρ could be estimated from this data. For example, Vembye, Weiss, and Bhat[46] used individual participant-level data from the Project STAR to estimate ρ and inform the choice of ρ in their systematic review regarding the effects of collaborative models of instruction on student achievement. Third, if reviewers have access to relevant meta-analytical pilot data containing studies reporting two outcome measures, then ρ could be roughly approximated by estimating the correlation between the pairs of effect sizes estimates from those studies that provide both types of outcomes measures.[47] In this case, it is recommended to obtain at least 10 such studies to be able to obtain a reliable estimate of ρ.[47] Independently of the used methods to obtain ρ, we suggest that reviewers conduct power analyses across a range of different assumptions about ρ to assess the impact of ρ on estimated power levels.

4 | EMPIRICAL EXAMPLE

4.1 | Replication materials

All R codes for replicating the below-presented power approximation examples are available on the Open Science Framework https://osf.io/vpnmb/. For plot generation, the *POMADE* package draws on the *ggplot2* R package.[48]

4.2 | Power example of the CHE-RVE model using relevant pilot data

We now illustrate the process of power analysis for meta-analysis of dependent effect sizes. Suppose that we are planning a meta-analysis about the effects of extending the school day on student achievement. To compute power for the overall average effect, we use pilot data from Vembye, Weiss, and Bhat's[46] (henceforth VWB23) meta-analysis on the effects of collaborative models of instruction on student achievement. We consider this study an appropriate source of pilot data because collaborative instruction interventions represent viable alternatives to increasing the length of the school day. From this systematic review and pilot data,[1] we can find all of the relevant parameters and quantities needed to conduct power analysis except for the smallest effect size of practical concern. As previously emphasized, the smallest effect size of substantial concern should be deduced from theoretical and practical considerations and not from universal guidelines.

VWB23 found a total of 76 studies eligible for meta-analysis, of which 82% of the effect sizes were adjusted for pretest measures. The database includes both CHEs dependence structures, supporting the use of the CHE-RVE model. Based on this information, we assume that we will find *76 studies* ± *10*, which we think is a plausible range because it falls within the average number of studies found in education and applied psychology.[3] In addition, VWB23 found a substantial amount of heterogeneity, with variance components (reported as SDs) of *0.25 SD* at the effect size level (ω) and *0.1 SD* at the study level (τ). VWB23 estimated $\rho \approx$ *0.7* from paired effect size estimates for studies both reporting STEM and Language Arts outcomes, as suggested by Kirkham et al.[47] From the VWB23 data, it is furthermore possible to obtain a vector of k_js, with $\bar{k}_j = 3.8$, ranging from 1 to 27, and a vector of cluster bias-corrected σ_j^2s aggregated to the study level. Cluster bias correction was needed in this case because 67 out of the 76 included studies did not adequately account for nesting of students within classes and schools. Because both collaborative models of instruction and increased instruction time are provided at the class level, it is important to account for clustering in such reviews.[32]

In this case, we define the smallest effect size of practical importance relative to the overall effect size of interventions with similar cost and resource intensity, such as co-teaching and class size reduction, both of which appear to have an overall average effect of approximately *0.1 SD*.[46,49] Consequently, we consider an overall average effect size falling below 0.1 to be irrelevant compared to these related interventions. With all the needed assumptions in place for power approximation of the mean effect size of the CHE-RVE model, power can be approximated from the `power_MADE()` function from the *POMADE* package. By using the sample characteristics and estimated parameters from VWB23 (i.e., empirically sampling k_js and σ_j^2s from the VWB23 data and specifying $J = 76$, $\tau = 0.1$, $\omega = 0.25$, $\rho = .7$), it appears that the CHE-RVE yields 76.1% power to detect $\mu = 0.1$.

4.3 | Number of studies needed to find the smallest effect of interest

The *POMADE* package also provides functions that allow researchers to answer questions concerning how many studies are needed to obtain a specified power level, given an effect size considered to be of practical interest and a prespecified level of statistical significance. This can be investigated via the `min_studies_MADE()` function. Based on the assumptions described above, we can see that it would require 84 studies to have 80% power to detect $\mu = 0.1$ under these conditions.

4.4 | Minimum detectable effect size

Another feature of the *POMADE* package is the `mdes_MADE()` function, which allows one to answer questions about the *minimum detectable effect size* (MDES) for a specified sample size, given preset levels of statistical significance and power as well as parameter values and

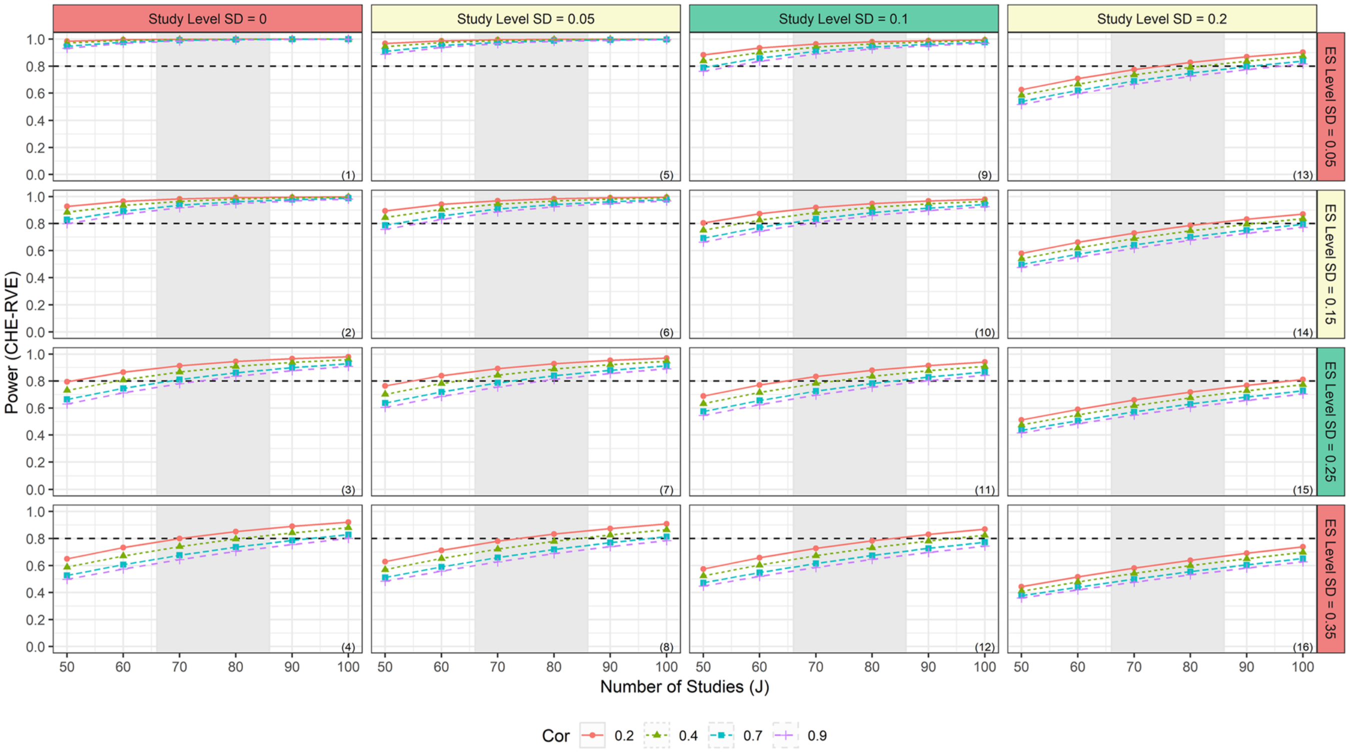

FIGURE 1 Traffic light power plot across J (CHE-RVE). Dashed lines indicate power of 80%. Shaded gray areas mark the range of studies expected to be found by the reviewers (in this case 76 ± 10). the colors of the strips indicate the reviewers' expectation of the likelihood of the given scenarios appearing in the dataset and analysis with green representing expected scenarious.

study characteristics. Using this function, we find that the smallest effect size detectable with 80% power under the given conditions is 0.105, which is quite close to the smallest effect size considered to be of practical relevance under the conditions given in the previous examples.

4.5 | Plotting

4.5.1 | Traffic light power plots

We acknowledge that it can be rather difficult to make definitive assumptions about the true model parameters and sample characteristics, including the final number of studies. For instance, our heterogeneity parameter assumptions were based on estimates from pilot data, and thus subject to some degree of uncertainty. Reporting only one power estimate can be misleading, even if the true model and data structure diverge only modestly from one's initial assumptions. To maximize the informativeness of the power approximations, we suggest accommodating the a priori uncertainty of the power approximations by reporting or plotting power estimates across a range of possible scenarios. Figure 1 depicts such a plot, in which power estimates are approximated across varying assumptions of τ, ω, ρ, and J. With such a plot, investigators can also illustrate the interval in which they expect the final number of studies to fall. This provides a means for reviewers to assess the consequences of the assumptions for the power level and determine under which scenarios the model power exceeds a target level. In this context, we applied the convention of setting 80% power as the minimum acceptable power level for model fitting, meaning that the Type I error is considered four times as serious as making a Type II error.[21]

To further augment and more clearly illustrate the assumptions put forward by the investigator, we suggest illustrating the likelihood of the reviewers' assumptions by coloring the strips of the facetted plots, with green indicating the expected scenario, yellow indicating other plausible scenarios, and red indicating other, even less likely scenarios.[2] Consequently, we coin this type of plot as a *traffic light power plot*. Traffic light assumptions should ideally be deduced from prior work related to one's topic. From illustrations such as Figure 1, it should be more clear to others, including funders, what they can expect in terms of power, while also acknowledging some degree of uncertainty in these estimates. We suggest approximating no more than four less likely assumptions to keep the scenarios depicted down to a manageable number. Investigators can make the power plot displayed in Figure 1 by using the `plot_MADE()` function in the

POMADE package, including for their own preferred values of the parameters τ, ω, ρ, and J.

When planning a meta-analysis, reviewers should take particular care to develop plausible assumptions about the likely magnitude of design parameters that have relatively strong influence on the anticipated power of the study. The sensitivity analysis depicted in traffic light plots such as Figure 1 are helpful in assessing such questions. Under the most likely conditions for within- and between-study heterogeneity, power for $J = 70$ studies ranges from 0.69 to 0.83 across the values of ρ. When $\tau = 0.1$ (the most likely scenario), $\rho = 0.7$, and $J = 70$ studies, power ranges from 0.61 to 0.91 across the values of ω. Finally, when $\omega = 0.25$ (the most likely scenario), $\rho = 0.7$, and $J = 70$ studies, power ranges from 0.57 to 0.81 across the values of τ. Thus, power is most sensitive to the assumption regarding ω, the degree of within-study heterogeneity. However, this pattern is specific to this particular example, including especially to our assumptions regarding the distribution of effect sizes per study. Under different assumptions, anticipate power may be relatively more sensitive to other design parameters, such as the correlation between effect size estimates ρ.

4.5.2 | Interpretation of power analyses

Substantively, Figure 1 illustrates the power of a test based on the CHE-RVE model to find $\mu = 0.1$ across the assumptions we made regarding the effect of increased instruction time on student achievement. The green strips indicate our expectation to find $\tau = 0.1$ and $\omega = 0.25$ based on the variance estimates from VWB23 (τ and ω are here reported as SDs so that they can be interpreted in the same unit as μ). Furthermore, the gray shades in Figure 1 depict our expectation to find ± 10 studies of what was found in VWB23, which also happens to be the mean number of studies reported in the Review of Education Research and Applied Psychology journals.[3] The four lines in the traffic light power plot indicate various assumptions about the common sample correlation among effect sizes coming from the same study, ρ. We assumed $\rho = 0.7$, and under the expected (green) scenario in panel (11) in Figure 1, power estimates range from ~70% power with 66 studies to ~80% power with 86 studies. Though power does not exceed 80% in all scenarios, we would still suggest proceeding with meta-analysis, considering that a minor reduction of the within-study variance would yield power above 80%. As can be seen in panel (7) in Figure 1, reducing ω with 0.1 SD would increase power by 10% or more and thus produce power above 80% across the likely range for the expected number of studies (J). A further implication of these results is that the investigators could consider whether they can tighten their selection criteria to reduce the within-study SD. For example, averaging results across subscale or subgroup results irrelevant to the main analyses of the given review might help to avoid artificial inflation of the within-study SD.

4.5.3 | Number of studies (J)

Besides power, the `plot_MADE()` function also includes the option to visualize how many studies are needed to detect a given effect size of practical concern across varying assumptions about τ, ω, and ρ. From plots like Figure 2, researchers can gain knowledge about the target range of the number of studies needed to detect the smallest effect size of practical concern. Furthermore, if this kind of analysis is conducted across multiple values of μ, the `plot_MADE()` function allows reviewers to visualize how the number of studies needed changes as a function of the smallest effect size of interest, as presented in Figure 3.

4.5.4 | Minimum detectable effect size

Finally, the `plot_MADE()` function can be used to understand and visualize how the minimum detectable effect size varies across the number of included studies and various model parameters, as presented in Figure 4. Concretely, Figure 4 provides a means for reviewers to understand what effect sizes can be detected under a range of different data and model assumptions. From Figure 4, it can, for instance, be seen that across all the different scenarios, reviewers can at minimum detect a moderate[24] effect, clearly justifying meta-analysis.

5 | UTILITY OF PROSPECTIVE POWER ANALYSIS FOR META-ANALYSIS

One of the major aims of a priori power analysis for meta-analysis is to shed light on the utility of a planned systematic review. Prospective power calculations can inform reviewers and funders if enough studies are available to find the smallest effect size of practical or substantial concern and thus whether the literature is mature enough for a meta-analysis. However, we must emphasize that power calculations should not be the sole or determinative factor when considering whether to undertake (or to fund) a synthesis, and reviewers should be careful abandoning meta-analysis based on power

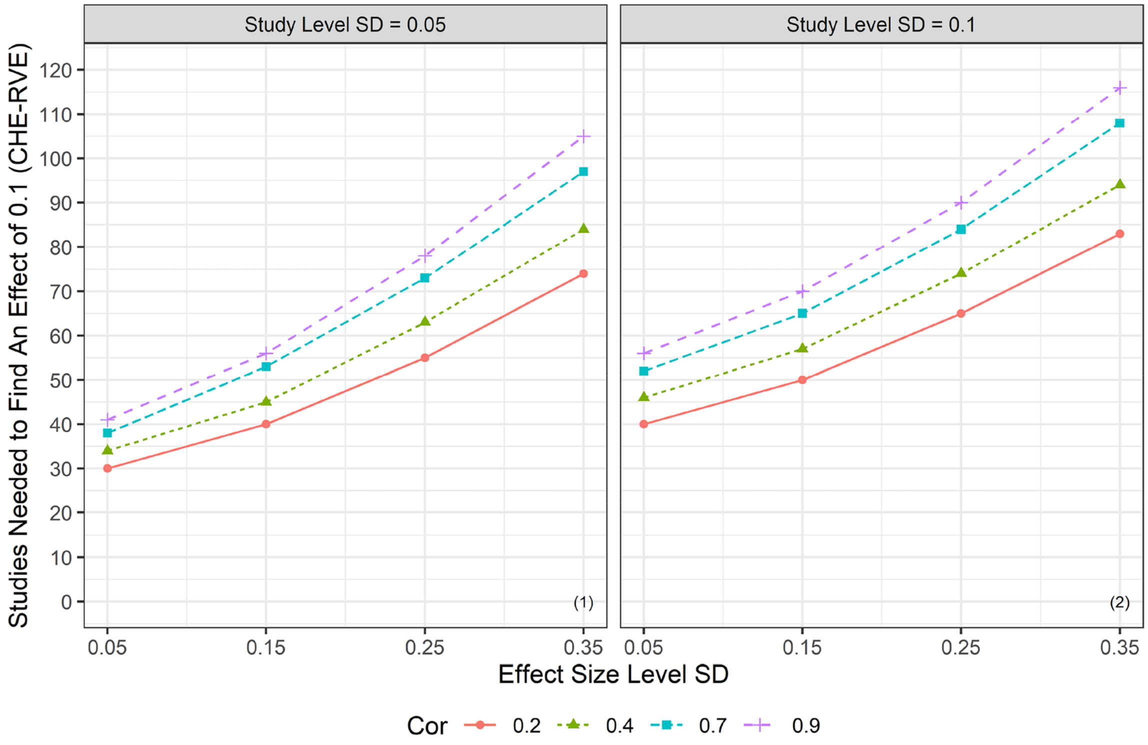

FIGURE 2 Studies needed to find $\mu = 0.1$ across varying values of τ^2 and ω^2 (CHE-RVE).

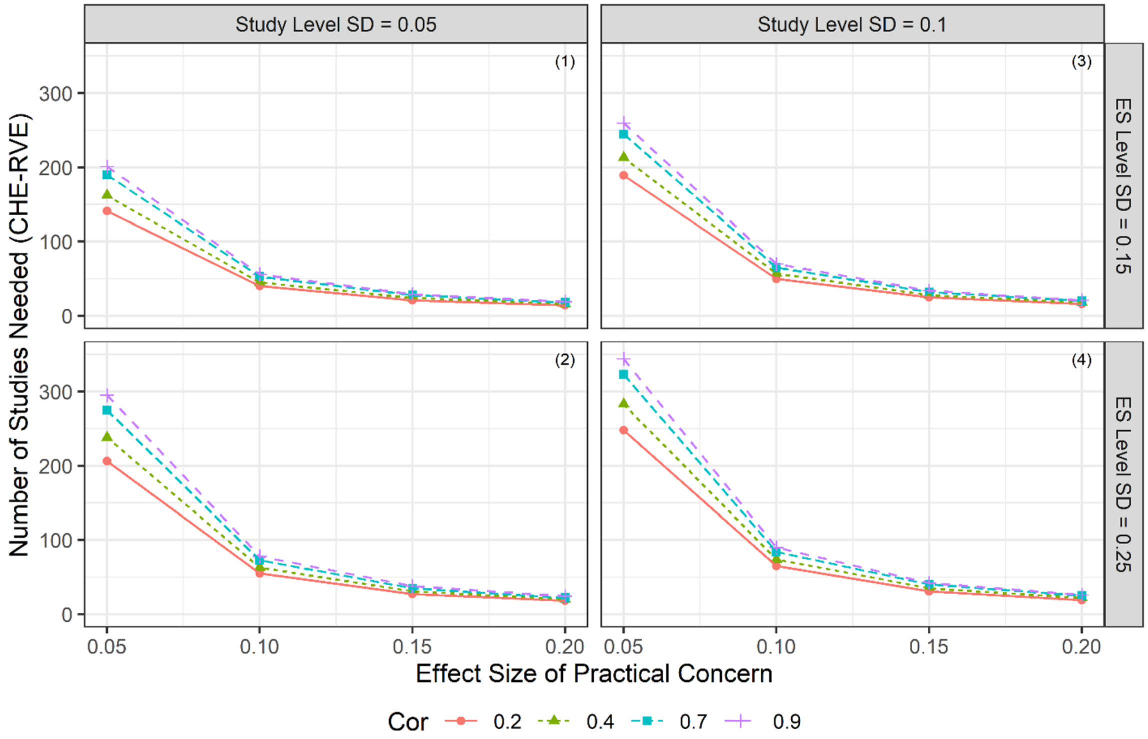

FIGURE 3 Number of studies needed as function of μ (CHE-RVE).

analyses conducted before the full literature search. Power pertains only to hypothesis-testing aims of a review, but syntheses might also have aims focused more on summary estimation or characterization of heterogeneity—or even on simply describing and organizing a literature. We agree with Valentine et al.[10] that reviewers should strive to apply some kind of meta-analytical approach, “[n]ot because it is ideal but rather

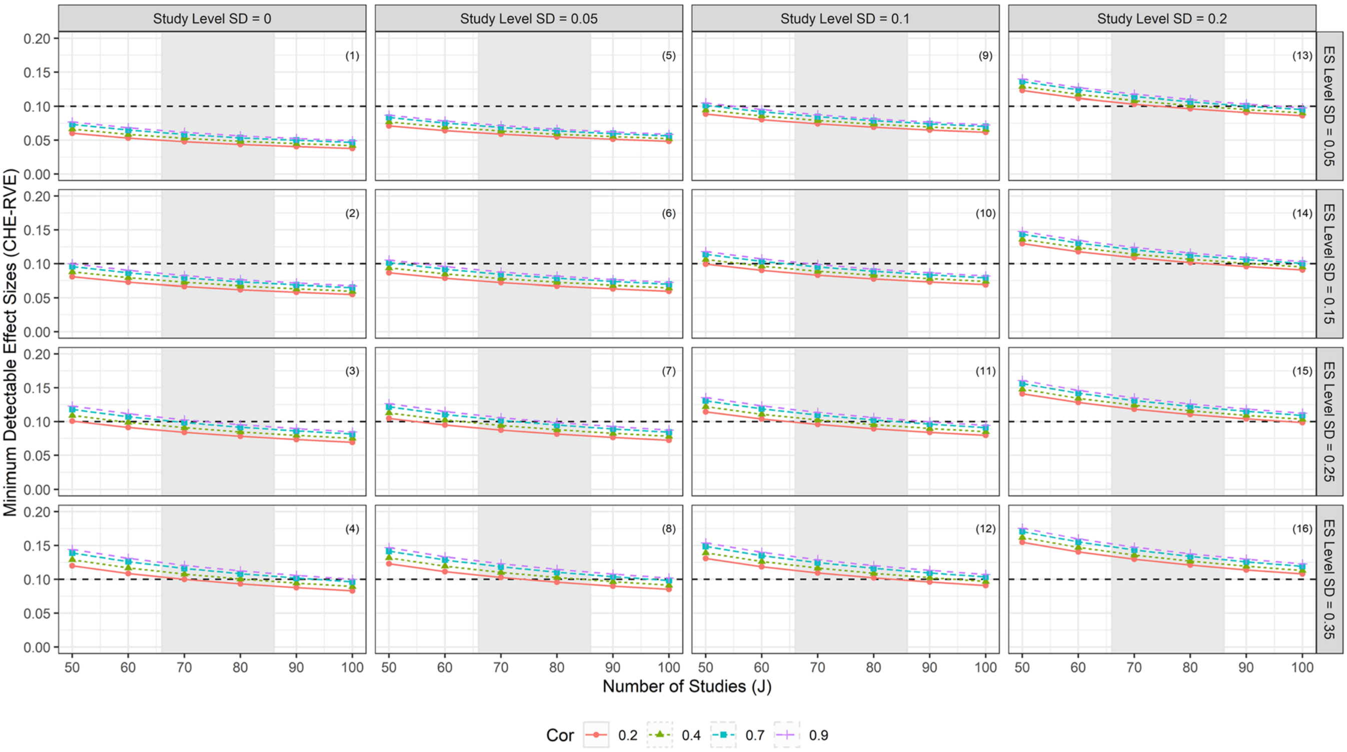

FIGURE 4 Minimum detectable effect size plot as function of J (CHE-RVE). Dashed lines indicate the smallest effect size of practical concern. Shaded gray areas mark the range of studies expected to be found by the reviewer.

because given the need for a conclusion (e.g., an administrator who needs to pick a program), it is a better analysis strategy than the alternatives."

A further reason to be cautious in using power calculations is that, as we have illustrated, the approximations involved require extensive assumptions that can be error-prone (and thereby misleading). For example, a larger number of eligible studies might be revealed to the reviewers during the literature search, such as through searches of gray literature databases.[10] As anecdotal evidence to support this possibility, the first author was a part of a review[46] in which the authors only expected to find 20 eligible studies but ended up finding 128 studies, with approximately 100 studies coming from gray literature searches.

If a reviewer finds that the planned meta-analysis will have lower power than anticipated, we recommend proceeding with the meta-analysis, and consider alternative modeling strategies[7,50,51] or other relevant quantitative alternatives.[10,52] If reviewers identify few studies, they could also consider using Bayesian meta-analytical techniques, using informed priors of μ.[51] This can potentially ease the interpretability of the meta-analysis results based on few studies. In the social and behavioral sciences, it is common to find many small studies that contribute a large number of effect size estimates to the overall database. In such cases, prospective power calculations can be informative about the consequences of including a large proportion of such studies on the within-study variance estimation in random-effects models. This information can suggest whether reviewers should consider alternative strategies such as averaging within-study results reported across subgroups and/or sub-scales irrelevant to the main analyses of the given review.[50] By reducing the number of imprecise effect sizes, it might be possible to avoid artificially inflating the within-study variance estimation and thereby gain power for their models. Further methodological investigation of such strategies strikes us as warranted.

Power analysis focuses exclusively on statistical significance testing and thus, to some degree, requires arbitrarily selected cutpoints for determining statistical significance and desired power levels. Consequently, we urge investigators to be careful in decisions about conducting a meta-analysis based on a priori power analyses unless the evidence is decisive. Future research could profitably concentrate on developing alternative methods that complement power analysis. One possibility is precision analysis,[53] which aims to approximate the number of studies needed to obtain confidence intervals of a certain width with a given probability. With precision analysis, reviewers would not need to premise the conduct of meta-analysis on a dichotomized choice of either having or not having adequate power to find the smallest effect of practical concern. Nevertheless, we still believe power analysis for meta-analysis of dependent effect sizes provides a means for reviewers to gain an a priori understanding of the given stage and maturity of the literature in point for review.

Finally, an over-arching benefit of conducting a priori power analysis is that it requires the reviewers to plan for and think carefully about the likely structure of their meta-analysis dataset and about the smallest effect size of practical interest. This might naturally yield a deeper understanding of the structure of the literature as well as the topic under review and thus encourage more fine-grained and content-relevant interpretations of the final meta-analysis results. However, it is important to note that prospective power analyses should not be compared to the final results because, *by definition*, they do not add any further information to the final results.[10]

6 | CONCLUSION

In this article, we have developed common guidelines for conducting power analysis for meta-analysis of dependent effect sizes and introduced the *POMADE* package for this purpose. Moreover, we have introduced new graphical tools for illustrating power approximations across a range of plausible scenarios. As is apparent from the above illustration, power approximations for meta-analysis will be more informative when based on pilot data from previous syntheses on a similar research topic. This circumstance is further impetus for the entire research synthesis community to embrace and follow open science and open data[54] policies, so that prospective power analyses can be conducted using well-justified, empirically supported assumptions. With this paper, we hope to have provided guidance needed for investigators to apply the power approximations as common practice for future systematic reviews involving meta-analysis. On this note, we invite readers to provide feedback on whether this guidance adequately meets the challenges that they encounter as well as on the utility of the POMADE software package.

AUTHOR CONTRIBUTIONS

Mikkel Helding Vembye: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; visualization; writing – review and editing; writing – original draft. **James Eric Pustejovsky:** Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; supervision; validation; visualization; writing – review and editing; writing – original draft. **Therese Deocampo Pigott:** Conceptualization; supervision; writing – original draft; writing – review and editing; resources.

FUNDING INFORMATION

The authors received no funding for this work.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/vpnmb/.

ORCID

Mikkel Helding Vembye https://orcid.org/0000-0001-9071-0724
James Eric Pustejovsky https://orcid.org/0000-0003-0591-9465
Therese Deocampo Pigott https://orcid.org/0000-0002-5976-246X

ENDNOTES

[1] Find data and background material for this study at https://osf.io/fby7w/.

[2] When needed this type of plot can be made color-blind friendly by for example using a gray-scale version with white indicating the expected scenario, light gray indicating other plausible scenarios, and dark gray indicating other, even less likely scenarios. We have written the plot_MADE() function so that the user can apply their preferred palette.

REFERENCES

1. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press; 1985.
2. Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychol Bull*. 1988;103(1):111-120. doi:10.1037/0033-2909.103.1.111
3. Tipton E, Pustejovsky JE, Ahmadi H. Current practices in meta-regression in psychology, education, and medicine. *Res Synth Methods*. 2019;10(2):180-194. doi:10.1002/jrsm.1339
4. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods*. 2010;1(1):39-65. doi:10.1002/jrsm.5
5. Van den Noortgate W, López-López J, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. *Behav Res Methods*. 2013;45(2):576-594. doi:10.3758/s13428-012-0261-6
6. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Meta-analysis of multiple outcomes: A multilevel approach. *Behav Res Methods*. 2014;47(4):1274-1294. doi:10.3758/s13428-014-0527-2
7. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: Expanding the range of working models. *Prev Sci*. 2021;23(1):425-438. doi:10.1007/s11121-021-01246-3

8. Hedges LV, Pigott TD. The power of statistical tests in meta-analysis. *Psychol Methods*. 2001;6(3):203-217. doi:10.1037/1082-989X.6.3.203
9. Hedges LV, Pigott TD. The power of statistical tests for moderators in meta-analysis. *Psychol Methods*. 2004;9(4):426-445. doi:10.1037/1082-989X.9.4.426
10. Valentine JC, Pigott TD, Rothstein HR. How many studies do you need?: A primer on statistical power for meta-analysis. *J Educ Behav Stat*. 2010;35(2):215-247. doi:10.3102/1076998609346961
11. Jackson D, Turner R. Power analysis for random-effects meta-analysis. *Res Synth Methods*. 2017;8(3):290-302. doi:10.1002/jrsm.1240
12. Vembye MH, Pustejovsky JE, Pigott TD. Power approximations for overall average effects in meta-analysis with dependent effect sizes. *J Educ Behav Stat*. 2023;48(1):70-102. doi:10.3102/10769986221127379
13. Zhang B, Konstantopoulos S. Statistical Power Analysis for Univariate Meta-Analysis: A Three-Level Model. *J Res Educ Eff*. 2024;1-28. doi:10.1080/19345747.2023.2290544
14. Vembye MH, Pustejovsky JE. *POMADE: Power for Meta-Analysis of Dependent Effects (R package version 0.2.0)*. CRAN. 2024. https://doi.org/10.32614/CRAN.package.POMADE
15. Fernández-Castilla B, Aloe AM, Declercq L, et al. Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behav Res Methods*. 2020;53(1):702-717. doi:10.3758/s13428-020-01459-4
16. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics*. 1946;2(6):110-114.
17. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat*. 1981;6(2):107-128. doi:10.2307/1164588
18. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods*. 2015;20(3):375-393. https://doi.org/10.1037/met0000011
19. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. *Res Synth Methods*. 2011;2(1):61-76. doi:10.1002/jrsm.35
20. Lakens D, Adolfi FG, Albers CJ, et al. Justify your alpha. *Nat Hum Behav*. 2018;2(3):168-171. doi:10.1038/s41562-018-0311-x
21. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge; 1988. doi:10.4324/9780203771587
22. Hattie J. *Visible Learning: A Synthesis of over 800 Meta-Analysis Relating to Achievement*. Routledge; 2009.
23. Lipsey MW, Puzio K, Yun C, et al. Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *Natl Cent Spec Educ Res*. 2012;1-46. https://files.eric.ed.gov/fulltext/ED537446.pdf
24. Kraft MA. Interpreting effect sizes of education interventions. *Educ Res*. 2020;49(4):241-253. doi:10.3102/0013189X20912798
25. Schäfer T, Schwarz MA. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Front Psychol*. 2019;10(813):1-13. doi:10.3389/fpsyg.2019.00813
26. Ahn S, Ames AJ, Myers ND. A review of meta-analyses in education: Methodological strengths and weaknesses. *Rev Educ Res*. 2012;82(4):436-476. doi:10.3102/0034654312458162
27. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11(1):160. doi:10.1186/1471-2288-11-160
28. White T, Noble D, Senior A, Hamilton KW, Viechtbauer W. *metadat: Meta-analysis datasets* (R package version 1.0-0). 2021. https://doi.org/10.32614/CRAN.package.metadat
29. Polanin JR, Espelage DL, Grotpeter JK, et al. Locating unregistered and unreported data for use in a social science systematic review and meta-analysis. *Syst Rev*. 2020;9(1):116. doi:10.1186/s13643-020-01376-9
30. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. 1st ed. John Wiley & Sons; 2009.
31. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Vol 1. 2nd ed. Sage; 2002.
32. Higgins JPT, Eldridge S, Li T. Including variants on randomized trials. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley; 2019:569-593. doi:10.1002/9781119536604
33. Hedges LV. Effect sizes in cluster-randomized designs. *J Educ Behav Stat*. 2007;32(4):341-370. doi:10.3102/1076998606298043
34. Hedges LV. Effect sizes in three-level cluster-randomized experiments. *J Educ Behav Stat*. 2011;36(3):346-380. doi:10.3102/1076998610376617
35. Taylor JA, Pigott TD, Williams R. Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educ Res*. 2021;51(1):72-80. doi:10.3102/0013189X211051319
36. Hedges LV, Hedberg EC. Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal*. 2007;29(1):60-87. doi:10.3102/0162373707299706
37. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the Health Survey for England 1994. *Am J Epidemiol*. 1999;149(9):876-883. doi:10.1093/oxfordjournals.aje.a009904
38. Verma V, Lee T. An analysis of sampling errors for the demographic and health surveys. *Int Stat Rev*. 1996;64(3):265-294. doi:10.2307/1403786
39. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev*. 2003;27(1):79-103. doi:10.1177/0193841X02239019
40. Tanner-Smith EE, Lipsey MW. Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis. *J Subst Abuse Treat*. 2015;51(1):1-18. doi:10.1016/j.jsat.2014.09.001
41. Dietrichson J, Bøg M, Filges T, Klint Jørgensen AM. Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Rev Educ Res*. 2017;87(2):243-282. doi:10.3102/0034654316687036
42. Linden AH, Hönekopp J. Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspect Psychol Sci*. 2021;16(2):358-376.
43. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34(6):984-998. doi:10.1002/sim.6381
44. Pigott TD. *Advances in Meta-Analysis*. Springer; 2012.
45. Fraley RC, Vazire S. The N-Pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical

power. *PLoS One*. 2014;9(10):e109019. doi:10.1371/journal.pone.0109019

46. Vembye MH, Weiss F, Bhat BH. The effects of co-teaching and related collaborative models of instruction on student achievement: A systematic review and meta-analysis. *Rev Educ Res*. 2023;1-47. doi:10.3102/00346543231186588

47. Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Stat Med*. 2012;31(20):2179-2195. doi:10.1002/sim.5356

48. Wickham H. *ggplot2: Elegant graphics for data analysis* (R package version 3.5.1). CRAN. 2016. https://doi.org/10.32614/CRAN.package.ggplot2

49. Filges T, Sonne-Schmidt CS, Nielsen BCV. Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Syst Rev*. 2018;14(1):1-107. doi:10.4073/csr.2018.10

50. Pustejovsky JE, Chen M. Equivalences between ad hoc strategies and meta-analytic models for dependent effect sizes. *J Educ Behav Stat*. 2023;1-31. doi:10.3102/10769986241232524

51. Valentine JC, Wilson SJ, Rindskopf D, et al. Synthesizing evidence in public policy contexts: The challenge of synthesis when there are only a few studies. *Eval Rev*. 2017;41(1):3-26. doi:10.1177/0193841X16674421

52. McKenzie JE, Brennan SE. Synthesizing and presenting findings using other methods. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley; 2019:321-347.

53. Rothman KJ, Greenland S. Planning study size based on precision rather than power. *Epidemiology*. 2018;29(5):599-603. doi:10.1097/EDE.0000000000000876

54. Moreau D, Gamble B. Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychol Methods*. 2020;27:426-432. doi:10.1037/met0000351

How to cite this article: Vembye MH, Pustejovsky JE, Pigott TD. Conducting power analysis for meta-analysis with dependent effect sizes: Common guidelines and an introduction to the POMADE R package. *Res Syn Meth*. 2024;1-17. doi:10.1002/jrsm.1752