2021 ■ Vol. 17 ■ no. 1

Calculating statistical power for meta-analysis using metapower

Jason W. Griffin [a]

[a]Department of Psychology, Pennsylvania State University

Abstract ■ Meta-analysis is an influential evidence synthesis technique that summarizes a body of research. Though impactful, meta-analyses fundamentally depend on the literature being sufficiently large to generate meaningful conclusions. Power analysis plays an important role in determining the number of studies required to conduct a substantive meta-analysis. Despite this, power analysis is rarely conducted or reported in published meta-analyses. A significant barrier to the widespread implementation of power analysis is the lack of available and accessible software for calculating statistical power for meta-analysis. In this paper, I provide an introduction to power analysis and present a practical tutorial for calculating statistical power using the R package *metapower*. The main functionality includes computing statistical power for summary effect sizes, tests of homogeneity, categorical moderator analysis, and subgroup analysis. This software is free, easy-to-use, and can be integrated into a continuous work flow with other meta-analysis packages in R.

Keywords ■ moderator analysis, power analysis, statistical software, systematic review, evidence synthesis. **Tools** ■ R.

jxg569@psu.edu

10.20982/tqmp.17.1.p024

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers ■ One anonymous reviewer.

Introduction

Meta-analysis is a powerful statistical tool widely used across a broad range of scientific disciplines to quantitatively summarize an area of research. By identifying, synthesizing, and summarizing empirical research findings, meta-analyses increase generalizability and improve effect size estimates of the existing literature. Evidence synthesis methods like meta-analysis are widely regarded as the highest form of scientific evidence, and routinely inform policy decisions, clinical practice, and evidence-based medicine (Gopalakrishnan & Ganeshkumar, 2013). Although meta-analyses are highly influential, this technique is resource intensive and time-consuming, often taking at least a year to complete (Borah, Brown, Capers, & Kaiser, 2017). In addition to logistical demands, conducting a meta-analysis before enough studies are available can result in inaccurate and misleading conclusions, especially when the number of studies is small (Jackson & Turner, 2017; Thorlund et al., 2011). Therefore, power analysis plays an important role in the planning stage and is necessary to determine the feasibility of a meta-analysis.

The goal of power analysis in primary and meta-analytic research is to determine the number of participants or number of studies, respectively, needed to have a reasonable chance at rejecting the null hypothesis given a statistical test. For meta-analysis, these statistical tests include estimating a summary effect size that is different than zero (e.g., size of association between variables, group differences on a variable), evaluating whether there are group differences in effect size between different types of studies (e.g., children vs. adult studies), or evaluating whether there are subgroup differences in effect size within studies (e.g., men vs. women). By postulating what we expect to find, it becomes possible to calculate statistical power - the probability of rejecting the null hypothesis when, in fact, the alternative hypothesis is true.

For primary research, power analysis is used at the planning and design stage of a study to determine the number of participants required to detect a substantive effect given an expected effect size. Widely considered an essential part of research design, *apriori* power analyses are

often required for federally funded research grants and randomized controlled trials (National Institute of Health, 2018). The inclusion of power analysis in evidence synthesis methods like meta-analysis is also critical because underpowered meta-analyses lack precision in estimating a summary effect size - especially when individual studies vary considerably - which can lead to conclusions that are incorrect (Thorlund et al., 2011; Pigott & Polanin, 2020; Jackson & Turner, 2017). Although the necessary underlying theoretical statistics, equations, and procedures have been articulated (Hedges & Pigott, 2001, 2004; Jackson & Turner, 2017; Pigott, 2012; Valentine, Pigott, & Rothstein, 2009), power analyses are rarely considered, conducted, or reported in published meta-analyses.

A significant barrier to the widespread implementation of power analysis in meta-analysis is the lack of easy-to-use software. While numerous software options like G*power have been developed to compute power calculations for primary research, allowing for widespread implementation of power analysis in primary research (Faul, Erdfelder, Lang, & Buchner, 2007), analogous software options do not exist for meta-analysis, despite similarity in procedure. This means that to compute statistical power for meta-analysis, researchers must manually perform the calculations, use an online calculator, or utilize a user-defined script (e. g., Cafri, Kromrey, & Brannick, 2009). Such resources can be limited in functionality and difficult to integrate into a reproducible work flow.

To fill this methodological gap, I developed *metapower*, an R package for computing statistical power for meta-analysis. This package supports power analysis for (1) summary effect sizes; (2) tests of homogeneity; (3) moderator analysis; and (4) subgroup analysis. Additionally, power calculations are available for fixed- and random-effects models and can accommodate multiple types of effect sizes (i. e., Cohen's *d*, correlation coefficient, and odds ratio). *metapower* was designed to be user-friendly (i. e., minimal coding) and accessible (i. e., free) to researchers with various degrees of expertise including students, principal investigators, applied researchers, non-statisticians, and those with little programming experience. To compliment this goal, I also developed a fully functional, web-based application for users unfamiliar with R (jason-griffin.shinyapps.io/shiny_metapower). In what follows, I overview the major components of power analysis for meta-analysis, provide guidance on how to make decisions about anticipated parameter values, and provide a step-by-step tutorial on how to conduct a power analysis for meta-analysis using *metapower*.

Power analysis for meta-analysis

Like traditional power analysis, computing statistical power for a meta-analysis requires making informed assumptions about expected findings. In primary research, this includes the effect size magnitude and sample size of an individual study. For meta-analysis, the unit of analysis is an individual study (rather than a participant); therefore, power is calculated based on expected values for effect size magnitude, sample size, the number of studies, and the amount of between-study variability. Generally speaking, power can be calculated with four values: effect size magnitude, study sample size, number of studies, degree of heterogeneity.

Study-specific effect sizes and variances

Meta-analyses can be conducted using different types of effect sizes, including Cohen's *d*, correlation coefficient, and odds ratio (Pigott, 2012). Importantly, different effect size metrics generally reflect specific types of research questions. For example, the mean difference between two independent groups can be evaluated with Cohen's *d*, whereas a correlation coefficient reflects the within-group association between two continuous variables. The logic for calculating statistical power is similar for each of these as the input data for meta-analysis are the study-specific effect sizes and variances. However, the distributional characteristics for some effect sizes are unfavorable for quantitative synthesis and must be transformed for meta-analysis and as a consequence, for power analysis as well. Cohen's *d* can be used directly in meta-analysis as computed with

$$ES_d = \frac{M_2 - M_1}{s_p} \tag{1}$$

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{ES_d^2}{2(n_1 + n_2)} \tag{2}$$

where M_2 and M_1 are the two group means and s_p is the pooled standard deviation. In contrast, the correlation coefficient, *r*, is not normally distributed; thus, meta-analysis of correlation coefficients use Fisher's r-to-z transformations and the respective variance given by

$$ES_{rz} = 0.5ln\left[\frac{1+r}{1-r}\right] \tag{3}$$

$$v_{rz} = \frac{1}{n-3} \tag{4}$$

Odds ratios are also not normally distributed and range from 0 to ∞. As a result, meta-analyses of odds ratios are conducted using the log odds ratio, $ES_{log(OR)}$, and respective variance, $v_{log(OR)}$, given by

$$ES_{log(OR)} = ln\left(\frac{ad}{bc}\right) \tag{5}$$

$$v_{log(OR)} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \quad (6)$$

Throughout the remainder of this paper, equations will refer to effect size, ES, and variance, v, whose values depend on the type of effect size as shown above.

Power for summary effect size

In meta-analysis, input data reflects study-specific effect sizes (ES_i) and variances (v_i) for *i*th study where $i = 1,\ldots,$ *k*, where *k* is the total number of studies. These input data are statistically combined to estimate a weighted summary effect size θ and variance $v_\bullet$. The goal of power analysis is to determine the probability of correctly rejecting the null hypothesis (e.g., $\theta = 0$) in favor of an expected alternative, where the alternative is based on what researchers expect to find (e.g., $\theta = 0.5$). Since providing guesses about each study-specific effect size and variance is nonviable, it is assumed that overall effect size, ES = ES_i, and overall variance, v = v_i, for *i*th study where $i = 1,\ldots,k$. Furthermore, since v depends on v_i and within-study sample sizes n_i, power calculations also assume that studies have the same sample sizes, such that n = n_i for *i*th study where $i = 1,\ldots,$ *k*. In other words, prior expectations should be based on a "typical" study that is most characteristic of the prospective meta-analysis (Hedges & Pigott, 2001; Jackson & Turner, 2017).

Fixed-effects model

Fixed-effects models assume that a single common effect size underlies all study-specific effect sizes in a meta-analysis. To calculate power under this assumption, we first posit a value for the expected overall effect size, ES, within-study sample size, n, and the total number of studies to be included in the meta-analysis, k (Hedges & Pigott, 2001). With these expected values, it is possible to derive an alternative distribution representing the expected outcome, which can be compared to the null distribution. This alternative is given by the non-centrality parameter, λ, which is based on the value of the expected summary effect size θ and variance $v_\bullet$. The weighted summary effect size reflects the expected magnitude of the effect size such that $\theta = ES$, whereas the weighted variance $v_\bullet = v/k$, where v reflects the the common variance (see Eq. 2, 4, & 6), and *k* is the total number of studies (Hedges & Pigott, 2001). With θ and $v_\bullet$, the non-centrality parameter λ can be calculated with

$$\lambda = \frac{\theta - 0}{\sqrt{v_\bullet}} \quad (7)$$

This non-centrality parameter can then be compared to the null distribution to derive the probability of rejecting the null hypothesis in favor of the expected alternative. For a summary effect size, the null hypothesis is that the summary effect size equals zero ($H_0 : \theta = 0$), and is based on a standard normal distribution (*M* = 0, *SD* = 1), whereas the expected alternative distribution has a mean equal to λ and variance 1. The statistical power for this test is given by calculating the area under the alternative distribution that exceeds the critical value of the null distribution written as

$$power = 1 - \Phi(c_\alpha - \lambda) \quad (8)$$

where $\Phi(x)$ is the cumulative distribution function, c_α is the specified critical value, and λ is the non-centrality parameter. For a two-tailed test, the area under the curve that is greater or less than $|c_{\alpha/2}|$ is written as

$$power = 1 - \Phi(c_{\alpha/2} - \lambda) + \Phi(c_{\alpha/2} - \lambda) \quad (9)$$

Random-effects model

Power analysis for fixed-effects models relies on the strict assumption that a single effect size underlies all individual studies included in the meta-analysis. This strong assumption is difficult to justify in the majority of cases because studies vary considerably with regard to their population, inclusion criteria, methodology, and measurement. Under random-effects models, power analysis is more complex because variation among effect sizes is the result of within-study variance, v, *and* the estimated between-study variability, τ^2. To account for this additional source of variation, the conventional approach has been to posit different values of τ^2 and incorporate into the weighted variance $v_\bullet = (v + \tau^2)/k$. However, τ^2 is a parameter that is estimated with some degree of uncertainty and the aforementioned method does not account for this uncertainty. Importantly, when τ^2 is estimated with large uncertainty (e.g., in the case of meta-analysis with few studies), statistical power is reduced compared to when τ^2 is estimated with greater certainty. Considering the median number of studies included in meta-analyses is estimated to be three (Davey, Turner, Clarke, & Higgins, 2011), accounting for the uncertainty of τ^2 is important for obtaining accurate power calculations (Jackson & Turner, 2017).

To account for uncertainty in the estimation of τ^2, Jackson and Turner (2017) developed an approach to compute statistical power for random-effects meta-analysis that only requires researchers to posit values for the summary effect size, within-study sample size, total number of studies, and degree of between-study variability (i. e., τ^2). However, numerical estimates of τ^2 are difficult to interpret by themselves especially *apriori*; therefore, it is recommended to think about between-study variability in terms of the percent of variation that is due to heterogeneity among effect sizes rather than sampling error (i. e., I^2). Values equal to 25%, 50%, and 75% are thought to reflect small, moderate, and large degrees of heterogeneity respectively (Higgins & Thompson, 2002) and can be written

as

$$I^2 = \frac{\tau^2}{\tau^2 + v} \tag{10}$$

With this information, the non-centrality parameter λ can be calculated (see Eq. 7) and the cumulative distribution function of the test statistic can be obtained (Jackson & Turner, 2017). Specifically, the cumulative distribution function of θ is given by

$$P(T \leq t) = \Gamma_1\left(\frac{df}{2}, \frac{df(1-I^2)}{2}\right)\Phi\left((t-\lambda)\sqrt{1-I^2}\right) + 2df\int_{\sqrt{(1-I^2)}}^{\infty} x\Phi\left(tx - \lambda\sqrt{1-I^2}\right)\chi^2_{df}(dfx^2)dx \tag{11}$$

where $I^2 = \tau^2/(v+\tau^2)$, $\lambda = ES/\sqrt{v/k}$ is the non-centrality parameter, and $\chi^2_{df}(x)$ is the probability density function of the chi distribution with degrees of freedom, $df = k-1$ (Jackson & Turner, 2017).[1] Since the probability of accepting the null hypothesis is given by $P(T \leq c_\alpha)$, power is given by subtracting this value from 1, which is simply written as

$$power = 1 - P(T \leq c_\alpha) \tag{12}$$

For a two-tailed test, the logic is similar such that we evaluate the area under the curve that is either greater or less than $|c_{\alpha/2}|$ written as

$$power = 1 - P(T \leq c_{\alpha/2}) - P(T \leq -c_{\alpha/2}) \tag{13}$$

Power for test of homogeneity

Meta-analysis software commonly reports a measure of homogeneity among the individual study effect sizes. The homogeneity statistic, Cochrane's *Q*, evaluates whether the amount of variation in individual effect sizes is greater than expected from just sampling error alone for fixed-effects models (Cochran, 1954). For random-effects models, *Q* evaluates whether between-study variability τ^2 is greater than 0. As a result, statistical power can be computed for these tests given expectations about the amount of variation within (fixed-effects) and between (random-effects) a group of studies.

Fixed-effects model

Computing power for a test of homogeneity requires an expectation of how much the individual effect sizes differ from the overall effect size (Pigott, 2012). This expectation can be instantiated as an alternative distribution, which can be compared to the null distribution to calculate the statistical power of the test. This alternative distribution is given by a non-centrality parameter, λ, written as

$$\lambda = \sum_{i=1}^{k} w_i(\theta_i - \theta)^2 \tag{14}$$

where θ is the overall effect size, θ_i are the study-specific effects sizes, and w is common inverse variance weight, $w = 1/v$. In this context, λ depends on the difference between the study-specific effect sizes and the overall effect size (i. e., $\theta_i - \theta$). Since this is unknown prior to data collection, values for this difference must be provided. Pigott (2012) shows this can be done by positing the average difference between the study-specific effect sizes and the overall effect size in terms of standard deviation units (i. e., $\sqrt{v_\bullet}$). In this case, λ can be rewritten replacing $\theta_i - \theta$ with the expected standard deviation, sd, which ultimately becomes just the square of the standard deviation shown here

$$\begin{aligned}\lambda &= \sum_{i=1}^{k} w(sd\sqrt{v_\bullet})^2 \\ &= kwv_\bullet(sd)^2 \\ &= \frac{kv_\bullet(sd)^2}{v} \\ &= \frac{k}{v} \times \frac{v}{k}(sd)^2 \\ &= sd^2\end{aligned} \tag{15}$$

The test of homogeneity is based on *Q*, which follows a chi-square distribution and evaluates the amount of variation across study-specific effect sizes (Hedges & Pigott, 2004). Here, the null hypothesis is that all effect sizes estimate a common effect ($H_0 : \theta_1 = \ldots = \theta_k = \theta$). An alternative is that the study-specific effect sizes differ by some value of sd, which is reflected in the computation of λ (see Eq. 15). Therefore, the statistical power for this test is the area under the alternative distribution with mean λ that that exceeds c_α from a non-central chi-square distribution with a non-centrality parameter λ and $k-1$ degrees of freedom written as

$$power = 1 - F(c_\alpha \mid k-1; \lambda) \tag{16}$$

[1] The lower incomplete gamma function is defined as $\Gamma_1(a,x) = \frac{1}{\Gamma(a)}\int_0^x t^{a-1}e^{-t}dt$.

Random-effects model

For random-effects models, the test of homogeneity evaluates whether there is significant between-study variability among the study-specific effect sizes (i. e., $\tau^2 \neq 0$). Given that τ^2 and I^2 are related, Eq. 10 can be rewritten as $\tau^2 = I^2 v/1 - I^2$ to derive values of τ^2 from different values of I^2 based on the expected degree of heterogeneity among effect sizes. Like fixed-effects models, the *Q* statistic has a chi-square distribution with $k-1$ degrees of freedom; however, when the null hypothesis is false, the *Q* distribution is a weighted combination of chi-square distributions, which must be approximated. Hedges and Pigott (2004) derived one way of approximating this distribution. First, the mean of the *Q* distribution is given by

$$\mu_Q = c\tau^2 + (k-1) \tag{17}$$

where c is derived from the fixed-effects weights given by

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \tag{18}$$

The variance of the *Q* distribution is given by

$$\sigma_Q^2 = 2df + 4\left(\sum w_i - \frac{\sum w_i^2}{\sum w_i}\right)\tau^2 + 2\left(\sum w_i^2 - 2\frac{\sum w_i^3}{\sum w_i} + \frac{(\sum w_i^2)^2}{(\sum w_i)^2}\right)\tau^2 \tag{19}$$

To compute the correct approximation with a central chi-square distribution with non-integer degrees of freedom, we must also compute r and s with

$$r = \sigma_Q^2/\mu_Q \tag{20}$$

$$s = 2(\mu_Q)^2/\sigma_Q^2 \tag{21}$$

For this statistical test, the null hypothesis is that there is zero between-study variability (i. e., $H_0 : \tau^2 = 0$). The alternative ($H_1 : \tau^2 > 0$) is derived based on the expected degree of heterogeneity (i. e., I^2). Therein, statistical power is the area under the alternative distribution that exceeds the critical value when $\tau^2 = 0$, given by

$$power = F(c_\alpha/r \mid s; 0) \tag{22}$$

where $F(x)$ is the cumulative distribution function of the central chi-square distribution with s degrees of freedom, and c_α is the $100(1-\alpha)$ percentile point of the chi-square distribution with $(k-1)$ degrees of freedom (Pigott, 2020).

Power for moderator analysis

Heterogeneity in meta-analysis is inevitable because studies are methodologically diverse. A recent study estimated that across 200 meta-analyses, including 12,065 effect sizes, the mean I^2 value was 74%, which reflects a high degree of heterogeneity (Stanley, Carter, & Doucouliagos, 2018). To investigate heterogeneity, moderator analyses are conducted to understand variation in the summary effect size. Moderator variables are defined for different groups of studies (e.g., studies that used different types of tasks). When the primary goal is moderator analysis, power analysis is esepcially relevant to study feasibility because moderator analyses require more studies than simply estimating a summary effect size.

For moderator analysis, the goal is to compare the effect sizes for between different types of studies *p* (e.g., children studies vs. adult studies; Hedges & Pigott, 2004; Pigott, 2012). In this case, the number of studies, *k*, can be divided into different groups of studies: $k = m_i + ... + m_p$, where *i* = 1,..., *p*. Here, m_i represents the number of studies that contribute to each level of a moderator variable. For power analysis, the individual values for m_i are unknown; therefore, it is commonly assumed that an equal number of studies contribute to each group (e.g., the number of studies are divided evenly among levels of the moderator variable). In meta-analysis, moderator analysis is, in essence, analogous to a one-way analysis of variance, where the null hypothesis is that the overall effect size is equal across all groups ($\theta_1 = \theta_2, ..., = \theta_p$), whereas the alternative is that at least one group differs (Pigott, 2012). Power can be computed for a between-groups omnibus test of homogeneity to detect this alternative hypothesis. This requires posited values for the number of groups defined by the moderator effect sizes for each group, within-study sample size, total number of studies, and degree of between-study variability.

To compute power, the first step is to specify the number of groups, p, defined by a moderator variable. Meta-analyses commonly evaluate moderator variables related to age group (e.g., children, adolescents, and adults). Since moderator analyses evaluate between-study differences in effect sizes, m_i designates the number of studies in the *i*th group so that $k = m_1 + m_2 + ... + m_p$. Next, all group effect

sizes θ_i must be specified for *i*th group where $i = 1, ..., p$, and the overall effect size (θ) should be calculated. For example, if is was expected that effect sizes were different for children, adolescents, and adults ($\theta_{children} = 0.2$, $\theta_{adolescents} = 0.4$, $\theta_{adults} = 0.8$), the overall effect size θ would be equal to (0.2 + 0.4 + 0.8)/3 = 0.47 across these three groups. We must also assume that that each group is represented equally across studies (Pigott, 2012). For example, for a meta-analysis of 30 studies, each group would be divided evenly among the total number of studies, *k*, where $m_{children} = m_{adolescent} = m_{adult} = 10$.

The next step is to compute the inverse-variance weights for each group separately computed with

$$w_i = \sum_{j=1}^{m_i} w_{ij} \tag{23}$$

where *i* = 1, ..., *p*, and *j* = 1,..., m_i, and $w_{ij} = 1/v_{ij}$. With these values, an expected alternative (i. e., at least one group effect size is different) can be derived by calculating the non-centrality parameter λ given by

$$\lambda = \sum_{i=1}^{p} w_i(\theta_i - \theta)^2 \tag{24}$$

where *p* is the number of groups, w_i are the summed weights, θ_i are the expected group effect sizes, and θ is the overall effect size. Subsequently, the power for this test is written as

$$power = F(c_\alpha \mid p - 1; \lambda) \tag{25}$$

where *F(x)* is the cumulative distribution of the non-central chi-square with $p - 1$ degrees of freedom and non-centrality parameter λ. Statistical power for this test reflects the area under the curve that exceeds the critical value c_α.

Random-effects model

The same sequence of steps applies to random-effects models with one exception. The between-study variability, τ^2 must be incorporated into the variance estimate. The conventional method for this is to represent the variance as the sum of the variance, v_{ij}^2, and between-study variability, τ^2 (Pigott, 2012). Here, the inverse variance weights can be written as

$$w_i = \sum_{j=1}^{m_i} \frac{1}{(v_{ij}^2 + \tau^2)} \tag{26}$$

Subsequently, the non-centrality parameter and power can be computed in the same way as the fixed-effects method (see Eq. 24 & 25).

Subgroup analysis

In addition to evaluating group differences *between* studies, group differences *within* studies are commonly evaluated with subgroup analysis. For instance, meta-analysts may be interested in how the summary effect differs among different subgroups of study samples, like men and women. To evaluate power for subgroup analysis, the basic logic is similar to that of power for moderator analysis (Pigott, 2020). The main difference is that the number of groups, *p*, now reflects the number of subgroups *within* a study. As a result, the common variance v is now calculated for both subgroups instead of the entire study sample. For parsimony, we make the simplifying assumption that the subgroups are represented equally in each study (*N* = 40; n_{men} = 20, n_{women} = 20). Additionally, since the subgroup differences are within-study comparisons, the number of studies *k* is not divided among groups as in the moderator analysis; therefore, $k = m_1 = m_2 = ... = m_p$. Aside from these specifications, power analysis for subgroups can be calculated using the same sequence as the moderator analysis (see Eq. 23-26).

Determining parameter values for power analysis

As outlined above, calculating statistical power is a complex procedure that requires postulations about expected values such as the effect size, study size, number of studies, and degree of statistical heterogeneity. Importantly, these values have a large impact on statistical power and should not be arbitrarily selected. Instead, these expected values should be informed decisions based on experience and previous literature. Furthermore, expected values are reasonable guesses which are not always correct; it is thus important to be transparent and open about the process of arriving at these expected values. In what follows, I provide guidance on making these informed decisions.

Effect Size Magnitude

The expected effect size has a significant impact on statistical power such that large effect sizes require fewer studies to obtain the same power as small effect sizes. Instead of selecting arbitraily, the expected effect size should be informed by likelihood constraints, previous literature, and experience. For example, although Cohen's *d* ranges from 0 to ∞, the average effect size in psychology is 0.4 with 87% of all effects being less than 0.8 (Cumming, 2014). Therefore, in most psychology applications, the suggested benchmarks of small (Cohen's *d* = 0.2), moderate (Cohen's *d* = 0.5), and large (Cohen's *d* = 0.8) are good starting points. However, the expected effect size should be directly informed by previous literature when possible.

Unlike primary research, previous literature is a pre-

requisite to meta-analysis; therefore, access to previous estimates of a particular effect from individual studies are always available. For example, while planning a meta-analysis, a brief examination of the literature can provide some estimates of a particular effect. Note that studies will not always report effect size information, but they can be calculated manually (see Eq. 1, 3, 5). If numerous studies report a Cohen's *d* between 0.2 and 0.6, 0.5 would be a reasonable guess for the expected effect size of a meta-analysis. In contrast, in some areas of psychology, like psychopharmacology, effect sizes can be considerably larger (Cohen's $d = 5$). It is also important to consider that the published literature tends to overestimate the true magnitude of effects due to publication bias, or the tendency for articles to be published based on statistical significance (Gelman & Carlin, 2014). In sum, researchers evaluating the feasibility of a prospective meta-analysis can make informed judgments based on previous estimates and general knowledge of the published literature.

Study Size

Power analysis also requires expected values of the "typical" sample size of a study to be included in the meta-analysis, which varies considerably across different research areas (e.g., Hedges & Pigott, 2001; Jackson & Turner, 2017). For this expected value, area-specific contextual knowledge can provide reasonable guesses about the typical study size. For example, for research questions that can be addressed with standard questionnaires in large online studies, study sizes may be considerably large (e.g., $N = 500$); however, research questions related to those with neurodevelopmental disorders are likely much smaller (e.g., $N = 40$). Furthermore, there can be considerable variability in study sizes among studies in a specific area of research. For example, studies of neurodevelopmental disorders are generally small (e.g., $N = 40$), but studies evaluating specific subgroups of the population (e.g., those with a co-morbid condition) or using a specific methodology (e.g., eye-tracking) can be even smaller (e.g., $N = 20$). Therefore, the expected study size should be highly specific to the meta-analytic research question.

Number of Studies

How many studies are needed to conduct a meta-analysis? The answer lies at the core of power analysis: to determine if a given research question has enough published studies to warrant conducting a meta-analysis. As shown in the previous section of the paper, the number of studies required for adequate meta-analytic power depends on the expected effect size, study size, Type 1 error probability, test directionality (1-tailed vs. 2-tailed), type of statistical test (e.g., summary effect size, moderator analysis), and model assumptions (fixed- vs. random-effects model). For instance, under a fixed-effects model, which assumes that variability in effect sizes is purely due to sampling error, the number of studies needed to conduct a meta-analysis that has more power than an individual study is two (Valentine et al., 2009). However, in psychology, the fixed-effects assumption is rarely justified and difficult to defend; therefore the random-effects model, which assumes that effect size variability is due to sampling variability and between-study differences, is widely used. Under the random-effects model, Jackson and Turner (2017) proposed a general rule of thumb that at least five studies are required for a meta-analysis that is more informative than the largest individual study of that meta-analysis.

Fortunately, before researchers conduct a meta-analysis, there is generally some notion regarding the size of the published literature. For instance, published narrative or systematic reviews of the literature make estimating the total number of published studies trivial. However, for newer areas of research, where fewer studies have been published, it can be difficult to anticipate the total number of published studies on a given research question. If a previous systematic review of the literature included 40 empirical studies, it is reasonable to assume that at least 40 studies will be included in a meta-analysis. However, in a new area of research where the meta-analyst is only aware of 5 empirical studies, and no previous reviews have been conducted, it may be reasonable to assume that 5-10 studies will be included in the meta-analysis. For randomized controlled trials, which are commonly registered publicly (e.g., clinicaltrials.gov), researchers can proactively search these databases to get an idea of how many studies would be included in a meta-analysis. Likewise, a preliminary search of journal databases can provide an approximation for the number of total studies. In sum, determining the size of a literature pertinent to a specific meta-analytic research question may require extensive knowledge of the field or a brief search of the literature to arrive at a reasonable guess for the anticipated number of studies.

Statistical Heterogenity

In addition to these expected quantities, statistical power is greatly impacted by between-study variation in effect sizes (i. e., τ^2). For example, more data are required for meta-analyses with highly variable effect sizes to achieve the same statistical power as those with highly homogeneous effect sizes. Theoretically, between-study variation can be incorporated to estimate statistical power, but values of τ^2 are difficult to interpret and anticipate. As a result, estimates of τ^2 can be estimated from another index of heterogeneity (i. e., I^2; Higgins & Thompson, 2002). The I^2

statistic is a simple and intuitive index of statistical heterogeneity that quantifies the percentage of variation across studies that is not due to sampling variability (see Eq. 10).

For fixed-effects models, all variation in effect sizes is attributed to sampling error and thus $I^2 = 0$. However, this assumption is only justified in very specific situations. In the wake of the replication crisis, there has been increased emphasis on conducting direct replications. Replication studies attempt to replicate the methodology and results of a previous study in a new sample. Here, it may be reasonable to assume that all studies were measuring a single common effect size, and calculate statistical power under a fixed-effects model (i. e., assuming that $I^2 = 0$). Another example may be the meta-analysis of multiple randomized controlled trials of the same treatment. It is common for there to be multiple studies evaluating the efficacy of a specific treatment for a specific medical condition; in these situations, a fixed-effects model may also be justified given the homogeneity in methodology and implementation.

The vast majority of meta-analyses are conducted under a random-effects model, which assumes that some percentage of between-study variation is not due to sampling error (i. e., $I^2 > 0$). Anticipating values of I^2 can be difficult, but standard conventions of I^2 values equal to 25%, 50%, and 75% reflect small, moderate, and large degrees of heterogeneity have been established (Higgins & Thompson, 2002). In addition, recent evidence across 200 meta-analyses in psychology, including 12,065 effect sizes, show that the average I^2 value was 74% (Stanley et al., 2018), indicating large heterogeneity among effect sizes was common. Thus, unless researchers have a strong prior belief that heterogeneity will be small, a conservative approach would be to expect a moderate to high degree of heterogeneity as reflected in I^2 values of 50% or 75%. In sum, all of aforementioned parameter values should be informed by previous literature when possible, but there are some general rules of thumb when informed decisions cannot be made.

Tutorial: Using metapower to calculate statistical power for meta-analysis

metapower is a freely available open source R package. All power analysis calculations were derived from the most recent theoretical statistics and methodology. This includes power analysis for summary effect sizes (Hedges & Pigott, 2001; Jackson & Turner, 2017), test of homogeneity (Pigott, 2012), moderator analysis (Hedges & Pigott, 2004), and subgroup analysis (Pigott, 2020). All source code is publicly available on github (github.com/jasonwgriffin/metapower). *metapower* can be downloaded in R (cran.r-project.org) or Rstudio (rstudio.com). Users must have R version 3.6.0 or later. Because *metapower* is hosted on the Comprehensive R Archive Network (CRAN.R-project.org/package=metapower), it can be downloaded and attached directly in Rstudio using

```
install.packages("metapower")
library(metapower)
```

Example Research Question

Researchers are often interested in using meta-analysis to quantify the group difference in some outcome (e.g., working memory, face recognition, processing speed) between two independent groups (e.g., clinical population vs. typically developing, men vs. women). Since this type of research question reflects the mean difference between two independent groups on some outcome measure, it would be most appropriate to compute power based on Cohen's *d*. For this example, we will assume a previous review of relevant literature included 20 studies, that had study sizes around 40, and routinely reported effect sizes around 0.4. In addition, the outcome was measured across a wide variety of tasks; therefore, there is likely to be considerable statistical heterogeneity across studies (e.g., $I^2 = 75\%$). With this contextual knowledge, we can compute statistical power given these expected values.

Summary effect size

To calculate statistical power for detecting a mean difference between between the two groups, use the `mpower()` function with the aforementioned expected values while specifying the effect size type (`es_type = d`) as shown below

```
my_power <- mpower(effect_size = 0.4,
                   study_size = 40,
                   k = 20 ,
                   i2 = 0.75,
                   es_type = "d")
summary(my_power)
```

```
##  Power Analysis for Meta-analysis
##
##  Effect Size Metric:                d
##  Expected Effect Size:              0.4
##  Expected Study Size:               40
##  Expected Number of Studies:        20
##
##  Estimated Power: Mean Effect Size
##
##  Fixed-Effects Model                0.9998643
##  Random-Effects Model (i2 = 75%):   0.7956667
```

The first section of the output shows information about the user-specified input parameters. Note that for Cohen's *d* the study size reflects the entire sample, which are di-

vided equally per group. The second section presents the summary of the power analysis. Statistical power is shown under a fixed- and random-effects model. In this example, there appears to be adequate power to detect a meaningful summary effect size (i. e., 99.99%) according to standard power conventions (i. e., 80%). Under a random-effects model (I^2 = 75%), statistical power is 79.57%, indicating even with the expectation of considerable between-study variation, the published literature is likely sufficiently mature to reliably detect a statistically significant difference between the two groups.

Test of homogeneity

To compute statistical power for the test of homogeneity, which is often conducted by default in meta-analysis software, use `homogen_power()` with the same expected parameter values as before

```
my_homogen_power <- homogen_power(
                        effect_size = 0.4,
                        study_size = 40,
                        i2 = .75,
                        k = 20,
                        es_type = "d")
summary(my_homogen_power)
```

```
##  Power Analysis for Test of Homogeneity
##  in Meta-analysis
##
##  Effect Size Metric:                d
##  Expected Effect Size:              0.4
##  Expected Study Size:               40
##  Expected Number of Studies:        20
##
##  Estimated Power: Test of Homogeneity
##
##  Fixed-Effects Model (SD = 1)        0.0717026
##  Fixed-Effects Model (SD = 2)        0.1612124
##  Fixed-Effects Model (SD = 3)        0.3672535
##  Fixed-Effects Model (SD = 4)        0.660215
##  Fixed-Effects Model (SD = 5)        0.8901375
##
##  Random-Effects Model (i2 = 75%):    1
```

The results of this power analysis are reported separately for fixed- and random-effects models. This is because under fixed-effects models, the test of homogeneity evaluates if all effect sizes measure the same underlying true effect. Here, statistical power is presented for different standard deviations among the overall and study-specific effect sizes. As shown for this example, if effect sizes are highly variable (i. e., $SD = 4$), the power for test of homogeneity is large (66.02%); however, if effect sizes are less variable (i. e., $SD = 1$), then power is considerably lower (7.17%). This is because power to detect heterogeneity is increased when study-specific effects sizes are on average 4 standard deviations from the overall mean effect size.

For random-effects models, this test evaluates if between-study variability is greater than zero. As shown, under assumptions of large heterogeneity ($I^2 = 75\%$) there is 100% power to detect significant heterogeneity (i. e., $\tau^2 > 0$). Taken together, this summary output allows users to view power calculations under different model assumptions, varying within-study homogeneity, and for user-specified heterogeneity.

Moderator analysis

For this example, researchers are interested in the group difference in some outcome between two independent groups, which is expected to be moderate in magnitude (Cohen's d = 0.4). However, they may also want to evaluate if this group difference is moderated by the type of stimuli (e.g., verbal vs. visual). Specifically, researchers anticipate that the effect size will be smaller for verbal (Cohen's d = 0.2) compared to visual (Cohen's d = 0.6) stimuli. Since the number of moderator categories is 2 (i. e., verbal, visual), it is assumed that these types of studies are represented equally (i. e., k_{verbal} = 20; k_{visual} = 20) among the total number of studies (k_{total} = 40). To calculate statistical power to detect the difference between these two tasks, use `mod_power()` with at least five arguments: (1) number of groups (`n_groups = 2`), (2) effect size magnitudes (`effect_sizes = c(.2,.6)`, (3) study size, (4) total number of studies, and (5) effect size metric.

```
my_mod_power <- mod_power(n_groups = 2,
                    effect_sizes = c(.2,.6),
                    study_size = 40,
                    k = 20,
                    i2 = .75,
                    es_type = "d")
summary(my_mod_power)
```

```
##  Power Analysis for Moderator Analysis:
##
##  Effect Size Metric:                d
##  Number of Categorical Groups:      2
##  Groups:
##  Expected Effect Sizes:             0.2 0.6
##  Expected Study Size:               40
##  Expected Number of Studies:        20
##
##  Esimated Power: Moderator Analysis
##
##  Fixed-Effects Model:               0.7997139
##  Random-Effects Model (i2 = 75%):   0.2882369
```

There is 79.97% power to detect a meaningful group

difference in effect sizes between verbal and visual task paradigms under a fixed-effects model. Under a random-effects model (I^2 = 75%), power is considerably lower (i. e., 28.82%). Therefore, if the primary goal of the meta-analysis was to test the difference between verbal and visual task paradigms, more studies would be needed to have reasonable power at detecting these between-study differences.

Subgroup analysis

Researchers may also wish to test within-study group differences or subgroup differences. For example, in addition to the task paradigm, researchers may expect group differences between men and women within a study. In this case, researchers may expect that the summary effect size is larger for men (Cohen's d = 0.7) than women (Cohen's d = 0.1). To compute power to detect these subgroup differences, use the expected parameter values from before such that the number of studies is 20, the study size is 20 (i. e., 10 men, 10 women), and the overall effect is 0.4. Also, we must specify that the number of subgroups is 2 (`n_groups = 2`) and specify both subgroup effect sizes (`effect_sizes = c(.7,.1)`) for men and women respectively. To do this, use `subgroup_power()` with the aforementioned values

```
my_subgroup_power <- subgroup_power(
               n_groups = 2,
               effect_sizes = c(.7,.1),
               study_size = 40,
               k = 20,
               i2 = .75,
               es_type = "d")
summary(my_subgroup_power)
```

```
##  Power Analysis for Subgroup analysis:
##
##  Effect Size Metric:                d
##  Number of Subgroups:               2
##  Groups:
##  Expected Effect Sizes:             0.7 0.1
##  Expected Study Size:               40
##  Expected Number of Studies:        20
##
##  Esimated Power to detect subgroup differences
##
##  Fixed-Effects Model:               0.987483
##  Random-Effects Model (i2 = 75%):   0.5558747
```

Under these assumptions, there is reasonable power to detect subgroup differences between men and women under a fixed-effects models (i. e., 98.75%). However, there is considerably less power under a random-effects model (i. e., 55.59%). Therefore, like the calculated power for moderator analysis, the literature may not be sufficiently mature to test these within-study group differences.

Visualizing power analysis

Power analysis require researchers to make numerous assumptions about effect sizes, sample sizes, number of studies, and degree of heterogeneity that they *expect* to find. However, this information can sometimes be very difficult to know before data collection, and even if researchers have a good estimation informed by experience and previous literature, these assumptions can often be wrong. As a result, it is extremely informative to evaluate power curves across a range of values so that researchers can make informed decisions about the feasibility of a prospective meta-analysis. To facilitate this, *metapower* uses four plotting functions for each type of power analysis described above.

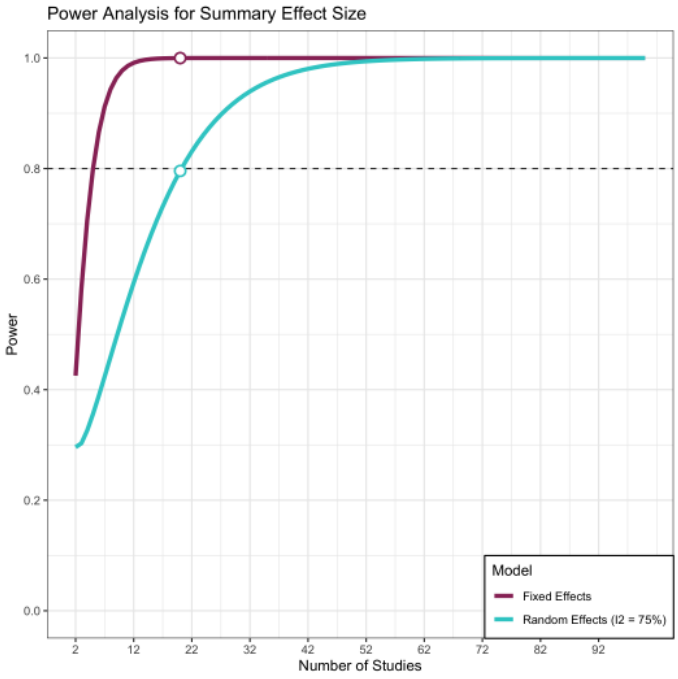
These functions include `plot_mpower()`, `plot_homogen_power()`, `plot_mod_power()`, and `plot_subgroup_power()`. To use these, simply wrap the power analysis object with these plotting functions (e.g., `plot_mpower(my_power)`). As shown in Figure 1, power curves under a fixed- and random-effects model for the summary effect size are presented as a function of the total number of studies. For the test of homogeneity, power curves are shown for fixed- and random-effects separately since the former tests within-study homogeneity, whereas the latter evaluates between-study variability (see Figure 2). Moderator and subgroup power analyses display power curves under fixed- and random-effects models for the respective between-study or within-study group differences (see Figures 3 and 4).

Visualization in power analysis is an essential tool for determining the feasibility of a meta-analysis. This is because statistical power is conditional on what we expect to find. However, our expectations can be wrong. In our example, we anticipated that the summary effect size would be 0.4, total number of studies would be 20, and average sample size would be 40. However, it is completely plausible that when conducting the meta-analysis, we find 40 studies. As shown in Figure 1, this would have a major impact on statistical power. By understanding statistical power across a range of possible outcomes, researchers can make informed decisions on the feasibility of a meta-analysis by being more or less conservative based on plausible range of outcomes.

Shiny Application

To maximize accessibility for users that are unfamiliar with the R programming environment, I developed a fully functional shiny application for *metapower* (https://jason-griffin.shinyapps.io/shiny_metapower/). Here, users

Figure 1 ■ Power curves generated from `plot_power()` for estimating a summary effect size. Power curves reflect statistical power as a function of the number of studies, *k*. By default, power curves are shown for the user-specified effect size under a fixed- and random-effects model. The range of values for the number of studies axis is 5 times that specified by the user. The point along the power curve reflects the current power estimate given the user-specified input parameters. The dashed horizontal line reflects 80% power.



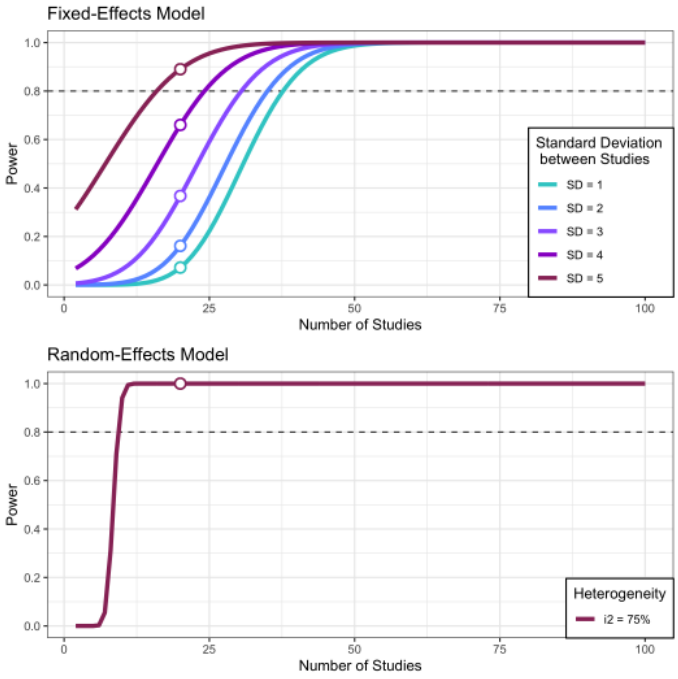
can access the web-based application to set parameters according to their research question and calculate statistical power for each of the statistical tests described here as well as visualize the respective power curves (see Figure 5).

Discussion

Power analysis plays an important role in determining the feasibility of prospective meta-analyses. Despite this, power analyses are rarely conducted in published meta-analyses. Two major barriers to the routine inclusion of power analyses in meta-analysis include: the complexity of calculating statistical power and the absence of available software for carrying out such power calculations. As demonstrated, power calculations are conditional on a number of different values, which include the total number of studies, study sizes, effect size magnitude, effect size metric, model assumptions, test directionality, Type 1 error probability, and degree of statistical heterogeneity. These values must be used in a complex series of formulas and equations for power to be calculated. *metapower* provides a tool for researchers to instantiate this complexity in just a single line of code.

In this tutorial, I have shown how *metapower* can be used to appropriately plan a meta-analysis by conducting an *apriori* power analysis without necessitating a deep knowledge of statistics or programming. Furthermore, *metapower* offers an effective and efficient solution to address the common question, "Is the literature large enough to warrant the time and effort it takes to conduct a meta-analysis?" It accomplishes this in a way that is computationally simple, easy-to-use, and entirely reproducible. *metapower* can also be integrated into a single workflow with other R packages. For example, numerous packages exist for the major steps of any meta-analysis, which include searching the literature, screening articles, and analyzing the data. *metapower* is another tool used to determine when a meta-analysis is warranted. As an added bonus, users unfamiliar with R can access a fully functional, web-based shiny application to implement the core functions of *metapower* using a graphical user interface. This shiny application requires no programming experience and can be opened in any web browser.

Figure 2 ■ Power curves generated from `plot_homogen_power()`. Power curves reflect statistical power as a function of the number of studies, *k*. The top panel displays power curves under a fixed-effects model. Power curves are shown for various levels of variation among effect sizes (i. e., SD among study-specific effect sizes). The bottom panel reflects the power curve under a random-effects model with the user-specified heterogeneity. The points along the power curve reflects the current power estimate given the user-specified input parameters The dashed horizontal line reflects 80% power.



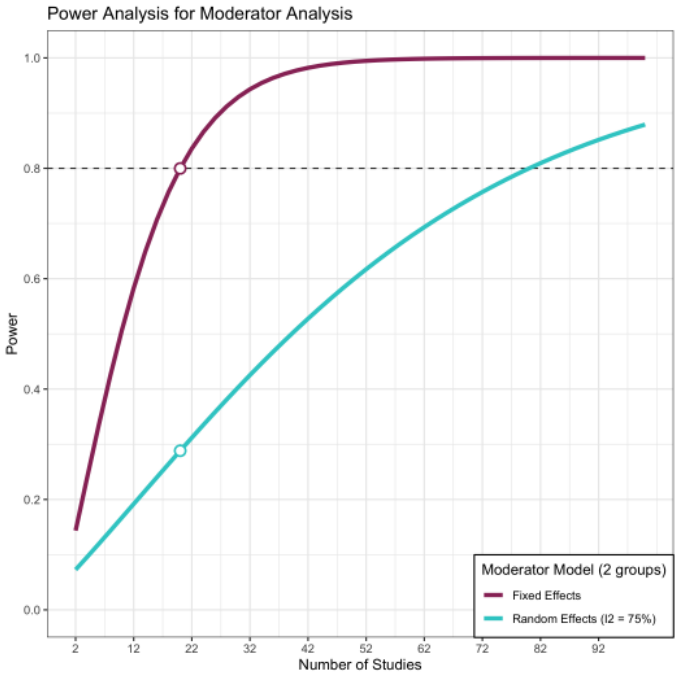
Limitations

Statistical power fundamentally depends on expecting the actual presence of an effect. That is, primary studies are conducted with the expectation of finding a significant difference between two groups or a significant association between two variables. However, in meta-analysis, we may actually expect that a summary effect size will not be statistically different than zero. In this way, meta-analyses sometimes demonstrate that an effect size is not distinguishable from zero despite individual studies suggesting otherwise. Currently, *metapower* is not capable of computing statistical power for tests of equivalence, which are are used to statistically reject the presence of an effect (Goertzen & Cribbie, 2010; Lakens, 2017; Lakens, Scheel, & Isager, 2018). Given that equivalence tests are simply a reformulation of the standard null-hypothesis significance testing framework, it is possible to calculate statistical power for these types of statistical tests for primary research (e.g., Shieh, 2016; Lakens, 2017). In theory, these power calculations could be adapted for meta-analysis in the future.

Similarly, statistical power is inextricably associated with specific statistical tests that are focused on rejecting the null hypothesis (e.g., $H_0 = 0$). However, researchers may not be interested in simply showing that an effect size is different than zero. For instance, it can be valuable to power a study capable of estimating an effect size within a specific margin of error. One reason for this would be that a clinically significant effect in some context would only be one that was at or above a certain value. In this case, it would be useful to power a meta-analysis for precision - as opposed to statistical power - in order to estimate a confidence interval that was above a certain value. Given that these methods are being developed in primary research (see Goulet-Pelletier & Cousineau, 2018), they may be adapted to meta-analysis in the future.

As it relates to moderator and subgroup analyses, *metapower* assumes that categorical groups (e.g., moderator or subgroup analyses) are equal in number of studies and sample size. For situations where there are unequal group sizes, the power estimates provided by *metapower* will overestimate statistical power. However, given that

Figure 3 ■ Power curves generated from `plot_mod_power()`. Power curves reflect statistical power as a function of the number of studies, *k*. Power curves reflecting the power to detect between-study differences (categorical moderators) among studies are presented under a fixed- and random-effects model. The points along the power curve reflects the current power estimate given the user-specified input parameters The dashed horizontal line reflects 80% power.



power analyses are conducted *apriori*, it is untenable to make guesses about how many studies will comprise a particular category. In addition, *metapower* is limited in that all calculations assume that each study contributes a single effect size to the meta-analysis. For meta-analyses that incorporate hierarchical structure (i. e., multiple effect sizes from a study), power cannot currently be calculated for these more complex data structures.

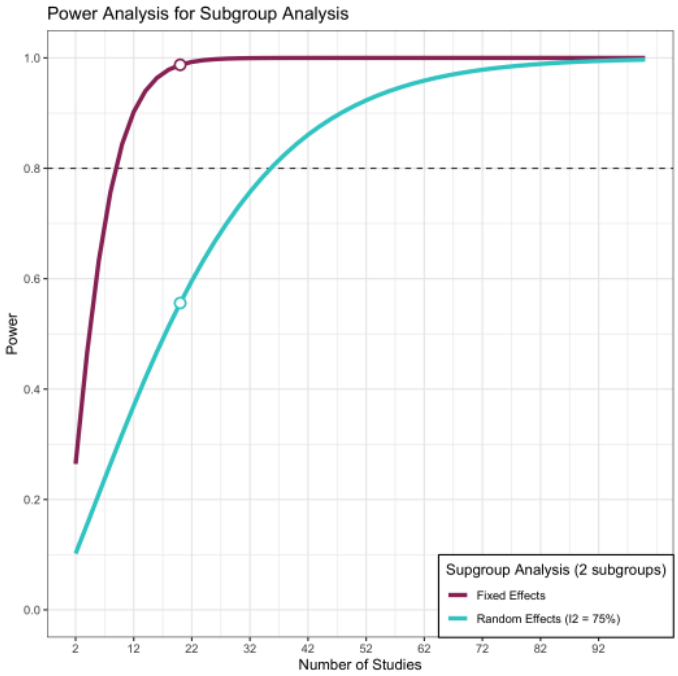
Future Development

metapower currently accommodates the three most commonly used effect sizes in meta-analysis (i. e., Cohen's *d*, correlation, and odds ratio); however, future versions of *metapower* plan to incorporate additional effect size metrics to handle these research designs (see Goulet-Pelletier & Cousineau, 2018). For example, *metapower* currently handles Cohen's *d* for between-subject designs, but future developments can include calculations for Cohen's *d* in one-sample or correlated measurements (e.g., Lakens, 2013). Additionally, multilevel modeling (e.g., two- and three-level models) is becoming an increasingly popular method for meta-analysis since it can include multiple effect sizes from source studies while modeling the statistical dependence among observations (Assink & Wibbelink, 2016). These models achieve more statistical power than traditional methods, but there is no agreed upon method for calculating statistical power for these more complex models. Future versions of *metapower* will incorporate these more advanced power caluclations as the methods become available and validated.

Conclusion

The influence of meta-analyses in the scientific community is ubiquitous. Given their impact, high-quality meta-analyses should be conducted when there are a sufficient number of published studies given a particular research question. To do this objectively, power analysis specific to the research question (summary effect size, moderator analysis, subgroup analysis) should be conducted during the planning phase of a prospective meta-analysis. To my knowledge, *metapower* is the first freely available and easy-to-use software package that allows researchers to do this.

Figure 4 ■ Power curves generated from `plot_subgroup_power()`. Power curves reflect statistical power as a function of the number of studies, *k*. Power curves reflecting the power to detect within-study differences (subgroups) among studies are presented under a fixed- and random-effects model. The points along the power curve reflects the current power estimate given the user-specified input parameters The dashed horizontal line reflects 80% power.



Authors' note

I would like to thank Flora Oswald and Courtney R. Gerver for their feedback on the manuscript.

References

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in r: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*(3), 154–174. doi:10.20982/tqmp.12.3.p154

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ Open*, *7*(2), e012545. doi:10.1136/bmjopen-2016-012545

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2009). A sas macro for statistical power calculations in meta-analysis. *Behavior Research Methods*, *41*(1), 35–46. doi:10.3758/brm.41.1.35

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129. Retrieved from http://www.jstor.org/stable/3001666

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:10 . 1177 / 0956797613504966

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*(1), 160–170. doi:10.1186/1471-2288-11-160

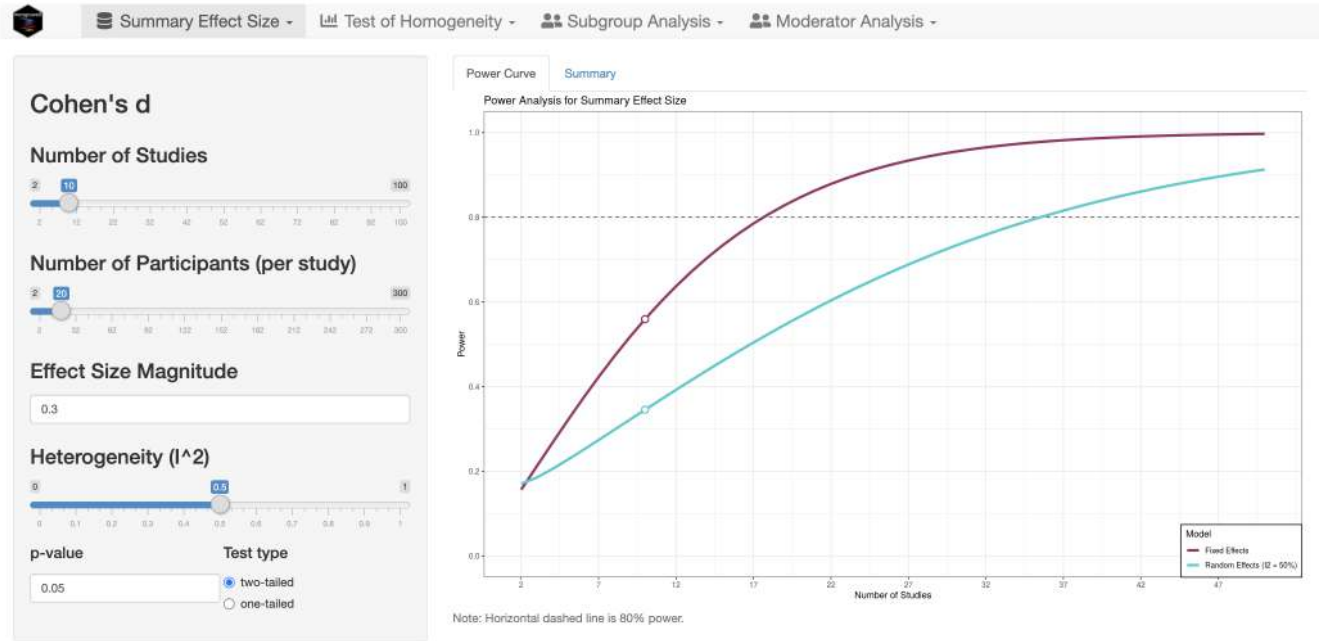
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. doi:10.3758/bf03193146

Gelman, A., & Carlin, J. (2014). Beyond power calculations. *Perspectives on Psychological Science*, *9*(6), 641–651. doi:10.1177/1745691614551642

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 527–537. doi:10.1348/000711009x475853

Gopalakrishnan, S., & Ganeshkumar, P. (2013). Systematic reviews and meta-analysis: Understanding the best

Figure 5 ■ Shiny application interface for *metapower*. The top menu tabs correspond to the different types of power analysis, including summary effect size, test of homogeneity, moderator analysis, and subgroup analysis. The left panel controls the expected input parameters. The right panel displays the power curves under the Power Curve panel (default), but can also view the summary statistical power results using the Summary tab.



evidence in primary healthcare. *Journal of Family Medicine and Primary Care*, *2*(1), 9–14. doi:10.4103/2249-4863.109934

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. doi:10.20982/tqmp.14.4.p242

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. doi:10.1037/1082-989x.6.3.203

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*(4), 426–445. doi:10.1037/1082-989x.9.4.426

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

Jackson, D., & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods*, *8*(3), 290–302. doi:10.1002/jrsm.1240

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas. *Frontiers in Psychology*, *4*, 863–893. doi:10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi:10.1177/2515245918770963

National Institute of Health. (2018). Enhancing reproducibility through rigor and transparency. Retrieved from https://grants.nih.gov/policy/reproducibility/index.htm

Pigott, T. D. (2012). *Advances in meta-analysis*. doi:10.1007/978-1-4614-2278-5

Pigott, T. D. (2020). Power of statistical tests for subgroup analysis in meta-analysis. In J. C. C. Ting, S. Ho, & (. D.-G. Chen (Eds.), *N* (pp. 347–368). Design: Springer International Publishing.

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. doi:10.3102/0034654319877153

Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PLOS ONE*, *11*(9), e0162093. doi:10.1371/journal.pone.0162093

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. doi:10.1037/bul0000169

Thorlund, K., Imberger, G., Walsh, M., Chu, R., Gluud, C., Wetterslev, J., … Thabane, L. (2011). The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis? a simulation study. *PLoS ONE*, *6*(10), e25491. doi:10.1371/journal.pone.0025491

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2009). How many studies do you need? *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247. doi:10.3102/1076998609346961

Citation

Griffin, J. W. (2021). Calculating statistical power for meta-analysis using metapower. *The Quantitative Methods for Psychology*, *17*(1), 24–39. doi:10.20982/tqmp.17.1.p024

Received: 11/11/2020 ∼ Accepted: 16/02/2021