A Practical Guide to Estimate Statistical Power in Meta-Analysis

Jin Liu
University of South Carolina

Yu Bao
James Madison University

Statistical power is important in meta-analysis as in primary studies to detect the existing treatment effects in target populations. The mathematical computation of the approximated power (i.e., analytical power) in meta-analysis has been developed. Alternatively, Monte Carlo simulation can be used to estimate statistical power. The current study was conducted to identify the discrepancy between analytical power and simulated statistical power for the hypothesis test of two group means. Results indicated noticeable discrepancies between analytical and statistical power under certain conditions. To conclude, recommendations for the average sample size and number of studies were provided to practical researchers who seek to meet the desired statistical power in meta-analysis. We provided `R` code for applied researchers to obtain power estimates through simulation.

Meta-analysis is a quantitative analysis method that synthesizes the results of multiple empirical studies on the same topic. Researchers combine the effect size estimates from a set of primary studies to obtain a common effect size estimate for a summary result (Hedges & Pigott, 2001). Meta-analysis has gained popularity in social sciences because it demonstrates the potential to overcome the shortcomings of a single primary study which may be limited in sample size, estimate precision, and generalizability (Ellis, 2010).

In inferential statistics, statistical power is defined as the probability of rejecting a false null hypothesis. A higher statistical power represents a higher probability of detecting statistical differences in hypothesis testing. Statistical power can be applied to diverse statistical tests, and researchers have expressed their concerns over power in meta-analysis. Cafri, Kromrey, and Brannick (2010) asserted that "power analysis is more important in meta-analysis because such studies quantitatively summarize a whole body of research and influence more on theory and practice."

Power in meta-analysis is influenced by multiple factors: population effect size, Type I error rate, sample size, and number of studies (Borenstein et al., 2011; Cohen, 1988; Ellis, 2010; Liu, 2013). Low statistical power in meta-analysis is a concern that has been widely discussed by considering the influencing factors (e.g., Cook & Hatala, 2014; Jackson & Tuner, 2017; Quintana, 2023; Valentine et al., 2010). The number of studies does not always increase statistical power and between-study variances should be considered with varied population effect sizes across studies (Cohen & Becker, 2003). Low statistical power was presented in regression models and rank correlation tests in small studies (Stern et al., 2000). Field (2001) investigated different meta-analytical models for correlation coefficient studies and concluded that power could be low with a small number of studies, sample sizes and/or population effect sizes. The results can be biased if researchers run an analysis with insufficient statistical power (Ellis, 2010). One difficulty in estimating power is to identify population effect size (Hedges & Pigott, 2001) as researchers rarely know the "true" population effect size in practice.

Hedges and Pigott (2001) used the averaged variance across studies to approximate the combined variance estimate to simplify the power computation. However, researchers rarely identify primary studies with equal variances in a research scenario. The power estimation accuracy in the approximate procedure has never been thoroughly vetted. An alternative way to estimate statistical power in meta-analysis is through simulation, which does not need to average variance across studies and is considered a more accurate method compared with the existing power approximation functions. However, practical researchers may be impeded by the required technical expertise in simulation compared to the use of power formulas. Therefore, the current study was conducted to show the discrepancy between analytical power and simulated statistical power under various conditions. Afterward, recommendations of average sample size and number of studies were provided to practical researchers who seek to meet the desired statistical power in meta-analysis. Finally, we provided `R` code to applied researchers to obtain power estimates through simulation. In the next section, we reviewed the key concepts of estimating power in meta-analysis.

Prospective and Retrospective Power

There are two main types of power analyses – prospective and retrospective power analysis (O'Keefe, 2007). A prospective power analysis is a part of research planning and is completed prior to implementing a study. It is normally used to estimate the required sample size by considering the other parameters in hypothesis testing. For instance, when researchers intend to conduct a replicated study, they need to search past research to determine the potential population effect sizes. Together with the pre-specified type I error rate and power, the necessary sample needed for error control is determined in the retrospective power analysis. A retrospective power analysis is conducted after a study. For example, low statistical power could be the reason for a research scenario in which we fail to reject a null hypothesis yet claim the existence of a treatment effect. Moreover, scholars are cautious about the retrospective power analysis and suggest that the power estimate is not accurate based on the effect size obtained from the sample due to the possible large sampling error. The retrospective power analysis should utilize the population effect sizes from previous studies of a similar nature (Thomas, 1997).

Fixed and Random Effects Models

A fixed-effects model assumes an equal population effect size across individual studies. By contrast, a random-effects model treats the population effect sizes from individual studies as a random sample of all possible effect sizes with an underlying distribution (e.g., normal distribution). In the fixed-effects model, the only reason that the effect size varies is the random error (within-study variances). In a random-effects model, the effect size can be influenced by the random error (within-study variances) and the effects of different studies (between-study variances).A fixed-effects model is designed to make inferences about a population from the sampled studies, referred to as a conditional inference, while a random-effects model is designed to draw inferences about a population that is larger than the sampled studies, referred to as an unconditional inference (Hedges & Vevea, 1998).

Effect Size

Researchers often use a *p*-value to determine whether to reject the null hypothesis, referred to as statistical significance. Effect size is used to measure the magnitude of an effect, which is referred to as practical significance. As suggested by Cohen (1990, page 1310), "the primary product of a research inquiry is one or more measures of effect size, not p values. "Effect size is important not only in primary studies but also in meta-analysis studies as scholars combine effect sizes from primary studies to obtain an overall estimate of the treatment effect. There are two major families of effect size: *d* (e.g., odds ratio, Cohen's *d*; differences between groups) and *r* (e.g., Pearson correlation, Cohen's *f*; measure of association) (Ellis, 2010). The current study focuses on two independent group tests that are applicable to research scenarios in experimental designs. In a meta-analysis, the test statistics is computed by the mean and the variance of the effect sizes from primary studies. One then can make a statistical conclusion about whether to reject or fail to reject the null hypothesis based on the test statistic.

Cohen's *d* is the effect size index of each study used to investigate the mean differences between two groups. The formula to calculate Cohen's *d* is:

$$d = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p} \quad (1)$$

In this formula $\bar{Y}_1$ and $\bar{Y}_2$ are the means for two groups, and s_p is the pooled standard deviation of two groups.

Note that the assumption of pooled standard deviation is not always met in practice, especially when the sample sizes of two groups are not equal. In addition, *d* tends to overestimate the population variance. The bias can be removed by Hedge's *g*, which weighs the standard deviation by its sample size (Hedges, 1981). It can then be converted from *d* by multiplying the following correction index:

$$J = 1 - \frac{3}{4df - 1} \quad (2)$$

where the *df* as the degree of freedom is equal to the overall sample size – 2.

$$g = J * d. \quad (3)$$

Analytical Statistical Power

One of the major meta-analysis methods was developed by Hedges and his colleagues (Hedges & Vevea, 1998). The analytical power formulas were developed by Hedges and Pigott (2001). The fixed and

random-effects models were discussed separately using these resources (Hedges & Pigott, 2001; Hedges &Vevea, 1998; Liu, 2013).

Fixed-effects Meta-analysis. The common effect size estimate for the *i*th individual study is equal to the standardized mean difference between the treatment condition and control condition (Cohen, 1988)

$$d_i = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p}. \quad (4)$$

In this formula $\bar{Y}_1$ and $\bar{Y}_2$ are the means for two groups, and s_p is the pooled standard deviation in a two independent sample t- test. The effect size estimate d_i corresponds to a population effect size of θ_i.

The t-statistics for the *i*th study t_i is computed as

$$t_i = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{\frac{1}{\bar{n}_{1i}} + \frac{1}{\bar{n}_{2i}}}} \quad (5)$$

where $\bar{n}_{1i}$ and $\bar{n}_{2i}$ are the treatment group size and the control group size of the *i*th study in a meta-analysis. The effect size for the *i*th study d_i can be expressed in terms of t_i,

$$d_i = t_i \sqrt{\frac{1}{n_{1i}} + \frac{1}{n_{2i}}}. \quad (6)$$

Effect size variances within each individual study (v_i) are approximately estimated by the pre-existing formulas in analytical and simulated power (Goulet-Pelletier et al., 2018; Hedges & Vevea, 1998; Hedges & Pigott, 2001). Previous research has addressed the major bias caused by using approximated variance formulas, especially in small samples (Goulet-Pelletier et al., 2018). The actual formula of effect size variance should be utilized in both procedures to provide more accurate estimates (Hedges, 1981).

$$v_i = \frac{v}{v-2} * \frac{2}{\tilde{n}}\left(1 + PES^2 * \frac{\tilde{n}}{2}\right) - \frac{PES^2}{J^2} \quad (7)$$

where v is the degree of freedom equal to the overall sample size – 2, $\tilde{n}$ is the harmonic mean of the sample sizes in two groups, PES is the standardized population effect size and J is the correction formula (Equation (2)). Equation (7) is used in the current study for both analytical and simulated statistical power to improve the estimation accuracy.

The corrected variance of Hedge's g is

$$v_{gi} = J^2 * v_i. \quad (8)$$

The null hypothesis for the population effect size for each individual study is $\theta_1 = \theta_2 \ldots = \theta_i = \cdots = \theta = 0$. The fixed-effects model then becomes

$$d_i = \theta + e_i \quad (9)$$

Where e_i has a mean of zero and a variance of v_i. The common effect size can be estimated by pooling the estimates from individual studies, where the effect size estimates from those studies are weighted by the sampling variances of individual studies. An effect size estimate from a study with a larger sample size will receive more weight because the estimate is more precise and with a smaller sampling variance. The weight w_i is the reciprocal of the variance term v_i ($w_i = 1/v_i$). The estimate of the common effect size $\hat{\theta}$ is the weighted average:

$$\hat{\theta} = \bar{d} = \sum_{i=1}^{I} w_i d_i \,/\, \sum_{i=1}^{I} w_i \quad (10)$$

The variance of the weighted average v is simply the reciprocal of the sum of weights.

$$v = 1/\sum_{i=1}^{I} w_i \quad (11)$$

An approximate *Z*-test can be used to test the null hypothesis in which the common effect size θ is zero, using the weighted average estimate.

$$Z = \frac{\bar{d} - 0}{\sqrt{v}} \quad (12)$$

The *p*-value in a two-sided test is the probability of obtaining a z statistic at least deviant from the center of the standard normal distribution as the computed one. A small *p*-value less than or equal to five percent will result in the rejection of the null hypothesis, which is followed by a pronouncement of a non-zero common effect size. A confidence interval can be computed to accompany the significance test for the common effect size.

The 95% confidence interval for the common effect size is estimated as:

$$\bar{d} \pm 1.96 * \sqrt{v}. \quad (13)$$

When the alternative hypothesis is true, the common effect size is equal to a non-zero constant θ_a. The Z statistics follows a non-central normal distribution Z' with a non-centrality parameter λ:

$$\lambda = \frac{\theta_a}{\sqrt{v_{\theta_a}}}. \tag{14}$$

The current procedure assumes a common variance in all studies ($\bar{v}_i$) to simplify the v for power computation because it can greatly simplify the variance formula. Variances of all the studies are considered to be approximately equal, that is, $v_1 = v_2 \ldots = v_i = \cdots = v_I$. It is noted that this an ideal assumption because the variances of individual studies are not identical. This approximation results in the underestimation of statistical power (Hedges & Pigott, 2001).

The variance v can be simplified to

$$v_{\theta_a} = \frac{\bar{v}_i}{I}, \tag{15}$$

where $\bar{v}_i$ is the average of the overall variance for all studies. $\bar{v}_i$can be computed by using the average sample sizes for n_{ei} and n_{ci} and the estimated $d_i = \theta_a$. The variance thus computed is an approximation of the actual variance (Hedges & Pigott, 2001),

$$\lambda = \frac{\theta_a}{\sqrt{v_{\theta_a}}} \approx \frac{\theta_a}{\sqrt{\frac{\bar{v}_i}{I}}} = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i}}. \tag{16}$$

In order to simplify the calculation, the treatment group and control group sample sizes are assumed to be equal ($\bar{n}_{1i} = \bar{n}_{2i} = n$):

$$\bar{v}_i \approx \frac{\bar{n}_{1i}+\bar{n}_{2i}}{\bar{n}_{1i}\bar{n}_{2i}} + \frac{\theta_a^2}{2(\bar{n}_{1i}+\bar{n}_{2i})}. \tag{17}$$

The non-centrality parameter in the meta-analysis can be changed to

$$\lambda = \frac{\sqrt{I}\theta_a}{\sqrt{\frac{2}{n}+\frac{\theta_a^2}{4n}}}, \tag{18}$$

where θ_a is the standardized mean difference common to all individual studies.The term $\theta_a^2/4n$ is very small, especially when the population effect size (θ_a) is small and the sample size for each group (n) is large. Dropping the negligible term in λ yields

$$\lambda \approx \sqrt{I}\theta_a\sqrt{\frac{n}{2}}. \tag{19}$$

The power function for the two-sided test is, therefore,

$$\begin{aligned} 1 - \beta &\approx P[|Z'(\lambda)| \geq Z_0] \\ &= 1 - \Phi(Z_0 - \lambda) + \Phi(-Z_0 - \lambda). \end{aligned} \tag{20}$$

Random-Effects Meta-analysis. In the random-effects model, the effect size estimates from individual studies have an underlying distribution. The effect size estimate d_i follows a normal distribution with the mean of θ_i and the variance of v_i, that is,

$$d_i = \theta_i + e_i. \tag{21}$$

The parameter θ_i has an underlying distribution with a mean θ and a variance of τ. It is assumed that the population effect sizes from individual studies follow a normal distribution. Unlike the fixed-effects model, the random-effects model suggests that the effect sizes bounce around the grand average effect size θ. Thus, d_i becomes

$$d_i = \theta + \alpha_i + e_i. \tag{22}$$

The random effect α_i is due to different individual studies with its variance τ.The random effect e_i is the sampling error of d_i with its variance of v_i.

The random-effects model can be reformulated so that the same procedure can be applied to the fixed-effects model. The random-effects α_i and $e_i$can be combined into a single error term e_i^*. Thus d_i becomes

$$d_i = \theta + e_i^*, \tag{23}$$

where,

$$e_i^* = \alpha_i + e_i, \tag{24}$$

$$v_i^* = Var(e_i^*) = v_i + \tau. \tag{25}$$

Now the random-effects model can be treated as a special case of the fixed-effects model with a more complex variance v_i^*. An approach that is similar to that used for the fixed-effects can be followed. The weight w_i^*in the random-effects model is the reciprocal of the variance term v_i^* ($w_i^* = 1/v_i^*$).

The weighted mean of the random-effects model can be computed:

$$\bar{d} = \frac{\sum_{i=1}^{I} w_i^* d_i}{\sum_{i=1}^{I} w_i^*}. \quad (26)$$

The variance of $\bar{d}$ is:

$$v^* = \frac{1}{\sum_{i=1}^{I} w_i^*} \quad (27)$$

where v_i is estimated as before. Hedge's g correction is used for the random-effects model and is similar to the fixed-effects model. The varince τ can be estimated, according to Hedges and Vevea (1998):

$$\tau = \frac{Q-(k-1)}{c} \quad (28)$$

where,

$$Q = \sum_{i=1}^{I} w_i \, (d_i - \bar{\mathrm{d}})^2, \quad (29)$$

$$c = \sum_{i=1}^{I} w_i - \frac{\sum_{i=1}^{I} {w_i}^2}{\sum_{i=1}^{I} w_i} 0. \quad (30)$$

An approximate Z test can be used to test the null hypothesis($\theta = 0$), based on the weighted average estimate:

$$Z = \frac{\bar{d}-0}{\sqrt{v^*}}. \quad (31)$$

A small *p*-value less than or equal to five percent will result in the rejection of the null hypothesis, which is followed by a declaration of a non-zero common effect size. A confidence interval can be computed to accompany the significance test for the common effect size. The 95% confidence interval for summary effect is estimated as

$$\bar{d} \pm 1.96 * \sqrt{v^*}. \quad (32)$$

Under the alternative hypothesis, the common (grand average) effect size is equal to a non-zero constant θ_a. The Z test follows a non-central normal distribution Z' with a non-centrality parameter λ,

$$\lambda = \frac{\theta_a}{\sqrt{{v^*}_{\theta_a}}}. \quad (33)$$

Conjectures are needed to approximate ${v^*}_{\theta_a}$.One assumes that the sample sizes are equal among individual studies, following Hedges and Pigott (2001). So one obtains $v_1^* = v_2^* \ldots = v_i^* \ldots = v_I^*$. The variance v^* can be computed as:

$${v^*}_{\theta_a} = \frac{\bar{v}_i^*}{I}. \quad (34)$$

Then the non-centrality parameter can be rewritten as

$$\lambda = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i^*}} \quad (35)$$

where the overall variance for all studies is equal to

$$\bar{v}_i^* = \bar{v}_i + \tau \approx \frac{\bar{n}_{1i}+\bar{n}_{2i}}{\bar{n}_{1i}\bar{n}_{2i}} + \frac{\theta_a^2}{2(\bar{n}_{1i}+\bar{n}_{2i})} + \tau. \quad (36)$$

The λ in the random-effects model is usually smaller compared with the non-centrality parameter in fixed-effects model. The ratio of $\bar{v}_i$ (within-study variance) and τ (between-study variance) can be denoted by $p = \tau/\bar{v}_i 0$. Thus the non-centrality parameter can be expressed in this way,

$$\lambda = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i+\tau}} = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i(1+p)}}. \quad (37)$$

After setting up the p ratio, λcan be calculated in the same way as in the fixed-effects model. Although the research findings from the random-effects model can be generalized within a broader context, the fixed-effects meta-analysis tends to have higher power than the random-effects meta-analysis as the between-study variance is under consideration. The power function for a two-sided test is the same as the fixed-effects model:

$$\begin{aligned} 1 - \beta &\approx P[|Z'(\lambda)| \geq Z_0] \\ &= 1 - \Phi(Z_0 - \lambda) + \Phi(-Z_0 - \lambda). \end{aligned} \quad (38)$$

Simulated Statistical Power

The Monte Carlo simulation studies involve generating data from computer programs to study the performance of the statistical estimates under different conditions (Hutchinson & Bandalos, 1997). The idea of simulating statistical power uses the same logic of hypothesis testing. If there is no real effect, researchers expect to retain the null hypothesis most of the time and control the Type I error. If there is a

real existing effect, researchers expect to have a high power to reject the null hypothesis. The simulation can be used to check the performance of Type I error and power by repeating the same statistical procedures many times under model assumptions. Following the typical analytical procedures of meta-analysis described above, the effect size can be simulated by assuming a fixed population effect size. The rejection rates can be computed by averaging the results obtained from the replications. In each replication, a *p*-value is retained to make a statistical decision (reject or retain the null hypothesis). Finally, all simulated *p*-values are stored in the output. To verify the Type I error, we used 5% as the alpha level to decide whether the population effect is nonzero. The same strategy applies to simulated statistical power or one minus the Type II error. The simulated statistical power equals the proportion of the rejected null hypotheses among all the simulated tests when the treatment effect is non-zero. Simulated statistical power has been used by researchers in meta-analysis and other research designs (Arnold et al., 2011; Field, 2001).

The power guidance of meta-analysis has been provided by researchers (Valentine et al., 2010) with analytical power functions. SAS macros or R functions have been developed to facilitate power computation (Cafri et al., 2009; Borenstein et al., 2011) in meta-analysis. However, the accuracy of such guidance remains unclear with the approximation of variance across studies as well as the unadjusted effect size and the effect size variances. Although power of other effect sizes, such as *r*, has been investigated using simulation in meta-analysis (Field, 2001). The study did not investigate the discrepancy between analytical and simulated power. Recently, researchers have noted a discrepancy between simulated and analytical power in meta-analysis in retrospective power analysis and have recommended to use simulated power as it is more accurate by considering the uncertainty of between-study variances using effect size of log-risk ratio (Jackson & Turner, 2017). However, no studies have been conducted to investigate simulated statistical power with the effect size *d* and to identify its discrepancy with analytical power. In this study, we addressed the influence of unbalanced design on power in meta-analysis. We also aimed to improve the accuracy of power functions by using Hedge's *g* and true effect size variances. Power guidance was provided after accurate power estimates were obtained.

The following simulation conditions were defined based on similar studies (e.g., Field, 2003) utilizing effect size *r* and researchers' pilot study:

(1) Sample size: The average sample size varied in different meta-analysis studies. In the current study, the sample size ranged from 30 to 100 (i.e., 30, 40, 50, 60, 80, and 100). The average sample size in the real meta-analysis is usually large, but this study was intended to check the influence of a small sample size, which was more likely to be associated with low statistical power. Thus, a sample size larger than 100 was not considered. In practice, the sample sizes among individual studies are unequal. Therefore, a truncated binomial distribution was used to generate integer positive numbers to meet the requirements of a sample size. The sample sizes for the two groups in an individual study were assumed to be equal (i.e., balanced design) initially and then the sample size of two groups varied based on a ratio of 1:2 to study the influence of unbalanced design on statistical power.

(2) Effect size: We used no effect (0), a small effect (0.1, 0.2, and 0.3), a moderate effect (0.5), and a large effect (0.8). These effect sizes were selected based on Cohen's guidelines (1988). In addition, small effect sizes were common in practice. Hattie (2009) synthesized over 800 meta-analyses related to achievement. The overall distribution of all the effect sizes indicated that many of the effect sizes were small, i.e., under 0.4 (72 out of 138 studies). Therefore, the population effect size in the lower range was studied more carefully.

(3) Number of studies: The number of studies ranged from 5 to 80 (i.e., 5, 10, 20, 50, 80). These numbers were chosen based on the real meta-analysis datasets. For instance, studies of children's self-conscious emotions (Else-Quest et al., 2012) had different numbers of studies with different emotion aspects ranging from 17 to 307. Different study numbers were used to cover most of the practical situations, and the number of studies higher than 80 were normally with satisfactory statistical power and were not considered in the simulation.

(4) Fixed- and random-effects: The fixed-effects model and two random-effects models were considered separately following these procedures. The population effect size across studies was the same in the fixed-effects model, while the population effect size of the random-effects model was assumed to follow a normal distribution with a mean of the average population effect size and a standard deviation of 0.1. In addition, the simulated and analytical power used the same

between-study and within-study variance ratio to guarantee the comparability of simulated and analytical power in the random-effects model.

The total simulated scenarios were based on these factors: 6 population effect sizes (0, 0.1, 0.2, 0.3, 0.5, 0.8), 6 average sample sizes (30, 40, 50, 60, 80, 100), 2 group size ratios (balanced or unbalanced), 5 studies (5, 10, 20, 50, 80) and 2 models (fixed-effects model, random-effects model). The Type I error rate was set to 0.05 and power of 0.8 was considered the optimal cut-off point. The study was conducted using R software (2021).

There were two layers of simulation to report the stability of the simulation results. First, in each meta-analysis, a *p*-value was saved. This process was repeated 1,000 times to obtain one estimate (number of runs that reject the null hypothesis/100) based on the assigned Type I error rate of 0.05. When the population effect size was not zero, the percentage of rejecting the null hypothesis was equal to statistical power. When the population effect size was zero, the percentage of rejecting the null hypothesis was equal to the actual Type I error rate. This process was repeated 1000 times to obtain standard error (i.e., standard deviation of the power or Type I error estimates).

Before power estimation, Type I error rates were examined with the population effect size of 0 for simulated statistical power. This check was necessary because the Type I error can affect Type II error and, in turn, the statistical power. After that, the analytical power was computed by the formulas in Hedges and Pigott (2001) and the simulated power was computed following the above-mentioned procedure. Analytical power and simulated power were saved within tables under different conditions described above. The standard error was only applicable to simulated power. The discrepancy between statistical and analytical power was interpreted by calculating differences between two estimates under the same condition. We expected to find that simulated power was higher than analytical power which underestimates statistical power (Hedges & Pigott, 2001).

Next, power guidance was provided based on the simulated statistical power in tables. Under various population effect sizes and average sample sizes per study, the minimum number of studies needed was suggested to reach the power of 0.8. Under various population effect sizes and number of studies, average sample sizes per study were suggested to reach the power of 0.8. The results for the balanced- and unbalanced-designs were displayed separately. Such guidance provided suggestions to practical researchers at the initial stage of planning a meta-analysis study. Finally, by utilizing real data resources in the meta-analysis, we illustrated how researchers could use the provided R code to estimate the statistical power of a meta-analysis with standard deviation.

Results

Discrepancy between Analytical and Simulated Power

The actual Type I error rate was checked through power simulation before investigating the discrepancy between simulated and analytical power. Using a nominal α=0.05, it was clear that Type I errors were under control and limited to the purported five percent under all conditions.

Table 1 to Table 4 show the simulated power values with standard errors and analytical power values across population effect size, average sample size, number of studies for fixed effects, and random effects models under balanced and unbalanced designs. The standard errors were minimal across all conditions indicating the stability of power estimation, so average power estimates were used in the following analysis.

First, the general patterns of statistical power were interpreted. Statistical power was understandably higher when the population effect size, the average sample size, and the number of studies were larger in general. When the population effect size was at 0.8, the simulated power and analytical power estimates were close to 1 without any discrepancy no matter what sample size, number of studies, or designs were used. In other words, the influence of other parameters became inconsequential under such conditions. However, this was not true for the average sample size or number of studies. The largest average sample size (100) itself was not enough to reach a power of 0.8 when the number of studies and the population effect size were small. Similar conclusions were obtained for the largest number of studies. As expected, fixed-effects models had higher power estimates than random-effects models when other parameters were held constant (Tables 1 and 2; Tables 3 and 4). Multiple conditions showed noticeable differences of higher than 0.1. The unbalanced design was associated with lower statistical power in fixed-effects models and random effects models across conditions compared to balanced design (Table 1 and Table 3; Table 2 and

Table 1. Statistical Power of the Fixed-Effects Model (Balanced Design).

Average Sample Size	Power Simulation					Power Function				
	Number of Studies					Number of Studies				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = 0.1										
30	.095 (.009)	.141(.011)	.233 (.013)	.488 (.015)	**.684 (.015)**	.089	.128	.209	.442	**.631**
40	.110 (.010)	.171 (.012)	.293 (.014)	.606 (.016)	.804 (.012)	.104	.159	.271	.568	.769
50	.125 (.010)	.201 (.013)	.352 (.015)	.703 (.015)	.883 (.010)	.119	.190	.331	.673	.861
60	.139 (.011)	.232 (.013)	.410 (.016)	.781 (.014)	.932 (.008)	.134	.221	.389	.756	.919
80	.171 (.012)	.293 (.014)	.516 (.015	.884 (.010)	.979 (.005)	.165	.282	.497	.871	.974
100	.201 (.012)	.352 (.015)	.608 (.015)	.943 (.007)	.994 (.003)	.195	.342	.593	.935	.992
Population Effect Size = 0.2										
30	.230 (.013)	.405 (.016)	**.681 (.015)**	.970 (.006)	1 (<.001)	.209	.367	**.629**	.952	1
40	.290 (.015)	.511 (.016)	.802 (.012)	.993 (.003)	1 (<.001)	.270	.478	.768	.989	1
50	.350 (.015)	.605 (.015)	.882 (.010)	.999 (.001)	1 (<.001)	.330	.575	.860	.998	1
60	.406 (.015)	.683 (.015)	.931 (.008)	1 (<.001)	1 (<.001)	.389	.659	.918	1	1
80	.513 (.016)	.804 (.013)	.978 (.005)	1 (<.001)	1 (<.001)	.496	.788	.974	1	1
100	.605 (.015)	.883 (.011)	.994 (.003)	1 (<.001)	1 (<.001)	.591	.872	.992	1	1
Population Effect Size = 0.3										
30	.442 (.016)	**.728 (.014)**	.952 (.007)	1 (<.001)	1 (<.001)	.402	**.678**	.928	1	1
40	.555 (.016)	.844 (.012)	.988 (.004)	1 (<.001)	1 (<.001)	.521	.812	.981	1	1
50	.651 (.014)	.914 (.009)	.997 (.002)	1 (<.001)	1 (<.001)	.623	.895	.995	1	1
60	.731 (.014)	.954 (.007)	.999 (.001)	1 (<.001)	1 (<.001)	.707	.943	1	1	1
80	.845 (.012)	.988 (.003)	1 (<.001)	1 (<.001)	1 (<.001)	.830	.985	1	1	1
100	.915 (.009)	.997 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	.906	.996	1	1	1
Population Effect Size = 0.5										
30	.850 (.012)	.989 (.003)	1 (<.001)	1 (<.001)	1 (<.001)	.806	.979	1	1	1
40	.935 (.008)	1.00 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	.913	1	1	1	1
50	.973 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.963	1	1	1	1
60	.989 (.003)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.985	1	1	1	1
80	.998 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.998	1	1	1	1
100	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1	1	1	1
Population Effect Size = 0.8										
30	.997 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.993	1	1	1	1
40	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.999	1	1	1	1
50	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1	1	1	1
60	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1	1	1	1
80	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1	1	1	1
100	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1	1	1	1

Table 4). Overall, the unbalanced design did not significantly influence the statistical power when other conditions were fixed.

The simulated power was systemically higher than the analytical power in most conditions, which aligned with the hypothesis of underestimation of the analytical power. Next, the discrepancy patterns in the fixed and random-effects models were discussed. The discrepancies between the simulated power and the analytical power for the fixed-effects model were generally minimal, as shown in Tables 1 and 3. Most of the discrepancies were equal to or less than 0.05. The discrepancy levels were similarly low in the random-effects models (Tables 2 and 4). The discrepancy of simulated and analytical power was larger for balanced designs compared with unbalanced designs in both fixed- and random-effects models. The random-effects model with balanced designs (Table 2) displayed the largest power discrepancies with six conditions above 0.05. The fixed-effects model with balanced designs (Table 1) had power discrepancies

Table 2. Statistical Power of the Random-Effects Model (Balanced Design)

Average Sample Size	Power Simulation					Power Function				
	Number of Studies					Number of Studies				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = 0.1										
30	.075 (.008)	.114 (.010)	.196 (.013)	.437 (.016)	**.634 (.015)**	.079	.112	.183	.397	**.581**
40	.089 (.009)	.140 (.011)	.249 (.014)	.547 (.016)	.753 (.014)	.089	.136	.233	.509	.714
50	.102 (.010)	.165 (.012)	.298 (.015)	.636 (.015)	.834 (.012)	.100	.159	.281	.604	.808
60	.115 (.010)	.191 (.012)	.346 (.015)	.708 (.014)	.890 (.010)	.110	.181	.327	.682	.872
80	.140 (.011)	.239 (.014)	.430 (.015)	.813 (.012)	.951 (.007)	.130	.224	.410	.796	.943
100	.164 (.012)	.282 (.014)	.502 (.015)	.878 (.010)	.977 (.005)	.149	.264	.483	.868	.974
Population Effect Size = 0.2										
30	.181 (.012)	.339 (.015)	**.612 (.016)**	.952 (.007)	.996 (.002)	.167	.305	**.555**	.924	.991
40	.232 (.013)	.433 (.016)	.734 (.014)	.986 (.004)	1 (.001)	.211	.395	.689	.977	.999
50	.280 (.014)	.513 (.016)	.820 (.012)	.996 (.002)	1 (<.001)	.255	.476	.786	.993	1.000
60	.326 (.015)	.584 (.015)	.878 (.010)	.999 (.001)	1 (<.001)	.296	.549	.854	.998	1.000
80	.407 (.015)	.696 (.015)	.943 (.008)	1 (<.001)	1 (<.001)	.373	.667	.933	1.000	1.000
100	.479 (.017)	.776 (.013)	.973 (.005)	1 (<.001)	1 (<.001)	.442	.755	.969	1.000	1.000
Population Effect Size = 0.3										
30	.359 (.015)	**.647 (.014)**	.922 (.009)	1 (.001)	1 (<.001)	.315	**.581**	.880	.999	1.000
40	.458 (.016)	**.767 (.013)**	.973 (.005)	1 (<.001)	1 (<.001)	.408	**.716**	.956	1.000	1.000
50	.543 (.016)	.848 (.012)	.991 (.003)	1 (<.001)	1 (<.001)	.493	.811	.984	1.000	1.000
60	.615 (.016)	.901 (.009)	.997 (.002)	1 (<.001)	1 (<.001)	.566	.876	.995	1.000	1.000
80	.725 (.014)	.957 (.007)	1 (.001)	1 (<.001)	1 (<.001)	.687	.947	.999	1.000	1.000
100	.801 (.013)	.981 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	.774	.977	1.000	1.000	1.000
Population Effect Size = 0.5										
30	**.762 (.013)**	.973 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	**.683**	.946	.999	1.000	1.000
40	**.866 (.010)**	.994 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	**.813**	.987	1.000	1.000	1.000
50	.924 (.008)	.998 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	.893	.997	1.000	1.000	1.000
60	.956 (.007)	1 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	.940	.999	1.000	1.000	1.000
80	.985 (.004)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.981	1.000	1.000	1.000	1.000
100	.994 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.994	1.000	1.000	1.000	1.000
Population Effect Size = 0.8										
30	.986 (.004)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.967	1.000	1.000	1.000	1.000
40	.997 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.994	1.000	1.000	1.000	1.000
50	.999 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.999	1.000	1.000	1.000	1.000
60	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1.000	1.000	1.000	1.000
80	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1.000	1.000	1.000	1.000
100	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1.000	1.000	1.000	1.000

above 0.05 in three conditions. The random-effects model with unbalanced designs (Table 4) had power discrepancies above 0.05 in one condition, while the fixed-effects model with unbalanced designs (Table 3) had power discrepancies below0.05 in all conditions. These conditions were highlighted in the tables. When the average sample size was small and the number of studies was limited, highest discrepancies in balanced designs for small population effect sizes were identified (Tables 1 and 3). For instance, simulated power was 0.53 higher than analytical power when the average sample size was 30 and the number of studies was 80 in both fixed- and random-effects models.

Overall, analytical power was close to the simulated power with acceptable discrepancies when the average sample size, the population effect size, and the number of studies varied in most conditions. No discussion of power discrepancies was necessary when power estimates were close to 1. A few conditions showed a discrepancy higher than 0.05.

Table 3. Statistical Power of the Fixed-Effects Model (Unbalanced Design)

Average Sample Size	Power Simulation					Power Function				
	Number of Studies					Number of Studies				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = 0.1										
30	.079 (.009)	.116 (.010)	.193 (.012)	.418 (.015)	.607 (.015)	.084	.119	.191	.402	.581
40	.094 (.009)	.145 (.011)	.250 (.013)	.537 (.016)	.741 (.014)	.097	.147	.246	.520	.720
50	.109 (.010)	.173 (.012)	.304 (.014)	.637 (.016)	.834 (.012)	.111	.174	.300	.622	.819
60	.122 (.011)	.202 (.014)	.360 (.015)	.720 (.015)	.896 (.010)	.124	.201	.353	.706	.886
80	.151 (.012)	.258 (.014)	.461 (.016)	.840 (.011)	.962 (.006)	.151	.256	.453	.830	.958
100	.179 (.012)	.313 (.015)	.551 (.016)	.912 (.009)	.987 (.004)	.179	.310	.544	.905	.985
Population Effect Size = 0.2										
30	.191 (.013)	.343 (.015)	.605 (.015)	.945 (.007)	.995 (<.001)	.191	.333	.580	.927	.991
40	.247 (.014)	.447 (.016)	.739 (.014)	.985 (.004)	.999 (<.001)	.245	.434	.718	.980	.999
50	.304 (.015)	.539 (.016)	.833 (.011)	.996 (.002)	1 (<.001)	.300	.527	.818	.995	1
60	.356 (.015)	.620 (.016)	.896 (.009)	.999 (.001)	1 (<.001)	.352	.608	.885	.999	1
80	.458 (.016)	.749 (.014)	.962 (.006)	1 (<.001)	1 (<.001)	.452	.739	.957	1	1
100	.548 (.016)	.840 (.012)	.987 (.004)	1 (<.001)	1 (<.001)	.542	.832	.985	1	1
Population Effect Size = 0.3										
30	.376 (.015)	.653 (.015)	.918 (<.001)	1 (<.001)	1 (<.001)	.365	.627	.898	.999	1
40	.487 (.015)	.786 (.013)	.975 (<.001)	1 (<.001)	1 (<.001)	.475	.767	.967	1	1
50	.585 (.015)	.872 (.010)	.993 (<.001)	1 (<.001)	1 (<.001)	.573	.858	.990	1	1
60	.669 (.015)	.926 (.008)	.998 (<.001)	1 (<.001)	1 (<.001)	.657	.917	.997	1	1
80	.794 (.013)	.977 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	.786	.973	1	1	1
100	.877 (.010)	.993 (.003)	1 (<.001)	1 (<.001)	1 (<.001)	.871	.992	1	1	1
Population Effect Size = 0.5										
30	.787 (.013)	.976 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	.760	.965	1	1	1
40	.900 (.009)	.996 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	.880	.993	1	1	1
50	.952 (.007)	.999 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	.943	.999	1	1	1
60	.979 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.974	1	1	1	1
80	.996 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.995	1	1	1	1
100	.999 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.999	1	1	1	1
Population Effect Size = 0.8										
30	.992 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.987	1	1	1	1
40	.999 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.998	1	1	1	1
50	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1	1	1	1	1
60	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1	1	1	1	1
80	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1	1	1	1	1
100	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1	1	1	1	1

Power Guide

Based on the above results, we proceeded next to develop a power guide to help researchers make practical decisions. We utilized simulated power results in this section due to the estimation accuracy in theory. Tables 5 and 6 were developed using the simulation codes to identify the average sample size and number of studies needed to receive a power of 0.8 in retrospective power analysis. The population effect size of 0.8 was not included since the estimates of power would likely be 1 in the most relevant conditions. Researchers need to acquire an average sample size of 80 when the population effect size is 0.3 and the number of studies is 5 in a fixed-effects model with balanced designs (Table 5). Researchers need to find 25 studies when the population effect size is 0.2 and the average sample size is 40 in a random-effects model with balanced designs (Table 6).

Next, we illustrated how to use both tables to guide the sample selection process in meta-analysis. At the beginning of a meta-analytic study, researchers may need to determine the number of studies and

Table 4. Statistical Power of the Random-Effects Model (Unbalanced Design)

Average Sample Size	Power Simulation					Power Function				
	Number of Studies					Number of Studies				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = 0.1										
30	.064 (.008)	.098 (.009)	.169 (.012)	.387 (.016)	.575 (.015)	.077	.108	.174	.375	.552
40	.078 (.008)	.123 (.010)	.219 (.013)	.495 (.016)	.701 (.014)	.086	.129	.219	.479	.682
50	.091 (.009)	.146 (.011)	.266 (.014)	.585 (.015)	.790 (.013)	.096	.150	.262	.570	.777
60	.103 (.009)	.170 (.012)	.311 (.014)	.659 (.015)	.855 (.011)	.105	.170	.304	.646	.845
80	.126 (.011)	.215 (.013)	.391 (.015)	.771 (.014)	.929 (.008)	.123	.209	.382	.763	.926
100	.149 (.011)	.256 (.013)	.461 (.016)	.845 (.011)	.965 (.006)	.140	.246	.451	.840	.964
Population Effect Size = 0.2										
30	.154 (.011)	.293 (.014)	.551 (.016)	.925 (.009)	.992 (.003)	.159	.287	.525	.905	.987
40	.202 (.013)	.384 (.016)	.679 (.015)	.976 (.005)	.999 (.001)	.199	.369	.654	.968	.998
50	.247 (.014)	.462 (.016)	.773 (.014)	.992 (.003)	1 (<.001)	.238	.445	.752	.989	1.000
60	.290 (.015)	.533 (.015)	.840 (.011)	.997 (.002)	1 (<.001)	.275	.514	.824	.997	1.000
80	.368 (.015)	.648 (.015)	.920 (.009)	1 (.001)	1 (<.001)	.347	.629	.912	1.000	1.000
100	.437 (.017)	.734 (.014)	.959 (.006)	1 (<.001)	1 (<.001)	.411	.718	.956	1.000	1.000
Population Effect Size = 0.3										
30	.310 (.014)	.585 (.015)	.886 (.01)	.999 (.001)	1 (<.001)	.297	.551	.857	.998	1.000
40	.406 (.015)	.714 (.014)	.955 (.006)	1 (<.001)	1 (<.001)	.382	.681	.941	1.000	1.000
50	.490 (.016)	.804 (.013)	.983 (.004)	1 (<.001)	1 (<.001)	.461	.778	.977	1.000	1.000
60	.563 (.015)	.867 (.011)	.993 (.003)	1 (<.001)	1 (<.001)	.531	.847	.991	1.000	1.000
80	.678 (.015)	.937 (.008)	.999 (.001)	1 (<.001)	1 (<.001)	.648	.928	.999	1.000	1.000
100	.761 (.014)	.970 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	.737	.967	1.000	1.000	1.000
Population Effect Size = 0.5										
30	**.703 (.014)**	.955 (.006)	1 (.001)	1 (<.001)	1 (<.001)	**.652**	.931	.999	1.000	1.000
40	.823 (.012)	.988 (.004)	1 (<.001)	1 (<.001)	1 (<.001)	.782	.980	1.000	1.000	1.000
50	.894 (.010)	.997 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	.867	.995	1.000	1.000	1.000
60	.936 (.008)	.999 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	.920	.999	1.000	1.000	1.000
80	.975 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.972	1.000	1.000	1.000	1.000
100	.990 (.003)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.990	1.000	1.000	1.000	1.000
Population Effect Size = 0.8										
30	.976 (.005)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.957	1.000	1.000	1.000	1.000
40	.995 (.002)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.990	1.000	1.000	1.000	1.000
50	.999 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	.998	1.000	1.000	1.000	1.000
60	1 (.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1.000	1.000	1.000	1.000
80	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1.000	1.000	1.000	1.000
100	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1 (<.001)	1.000	1.000	1.000	1.000	1.000

average sample sizes that are needed to achieve the optimal statistical power. One difficulty in power estimation is to identify population effect size, which is unknown to researchers in practice. Instead, researchers could refer to prior meta-analysis studies on similar topics to obtain a general idea of population effect size. A similar example that summarized meta-analysis results of previous research on various topics related to student achievement can be found in Hattie (2009).

Another example could be if researchers are interested in studying the summative effects of distance education after the outbreak of COVID-19. Distance education has a small effect size of 0.09. An average sample size of more than 1000 is needed if only a small number of studies are available (5 or 10) in the random-effects model. On the other hand, parental involvement is of interest since students need parental support with distance education. With a moderate effect size of 0.51, an average sample size of 40 is enough if we have five or more studies, no matter what models we use. We suggest that researchers go through both tables under different conditions to obtain a general idea of what samples are needed to achieve a statistical power of 0.8.

Table 5. Sample Size Needed to Receive Power of 0.8.

Fixed Effects Model										
Population	**Balanced Design Number of Studies**					**Unbalanced Design Number of Studies**				
Effect Size	5	10	20	50	80	5	10	20	50	80
0.5	30	< 30	< 30	< 30	< 30	40	< 30	< 30	< 30	< 30
0.3	80	40	<30	< 30	< 30	100	50	<30	< 30	< 30
0.2	160	80	40	< 30	< 30	180	100	50	< 30	< 30
0.1	640	320	160	80	40	720	360	240	100	50
Random Effects Model										
Population	**Balanced Design Number of Studies**					**Unbalanced Design Number of Studies**				
Effect Size	5	10	20	50	80	5	10	20	50	80
0.5	40	<30	<30	<30	<30	40	<30	<30	<30	<30
0.3	100	50	<30	<30	<30	120	50	<30	<30	<30
0.2	300	120	50	<30	<30	340	140	60	<30	<30
0.1	>1000	>1000	300	80	50	>1000	>1000	320	90	60

Table 6 Number of Studies Needed to Receive Power of 0.8

Fixed Effects Model												
Population	**Balanced Design Average Sample Size**						**Unbalanced Design Average Sample Size**					
Effect Size	30	40	50	60	80	100	30	40	50	60	80	100
0.5	5	<5	<5	<5	<5	<5	6	<5	<5	<5	<5	<5
0.3	12	10	8	6	5	<5	17	11	9	7	6	<5
0.2	30	20	16	14	10	8	35	25	19	16	12	9
0.1	115	80	65	55	40	35	150	95	75	65	50	40
Random Effects Model												
Population	**Balanced Design Number of Studies**						**Unbalanced Design Number of Studies**					
Effect Size	30	40	50	60	80	100	30	40	50	60	80	100
0.5	6	<5	<5	<5	<5	<5	7	5	<5	<5	<5	<5
0.3	14	11	9	8	6	5	16	13	10	9	7	6
0.2	35	25	20	17	13	11	40	30	25	19	15	12
0.1	120	90	75	65	50	45	135	105	85	70	55	45

Estimating Statistical Power Using `R` codes

If researchers have obtained a sample data file, they may be concerned as to whether or not the available data resource has optimal statistical power. The sample dataset was taken from studies on gender differences in mental rotation and cognitive abilities (Voyer, 2011). Prior research had well documented that men were better at mental rotation and cognitive abilities, as compared with women. Six studies were included in the meta-analysis to examine the gender differences in mental rotation tasks with long time limits. Table 7 shows the sample size of two groups in each study. We did not provide sample standardized effect size estimates as population effect size is needed for power estimation. Gender differences were very small in math &science with an average effect size of 0.12 (Hattie, 2009). We obtained simulated power estimates with the `R` codes provided in the Appendix. The statistical power using the fixed-effects model was 0.62 with a standard deviation of 0.02; the statistical power using the random-effects model was 0.42 with a standard deviation of 0.02. More studies might be needed before researchers can conduct the meta-analysis to increase statistical power. Researchers can use the codes in the Appendix to estimate statistical power for their own studies.

Discussion

Meta-analysis has been used for several decades to synthesize research results of similar nature. There has been an increasing interest in using meta-analysis because it enables researchers to reconcile inconsistent findings from small studies on the same topic and reach a definitive answer pertaining to the research question of interest.

The current study investigated the discrepancy between the simulated power and the analytical approximated power for the Hedge's *g* (corrected from Cohen's effect size *d*) and the true formula of effect size variances under various conditions (i.e., sample size, design balance, number of studies, population effect size and models). The findings can potentially inform educational researchers about the accuracy of statistical power in a planned meta-analysis.

No prior studies have been conducted to analyze the discrepancies between analytical and simulated statistical power. Therefore, we interpreted the results based on what we found from the results. The power discrepancies between simulated and analytical power were below 0.05 in most conditions. Certain conditions had noticeable discrepancies. In addition, the unbalanced design seems to have had less discrepancy with the balanced design in both types of models. The possible explanation is that the process of randomizing sample sizes between two groups in an unbalanced design could lead to very small sample sizes, which may decrease statistical power. Therefore, smaller discrepancies were observed.

In the process of computing the analytical statistical power for the random-effects model, we used the ratio of between and within-study effect size variances generated from the simulation to improve the comparability between analytical and simulated statistical power. This could explain why minimal discrepancies were observed in most of the random-effects model conditions. Typically, researchers assume a ratio of between and within-study variances in analytical power (Liu, 2013), which could influence the accuracy of the power estimates. Therefore, simulation is necessary to obtain relatively accurate power estimates for the random-effects model even when we utilized the analytical procedures.

The accuracy of analytical power procedures has improved because Hedge's *g* and true effect size variance formula of each study were utilized. This information helps to provide more accurate results, especially for studies with small sample sizes (Goulet-Pelletier et al., 2018). Finally, we recommend the usage of simulated power to increase cost-effectiveness levels in meta-analysis. As analytical power underestimates power (Hedges & Pigott, 2001), researchers only need a smaller number of studies or smaller average sample sizes to reach optimal simulated statistical power. In addition, the stability of power estimates can only be obtained in simulated statistical power.

In order to apply the study results in a practical meta-analysis, the first decision is to select an appropriate model. The literature review reveals that the random-effects models have become increasingly popular (Hall & Brannick, 2002). As cited by Field (2001), it is more likely to have datasets with varied effect sizes across studies. The assumption of fixed population effect size is tenable only when researchers do not intend to generalize the results beyond the datasets. For example, if the researchers include most of the representative datasets in their meta-analysis, they do not need to generalize the results. When the population effect sizes vary by study, the random-effect model should be used. Otherwise, the Type I error rate is not controlled properly (Liu, 2015). This is especially true when there is a large amount of variation in the effect sizes among studies. Researchers may choose a fixed-effects model or random-effects model by calculating the *Q* statistics, which can be used as a reference to decide if the population effect sizes are fixed across studies. However, it should be considered in conjunction with other criteria, such as the generalizability of the meta-analysis results. Researchers could also opt to conduct a power analysis, using both fixed and random-effects models. By doing so, they can make an informed decision if they are not sure about the heterogeneity of the dataset.

One difficulty in power analysis is the correct estimation of the population effect size. In theory, researchers cannot obtain a 100% accurate population effect size, but a relatively accurate estimate can be obtained. There are reference books and research articles on different research topics. The current study referred to Hattie's (2009) book, which synthesized over 800 meta-analysis studies related to student achievement. The effect size varies greatly, ranging from negative values to large positive values as indicated by the author. Estimating the population effect size from the sample dataset may either cause an underestimation or overestimation of power if the samples are biased. An alternative approach is to report the confidence interval of the effect size estimates from the dataset. The upper and lower bounds can be used to calculate the confidence interval for the statistical power.

Thirdly, researchers can easily obtain sample sizes of individual studies and the number of studies, as long as they have access to the primary research articles. Researchers can have a general idea of the average sample size and the number of studies they need to obtain after they select a model and estimate the population effect size. As suggested, power is more of a concern for studies with small population effect sizes. Under such situations, researchers should plan to obtain more studies or utilize large-scale studies to increase power. Finally, they also need to consider the influence of an unbalanced design on power by calculating the average sample size ratio between the two groups.

The following are recommendations to be considered. If researchers are certain about the large population effect size in a meta-analysis project (0.8 or above), researchers are likely to attain sufficient statistical power no matter what other parameters they have in their studies. They do not need to consider the probability of making Type II errors. Low statistical power may be a concern when the population effect size is 0.5 or below. The simulated power is recommended on top of analytical power. Practical researchers could refer to Tables 6 and 7 to obtain the guidance of the required average sample size or number of studies under different population effect sizes. Finally, simulation methods offer a flexible tool to estimate statistical power for evaluating study designs in social science research (Arnold et al., 2011). R codes are available for researchers if they have access to the data files which are needed to obtain an accurate power estimation with standard deviations.

The following research limitations have been identified. The effect sizes are generated from the *t* distribution. The assumptions of the *t* distribution may not hold true for a small sample size under 30. Thus, the results of an average sample size of fewer than 30 were not considered in the current study. It is hoped that the current study will motivate further research which is aimed at examining statistical power in more complicated meta-analyses. Further research can examine statistical power in testing moderator effects. For example, there are differences in math achievement between female and male students, but such differences may depend on grade levels. The moderating effect of grade levels on gender differences can be of great interest, and so is the statistical power for testing the moderating effect.

This study provides a brief guide for researchers who are interested in estimating the statistical power of meta-analysis. Simulation was introduced as an alternative and accurate way to estimate statistical power. The discrepancies between analytical and simulated statistical power were noted. General power guidance and R code is provided for practical researchers who want to obtain accurate power estimates.

References

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology*, *11*(1), 1-10. . https://doi.org/10.1186/1471-2288-11-94

Borenstein, M., Hedges, L. V., Higgins, J.P.T, & Rothstein, H.R. (2011). *Introduction to meta-analysis*. Chichester, U.K.: John Wiley & Sons.

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. https://doi.org/10.3758/brm.41.1.35

Cafri, G., Kromrey, J. D., & Brannick, M.T. (2010). A Meta-Meta-Analysis: Empirical Review of Statistical Power, Type I Error Rates, Effect Sizes, and Model Selection of Meta-Analyses Published in Psychology, *Multivariate Behavioral Research*, *45*(2), 239-27. https://doi.org/10.1080/00273171003680187

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). "Things I have learned (so far)," *American Psychologist*, 45(12): 1304-1312. https://doi.org/10.1037//0003-066x.45.12.1304

Cook, D.A., & Hatala, R. (2014). Got power? A systematic review of sample size adequacy in health professions education research. *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-014-9509-5

Ellis, P.D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge University Press.

Else-Quest, N. M., Higgins, A., Allison, C., & Morton, L. C. (2012). Gender differences in self-conscious emotional experience: a meta-analysis. *Psychological Bulletin*, *138*(5), 947-981. https://doi.org/10.1037/a0027930

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*(2), 161–180. https://doi.org/10.1037/1082-989X.6.2.161

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(*2*), 105-124. https://doi.org/10.1207/s15328031us0202_02

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

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*(*2*), 377-389. https://doi.org/10.1037//0021-9010.87.2.377

Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. *New York: Routledge*.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(*2*), 107-128. https://doi.org/10.3102/10769986006002107

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

Hedges, L. V., & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. https://doi.org/10.1037//1082-989x.3.4.486

Hutchinson, S.R. & Bandalos, D.L. (1997). A guide to Monte Carlo simulations for applied researchers. *Journal of Vocational Education Research, 22(4),* 233-245.

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

Liu, J. (2015). *Statistical Power in Meta-Analysis.* (Doctoral dissertation). Retrieved from https://scholarcommons.sc.edu/etd/3221

Liu, X. (2013). *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques*. New York, NY: Routledge.

O'Keefe, D. J. (2007). Brief report: Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, *1*(*4*), 291-299. https://doi.org/10.1080/19312450701641375

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Quintana, D. S. (2023). A guide for calculating study-level statistical power for meta-analyses. *Advances in Methods and Practices in Psychological Science*, *6*(*1*). https://doi.org/10.1177/25152459221147260

Sterne, J.A., Gavagham, S., & Egger, M. (2000). Publication and related bias in meta-analysis" Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*. 53. 1119-1129.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology*, *11*(*1*), 276-280. https://doi.org/10.1046/j.1523-1739.1997.96102.x

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(*2*), 215-247. https://doi.org/10.3102/1076998609346961

Voyer, D. (2011). Time limits and gender differences on paper and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin Review.* 18. 267–277. https://doi.org/10.3758/s13423-010-0042-0

Send correspondence to: Jin Liu
University of South Carolina
Email: jin.liu@uta.edu

Appendix

Fixed- Effects Model

```
rm(list=ls())
##### Note: Practical Researchers need to input their sample sizes of
each study ##### number of studies and population effect size.
#sample size
N1<- c(53,117,153,63,431,29) N2<- c(32,97,106,43,312,48)
#Number of studies
I<- 6
# Set Type I error rate as .05
alpha <- 0.05
# number of simulation iterations (fixed)
sims <- 1000
#Population effect size (set as 0,.1,.2,.3,.4,.5)
PES<-0.12
#number of replications
nrep <- 100
# define the seed number
set.seed(xxx)
############################################################
main<-function(N1,N2, I, PES, sims, alpha)
{
# set up the output significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))
#Simulation loop for (i in 1:sims){
# In each simulation run, perform the meta-analysis
Nvary<-N1+N2
# Sample size between two groups in each study are equal
tdist <- rt(I,Nvary-2,PES/sqrt(1/N1+1/N2))
d<-tdist*sqrt(1/N1+1/N2) J<-1-(3/(4*(Nvary-2)-1))
ES<- d*J
PESg<-PES*J
#Calculate the Z-test statistics - get combined effect size and
variance of all studies
Variancewithin<-((Nvary-2)/(Nvary-4))*(Nvary/(N1*N2))+((Nvary-
2)/(Nvary-4))*PESg*PESg-((PESg*PESg)/(J*J))
Weight<-1/Variancewithin SumWeight<-sum(Weight) SumWd<-sum(Weight*ES)
WeightedD<- SumWd/SumWeight SEM<-sqrt(1/SumWeight)
Zstat<- WeightedD/SEM

#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(Zstat))
significant.experiments[i] <- ifelse(p.value[i] <= alpha,1,0)
{
prob<- mean(significant.experiments)
out <- prob
out
{
power.array <-rep(NA,nrep) for(r in 1:nrep){
set.seed(r)
power.array[r] = main(N1,N2,I, PES, sims, alpha)
(power_fix_m = round(mean(power.array),digits=3))
(power_fix_sd = round(sd(power.array),digits=3))
```

Random Effects Model

```
rm(list=ls())
##### Note: Practical Researchers need to input their sample sizes of
each study
##### number of studies and population effect size.
#sample size
N1 <- c(53,117,153,63,431,29)
N2 <- c(32,97,106,43,312,48)
#Number of studies
I <- 6
# Set Type I error rate as .05
alpha <- 0.05
# number of simulation iterations (fixed)
sims <- 1000
#Population effect size (set as 0,.1,.2,.3,.4,.5)
PES <- 0.12
#number of replicatons
nrep <- 100
# define the seed number
set.seed (xxx)
############################################################
main<-function(N1,N2, I, PES, sims, alpha)
{
# set up the output significant.experiments
<- rep(NA, sims) p.value<-as.numeric(rep(NA,sims))
#Simulation loop
for (i in 1:sims){
# In each simulation run, perform the meta-analysis
Nvary<-N1+N2
# Simulate the effect size using t distribution
### Vary the population effect size of each study to meet the random-
effects model assumption
PESVARY<-rnorm(I,PES,0.1)
# Sample size between two groups in each study are equal
tdist <- rt(I,Nvary-2,0.5*PESVARY/sqrt(1/Nvary))
d<-tdist*2*sqrt(1/(Nvary-2)) J<-1-(3/(4*(Nvary-2)-1)) ES<- d*J
PESg<-PESVARY*J
#Calculate the Z-test statistics - get combined effect size and
variance of all studies
Variancewithin<-((Nvary-2)/(Nvary-
4))*(4/Nvary)*(1+PESg*PESg*Nvary*0.25)-((PESg*PESg)/(J*J))
Weight<-1/Variancewithin SumWeight<-sum(Weight)
SumWd<-sum(Weight*ES)
SumWdsquare<-sum(Weight*ES*ES)
SumWsquare<-sum(Weight*Weight)
Qstat<- SumWdsquare-(SumWd*SumWd)/SumWeight
Cstat<-SumWeight-(SumWsquare/SumWeight)
df<- I -1
#Use if function to define Tsquare(between-study variance)
if(Qstat-df>0){Tsquare<-(Qstat-df)/Cstat} else {Tsquare<-0}
BetweenStudyVariance<-rep(Tsquare,I)
VarianceTotal<- BetweenStudyVariance+ Variancewithin
WeightRandom<- 1/VarianceTotal
SumWeightRandom<-sum(WeightRandom)
```

```
SumWeightRandomd<-sum(WeightRandom*ES)
WeightdRandom<- SumWeightRandomd/SumWeightRandom
SEMRandom<-sqrt(1/SumWeightRandom)
ZstatRandom<- WeightdRandom/SEMRandom
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandom))
significant.experiments[i] <- ifelse(p.value[i] <= alpha,1,0)
}
prob<- mean(significant.experiments)
out <- prob
#names(out) <- c("Real Type I error rate & Power") out
}
power.array <-rep(NA,nrep) for(r in 1:nrep){
set.seed(r)
power.array[r] = main(N1,N2,I, PES, sims, alpha)
}
(power_random_m = round(mean(power.array),digits=3))
(power_random_sd = round(sd(power.array),digits=3))
```