The Unappreciated Heterogeneity of Effect Sizes:

Implications for Power, Precision, Planning of Research, and Replication

David A. Kenny
University of Connecticut

Charles M. Judd
University of Colorado Boulder

David A. Kenny is a member of Department of Psychological Sciences and can be reached at david.kenny@uconn.edu and Charles M. Judd is a member of Department of Psychology and Neuroscience and can be reached at charles.judd@colorado.edu. A prior version of this paper was posted at https://osf.io/b6dtc/.

Abstract

Repeated investigations of the same phenomenon typically yield effect sizes that vary more than one would expect from sampling error alone. Such variation is even found in exact replication studies, suggesting that it is not only due to identifiable moderators but also to subtler random variation across studies. Such heterogeneity of effect sizes is typically ignored, with unfortunate consequences. We consider its implications for power analyses, the precision of estimated effects, and the planning of original and replication research. With heterogeneity, the usual power calculations and confidence intervals are likely misleading, and the preference for single definitive large-*N* studies is misguided. Researchers and methodologists need to recognize that effects are often heterogeneous and act accordingly.

When multiple studies of the same effect exist, one expects variation in the estimated effect sizes because of sampling error, even if they are all estimating the same true effect. Typically, however, in meta-analyses, there exists more variation in effect sizes than can be attributed to sampling error alone, leading to the conclusion of heterogeneous effect sizes across those studies. As we shall show, heterogeneity of effect sizes is even found in studies that are designed to be, as close as possible, replications of each other. The literature on power and research design has generally assumed that there is a single effect size in a given domain that does not vary. This failure to recognize that effect sizes are heterogeneous has led to unfortunate conclusions about statistical power and the design of research. Our goal is to make clear the implications of varying effect sizes for the planning and conduct of research.

To index heterogeneous effect sizes, one can estimate their true standard deviation, over and above what might be expected from sampling error. For instance, if the effect size is a Cohen's *d*, its true standard deviation might be denoted as σ_{δ}. We subscript with δ, rather than *d*, to make clear that we are referring to variation in effect sizes after removing sampling error. Meta-analysts typically test whether this quantity differs from zero using a chi-square test of homogeneity, symbolized as *Q* (Cochran, 1954). Not always reported is the estimated true variance, generically called $\hat{\tau}^2$ in the meta-analysis literature.

Meta-analysts consistently find evidence of heterogeneity. Richard, Bond, and Stokes Zoota (2003) conducted a meta-analysis of meta-analyses in 18 different domains of social psychology (a total of 322 meta-analyses summarizing 33,912 individual studies). They reported an average effect size estimate (*r*) of .21 and an average true

standard deviation estimate of those effect sizes of .15. More recently and more broadly, van Erp, Verhagen, Grasman, and Wagenmakers (2017) have made available a database of every meta-analysis published in the *Psychological Bulletin* from 1990 to 2013. The average value of τ estimates for studies using *d* or *g* is 0.24 (189 meta-analyses) and for *r* it is .13 (502 meta-analyses). Moreover, van Erp et al. report that 96 percent of meta-analyses with 60 or more studies find some level of heterogeneity. Finally a recent survey of 200 meta-analyses (Stanley, Carter, & Doucouliagos, 2017) found that study heterogeneity was on average about three times larger than sampling error.

Heterogeneity of effect sizes in meta-analyses is hardly surprising, because typically many different kinds of studies are included in a meta-analysis and important moderators may exist that affect the effect sizes. In other words, most meta-analyses are estimating effect sizes for different effects, rather than a single one. Additionally, there are other factors that potentially bias effect size estimates in meta-analyses: file drawer problems, *p*-hacking strategies, and publication biases. Likely these factors also affect estimates of heterogeneity.

Thus, there are good reasons to be cautious about estimating heterogeneity of effect sizes from meta-analyses. Fortunately, given the current interest in replications, there are the Many Labs project of Klein et al. (2014) and Registered Replication Reports (RRR) proposed by Simons, Holcombe, and Spellman (2014). These permit us to avoid the issues mentioned above because they involve pre-registered replications, with multiple studies all using the same materials and procedure, same analysis method, and same outcome measure.

The Many Labs project tested 16 different effects across 36 independent samples totaling 6,344 participants. Two of the effects had average effect sizes not significantly different from zero and both showed zero study variation. Of the remaining thirteen effect sizes that used *d*, their heterogeneity[1] was significantly greater than zero in 8 cases with an average standard deviation for the 13 studies of 0.21. Effect size variation was highly correlated with the average effect size, $r = .86$. Moreover, typically the standard deviation of the true effect sizes was about 25 percent of the average value of the study *d*'s.

So far, there are six completed RRR studies. However, most of the studies have small effects and in several, they are not significantly different from zero. Given the small levels of heterogeneity found with weaker effects in the Many Labs project, it is then not surprising that the effect sizes in these studies generally, though not always (e.g., Eerland et al. (2016) and Hagger et al. (2016)), have relatively small levels of variation.

We are not the only ones to remark on the rather surprising finding of heterogeneity in studies that are basically all the same:

> In sum, in large scale replication projects such as Many Labs and RRRs, we should for substantive reasons (i.e., protocols designed to eliminate heterogeneity) and statistical reasons (i.e., estimators and significance tests that perform poorly in a manner that falsely suggests homogeneity)—expect to observe little to no heterogeneity. The very fact we observe a nontrivial degree of it is compelling evidence that heterogeneity is not only the norm but also cannot be avoided in psychological research—even if every effort is taken to eliminate it (p. 9, McShane, Tackett, Böckenholdt, & Gelman, 2018).

McShane et al. (2018) further characterize it as “astounding” that effect size heterogeneity is found in these studies (p. 7).

When there are non-zero effects, why might there be heterogeneity of effect sizes even in these highly controlled circumstances? Obviously, there persist moderators that may be responsible for continuing heterogeneity. Even when studying the same effect in highly controlled situations using standardized procedures, there are variations in experimenters, participant populations, history, location, and many other factors that may be subtle or *hidden moderators.* Likely, the list of such hidden moderators is long and perhaps unknowable in its entirety. Ultimately, effect size variation may simply be due to random factors that we can never completely specify or control. At a later point we discuss this possibility in greater detail. Regardless of whether heterogeneity is due to measurable moderators, hidden moderators, or random sources, the effects of heterogeneity have not been fully appreciated in the literature.

Within the meta-analysis literature, the recognition of heterogeneity has led to the development of procedures for random effects meta-analyses (Hedges & Vevea, 1998). However, the implications of heterogeneity, outside of the methodological literature on how to conduct meta-analyses, have not been fully explored. The goal of this paper is to examine the implications of effect size heterogeneity for power analysis, the precision of effect estimation, and the planning of both original and replication research. Some of these implications are more easily dealt with than others. We do not having definitive answers for every issue that we raise. Our ultimate goal is to begin an informed discussion of the problems posed by heterogeneity, rather than naïvely assuming it simply does not exist.

Before we begin, we introduce notation and simplifying assumptions. Consider an effect that is measured using Cohen's *d*. We assume that all studies utilize two equal-sized independent groups of participants and their standardized mean difference yields the effect size estimate. We assume each study has a total of *N* persons with *N*/2 or *n* persons in each condition. The effect size estimate from the i^{th} individual study is d_i and it is an estimate of the true effect size for that study, δ_i. Across studies, there is a distribution of these true effect sizes, with a mean, denoted as μ_δ, and a standard deviation, denoted as σ_δ. To be clear, σ_δ refers to the standard deviation after sampling error has been removed and is denoted as τ in the meta-analysis literature. For a particular study *i*, the effect size d_i, has two parts: its true effect size, δ_i, and its sampling error, $d_i - \delta_i$. We assume that in any particular literature, we have a random sample of the population of all methodologically sound studies, and thus this sample provides estimates of both μ_δ and σ_δ. At a later point, we discuss difficulties underlying these assumptions.

In the next two sections, we discuss two under-appreciated consequences of the heterogeneity of effect sizes: the statistical power of detecting a significant effect size and the precision of any effect size estimate. We then turn our attention to the measurement of heterogeneity and a further discussion of the sources of heterogeneity. In the final section of the paper, we talk about the implications of heterogeneity for replication research and the planning of original research.

Power Given Heterogeneity

In a conditional power analysis in which the effect size is known and not an estimate (Maxwell, Lau, & Howard, 2015), one computes the power for a given effect from that effect size. This conventional analysis, where the only variability derives from

sampling error, assumes there is one true effect size. However, if there is effect size variation over and above sampling error, then the effect size does not take on one value, but rather is a random variable with a mean of μ_δ and a standard deviation of σ_δ.

We initially presume that the distribution of effect sizes is normal. We are well aware that this is an assumption that may be problematic. Accordingly, we return to it later and discuss an alternative. For now, however, making this assumption is a good way to introduce the subject. We adapt and extend the McShane and Böckenholdt (2014) method to determine power given heterogeneity. They also assumed normality. We consider power for the test of a single true effect size, δ_i, as well as power for the test of the mean of effect sizes, μ_δ. Note that when there is homogeneity, $\sigma_\delta = 0$, the power of these two tests is the same.

We first consider the null hypothesis that the true effect size for a given study δ_i is zero. Using the central *t* distribution with *df* degrees of freedom, the critical value, $t_{c(df)}$, can be determined such that $P(-t_{c(df)} > t\ ) + P(t_{c(df)} < t) = \alpha$. For instance, a study with a 50 persons in each group ($N = 100$) has a $t_{c(98)}$ of 1.9845 for a two-sided alpha of .05. If, however, the interest is in testing the null hypothesis that μ_δ is zero with σ_δ being greater than zero and $t_{c(df)}$ is still used as the critical value, the effective alpha is larger than the original alpha. Because of variation in δ_i values, sometimes the d_i for a particular study would be much larger than zero and at other times it would be much smaller than zero, even when the null hypothesis that μ_δ is zero is true. To determine the effective alpha, the standard error for d_i needs to include heterogeneity, making it equal to $\sqrt{\frac{4}{N} + {\sigma_\delta}^2}$, not

$2/\sqrt{N}$. If we denote $q = \sqrt{\frac{\frac{4}{N}}{\frac{4}{N}+ {\sigma_\delta}^2}}$, then the effective alpha or α_e equals $P(-qt_{c(df)} > t)$ + $P(qt_{c(df)} < t) = \alpha_e$, where $t_{c(df)}$ is the critical value for α with zero heterogeneity. For an example, with a μ_δ of zero, 50 units in each group, a σ_δ of 0.2, and α of .05, then q equals 0.7071, which results in an effective alpha of $P[(0.7071)(-1.9845) > t)$ + $P((0.7071)(1.9845) < t] = .162$.

We now turn to power, no longer assuming the true effect size is zero. In order to estimate power, we must determine the appropriate non-centrality parameter for the non-central t distribution. Without heterogeneity, the non-centrality parameter is defined as the true effect size or δ_i divided by its standard error, $2/\sqrt{N}$. With this non-centrality parameter and alpha, one computes $P(-t_{c(df)} > t') + P(t_{c(df)} < t')$, where t' is distributed as a non-central t with $N - 2$ degrees of freedom and a non-centrality parameter of μ_δ divided by $2/\sqrt{N}$. However with heterogeneity,[2] the standard error is $\sqrt{\frac{4}{N} + {\sigma_\delta}^2}$. Thus with heterogeneity, the non-centrality parameter equals μ_δ divided by $\sqrt{\frac{4}{N} + {\sigma_\delta}^2}$. With this non-centrality parameter and effective alpha, one computes $P(-qt_{c(df)} > t') + P(qt_{c(df)} < t')$, where t' is distributed as a non-central t with $N - 2$ degrees of freedom and a non-centrality parameter of μ_δ divided by $\sqrt{\frac{4}{N} + {\sigma_\delta}^2}$. For the example study, μ_δ is assumed to equal 0.3, σ_δ to be 0.2 with 50 persons in each group. The critical value is multiplied by 0.7071 to yield 1.4032 and the non-centrality parameter is 0.3/0.282843 = 1.06066. Power is estimated as .375, as opposed to the .318 value with zero heterogeneity. The R function in the Appendix can be used to estimate power given

heterogeneity. Additionally, a web-based program is available at https://davidakenny.shinyapps.io/SVPower. It is also possible to modify a conventional power program (e.g., G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) to obtain estimate of power with heterogeneity, using the effective alpha and the adjusted non-centrality parameter.

To be clear, we are computing power for a fixed value of the effect size, either δ_i or μ_δ. Several others (Anderson, Kelley, & Maxwell, 2017; Biesanz & Schrager, 2010; Dallow & Fina, 2011; Gillett 1994; 2002; Perugini, Gallucci, & Costantini, 2014; Schönbrodt & Wagenmakers, 2018) have suggested computing power across a distribution of effect sizes due to uncertainty in the estimate of the effect sizes (usually sampling error). See Anderson and Maxwell (2017) for a discussion of some of these methods. Moreover, Du and Wang (2016) presented a Bayesian method of power analysis which allows for heterogeneity, where the exact value of heterogeneity is not known but is assumed to have a prior distribution. Here, we are allowing for heterogeneity with a known value.

Table 1 presents power estimates for a range of values of μ_δ (0.0, 0.2, 0.5, and 0.8, corresponding to no, small, medium and large average effects), a range of values of the standard deviation of effect sizes, σ_δ (0.000, 0.050, 0.125, and 0.200), and a range of values of sample sizes, N (100, 200, 500, 1000, and ∞) with alpha[3] set at .05. (Recall that N is the total sample size of the study.) The standard deviations of the effect sizes, σ_δ, are chosen to be equal to 25 percent of small, medium and large effects based on statistics presented earlier from past meta-analyses and organized replication endeavors. Note that when σ_δ is equal to 0.0, there is no effect size heterogeneity and the results in the table

are consistent with a conventional power analysis. The null hypothesis is that μ_δ equals zero, but when σ_δ is zero, the null hypothesis is also that δ for a particular study is zero. For each combination of values, two probabilities are given: The one denoted by "+" is the probability of rejecting the null hypothesis when d_i is positive (i.e., in the same direction as μ_δ), and the one denoted by "–" is the probability of rejecting the null hypothesis when d_i is negative (i.e., in the opposite direction).

Examining first the results from conventional power analyses ($\sigma_\delta = 0.0$), the power of finding a positive effect increases as the effect size and sample size increase. In addition, there is almost no chance of finding a significant negative effect. If we compare these results to what we find when there is heterogeneity, there are several dramatic differences.

Next, we consider the case when the null hypothesis is true, i.e., $\mu_\delta = 0.0$. Note that with heterogeneity, although the average effect size is zero, any sample value is likely non-zero and may well be statistically significant. We see that with increasing heterogeneity and sample sizes, the number of Type I errors increases, well above the nominal alpha value of .05. In fact when $\sigma_\delta = 0.2$ and $N = 1{,}000$, the probability of making a Type I error is over 50 percent! It is important to be clear about what we mean by this Type I error. We are here talking about making an error in the conclusion that the mean of the population of effect sizes is different from zero. We are not talking about an error in concluding that the true effect size for a particular study δ_i is different from zero.

In the remainder of this section, we consider the power when there is a non-zero true effect. There are several results from Table 1 worth noting here. First, whenever conventional power analyses yields a power value less than .50, the estimate that allows

for heterogeneity is greater than the estimate based on the absence of heterogeneity. For instance, when $\mu_\delta = 0.2$ and $N = 100$, power is .166 with no heterogeneity and rises to .245 when σ_δ is 0.2. The increase in power is due to the fact that if we assume the true effect sizes are normally distributed, there is an asymmetry in the power function. To illustrate what this means, suppose that the effect size is overestimated or underestimated by 0.1. When it is overestimated, there is bigger boost in power (.318, a boost of .152) than when it is underestimated by the same amount (.071, a loss of .095). As a result, half the time the value of δ_i is larger than 0.2, its mean, which results in a net increase in power. To our knowledge, no one has previously noted that heterogeneity can increase power.

Second, when conventional power analyses yield values greater than .50, the opposite happens: Power declines once an allowance is made for heterogeneity. For instance, when $\mu_\delta = 0.5$ and $N = 200$, power is an impressive .940 with no heterogeneity but sinks to .817 when σ_δ is 0.2. When power is greater than .50, the asymmetry works in the opposite direction: If the effect size is overestimated by 0.1, there is smaller boost in power (.988, a boost of .048) than the loss when it is underestimated by 0.1 (.804, a loss of .136). This lowering of power due to heterogeneity has been noted previously (McShane & Böckenholdt, 2014).

Third, Table 1 gives the probability of finding a significant effect in the negative direction, opposite in sign from the average effect. When there is no study variation, this probability is negligible. However, with increasing heterogeneity, not too surprisingly this probability increases. What is surprising is that this probability actually increases as the sample size increases. For instance, with an N of 1,000, $\mu_\delta = 0.2$, and σ_δ of 0.2, there

is nearly a 6 percent chance of finding a significant result in the opposite direction from the average effect size. To our knowledge, this point has not been previously noted in the literature.

Fourth, related to the previous point, there are non-obvious results as *N* gets large. As expected, when there is no variation in effect size, power goes to a value of one with increasing sample sizes. However, with study variation, we see that power to detect an effect in the same direction as the average effect size goes to a value less than one; how much less than one depends on the ratio of σ_δ to μ_δ. As this ratio gets larger, there is a greater chance of obtaining a significant negative effect and this leads to a decrease in power for detecting a significant positive effect. For instance, given $\mu_\delta = 0.15$ and $\sigma_\delta = 0.2$, power in the same direction as the average effect size never reaches the traditionally desired level of .80; as *N* increases it asymptotes at .773.

To summarize, given effect heterogeneity, the power in testing an effect in any particular study is different from what conventional power analyses suggest, and the extent to which this is true depends on the magnitude of the heterogeneity. Whenever a conventional power analyses yields a power value less than .50, an estimate that allows for heterogeneity is greater; and when a conventional analysis yields a power value greater than .50, the estimate given heterogeneity is less.

Second, given some heterogeneity and a small to moderate average effect size, there is a non-trivial chance of finding a significant effect in the opposite direction from the average effect size reported in the literature. Perhaps, even more surprisingly, the power to detect an effect in the wrong direction (e.g., μ_δ is positive, but the test shows a significant negative effect) is non-trivial. This probability increases as *N* increases.

Non-normal Distribution of True Effect Sizes

The values in Table 1 make the strong assumption that true effect sizes are normally distributed. We note that the standard method of computing confidence intervals and *p* values for the average effect sizes in random effects meta-analyses also assumes normally distributed effect sizes. There are, however, some compelling reasons to believe that true effect sizes are not normally distributed. For instance, if the average true effect is positive, it may be implausible that some effects are in fact negative. An alternative and perhaps more reasonable position is that, given a positive effect size, the lower limit is zero. Thus, some studies have larger effect sizes and others have smaller ones, but they are all non-negative. There has been some work on specifying alternatives to the normal distribution for random effects (Lee & Thompson, 2007; Yamaguchi, Maruo, Partlett, & Riley, 2017).[4] One candidate distribution that we have explored is a log-normal one. To illustrate the difference between the two distributions, Figure 1 compares the normal and log-normal distributions of δ_i, each with a μ_δ of 0.2 and a σ_δ of 0.2. In this example, negative values occur about 17 percent of the time in the normal distribution but never in the log-normal. With sampling error, there could be small and infrequent negative effects, but they would arise solely from sampling error. It should also be noted that with a μ_δ of 0.0 for the log-normal distribution, there can be no heterogeneity, because all effect sizes must be positive.

Table 2 presents the power estimates for three effect sizes – 0.2, 0.5, and 0.8 – at four levels of heterogeneity – 0.0, 0.050, 0.125, and 0.200, using the log-normal distribution. Alpha is set to .05 for all analyses. All of the computations use numerical

integration and were done using the app SVPower, which is available at https://davidakenny.shinyapps.io/SVPower/.

Under the assumption that effect sizes are normally distributed, we showed that when a conventional power analysis yields a value of power less than .5, then power would be greater if one assumes heterogeneous effect sizes. This reverses, however, when the conventional analysis yields power values above .5. For the log-normal distribution of effect sizes, the point at which power is the same in both conventional and heterogeneous analyses is below .5. Thus, with the log-normal effect sizes, it is more likely that heterogeneity lowers estimated power.

When power is high in conventional analysis, power declines with heterogeneity. This decline is greater for the log-normal distribution than it is for the normal distribution of effect sizes. The good news is that the probability of finding negative effects, i.e., effects in the direction opposite in sign to μ_δ are very rare, pretty much paralleling that found in conventional analyses. Note here, unlike with the normal distribution of random effects, when these negative effect occur, they are Type I errors in that the true effects are only positive. Additionally, with the log-normal distribution, the effective alpha is no different from the nominal alpha value.

Precision in Estimating Effect Sizes Given Heterogeneity

The effect size in a study provides an estimate of the true effect size. Its standard error permits estimation of the confidence interval for that true effect size. Assuming a fixed effect size, the standard error derives solely from the sampling error within a study. For μ_δ, this can be closely approximated[5] by $2/\sqrt{N}$. If there is study variation, this is the standard error for δ_i, the true effect size for the particular study, and not μ_δ, the mean of

all possible effect sizes. In the presence of effect size variation, the proper standard error for μ_δ is $\sqrt{4/N + \sigma_\delta{}^2}$.

Table 3 presents the 95% confidence interval for μ_δ, given an estimated *d* from a study with the indicated *N* (total sample size), assuming varying degrees of known heterogeneity of effect sizes assumed to have a normal distribution. The values in this table indicate unsurprisingly that the confidence interval becomes narrower as the study *N* increases. They also show, again perhaps unsurprisingly, that as effect heterogeneity increases, the confidence interval for the true effect size becomes wider. This difference can be quite dramatic. Looking at the last row of Table 3, the width of confidence interval with no heterogeneity, a large effect size of 0.8, and a sample size of 1,000 is 0.248, a value much narrower than that with smaller sample sizes, but still relatively wide. However, if σ_δ is 0.2, the width of the interval widens by over a factor of three, to 0.822. With large effect sizes and sample sizes, we might have high power with heterogeneity, but we still have quite a bit of uncertainty about the size of the average true effect.

The confidence intervals in Table 3 assume a single study. Both Maxwell et al. (2015) and Shrout and Rodgers (2018) have argued that when conducting replication studies it may make sense to conduct multiple such studies to narrow the confidence interval. These multiple studies are assumed to be a random sample from the population of possible studies that could be run. If multiple studies were run, all estimating μ_δ, then the confidence interval for the average effect size decreases as a function of essentially pooling the observations from all studies into a single standard error, the approximate

formula being $\sqrt{(4/N + {\sigma_\delta}^2)/k}$ where *k* is the number of studies. In Table 4, we present the confidence intervals for μ_δ if five studies were run, all examining an effect in the same domain but with heterogeneity in effect sizes as indicated by the value of σ_δ.

To see the precision benefits of running five studies, as opposed to one, let us first compare the confidence intervals for the first columns in Tables 3 and 4, where there is no heterogeneity of effect sizes, i.e., $\sigma_\delta = 0.0$. If one runs a single study, with an *N* of 100 there is considerably less precision than if one runs five such studies, each with an *N* of 100. In fact, the confidence interval is exactly the same with one study having an *N* of 500 as for five studies each with an *N* of 100.

Importantly, however, if there is effect size heterogeneity, then there are substantial precision benefits that accrue from multiple smaller studies compared to a single large study. Compare again the rows in Table 3 where *N* equals 500 with the rows in Table 4 where the *N* equals 100 in each study, for a combined *N* across five studies of 500. If there is effect size heterogeneity, the confidence interval for μ_δ is substantially narrower with five studies, each with an *N* of 100, than for a single study with an *N* of 500. Parallel conclusions are found when comparing *N* = 1,000 in Table 3 to *N* = 200 in Table 4.

Note too that although very small levels of heterogeneity have relatively small if not trivial effects on power, they can have rather dramatic effects on precision. Consider a pooled effect based on five studies. Given a heterogeneity value of only 0.05, the confidence interval is 29 percent wider than it is if there is no heterogeneity.

Many analysts recommend what might be called a *one-basket strategy*. They put all their eggs in the one basket of a very large *N* study. It is also now common for

psychologists to dismiss a study as having too small a sample size and pay attention to only large *N* studies. As Tables 3 and 4 make clear, if effect sizes vary across studies, such a strategy is misguided. Clearly, one large *N* study is better than one small *N* study, but given the same total *N* and *heterogeneity*, multiple studies are better than a single study. This preference for many smaller *N* studies presumes that the studies include all such studies and not a non-random subset of small *N* studies that are published (Slavin & Smith, 2009).

The results in Tables 3 and 4 presume that the distribution of effect sizes are normal, which is the standard assumption made in random effects meta-analyses. If however, the distribution of effect sizes is positively skewed with zero as the lower limit (as in the log-normal distribution that we considered), the confidence interval would be asymmetric, with larger upper and lower limits than the values in Tables 3 and 4. We urge methodologists to work on the problem of determining the confidence intervals with non-normal heterogeneity.

Knowing the Magnitude of Heterogeneity

We have just seen that the degree of heterogeneity in effect sizes has substantial consequences for statistical power and precision. We have so far assumed that the magnitude of heterogeneity is known. As discussed by McShane and Böckenholdt (2014), knowing exactly the magnitude of σ_δ is difficult. We might use estimates from prior research, but simulation studies (e.g., Chung, Rabe-Hesketh, & Choi, 2013) have shown that estimates of heterogeneity are not very accurate, especially when the number of studies is small.

Power analyses always rest on a series of informed guesses. To conduct a conditional power analysis, we start with an informed guess of the effect size. Similarly, in the presence of heterogeneity of effect sizes, an informed guess for that heterogeneity is also needed.

How might a researcher make an informed guess? One might surmise that research domains with larger average effect sizes have larger effect size variances, consistent with what we reported earlier for the Many Labs project. There, heterogeneity averaged roughly one quarter the effect size. Following the original suggestion by Pigott (2012), McShane and Böckenholdt (2014) suggest using 0.10 for small heterogeneity, 0.20 for medium, and 0.35 for large. We suspect these estimates are a bit large, given the various biases in the published literature that we mentioned earlier. Accordingly, we used values of .050, .125, and .200 as representative in the power and precision results that we gave earlier. We would hope that there would be an evolving discussion of what value to use for heterogeneity in power analyses. We feel strongly that zero should no longer be the default value.

For precision, knowing the value of σ_δ is more problematic as it needs to be integrated with other statistical information (i.e., the amount of sampling error within studies). Even if we have multiple studies and so have a statistical estimate of heterogeneity, that estimate has a great deal of sampling error. One could just guess at the value and treat it as a population value. Alternatively, a Bayesian analysis, as outlined by McShane and Böckenholdt (2014), Maxwell et al. (2015) and Du and Wang (2016), might be attempted, perhaps using the van Erp et al. (2017) database to create a prior distribution of effect sizes.

There are certainly difficulties of knowing the extent to which there is effect size variance in a given domain. That said, we strongly feel those difficulties are no excuse for just assuming that it is zero. Effect size variation is both widespread and consequential. If researchers wish to ignore heterogeneity, something we hope does not happen, they need to state explicitly that power estimates and confidence intervals are based on the assumption of zero heterogeneity.

The Different Sources of Effect Size Heterogeneity

Earlier we discussed the reasons why there may be effect size heterogeneity. Here we want to review and amplify what we said there. Meta-analyses typically anticipate and attempt to document important moderators of effect sizes. That is, they often hypothesize known factors that can account for variations in effect sizes and conduct analyses to confirm those hypotheses (Tackett et al., 2017). However, typically there persists residual heterogeneity even after accounting for such anticipated moderators. Additionally, as we discussed, even when studies are conducted using standardized procedures and measures, effect size heterogeneity typically persists (see McShane, Tackett, Böckenhold, & Gelman, in press). Thus, there are *hidden moderators* that are likely responsible in part for heterogeneity.

We suggested that the list of such hidden moderators is likely long and its complete contents perhaps ultimately unknowable. To elaborate on that a bit, we believe that there are likely many randomly varying factors that may be responsible for effect size heterogeneity, as we move from study to study, and try as hard as we might, we will never identify all of them. Consider an analogy with random variance associated with participants in how they respond to some treatment. Participants' responses typically

vary for a variety of potentially knowable reasons that might be measured and controlled. However, over and above these, there is also simply random variance in people's responses that we probably will never fully understand or explain. The same holds true, we believe, for effects shown in different studies searching for a common effect. There is random variation due to study in the effects produced and this is not entirely reducible to a finite set of effect moderators. In essence, we are saying that some hidden moderators will always remain hidden.

Besides known moderators and hidden moderators, it might be that case that some of the variation is in principle unknown and so random. Perhaps, Einstein's belief (Einstein & Born, 2006) that "God does not play dice," is wrong, and studies vary for reasons that will never be completely understood. Whether due to fundamental randomness or the complexity of moderation effects, we need to accept the conclusion that one should anticipate heterogeneity even in very highly controlled settings and replication efforts.

Others who have considered the heterogeneity of effect sizes in meta-analyses and replication efforts (McShane et al., in press; Stroebe & Strack, 2014; Tackett et al., 2017) have largely assumed that there are a set of moderators responsible and that, once these are identified and theory refined, effects should be much more homogeneous and replicable, given appropriate control of those moderators. Although we agree that moderators that could potentially be identified and controlled are in part responsible for effect size heterogeneity, we seriously doubt that researchers would typically eliminate heterogeneity by controlling for a few, or even a lot, of such moderators. There will often exist perturbations from study to study that cannot be fully accounted for. The hope of

controlling for everything that might potentially affect the magnitude of a studied effect seems to us overly optimistic. We welcome such optimism, but we need in the meantime to be prepared for the possibility of randomness.

Our belief that random variation in effect sizes exists from study to study is in part responsible for our focus on inferences about μ_δ, the average effect size across a series of studies. An alternative view, suggested by a reviewer, is that perhaps the majority of studies in a domain are poorly done and have varying effect sizes around zero, while one or two studies, more appropriately conducted, have a true effect size that is different from zero. In this case, one could argue that one should be primarily interested in the true effect size, δ_i, estimated by these particular studies (assuming no variation in their true effect sizes). This view presumes that there is some underlying important moderator(s) that varies across the two sets of studies, those with a true effect size of zero and those few where this is not the case. This possibility demands that one attempts to identify such a moderator. Surely one would not want to look after the fact across studies and decide which ones estimate the "real effect" and which ones do not. A perspective that allows for random variation in studies and in their effects avoids this danger.

One might question the idea of studies as random, as they are surely not *randomly* sampled from some known population. We would suggest, however, that just as participants in most experiments are not randomly sampled, yet appropriately treated as random, so too studies should be considered as random. Particularly in situations when different investigators use the same materials, procedures, measures, and analyses, it

seems reasonable to consider the set of studies as a random sample from the population of replication studies that might have been done.

Care needs to be taken to account for any know non-random processes, e.g., publication bias. For instance, a reviewer has pointed out that several studies (e.g., Slavin & Smith, 2009) have reported a negative correlation between effect size and sample size, presumably due in part to the fact that published small *N* studies require larger effect sizes than large N studies. Thus, because of publication bias, sample size may create artificial heterogeneity even when there is no heterogeneity.

Planning and Replicating Research Given Heterogeneity

We have shown that effect size heterogeneity has important consequences for statistical power and for the precision of effect size estimates. These consequences deserve attention in planning research. We first explore these consequences in planning new research to demonstrate an effect. We then turn to implications for replication research, a topic that is particularly important in the context of recent concerns about replicability (e.g., Open Science Collaboration, 2015).

Research to Demonstrate an Effect

Conventional wisdom suggests that one is generally better off doing a single very large study to demonstrate an effect rather than doing a series of smaller and more modest studies. The results we have shown in the discussion surrounding Tables 3 and 4 lead us to take issue with this conventional wisdom.

We consider a single study with a modest number of participants.[6] Anticipating an average effect size, μ_δ, of 0.4 and 154 participants, the conventional conditional power estimate is .694, which is not very good. Even worse, if we allow for heterogeneity in

effect sizes of 0.20, the power is only .625 assuming a normal distribution of effect sizes, and only .600 assuming a log-normal distribution. We are faced with a dilemma. Conventional advice is that one should conduct only high-powered studies. However, with heterogeneity, any given study, no matter how large its sample size, might be far away from the mean of the effect sizes. Moreover, given heterogeneity, the power of any given study is not as great as might be thought. What then is the alternative? We see it as conducting a series of studies, each of which might be only moderately powered, but the combination of those studies would have decent power.

For instance, let us return to the case in which μ_δ is 0.4 and σ_δ is 0.20, and 7 studies have been conducted, each with a sample size of 154. The power of finding a significant effect in any one study, given a normal distribution of effect sizes, is only .625, making the power of finding all seven tests significant only .037. To test the null hypothesis that μ_δ is zero, one conducts a random effects meta-analysis of the seven studies (Maxwell et al., 2015). Denoting n_P is the number of persons per study and k the number of studies and setting $k = 7$, $n_P = 154$ ($N = 1{,}078$), and $\sigma_\delta = 0.20$, the power of a one-sample t-test of mean d or $\bar{d}$ is .926, again assuming normality. Note that given $\sigma_\delta = 0.20$, the standard error of $\bar{d}$ for 7 studies each with $N = 154$ is half the size the standard error of d with one study with 154 times 7 participants. Thus, although the power of any one study is not very impressive, the power of the test of the mean is quite acceptable. Additionally, across studies one can critically examine heterogeneity and begin to test factors responsible for variation in effect sizes.

However, if there are few studies, less than five, a random effects meta-analysis is impractical as there are too few studies to have a reliable estimate of the variance of

effect sizes. Our earlier discussion of how to determine the level of heterogeneity applies. However, just pretending that there is no heterogeneity should not be seen as a defensible option. Possibly a failsafe heterogeneity value could be determined. That is, we could compute how large heterogeneity would have to be to turn the significant pooled effect into a value that is no longer significant.

We have suggested that multiple smaller studies are preferable to a single large one, given effect size heterogeneity. However, what exactly does it mean to conduct multiple smaller studies? Clearly, it would not do to conduct one large study, say with an *N* of 1,000 and break it up, acting as if one had done five studies each with an *N* of 200. Conducting multiple studies must allow for the existing effect size heterogeneity, which, as we have already discussed, accrues randomly from a multitude of sources, including experimenters, samples, and so forth. The point is simply that we are better served by a number of studies that permit one to examine the existing variability of effect sizes in a domain. This is obviously particularly true if the primary interest is in examining factors moderating some effect. Then a series of smaller studies, varying such moderators systematically and insuring they are individually adequately powered, makes most sense.

Research to Replicate an Effect

We are all aware that concerns have lately been raised about the replicability of effects in psychology (Ioannides, 2005). In one well-publicized examination of replicability (Open Science Collaboration, 2015), 100 published psychology studies were each replicated one time. The results were interpreted to be relatively disturbing, as less than half of the studies were successfully replicated.

What can be learned from a single replication study? Table 1 can help provide an answer. Imagine that the initial study to be replicated yields an estimated effect size of 0.5. In an effort to conduct the replication with sufficient power, we assume that μ_δ, the true mean effect size, is 0.5, and we plan on a sample size of 200. This gives rise to an estimate of .94 power based on a conventional conditional power analysis. If the study fails to replicate, it seems reasonable to question the initial study result.

Let us, however, assume heterogeneity of effect sizes in the effect to be replicated, with σ_δ equal to 0.2. In this case, then the actual power is much less than .94, roughly about .82. Thus, over 20 percent of the time the study would fail to replicate. There is even a chance, albeit a very small one, of finding a significant effect in the opposite direction from the original effect, assuming a normal distribution of effect sizes.

In fact, the power in the case we have just explored is certainly worse than we have portrayed it, for two reasons. First, we assumed that μ_δ is the same value as the effect size that we estimated in the original study. However, that initial effect size has sampling error in it that has not been factored in (Anderson & Maxwell, 2017; Maxwell et al., 2015). Second, over and above the sampling error in the original effect size estimate, due to publication biases the actual true effect size is likely smaller than the typical reported estimated effect size (Anderson et al., 2017; Yuan & Maxwell, 2005). Recently, Hawkins et al. (2018) found that the replicated effect size is 60 percent of the original. Not surprisingly, the sampling of extreme scores, i.e., an effect size sufficient for publication, results in a smaller effect size for a replication study through regression toward the mean. Moreover, heterogeneity heightens the effects of publication bias because it makes more extreme positive effect sizes more likely.

In the presence of heterogeneity, our results show that power is not nearly as high as it would seem and that even large *N* studies may have a non-trivial chance of finding a result in the opposite direction from the original study. This makes us question the wisdom of placing a great deal of faith in a single replication study. The presence of heterogeneity implies that there is a variety of true effects that could be produced.

Additionally, the presence of heterogeneity makes us question the common practice of seeing whether zero is in the confidence interval of the difference between the effect in the original study and the effect in the replication study.[7] Doing so presumes that the only source of variance between the two studies is sampling error. However, given heterogeneity, the width of the confidence interval would be greater than that based solely on sampling error. For instance, consider two studies each with an *N* of 200 and estimated effect sizes of 0.60 and 0.15. The 95 percent confidence interval for the difference between these two effect sizes, assuming no heterogeneity, is from 0.053 to 0.847. Because this interval does not include zero, it appears that the two studies are statistically different. However, if we allow for heterogeneity with $\sigma_\delta = 0.15$, the confidence interval actually goes from -0.044 to 0.944, which now includes zero. Ignoring study variation leads to too narrow a confidence interval and sometimes the mistaken conclusion that the original and replication study results have produced inconsistent results.

Part of the recent focus on replication is based on the implicit belief that if procedures could be fully standardized, the only difference between study effects would then be sampling error. Such a view is likely mistaken (Maxwell et al., 2015). Even in a well-conducted replication, there are still many factors that may lead to effect

heterogeneity. For instance, studies are conducted in different locations, with different experimenters, in different historical moments, and with different non-randomly selected participants. All of these, and a variety of other randomly varying factors, likely lead to heterogeneity, a result confirmed by the Many Labs project of Klein et al. (2014). And this heterogeneity leads to concerns about the utility of any single replication study.

In their classic paper on "The Law of Small Numbers," Tversky and Kahneman (1971) described an experimenter who does the same study twice and in the first study, he or she obtains a significant effect, whereas in the second study the effect is no longer significant. When asked what they would do if faced with this situation, a plurality of psychologists said they would "try to find an explanation for the difference between the two groups" (p. 27). Perhaps even more perplexing, we have shown that the second study may even come up with a significant effect in the opposite direction from the first. The second study, does not necessarily "disconfirm" the first; rather it may well lead to the conclusion of considerable random variance in the effect in question.

Conclusion

Effect size heterogeneity is found nearly everywhere in science. However, in power analyses, computing confidence intervals, and the planning of research, researchers often act as if the results of studies are homogeneous. We have shown that heterogeneity leads to both lower and higher power than expected, possibly sometimes a finding in the "wrong" direction, and the conclusion that multiple smaller studies are preferable to a single large one. All of this leads to very different ideas about the conduct of research and the quest to establish the true effect in the presence of random variation. Replication research, it seems to us, should search to do more than simply confirm or

disconfirm earlier results in the literature. Replication researchers should not strive to conduct the definitive large *N* study in an effort to establish whether a given effect exists or not. The goal of replication research should instead be to establish both typical effects in a domain and the range of possible effects, given all of what Campbell called the "heterogeneity of irrelevancies" (Cook, 1990) that affect studies and their results. Many smaller studies that vary those irrelevancies likely serve us better than one single large study. Moreover, in this era of increasing preregistration and collaborative research efforts, multiple studies by different groups of researchers is increasingly feasible. For instance, Psychological Science Accelerator is a network of over 300 laboratories collaborating to collect large-scale international samples of psychological data.

Most researchers tend to believe that in any given domain, when evaluating any given effect, there really is only one effect and one should strive to uncover it in studies that are undertaken. It can be disconcerting, at best, to believe that there really is a variety of effects that exist and that might be found. However, that is what it means to have study variation in effect sizes and, as we emphasized early on, that is what we typically find. As a field, we need to begin to understand what it means for effects to vary and figure out how to include such heterogeneity in both analysis of data and the planning of research.

Some might believe that heterogeneity is a sign simply of poorly conceived and executed studies. Certainly effect size heterogeneity can be induced if some studies in a domain are well done and others not. However, we are convinced heterogeneity is not solely a result of poorly conducted studies. Heterogeneity is induced by randomly

varying factors that affect the magnitude of true effect sizes in any given domain, even in well-conducted studies.

We have raised the issue of heterogeneity and explored some of its implications, while nevertheless highlighting some difficult issues that require further attention. These include the nature of the underlying distribution of effect sizes, how to estimate their variability, and how much heterogeneity should be expected. All three of these issues are difficult ones, but they require intensive study by methodologists. Finally, we have limited our discussion of effects sizes to *d*; a full treatment of the topic would require extending the discussion to other effect size measures, e.g., correlations and odd ratios.

These issues notwithstanding, we firmly believe that we need to accept and, in fact, embrace heterogeneity (McShane et al., 2018). If there truly exist multiple effect sizes in a given domain, then power analyses and confidence intervals need to allow for that. Moreover, research should also examine that variability, and the factors that can partly explain it, rather than focusing solely on whether an effect exists or does not.

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28,* 1547-1562.

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "Replication Crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research, 52,* 305-324.

Biesanz, J. C., & Schrager, S. M. (2010). *Sample size planning with effect size estimates*. Unpublished paper, University of British Columbia.

Chung, Y., Rabe-Hesketh, S., & Choi, I-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine, 32,* 4071-4089.

Cochran, W. G. (1954). The combination of estimates from different experiments. Biometrics, 10, 101–129.

Cook, T. D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (DHHS Publication No. PHS 90-3454, pp. 9–31). Rockville MD: Department of Health and Human Services.

Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics, 10,* 311–317.

Du, H., & Wang, J. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research, 51*, 589-605.

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J., Aucoin, P., Berger, S., Birt, A., Capezza, N., Carlucci, M., *et al.* (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11,* 158–171.

Einstein, A., & Born, N. (2005). *Born-Einstein letters: 1916-1955*. London UK: Palgrave-MacMillan.

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

Gillett, R. (1994). An average power criterion for sample size estimation. *The Statistician, 43*, 389-394.

Gillett, R. (2002). The unseen power loss: Stemming the flow. *Educational and Psychological Measurement, 62*, 960-968.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., *et al.* (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11,* 546–573.

Hawkins, R. X. D. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science, 1*, 7-18.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine, 31*, 3328-3336.

Hedges, L. V., & Olkin, I. (1985) *Statistical methods for meta-analysis*. London: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, *2*(8), e124.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al., (2014). Data from investigating variation in replicability: A "many labs" replication project. *The Journal of Open Psychology Data, 2*, DOI: http://dx.doi.org/10.5334/jopd.ad.

Lee, K. J., & Thompson, S.G. (2007). Flexible parametric models for random-effects distributions. *Statistics in Medicine, 27*, 418–434.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 79*, 487-498.

McShane, B. B., & Böckenholdt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Sciences, 9*, 612-625.

McShane, B. B., Tackett, J. L., Böckenholdt, U., & Gelman, A. (2018). Large scale replication projects in contemporary psychological research. *American Statistician*, in press.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science, 349* (6251), aac4716.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9,* 319-332.

Pigott, T. (2012). *Advances in meta-analysis*. New York: Springer.

Richard, F. D., Bond, C. F, Jr., Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331-363.

Schönbrodt, F., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25,* 128-142.

Shrout, P., & Rodgers, J. (2018). Research and statistical practices that promote accumulation of scientific findings. *Annual Review of Psychology*, *69,* 487–510.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science, 9*, 552-555.

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31,* 500-506.

Stanley, T. D., Carter, E. C., & Doucourliagos, H. (2017). What *meta-analyses reveal about the replicability of psychological research*. Unpublished paper, Deakin Laboratory.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L, Krueger, R. F, Miller, J. D., Oltmans, T. F. & Shrout, P. E. (2017) It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science 12,* 742–56.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105-110.

van Erp, S., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990-2013. *Journal of Open Psychology Data, 5,* 4.

Yamaguchi, Y., Maruo, K., Partlett, C., & Riley, R. D. (2017). A random effects meta-analysis model with Box-Cox transformation. *BMC Medical Research Methodology, 17,* 109.

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30,* 141-167.

Table 1

Power for a positive "+" and a negative effect "–" given an effect size (μ_δ), study variation (σ_δ), total sample size (*N)* and an alpha of .05

		σ_δ							
		0. 000		0.050		0.125		0.200	
μ_δ	*N*	+	–	+	–	+	–	+	–
0.0	100	.025	.025	.029	.029	.048	.048	.082	.082
	200	.025	.025	.032	.032	.071	.071	.128	.128
	500	.025	.025	.043	.043	.127	.127	.211	.211
	1,000	.025	.025	.062	.062	.188	.188	.277	.277
0.2	100	.166	.002	.173	.002	.205	.006	.245	.018
	200	.290	.000	.301	.001	.339	.006	.374	.026
	500	.607	.000	.594	.000	.563	.007	.544	.043
	1,000	.885	.000	.827	.000	.706	.010	.641	.061
	∞	1.000	.000	1.000	.000	.945	.055	.841	.159
0.5	100	.697	.000	.692	.000	.669	.000	.643	.001
	200	.940	.000	.929	.000	.879	.000	.817	.001
	500	1.000	.000	.999	.000	.983	.000	.931	.001
	1,000	1.000	.000	1.000	.000	.996	.000	.963	.001
	∞	1.000	.000	1.000	.000	1.000	.000	.994	.006
0.8	100	.977	.000	.974	.000	.956	.000	.922	.000
	200	1.000	.000	1.000	.000	.997	.000	.983	.000
	500	1.000	.000	1.000	.000	1.000	.000	.998	.000
	1,000	1.000	.000	1.000	.000	1.000	.000	.999	.000
	∞	1.000	.000	1.000	.000	1.000	.000	1.000	.000

[a]Probability of detecting a positive effect, i.e., one consistent with the average effect size.

[b]Probability of detecting a negative effect, i.e., one inconsistent with the average effect size.

Table 2

Power given the log-normal distribution of effect sizes for a positive "+" and a negative effect "–" given an effect size (μ_δ), study variation (σ_δ), total sample size (N), and an alpha of .05

		σ_δ							
		0. 000		0.050		0.125		0.200	
μ_δ	N	+	–	+	–	+	–	+	–
0.2	100	.166	.002	.173	.002	.194	.004	.200	.006
	200	.290	.000	.300	.001	.309	.002	.295	.004
	500	.607	.000	.589	.000	.519	.001	.456	.002
	1000	.885	.000	.828	.000	.692	.000	.591	.001
	∞	1.000	.000	1.000	.000	1.000	.000	1.000	.000
0.5	100	.697	.000	.691	.000	.665	.000	.628	.000
	200	.940	.000	.930	.000	.884	.000	.825	.000
	500	1.000	.000	.999	.000	.991	.000	.965	.000
	1000	1.000	.000	1.000	.000	1.000	.000	.994	.000
	∞	1.000	.000	1.000	.000	1.000	.000	1.000	.000
0.8	100	.977	.000	.974	.000	.958	.000	.930	.000
	200	1.000	.000	1.000	.000	.998	.000	.992	.000
	500	1.000	.000	1.000	.000	1.000	.000	1.000	.000
	1000	1.000	.000	1.000	.000	1.000	.000	1.000	.000
	∞	1.000	.000	1.000	.000	1.000	.000	1.000	.000

[a]Probability of detecting a positive effect, i.e., one consistent with the average effect size.

[b]Probability of detecting a negative effect, i.e., one inconsistent with the average effect size.

Table 3

95 Percent Confidence Interval for the Effect with Lower (L) and Upper (U) Limits from a Single Study for Different Sample Sizes (n), Effect Sizes (μ_δ), and Level of Heterogeneity (σ_δ)

`

		σ_δ							
		0.000		0.050		0.125		0.200	
μ_δ	N	L	U	L	U	L	U	L	U
0.2	100	-0.192	0.592	-0.204	0.604	-0.262	0.662	-0.354	0.754
	200	-0.077	0.477	-0.094	0.494	-0.170	0.570	-0.280	0.680
	500	0.025	0.375	-0.001	0.401	-0.101	0.501	-0.229	0.629
	1,000	0.076	0.324	0.042	0.358	-0.075	0.475	-0.211	0.611
0.5	100	0.108	0.892	0.096	0.904	0.038	0.962	-0.054	1.054
	200	0.223	0.777	0.206	0.794	0.130	0.870	0.020	0.980
	500	0.325	0.675	0.299	0.701	0.199	0.801	0.071	0.929
	1,000	0.376	0.624	0.342	0.658	0.225	0.775	0.089	0.911
0.8	100	0.408	1.192	0.396	1.204	0.338	1.262	0.246	1.354
	200	0.523	1.077	0.506	1.094	0.430	1.170	0.320	1.280
	500	0.625	0.975	0.599	1.001	0.499	1.101	0.371	1.229
	1,000	0.676	0.924	0.642	0.958	0.525	1.075	0.389	1.211

Table 4

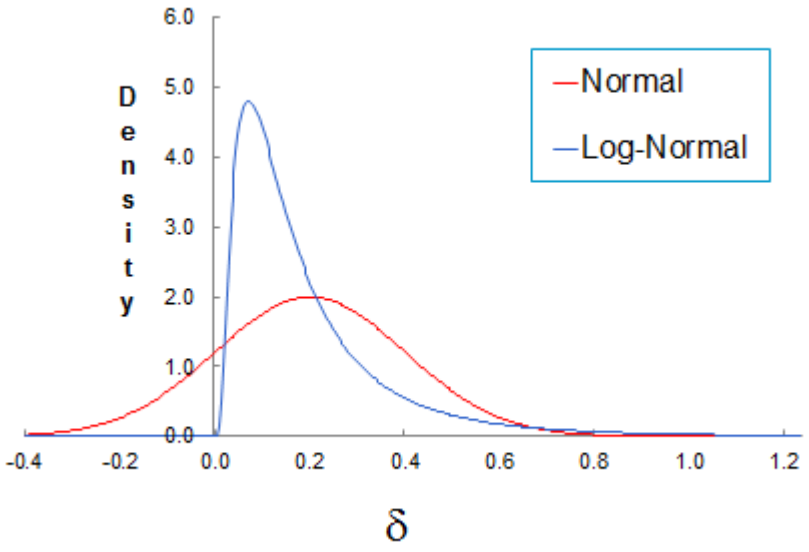
95 Percent Confidence Interval for the Mean Effect with Lower (L) and Upper (U) Limits from Five Studies for Different Sample Sizes (n), Effect Sizes (μ_δ), and Level of Heterogeneity (σ_δ)

`

		σ_δ							
		0.000		0.050		0.125		0.200	
μ_δ	N	L	U	L	U	L	U	L	U
0.2	100	0.025	0.375	0.019	0.381	-0.007	0.407	-0.048	0.448
	200	0.076	0.324	0.069	0.331	0.035	0.365	-0.015	0.415
	500	0.122	0.278	0.110	0.290	0.065	0.335	0.008	0.392
	1,000	0.145	0.255	0.129	0.271	0.077	0.323	0.016	0.384
0.5	100	0.325	0.675	0.319	0.681	0.293	0.707	0.252	0.748
	200	0.376	0.624	0.369	0.631	0.335	0.665	0.285	0.715
	500	0.422	0.578	0.410	0.590	0.365	0.635	0.308	0.692
	1,000	0.445	0.555	0.429	0.571	0.377	0.623	0.316	0.684
0.8	100	0.625	0.975	0.619	0.981	0.593	1.007	0.552	1.048
	200	0.676	0.924	0.669	0.931	0.635	0.965	0.585	1.015
	500	0.722	0.878	0.710	0.890	0.665	0.935	0.608	0.992
	1,000	0.745	0.855	0.729	0.871	0.677	0.923	0.616	0.984

Figure 1

Normal and Log-normal Distributions with μ_δ = 0.2 and σ_δ = 0.2



Appendix

R Function for Computing Power Given Heterogeneity

```
powj=NULL
#  The arguments are
#      the mean of the deltas (mu_delta),
#      the standard deviation of the deltas (heterogeneity: sigma_delta);
#  total sample size (cell size times two: N), and alpha (alph).
pow_ad  = function (mu_delta,sigma_delta,N,alph)
{
q=sqrt((4/N)/(4/N+sigma_delta^2))
z2T = qt(1-alph/2.,N-2)
cr_t=z2T*q
alph_a=(1-pt(cr_t,N-2))*2
ncp = (mu_delta/sqrt(4/N+sigma_delta^2))
powj[1]=  1 - pt(cr_t,N-2,ncp)
powj[2]=   pt(-cr_t,N-2,ncp)
powj[3]=   2*pt(-cr_t,N-2,0)
return(powj)
}
powj =pow_ad(.3,.20,100, .05)
pp1 = paste0("Power of a positive effect is ",round(powj[1],3),"."); pp1
pp2 = paste0("Power of a negative effect is ",round(powj[2],3),"."); pp2
pp3 = paste0("Effective alpha is ",round(powj[3],3),"."); pp3
```

Footnotes

[1]The estimated tau values are not included in the text but in supplementary files located at https://osf.io/43a8t/.

[2]One might argue that it might be more appropriate not to use the effective alpha when there is heterogeneity, but rather should use some predetermined value, e.g., .05 or .005. Such an approach can leadsto very low levels of power. For the example that follows later in the paragraph, if the effective alpha were set to .05, then the original alpha would need to be re-set to .006, and the power would be just .182.

[3]Some readers might be interested in knowing the effects of a more stringent alpha, e.g., .005, on power. For sample sizes resulting in 80 percent power, moving from .05 to .005, heterogeneity creates an increase in the decline in power.

[4]Non-normal distributions have been regularly assumed for level-one variances in multilevel models (e.g., Hedeker, Mermelstein, & Demirtas, 2012).

[5]Hedges and Olkin (p. 86, 1985) give the standard deviation for *d* as $\sqrt{\frac{4}{N} + \frac{\delta^2}{2N}}$, where δ is the population value of *d* for the study and *N* is the study sample size with the assumption that $n_1 = n_2$. This formula is used throughout this paper, but because δ varies due to heterogeneity, the second term is not a constant and so it is dropped. Doing so results in an underestimation of the standard deviation of *d*, but nonetheless the points

that make about the effect of increasing heterogeneity, sample size, and number of studies still hold.

[6]It is possible to estimate the optimal number of studies and optimal number of participants per study needed to minimize the standard error of the effect, given level of resources. To obtain these values, we would apply the standard formulas used in cluster sampling (Sudman, 1976) using the costs per study and per participant. Alternatively, a multilevel power analysis could be undertaken treating studies as level two and participants as level one.

[7]Sometimes researchers mistakenly check to see if the original effect size is in the confidence interval of the replication effect size. Such a practice is flawed because it ignores that the original effect has sampling error (Maxwell et al., 2015).