American Psychological Association

Psychological Methods

1082-989X/19/$12.00

2019, Vol. 24, No. 5, 578–589
http://dx.doi.org/10.1037/met0000209

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

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 because of 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 and an interest in generalizing to a population of studies, 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 plan accordingly.

Translational Abstract

Repeated investigations of the same phenomenon typically yield somewhat different results that vary more than one would expect from the fact that the investigations have different participants. Even when the very same phenomenon is studied, such variation is found, which implies there might be subtle random variation across studies. Such heterogeneity of effects is typically ignored, with unfortunate consequences. We consider its implications for determining the chances that the study would have a statistically significant effect, the statistical precision of estimated effects, and the planning of original and replication research. Researchers and methodologists need to recognize that effects are often heterogeneous and plan accordingly.

Keywords: power, replication, meta-analysis, precision, heterogeneity

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 *SD*, over and above what might be expected from sampling error. For instance, if the effect size is a Cohen's *d*, its true *SD* 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 χ^2 test of homogeneity, symbolized as *Q* (Cochran, 1954). (Several studies, e.g., Chung, Rabe-Hesketh, and Choi [2013], have shown that this test has relatively low power to detect heterogeneity.) 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 *SD* 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

This article was published Online First February 11, 2019.

David A. Kenny, Department of Psychology, University of Connecticut; Charles M. Judd, Department of Psychology and Neuroscience, University of Colorado Boulder.

A prior version of this article was posted at https://osf.io/b6dtc/ and on David A. Kenny's Web page. We especially thank Gary McClelland, Scott Maxwell, Christopher Rhoads, Blakeley McShane, and Felix Schönbrodt, as well as Deborah Kashy, Joshua Correll, Vincent Yzerbyt, Simine Vazire, Blair Johnson, Betsy McCoach, Lijuan Wang, and Thomas Ledermann who provided us with helpful feedback.

Correspondence concerning this article should be addressed to David A. Kenny, Department of Psychology, University of Connecticut, Unit 1020, Storrs, CT 06269-1020. E-mail: david.kenny@uconn.edu

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% of meta-analyses with 60 or more studies find some level of heterogeneity. Finally, a recent survey of 200 meta-analyses (Stanley, Carter, & Doucouliagos, 2018) found that 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. Undoubtedly, these factors also affect estimates of heterogeneity. Indeed McShane, Böckenholt, and Hansen (2016) (see also Augusteijn, van Aert, & van Assen, 2018) conclude that

> Publication bias distorts meta-analytic estimates of both the population average effect size and the degree of heterogeneity. Estimates of the former are typically biased upward, thus giving the false impression of large effect sizes, whereas estimates of the latter are typically biased downward, thus giving the false impression of homogeneity (p. 732).

Thus, there are good reasons to be cautious about estimating heterogeneity of effect sizes from meta-analyses. Fortunately, given the current interest in replication and the controlled replication studies conducted under the Many Labs 1 project of Klein et al. (2014), Registered Replication Reports (RRR) of Simons, Holcombe, and Spellman (2014), a meta-analysis of "close" replications by Linden and Hönekopp (2018), and the Many Labs 2 project of Klein et al. (2018), one can begin to estimate heterogeneity even in studies that all use the same designs, materials, analytic methods, and outcomes. Additionally, these replication studies permit one to estimate it in the absence of publication biases.

The Many Labs 1 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 studies, 13 used d and their heterogeneity[1] was significantly greater than zero in 8 cases, despite the low power of the Q test, with an average SD for the 13 studies of 0.21. Effect size variation was highly correlated with the average effect size, $r = .86$. Moreover, typically the SD of the true effect sizes was about 25% 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 1 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.

Linden and Hönekopp (2018) located 40 meta-analyses of what they called "close replications" (that largely includes Many Labs 1 and the RRRs) and found relatively low levels of heterogeneity, the average value of $\hat{\tau}$ being just 0.08. However, the average effect size was also relatively small being just 0.24, and the correlation between effect size and heterogeneity in these 40 meta-analyses was .70.

The Many Labs 2 project (Klein et al., 2018) tested 28 effects in about 60 laboratories and 15,305 total participants from 36 countries and territories. As with prior studies, they found very little heterogeneity when effect sizes were near zero. For the 12 studies with average $d > 0.20$, the average value of $\hat{\tau}$ was .072.

We are not the only ones to remark on the rather surprising finding of heterogeneity in studies that are essentially 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, Böckenholt, & Gelman, in press).

When there are nonzero 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 because of random factors that we can never completely specify or control. At a later point, we discuss this possibility in more detail. Regardless of whether heterogeneity is because of 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 article 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 are others. We do not have 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 simply acting as if it 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 ith 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

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

distribution of these true effect sizes, with a mean, denoted as μ_δ, and a *SD*, denoted as σ_δ. To be clear, σ_δ refers to the *SD* 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–δ_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 sections, we discuss two underappreciated 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 article, we discuss 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. In conventional power analysis, the null hypothesis is that the effect size for the planned study is zero, that is, $\delta_i = 0$. Alternatively, the interest might be in the average of all possible effect sizes; the effect size is a random variable with a mean of μ_δ and a *SD* of σ_δ, making the null hypothesis $\mu_\delta = 0$.

McShane and Böckenholt (2014) developed a method to determine power given heterogeneity using the *Z* distribution, and we adapt and extend their method to determine power given heterogeneity. Like them, 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 consider power for the test of the mean of the true effect sizes, μ_δ.

To develop the rationale for our power computations, 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*, 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 α of .05. Alternatively, if the interest is in testing the null hypothesis that μ_δ is zero and $t_{c(df)}$ is still used as the critical value, the effective α is larger than the nominal α because of variation in the δ_i values. Sometimes the δ_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 α given heterogeneity and *N*, the *SE* needs to include heterogeneity as well as sampling error, 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 α 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 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 α of $P[(0.7071)(-1.9845) > t)] + P[(0.7071)(1.9845) < t] = .162$.

We now turn to power, no longer assuming that the true effect size is zero. To estimate power for the test that $\delta_i = 0$, we must determine the appropriate noncentrality parameter for the noncentral *t* distribution. The noncentrality parameter is defined as the study's true effect size or δ_i divided by its *SE*, $2/\sqrt{N}$. With this noncentrality parameter and α, one computes $P(-t_{c(df)} > (t')) + P(t_{c(df)} < t')$, where t' is distributed as a noncentral *t* with $N - 2$ *df* and a noncentrality parameter of δ_i divided by $2/\sqrt{N}$. Alternatively, if the null hypothesis is that $\mu_\delta = 0$, the *SE* is $\sqrt{\frac{4}{N}+\sigma_\delta^2}$. Thus, with heterogeneity, the noncentrality parameter equals μ_δ divided by $\sqrt{\frac{4}{N}+\sigma_\delta^2}$. With this noncentrality parameter and effective α, one computes $P(-qt_{c(df)} > (t')) + P(qt_{c(df)} < t')$, where t' is distributed as a noncentral *t* with $N - 2$ *df* and a noncentrality 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 noncentrality 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 estimates of power with heterogeneity, using the effective α and the adjusted noncentrality parameter.

To be clear, we are computing power for a fixed value of the effect size, either δ_i or μ_δ. Several others (McShane & Böckenholt, 2016; Perugini, Gallucci, & Costantini, 2014) have suggested computing power across a distribution of effect sizes because of uncertainty in the estimate of the effect sizes (usually sampling error).[2] 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, heterogeneity is assumed to have 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 *SD* 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 α set at .05. (Recall that *N* is the total sample size of the study.) The *SD*s of the effect sizes, σ_δ, are chosen to be equal to 25% 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 of δ. For each combination of values, two probabilities are given: The one denoted by "+" is the probability of reject-

[2] McShane and Böckenholt (2016) make the interesting and important point (see their Footnote 1) that methods to correct for study sampling error in the estimate of δ for a power analysis, can also be used to correct for heterogeneity in effect sizes. They develop a method to determining the requisite sample size for desired power given sampling error in the effect estimate based on the *Z* distribution. The method we have developed here for determining power given heterogeneity using the *t* distribution can be adapted to additionally handle sampling error in the estimate of μ_δ by replacing σ_δ by $\sqrt{\sigma_\delta^2 + s_d^2}$, where s_d^2 is the sampling variance in the estimate of μ_δ.

Table 1
Power[a] for a Positive Effect "+" (in the Same Direction as the Average Effect) and a Negative Effect "−" (in the Opposite Direction as the Average Effect) Given an Effect Size (μ_δ), Study Variation (σ_δ), Total Sample Size (N), and an α of .05

		σ_δ							
		.000		.050		.125		.200	
μ_δ	*N*	+	−	+	−	+	−	+	−
.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
.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
.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
.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] When $\mu_\delta = 0$, the table entry is not power, but the probability of making a Type I error.

ing 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. We shall see that these conclusions change when there is heterogeneity.

Next, we consider the case when the null hypothesis is true, that is, $\mu_\delta = 0.0$. Note that with heterogeneity, although the average effect size is zero, any sample value is likely nonzero and may well be statistically significant. We see that with increasing heterogeneity and sample sizes, the probability of a Type I error increases, well above the nominal α value of .05. In fact when $\sigma_\delta = 0.2$ and $N = 1{,}000$, the probability of making a Type I error is over 55%! It is important to be clear about what we mean by this Type I error. We are referring to 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. Note that these probabilities become the effective α in determining power when μ_δ is no longer zero.

In the remainder of this section, we consider the power when there is a nonzero 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 .168 with no heterogeneity and rises to .263 when σ_δ is 0.2.[3] The increase in power is because of the fact that if we assume that 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 of 0.2 is overestimated or underestimated by 0.1. When it is overestimated, there is bigger boost in power (.318, a boost of .150) than when it is underestimated by the same amount (.078, a loss of .090). Because the gain in power when δ_i is larger than 0.2 is greater than the loss in power when it smaller, there is a net increase in 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 .818 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 because of heterogeneity has been noted previously by McShane and Böckenholt (2014).

Third, as already noted, the "–" values in Table 1 give 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 a 6% chance of finding a significant result in the opposite direction from the average effect size. Of interest

[3] The values that we give here are the sum of the "+" and "–" values in the table. Thus, we define power as the probability of rejecting the null hypothesis regardless of whether the sampled d_i is in the same or opposite direction of the value of μ_δ.

to the authors, Linden and Hönekopp (2018) estimate that the true effect or δ_i is in the opposite direction from μ_δ for about 9% of the studies that they included in their review of "close" replication studies. Additionally, the Many Labs 2 project of Klein et al. (2018) reports a number of replication studies with significant effects in the opposite direction from the average effects found.

Fourth, related to the previous point, there are nonobvious 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 (see McShane & Böckenholt, 2016). 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 nontrivial 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 nontrivial. This probability increases as N increases.

Nonnormal 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 (CIs) 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 δ_i values 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 nonnegative. There has been some work on specifying alternatives to the normal distribution for random effects (Lee & Thompson, 2008; Yamaguchi, Maruo, Partlett, & Riley, 2017). Nonnormal distributions have been regularly assumed for level-one variances in multilevel models (e.g., Hedeker, Mermelstein, & Demirtas, 2012).

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,

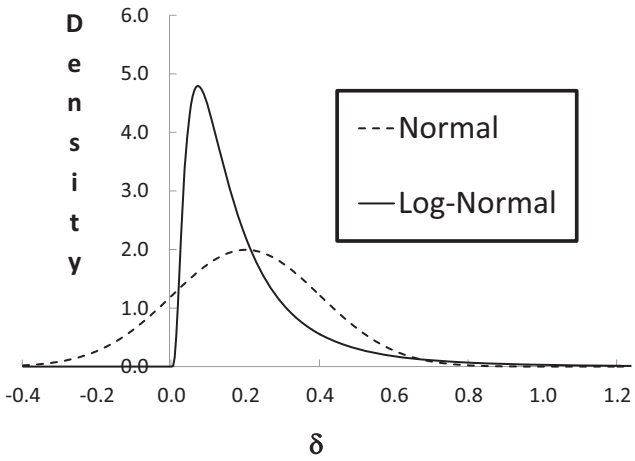

Figure 1. Normal and Log-normal distributions with $\mu_\delta = 0.2$ and $\sigma_\delta = 0.2$.

negative values occur about 17% 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 nonnegative. We note that there appears to be some empirical evidence from both Many Labs 1 and 2, as well as Linden and Hönekopp that when the average effect size is zero there is little or no evidence of heterogeneity.

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. For the values in Tables 1 and 2, 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, that is, 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 α is no different from the nominal α value.

Table 2
Power Given the Log-Normal Distribution for a Positive Effect "+" (in the Same Direction as the Average Effect) and a Negative Effect "−" (in the Opposite Direction as the Average Effect) Given an Effect Size (μ_δ), Study Variation (σ_δ), Total Sample Size (N), and an α of .05

		σ_δ							
		.000		.050		.125		.200	
μ_δ	*N*	+	−	+	−	+	−	+	−
.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
.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
.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

Precision in Estimating Effect Sizes Given Heterogeneity

The effect size in a study provides an estimate of the true effect size. Its *SE* permits estimation of the CI for that true effect size. Assuming a fixed effect size, the *SE* derives solely from the sampling error within a study. For δ_i, this can be approximated[4] by $2/\sqrt{N}$. If there is study variation, this is the *SE* for δ_i, the true effect size for the particular study, and not μ_δ, the mean of all possible effect sizes. However, if the interest is in the population of effect sizes or μ_δ, the approximate error is $\sqrt{4/N + \sigma_\delta^2}$.

Table 3 presents the 95% CI for μ_δ, given an estimated d_i 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 CI becomes narrower as the study *N* increases. They also show, again perhaps unsurprisingly, that as effect heterogeneity increases, the CI for the true effect size becomes wider. This difference can be quite dramatic. Looking at the last row of Table 3, the width of CI 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 CIs 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 CI. 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 CI for the average effect size decreases as a function of essentially pooling the observations from all studies into a single *SE*, the approximate formula being $\sqrt{(4/N + \sigma_\delta^2)/k}$ where *k* is the number of studies. In Table 4, we present the CIs 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 CIs for the first columns in Tables 3 and 4, where there is no heterogeneity of effect sizes, that is, $\sigma_\delta = 0.0$. 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 approximate CI is exactly the same with one study having an *N* of 500 as for five studies each with an *N* of 100.

More important, however, if the interest is generalizing over the population of studies with heterogeneity, then there are substantial precision benefits that accrue from multiple smaller studies compared with 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 CI 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 and *N* of 1000, the approximate CI is 27.5% wider than it is if there is no heterogeneity.

[4] Hedges and Olkin (1985, p. 86) give the *SD* 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 article, but because δ varies because of heterogeneity, the second term is not a constant and so it is dropped. Doing so results in an underestimation of the *SD* of *d*, but nonetheless the points that make about the effect of increasing heterogeneity, sample size, and number of studies still hold.

Table 3
Lower (L) and Upper (U) Limits of a 95% Confidence Interval of an Effect From a Single Study for Different Sample Sizes (N), Effect Sizes (μ_δ), and Level of Heterogeneity (σ_δ)

		σ_δ							
		.000		.050		.125		.200	
μ_δ	*N*	L	U	L	U	L	U	L	U
.2	100	−.192	.592	−.204	.604	−.262	.662	−.354	.754
	200	−.077	.477	−.094	.494	−.170	.570	−.280	.680
	500	.025	.375	−.001	.401	−.101	.501	−.229	.629
	1,000	.076	.324	.042	.358	−.075	.475	−.211	.611
.5	100	.108	.892	.096	.904	.038	.962	−.054	1.054
	200	.223	.777	.206	.794	.130	.870	.020	.980
	500	.325	.675	.299	.701	.199	.801	.071	.929
	1,000	.376	.624	.342	.658	.225	.775	.089	.911
.8	100	.408	1.192	.396	1.204	.338	1.262	.246	1.354
	200	.523	1.077	.506	1.094	.430	1.170	.320	1.280
	500	.625	.975	.599	1.001	.499	1.101	.371	1.229
	1,000	.676	.924	.642	.958	.525	1.075	.389	1.211

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 if the interest is in μ_δ. 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 nonrandom 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 is 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 CI 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 CIs with nonnormal heterogeneity.

Determining 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 Böckenholt (2014), knowing exactly the magnitude of σ_δ is difficult. We might use estimates from prior research, but simulation studies (e.g., Chung et al., 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.

Table 4
Lower (L) and Upper (U) Limits of a 95% Confidence Interval of the Mean Effect From Five Studies for Different Sample Sizes (N), Effect Sizes (μ_δ), and Level of Heterogeneity (σ_δ)

		σ_δ							
		.000		.050		.125		.200	
μ_δ	*N*	L	U	L	U	L	U	L	U
.2	100	.025	.375	.019	.381	−.007	.407	−.048	.448
	200	.076	.324	.069	.331	.035	.365	−.015	.415
	500	.122	.278	.110	.290	.065	.335	.008	.392
	1,000	.145	.255	.129	.271	.077	.323	.016	.384
.5	100	.325	.675	.319	.681	.293	.707	.252	.748
	200	.376	.624	.369	.631	.335	.665	.285	.715
	500	.422	.578	.410	.590	.365	.635	.308	.692
	1,000	.445	.555	.429	.571	.377	.623	.316	.684
.8	100	.625	.975	.619	.981	.593	1.007	.552	1.048
	200	.676	.924	.669	.931	.635	.965	.585	1.015
	500	.722	.878	.710	.890	.665	.935	.608	.992
	1,000	.745	.855	.729	.871	.677	.923	.616	.984

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 1 project. There, heterogeneity averaged roughly one quarter the effect size. Following the original suggestion by Pigott (2012); McShane and Böckenholt (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 0.050, 0.125, and 0.200 as representative in the power and precision results that we gave earlier. We expect 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 Böckenholt (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 CIs 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 (Linden & Hönekopp, 2018). Additionally, as we discussed, even when studies are conducted using standardized procedures and measures, effect size heterogeneity typically persists (see McShane et al., 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 because of 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, 2005) that "God does not play dice," is wrong, and studies vary for reasons that will never be completely understood. Whether because of 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 eliminate heterogeneity by controlling for a few, or even a lot, of such moderators. There will 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 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 nonrandom processes, for example, publication bias. For instance, 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 can create artificial heterogeneity even when there is none.

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 of μ_δ. 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.[5] Anticipating an average effect size, μ_δ, of 0.4 and 154 participants, the conventional conditional power estimate is .694 for δ_i, which is not very good. Even worse, if we allow for heterogeneity in effect sizes of 0.20, the power of the test for μ_δ is only .628 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 seven 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 .628, 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 as the number of persons per study and setting the number of studies or *k* to 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 *SE* of $\bar{d}$ for 7 studies each with $N = 154$ is about half the size the *SE* 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, even if the underpowed *Q* test is "not significant." 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. 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% 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,

[5] It is possible to estimate the optimal number of studies and optimal number of participants per study needed to minimize the *SE* 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 1.

2017; Maxwell et al., 2015). Second, over and above the sampling error in the original effect size estimate, because of 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% of the original, whereas Klein et al. (2018) find it at 25%. Not surprisingly, the sampling of extreme scores, that is, 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 nontrivial 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 are 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 CI of the difference between the effect in the original study and the effect in the replication study.[6] Doing so presumes that the only source of variance between the two studies is sampling error. However, given heterogeneity, the width of the CI would be greater than that based on sampling error alone. For instance, consider two studies each with an *N* of 200 and estimated effect sizes of 0.60 and 0.15. The 95% CI 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 CI actually goes from −0.044 to 0.944, which now includes zero. Ignoring study variation leads to too narrow a CI 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. We believe that such a view is 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 nonrandomly selected participants. All of these, and a variety of other factors, likely lead to heterogeneity. 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.

Perhaps the inevitability of effect heterogeneity begins to lead to a new view of how one does research and the nature of scientific conclusions. Rather than seeking out the "true" effect and demonstrating it once and for all, one approaches science from a contextual perspective with a goal of understanding all the complexities that underlie effects and how those effects vary. To quote from Gelman (2015)

> Once we accept that treatment effects vary, we move away from the goal of establishing a general scientific truth from a small experiment, and we move toward modeling variation. . . . We move away from is-it-there-or-is-it-not-there to a more helpful, contextually informed perspective (p. 636).

More important, we are not saying that the search to establish effects in psychology is misguided. There are indeed "true" effect out there, documented by research and yet to be discovered. However, there are a myriad of factors, which cause the magnitude of those effects, when they exist, to vary.

Conclusion

Effect size heterogeneity is found nearly everywhere in science. However, in power analyses, computing CIs, 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 350 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

[6] 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).

out how to include such heterogeneity in both analysis of data and the planning of research.

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, for example, correlations and odd ratios. Work by McShane and Böckenholt (2016), which discusses several other effects sizes besides *d*, would almost certainly be relevant.

These issues notwithstanding, we firmly believe that we need to accept and, in fact, embrace heterogeneity (McShane et al., in press). If there truly exist multiple effect sizes in a given domain, then power analyses and CIs 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., & 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. http://dx.doi.org/10.1080/00273171.2017.1289361

Augusteijn, H. E. M., van Aert, R. C. M., & van Assen, M. A. L. M. (2018). The effect of publication bias on the assessment of heterogeneity. *Psychological Methods, 37,* 2547–2666. http://dx.doi.org/10.1002/sim.7665

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. http://dx.doi.org/10.1002/sim.5821

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129. http://dx.doi.org/10.2307/3001666

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.

Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research, 51,* 589–605. http://dx.doi.org/10.1080/00273171.2016.1191324

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P. A., . . . Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11,* 158–171. http://dx.doi.org/10.1177/1745691615605826

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management, 41,* 632–643. http://dx.doi.org/10.1177/0149206314525208

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11,* 546–573. http://dx.doi.org/10.1177/1745691616652873

Hawkins, R. X. D., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., . . . Frank, M. C. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science, 1,* 7–18. http://dx.doi.org/10.1177/2515245917740427

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. http://dx.doi.org/10.1002/sim.5338

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. http://dx.doi.org/10.1037/1082-989X.3.4.486

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Data from investigating variation in replicability: A "many labs" replication project. *The Journal of Open Psychology Data, 2,* p. e4. http://dx.doi.org/10.5334/jopd.ad

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science, 1,* 443–490. http://dx.doi.org/10.1177/2515245918810225

Lee, K. J., & Thompson, S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine, 27,* 418–434. http://dx.doi.org/10.1002/sim.2897

Linden, A. H., & Hönekopp, J. (2018). *Heterogeneity in the results of close and conceptual replications: Implications for scientific progress and practical applications*. Unpublished paper, Northumbria University.

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, 70,* 487–498. http://dx.doi.org/10.1037/a0039400

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science, 9,* 612–625. http://dx.doi.org/10.1177/1745691614548513

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21,* 47–60. http://dx.doi.org/10.1037/met0000036

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11,* 730–749. http://dx.doi.org/10.1177/1745691616662243

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9,* 319–332. http://dx.doi.org/10.1177/1745691614528519

Pigott, T. (2012). *Advances in meta-analysis*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2278-5

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. http://dx.doi.org/10.1037/1089-2680.7.4.331

Shrout, P., & Rodgers, J. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69,* 487–510. http://dx.doi.org/10.1146/annurev-psych-122216-011845

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. http://dx.doi.org/10.1177/1745691614543974

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. http://dx.doi.org/10.3102/0162373709352369

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

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9,* 59–71. http://dx.doi.org/10.1177/1745691613514450

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

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12,* 742–756. http://dx.doi.org/10.1177/1745691617690042

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110. http://dx.doi.org/10.1037/h0031322

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. E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30,* 141–167. http://dx.doi.org/10.3102/10769986030002141

Appendix

R Function for Computing Power Given Heterogeneity

```
powj = NULL
# The arguments are
#  The mean of the deltas (mu_delta),
#  The SD of the deltas (heterogeneity: sigma_delta),
#  Total sample size (cell size times two: N),
#  The nominal 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
```

Received May 11, 2018
Revision received November 26, 2018
Accepted December 24, 2018 ■