# 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. We especially thank Gary McClelland, Christopher Rhoads, Blakeley McShane, and Felix Schönbrodt, as well as Scott Maxwell, Deborah Kashy, Joshua Correll, Vincent Yzerbyt, Simine Vazire, Blair Johnson, Betsy McCoach, and Thomas Ledermann who provided us with helpful feedback. David A. Kenny came up with the original idea for the paper. Each participated nearly equally in the preparation of the paper.

## 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. Although such heterogeneity of effect sizes is well documented, its implications for power analyses, the precision of effects, conducting replication studies, and the planning of research are not fully appreciated. In the presence of heterogeneity, there exists a range of effects that research studies might uncover, possibly including effects that go in the opposite direction from the average effect reported in the literature. Additionally, the usual power calculations and confidence intervals are 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.

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.

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 $\sigma_{\delta}$. We subscript with $\delta$, 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*. Not always reported is the estimated true variance, generically called $\tau^2$ 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 (*r*) of .21 and an average true standard deviation of those effect sizes of .15. More recently and more broadly, van Erp, Verhagen, Grasman, and Wagenmakers (in press) have made available a database of every meta-analysis published in the *Psychological Bulletin* from 1990 to 2013. The average value of tau of studies using *d* or *g* is 0.24 (189 meta-analyses) and for *r* it is .13 (502 meta-analyses). Moreover, 96 percent of meta-analyses with 60 or more studies find some level of heterogeneity.

Numerous issues potentially bias effect size estimates in reported meta-analyses: file drawer problems, *p*-hacking strategies, and publication biases. Likely, these factors also affect estimates of heterogeneity and, accordingly, it is sensible to be cautious about estimates of heterogeneity from meta-analyses. Fortunately, given the current interest in replications, there is 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 get around the publication biases 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. Thus, it is possible to find effect size heterogeneity without the biases inherent in meta-analyses of published studies and in studies that are procedurally very similar.

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 was significantly greater than zero in 8 cases with an average standard deviation for the 13 studies of 0.21. Study variation was highly correlated with effect size *d*, $r = .86$. Moreover, typically study variation was about 25 percent of the value of the study's *d*.

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 the weak effect sizes from the Many Labs project, it is then not surprising that the effect sizes in these studies generally, though not always, have relatively small levels of variation.

Finding that effect sizes in a domain vary over and above what might be expected from sampling error implies that there is a range of true effect sizes that exist, rather than a single one. 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, replication studies, and the conduct of research.

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, $\delta_i$. Across studies, there is a distribution of these true effect sizes, with a mean, denoted as $\delta$, and a standard deviation, denoted as $\sigma_\delta$. To be clear, $\sigma_\delta$ refers to the standard deviation after sampling error has been removed and is denoted as $\tau$ in the meta-analysis literature. For a particular study $i$, the effect size $d_i$, has two parts: its true effect size, $\delta_i$, and its sampling error, $d_i - \delta_i$. In sum, there is a mean effect size ($\delta$), and there are multiple *true* effect sizes rather than a single one. We assume that in any particular literature, we have a sample of all methodologically sound studies, and thus this sample provides estimates of both $\delta$ and $\sigma_\delta$. In Supplemental Material (davidakenny.net/doc/KJ_Sup17.pdf), we discuss the difficulties underlying these assumptions.

In the next two sections, we discuss two under-appreciated consequences of 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 moderators of heterogeneity. In the final section of the paper, we talk about the implications of heterogeneity on replications and the planning of 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, treats the effect size as fixed. 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.

We initially presume that the distribution of effect sizes is normal. At a later point, we consider other distributions that might also be reasonable. McShane and Böckenholdt (2014) have developed a straightforward way to determine the necessary sample size for a fixed level of power and known value of heterogeneity using the normal distribution. We have adapted their procedure for power computations using the *t* distribution. One first determines the critical *t* for a given sample size and alpha. One then multiples both this critical *t* and *d* by $\sqrt{\frac{\frac{4}{N}}{\frac{4}{N}+ {\sigma_\delta}^2}}$. With this adjusted critical *t*, one determines the new alpha and using that new alpha and adjusted *d*, one conducts a conventional power analysis. A web-based program is available at https://davidakenny.shinyapps.io/SVPower. To be clear, we are computing power for a fixed

value of $\delta$ and $\sigma_\delta$. Several others (Biesanz & Schrager, 2010; Dallow & Fina, 2011; Gillett 1994; 2002; Perugini, Gallucci, & Costantini, 2014; Schönbrodt & Wagenmakers, 2017) 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. Here we are considering only the variability in effect sizes due to heterogeneity.

Table 1 presents power estimates for a range of positive values of $\delta$ (0.2, 0.5, and 0.8, corresponding to small, medium and large average effects), a range of values of the standard deviation of effect sizes, $\sigma_\delta$ (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[1] set at .05. The standard deviations of the effect sizes, $\sigma_\delta$, are chosen to be equal to 25 percent of small, medium and large effects based on past meta-analyses and organized replication endeavors. Note that when $\sigma_\delta$ is equal to 0.0, there is no effect size heterogeneity and the results in the table are consistent with a conventional power analysis. For each combination of values, two probabilities are given: The one denoted by "+" is the probability of rejecting the null hypothesis when $\delta_i$ is positive (i.e., in the same direction as $\delta$), and the one denoted by "–" is the probability of rejecting the null hypothesis when $\delta_i$ 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.

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 $\delta = 0.2$ and $N = 100$, power is .166 with no heterogeneity and rises to .245 when $\sigma_\delta$ is 0.2. The increase in power is due to the fact that we assume the true effect sizes are normally distributed, and there is an asymmetry in the power function: If the effect size is overestimated by 0.1, there is bigger boost in power (.318, a boost of .152) than the loss when it is underestimated by 0.1 (.071, a loss of .095). As a result, half the time the value of $\delta_i$ is larger than 0.2, its mean, which results in 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 $\delta = 0.5$ and $N = 200$, power is an impressive .940 with no heterogeneity but sinks to .817 when $\sigma_\delta$ is 0.2. When power is greater than .50, the asymmetry works in the opposite direction: It 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).

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, $\delta = 0.2$, and $\sigma_\delta$ of 0.2, there is nearly a 6 percent chance of finding a significant result in the opposite direction from the average effect size.

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 $\sigma_\delta$ to $\delta$ ratio. As it gets larger, there is 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 $\delta = 0.15$ 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., $\delta$ is positive, but the test shows a significant negative effect) is non-trivial. This probability increases as *N* increases.

## 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 an 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 $\delta$, this can be closely approximated[2] by $2/\sqrt{N}$. If there is study variation, this is the standard error for $\delta_i$, the true effect size for the particular study, and not $\delta$, the mean of all possible effect sizes. In the presence of effect size variation, the proper standard error for $\delta$ is $\sqrt{4/N + {\sigma_\delta}^2}$.

Table 2 presents the 95% confidence interval for $\delta$, given an estimated *d* from a study with the indicated *n*, assuming varying degrees of known heterogeneity. 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 2, 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 $\sigma_\delta$ 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 2 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. If multiple studies were run, all estimating the same average effect, 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 3, we present the confidence intervals for $\delta$ 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 $\sigma_\delta$.

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 2 and 3, 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 2 where *N* equals 500 with the rows in Table 3 where the *N* equals 100 in each study, for a combined *N* across the studies of 500. If there is effect size heterogeneity, the confidence interval for $\delta$ 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 2 to *N* = 200 in Table 3.

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 2 and 3 make clear, if effect sizes vary across studies, such a strategy is misguided. Certainly, our point is not that small *N* studies are better than large *N* studies, but rather that large *N* studies are not as informative as we might think, given effect size heterogeneity.

## Knowing the Magnitude and Distribution 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. However, knowing exactly the magnitude of $\sigma_\delta$ 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. 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.

For precision, knowing the value of $\sigma_\delta$ 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) and Maxwell et al. (2015), might be attempted, perhaps using the van Erp et al. (2017) database to create a prior distribution.

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 no heterogeneity.

Following McShane and Böckenholdt (2014), we have assumed 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 good reasons to think that the true effect sizes may not be normally distributed. For instance, if the average true effect is positive, it may be implausible that some effects are in fact negative. A more reasonable position might be 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 positive. There has been some work on specifying alternatives to the normal distribution for random effects (Lee & Thompson, 2007). One candidate that we have explored is a log-normal one. The advantage of such a distribution is that it has only positive values, and so finding negative effect sizes can only happen because of sampling error. A table of power results using the log-normal distribution, paralleling Table 1, is available in the Supplemental Material.

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 size, 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 would lower the estimated power.

When power is large in conventional analysis, power declines with heterogeneity. This decline is greater for the log-normal distribution than what is found for the normal distribution of effect sizes. The good news is that finding negative effects, i.e., effects in the direction opposite to the average effect size, are very rare, pretty much paralleling that found in conventional analyses. Note here, unlike with the normal distribution of random effects, these are Type I errors in that the only true effects are positive.

Although it is relatively simple to determine power allowing for non-normal distributions of effect sizes, it would appear to be a much more difficult problem to allow for non-normal effect sizes in the computation of confidence intervals for an effect size

as well as in the estimation of such heterogeneity in random effects meta-analyses. We encourage work on both of these problems.

## The Origins of Effect Size Heterogeneity

Consistent with the considerable evidence for the heterogeneity of effect sizes, there are good reasons why such heterogeneity is to be expected. First and most obviously, study characteristics affect the magnitude of effects that are likely to be found even in a well-defined and limited research domain. Many factors vary between studies, such as sample characteristics, particular measurement techniques, investigators, locations, and historical events; all of these may be moderators of the effect studied.

Researchers certainly anticipate and know about some potential moderators of effect size. Thus, they may hypothesize that for their specific manipulation or their specific sample, the effect size might be larger or smaller than that generally reported in the literature. In fact, researchers may be primarily interested in estimating the effect size at particular levels of some moderator rather than in figuring out the average or typical effect size. Certainly, such known moderators are in part responsible for the level of heterogeneity typically found in meta-analyses. However, we have also seen that in more controlled replication efforts, where many of these moderators are presumably controlled, heterogeneity of some magnitude generally persists.

Accordingly, we believe that even when we exhaust all the factors that we can think of that may moderate the effect, there likely remains additional purely random heterogeneity. Consider an analogy with random variance associated with participants in how they respond to some treatment. Participants 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 are unlikely ever to explain completely. 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.

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 always treated as random, so too should studies be considered as random

## 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 2 and 3 lead us to take issue with this conventional wisdom.

We present formulas in the online Supplemental Material for the estimation of how many studies and how many participants per study are needed to minimize the

standard error of the effect, and so the width of the confidence interval, given available resources. What is likely to result is several studies with fewer participants than what we normally might think would be enough.

To illustrate, in one example discussed in the Supplemental Material, we start with a single study with a modest number of participants. Anticipating an effect size, $\delta$, 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. 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 $\delta$ is 0.4 and $\sigma_\delta$ 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 is only .625, making the power of finding all seven tests significant only .037. To test the null hypothesis that $\delta$ 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 $k = 7$, $n_P = 154$, and $\sigma_\delta = 0.20$, the power of a one-sample $t$-test of mean $d$ is .926. Note that given $\sigma_\delta = 0.20$, the standard error of average $d$ of 7 studies 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. One could determine a failsafe heteogeneity value. That is, we could compute how large heterogenity would have to be turn the significant pooled effect into a value that is no longer significant. Assuming a positive average effect size *d* with a standard error of $s_d$, we compute the square root of $\sqrt{k\left\{\frac{d^2}{Z_{1-\alpha/2}{}^2} - s_d{}^2\right\}}$ to obtain the failsafe heterogeneity value. For instance, if there are 4 studies with a *d* of .3, an alpha of .05 making $Z_{\alpha/2}$ equal to 1.96, and $s_d$ of .08, then the failsafe heterogeneity value would be 0.26. Thus, if the heterogeneity is less than or equal 0.26, the *d* would still be significant. Note that this procedure can be used even if the *d* is from a single study.

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 1000 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**

Recently concerns have 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 $\delta$, 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 $\sigma_\delta$ 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 will 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.

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 $\delta$ is the same as effect size 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.

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.[3] 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 study variance of $\sigma_\delta = 0.15$, 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 what we study. However, in power analyses, computing confidence intervals, and the planning of research, the field

acts 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.

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.

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. 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. 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., & 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.

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.

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.

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*(4), 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.

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.

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. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, in press.

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

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.

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*, in press.

Table 1

Power for a positive "+" and a negative effect "–" given an effect size ($\delta$), study variation ($\sigma_\delta$), total sample size ($N$) and an alpha of .05

| | | $\sigma_\delta$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0. 000 | | 0.050 | | 0.125 | | 0.200 | |
| $\delta$ | $N$ | + | – | + | – | + | – | + | – |
| 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

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 (*d*), and Level of Heterogeneity ($\sigma_\delta$)

` $\sigma_\delta$

| | | 0.000 | | 0.050 | | 0.125 | | 0.200 | |
|---|---|---|---|---|---|---|---|---|---|
| *d* | *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 3

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 (*d*), and Level of Heterogeneity ($\sigma_\delta$)

`

| | | $\sigma_\delta$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.000 | | 0.050 | | 0.125 | | 0.200 | |
| *d* | *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 |

## Footnotes

[1]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.

[2]Hedges and Olkin (p. 86, 1985) give the standard deviation for *d* as $\sqrt{\frac{4}{n} + \frac{\delta^2}{2n}}$, where $\delta$ 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 is the formula that we use throughout this paper, but in cases where $\delta$ is unknown, as it is here, or varies, we drop the second term.

[3]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).