You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic

Blakeley B. McShane and Ulf Böckenholt

Perspectives on Psychological Science 2014 9: 612
DOI: 10.1177/1745691614548513

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(6) 612–625

DOI: 10.1177/1745691614548513
pps.sagepub.com
SAGE

You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic

Blakeley B. McShane and Ulf Böckenholt
Marketing Department, Kellogg School of Management, Northwestern University

Abstract
Statistical power depends on the size of the effect of interest. However, effect sizes are rarely fixed in psychological research: Study design choices, such as the operationalization of the dependent variable or the treatment manipulation, the social context, the subject pool, or the time of day, typically cause systematic variation in the effect size. Ignoring this between-study variation, as standard power formulae do, results in assessments of power that are too optimistic. Consequently, when researchers attempting replication set sample sizes using these formulae, their studies will be underpowered and will thus fail at a greater than expected rate. We illustrate this with both hypothetical examples and data on several well-studied phenomena in psychology. We provide formulae that account for between-study variation and suggest that researchers set sample sizes with respect to our generally more conservative formulae. Our formulae generalize to settings in which there are multiple effects of interest. We also introduce an easy-to-use website that implements our approach to setting sample sizes. Finally, we conclude with recommendations for quantifying between-study variation.

Keywords
power, sample size, between-study variation, heterogeneity, effect size, statistical significance

The validity of research in the biomedical and social sciences is under intense scrutiny at present. A particular area of focus is on a widespread failure to replicate prior findings that some have labeled the replicability crisis (Brodeur, Le, Sangnier, & Zylberberg, 2012; Francis, 2013; Ioannidis, 2005; Yong, 2012). This problem has gained increasing recognition in psychology (Fanelli, 2009; Nosek & Lakens, 2014; Pashler & Wagenmakers, 2012), and, indeed, several prominent findings (Bargh, Chen, & Burrows, 1996; Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trotschel, 2001; Bem, 2011) have notoriously failed to replicate.

As a consequence of this crisis, interest in how to plan and conduct replications has increased (Asendorpf et al., 2013; Brandt et al., 2014; Open Science Collaboration, 2012; Klein et al., 2014; Pashler & Wagenmakers, 2012). Considerable attention has been devoted to factors that can cause effect sizes to vary across studies. For example, it has been shown that so-called questionable research practices can have a drastic impact on reported p values and thus effect sizes (Simmons, Nelson, & Simonsohn, 2011). Although this observation is important, it is clear that questionable research practices are not the only factors contributing to between-study variation in effect sizes. Another critical source of between-study variation is what can be broadly termed *method factors*, that is, anything pertaining to the implementation of a study that is not directly related to the theory under study (e.g., seemingly major factors, such as the operationalization of the dependent variable or the treatment manipulation, but also seemingly minor factors, such as the social context, the subject pool, or the time of day; for a comprehensive list, see the Replicability and Meta-Analytic Suitability Inventory of Brown et al., 2014, this issue, who use the term *sampling decisions* for what we term *method factors*.). The between-study variation in effect sizes resulting from such method factors can have dramatic and difficult to foresee effects on the outcome of a study and thus should be explicitly considered in planning a

Corresponding Author:
Blakeley B. McShane, Marketing Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Rd., Evanston, IL 60208-2001
E-mail: b-mcshane@kellogg.northwestern.edu

replication study so as to mitigate the likelihood of replication failure.

The reason between-study variation in effect sizes (also known as effect size heterogeneity or more simply as heterogeneity) complicates matters for researchers attempting replication is that it affects how likely one is to obtain a statistically significant estimate when the effect exists (i.e., statistical power). Best practices dictate setting sample sizes to achieve some prespecified level of power (typically 80%; Cohen, 1992). However, standard power formulae do not account for effect size heterogeneity (i.e., they assume it is zero) thereby resulting in assessments of power that are too optimistic—particularly when effect sizes are small to moderate. This unwarranted optimism causes sample sizes to be set too low and, thus, replication attempts to fail at a greater than expected rate.

We believe that ignoring effect size heterogeneity may add substantially to current difficulties in replicating psychological research. We illustrate why by first showing that psychological research often involves considerable heterogeneity and then showing the consequences of this heterogeneity for study planning. In particular, we demonstrate that data on close replications from the "Many Labs" Replication Project (Klein et al., 2014) and more general replications of the choice overload effect (i.e., that an increase in the number of options from which to choose can lead to adverse consequences such as a decrease in the likelihood of making a choice or the satisfaction with a choice; Iyengar & Lepper, 2000) exhibit substantial between-study variation.[1] We then show that this between-study variation means that sample sizes for future replications need to be set considerably higher than indicated by standard formulae to achieve adequate statistical power; in some cases, the impact of heterogeneity is so large that even sample sizes in the thousands do not provide sufficient power. We also demonstrate that the consequences of between-study variation are particularly significant for small to moderate effect sizes and when there are multiple effects of interest in a given study. To aid researchers in sample size planning, we provide power formulae that account for heterogeneity, and we suggest that sample sizes be set to achieve adequate power with respect to our generally more conservative formulae; these formulae are implemented on an easy-to-use website that we have created to facilitate the immediate assessment of the impact of effect size heterogeneity on replicability (see the Discussion section for details). Finally, we provide specific recommendations for quantifying between-study variation. These recommendations highlight the need to extend empirical findings to include information about the differing levels of heterogeneity that are observable across domains.

Why Heterogeneity Matters for Power Analyses

In a recent article, Cumming (2014) discussed the "dance of the confidence intervals," that is, how the point estimates and 95% confidence intervals from a set of replication studies tend to "bounce around":

> [When studies] all estimate the same population mean, μ . . . the bouncing around . . . should match what we expect simply because of sampling variability. If there is notably more variability than this, we can say the set of studies is heterogeneous, and there may be one or more moderating variables that affect the effect size [μ]. (p. 22)

Effect size heterogeneity—extra variability or bounce in the dance of the confidence intervals, to use the language of Cumming (2014)—has long been regarded as important for more general (i.e., systematic or conceptual) replications in psychological research. For instance, a meta-analysis of 17 general replications of the choice overload effect (see Appendix A for data) yields $I^2 = 78\%$ (i.e., more than three quarters of the variability in these 17 studies is due to heterogeneity—a large amount). Though substantial heterogeneity is unsurprising in the context of more general replications, there is mounting evidence of heterogeneity even under conditions that are nearly ideal for replication. For example, consider the Many Labs project that provides 16 estimates of 13 classic and contemporary effects in psychology from 36 independent samples totaling 6,344 subjects. Despite the fact that each of the 36 labs involved in the Many Labs project used identical materials and that these materials were administered through a web browser to minimize lab-specific effects, random effects meta-analyses conducted by the Many Labs authors yield nonzero estimates of heterogeneity for all 14 of the effects they found to be non-null (they studied 16 effects in total, but 2 were found to be null). Further, the average I^2 across these 14 studies was 40%: Lab-specific method factors account for nearly half of the total variability of the studies on average (see Table 3 of Klein et al., 2014).

Given these results, it is clear that substantial heterogeneity can occur even under conditions that are nearly ideal for replication and without questionable research practices: In the Many Labs studies, it was caused exclusively by as yet unidentified (and potentially unidentifiable) method factors specific to each of the 36 labs participating in the project. Consequently, it seems reasonable to conclude that some degree of effect size heterogeneity is likely to be present in much psychological research.

Effect size heterogeneity is caused by moderating variables (i.e., what we term method factors). When these

moderating variables can be identified (e.g., large effect for male subjects and small effect for female subjects), heterogeneity can be explained and controlled for (e.g., by controlling for sex in the study design and analysis). However, moderators are often hard to identify—particularly when a research area is new or when a set of studies consists of close replications (e.g., the Many Labs studies). We therefore suggest that researchers explicitly account for heterogeneity in study planning—in particular in setting sample sizes to achieve adequate statistical power—rather than assuming, as is typical, that heterogeneity is zero.

Statistical power is the probability of rejecting the null hypothesis when it is false. More formally, in most settings in psychology, statistical power is the probability that, if the true effect size is μ and $\mu \neq 0$, then a planned study will reject the sharp point null hypothesis H_0: $\mu = 0$ at size α (typically $\alpha = .05$). When the true effect size μ is on a standardized scale, such as the Cohen's *d* scale (which we assume throughout but relax in Appendix B), power is a function of μ, α, and the sample size. Consequently, the sample size can be set to achieve a desired level of power given μ and α, and best practices dictate planning studies that are adequately powered (typically at 80%; Cohen, 1992).

When each potential study has the same effect size μ, as would be the case for exact replications, there is no effect size heterogeneity. However, because replications in psychology are never exact (Brandt et al., 2014; Rosenthal, 1991; Tsang & Kwan, 1999), as evidenced by the choice overload and Many Labs results presented earlier, there is no single μ. Instead, each potential study has its own effect size μ_i that differs from the overall average effect size μ. Heterogeneity, denoted τ^2, quantifies the variance of the μ_i around μ, and it expresses the inherent variability in effect sizes that is observed when not all method factors are known and controlled for.

When effect size heterogeneity is present (i.e., when $\tau^2 > 0$), the variability of the sampling distribution of a replication study's effect size estimate (and consequently of any associated test statistic) will be greater than that assumed by standard null hypothesis significance tests. In the language of Cumming (2014), there will be extra bounce in the dance of the confidence intervals, but the usual amount of bounce will be assumed in the significance test. This assumption is in expectation generally quite optimistic, and thus power is overstated for a given sample size thereby leading sample sizes to be set too low and studies to be underpowered.

We derive new power formulae that account for this scenario. Whereas in standard power formulae power is calculated as a function of μ, α, and the sample size (assuming τ^2 is zero), in our formulae, it is calculated as a function of μ, τ^2, α, and the sample size (see Appendix B for details; our formulae nest the standard ones in that they reduce to them when τ^2 is set to zero). We can then set the sample size to achieve a desired level of power given μ, τ^2, and α using these new formulae.

In standard power formulae, the effect size μ is taken as a known input (α and the sample size are set by the researcher). In our formulae, the effect size heterogeneity τ^2 is also analogously taken as a known input. When prior studies in a research domain have been conducted, reasonable values for μ and τ^2 can be obtained by conducting a random effects meta-analysis of them (Cooper, Hedges, & Valentine, 2009; Cumming, 2014; Hunter & Schmidt, 2000); when they have not been conducted, we suggest conducting a sensitivity analysis across a range of reasonable values of μ and τ^2. In the sequel, we assume μ and τ^2 are given but return to their specification and estimation in the Discussion section.

Analyses of Single Effects

In this section, we consider the power of the single effect of interest obtained from a two-condition study. Using a hypothetical example, we examine how both power and sample size requirements vary as a function of μ and τ^2. We then present data from the Many Labs replications as well as from studies of the choice overload effect and calculate the sample size required for a future study in these domains to achieve adequate power.

Hypothetical example

Consider a simple two-condition, between-subjects experiment with equal sample size *n* in each condition in which the standardized difference between the means of the observations in each condition is given by μ (i.e., the effect size μ is on the Cohen's *d* scale). Cohen (1992) defined small, medium, and large effect sizes in psychology as $\mu = 0.2$, 0.5, and 0.8, respectively, and we consider each of these in turn. Standard power formulae (Faul, Erdfelder, Lang, & Buchner, 2007) require sample sizes of 310, 51, and 21 subjects per condition for these respective effect sizes to achieve 80% power for a one-sided test with $\alpha = .05$ (we use one-sided tests because replication requires matching the direction of the effect obtained in prior studies).

Now, suppose there is independent condition-specific heterogeneity τ^2. Pigott (2012) provided guidance on the typical degree of between-study variation, or heterogeneity, in psychology by relating it to the degree of within-study variation (i.e., sampling variation). In particular, she has defined a small amount of heterogeneity to be equal to one third the within-study variation, a medium amount of heterogeneity to be equal to the within-study variation, and a large amount of heterogeneity to be

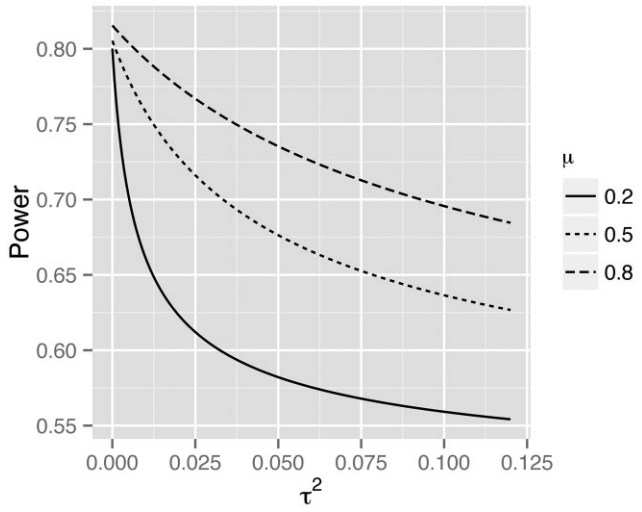

Fig. 1. Power at standard sample size. 80% power is achieved at the standard sample size when heterogeneity τ^2 is zero, but power diminishes as it increases.

equal to three times the within-study variation. In our setting, her framework equates the condition-specific within-study variation to the standard error of the sample mean; because the effect size μ is on the standardized Cohen's *d* scale, we can without loss of generality assume that the observation-level variance is one, and thus the standard error of the sample mean is one over the sample size. As an example, when the sample size is $n = 25$ subjects per condition, a small amount of heterogeneity would be $\tau^2 = 1/3 \times 1/25 = \sim.01$, a moderate amount of heterogeneity would be $\tau^2 = 1/25 = .04$, and a large amount of heterogeneity would be $\tau^2 = 3 \times 1/25 = .12$. Consequently, we examine how power and the sample size requisite for adequate power vary as heterogeneity τ^2 ranges from 0 (i.e., no heterogeneity) to .12 (i.e., large heterogeneity).

In Figure 1, we present power as a function of the effect size μ and condition-specific heterogeneity τ^2 when the sample size is set to value indicated by standard power formulae (i.e., 310, 51, or 21 subjects per condition for small, medium, and large effect sizes respectively). As can be seen, when heterogeneity is zero, 80% power is achieved, as would be expected on the basis of standard power formulae. However, power decreases as heterogeneity increases, and this is most pronounced for when the effect size is small. Indeed, the impact of heterogeneity on power depends strongly on the effect size. When the effect size is large, even relatively large amounts of heterogeneity have only a modest impact on power: Power remains at about 75% when heterogeneity is moderate (i.e., $\tau^2 = .04$) and drops to only just below 70% when heterogeneity is large (i.e., $\tau^2 = .12$). On the other hand, when the effect size is small, even small amounts of heterogeneity cause power to drop dramatically: Power is only about 65% when $\tau^2 = .01$. When the effect size is small, and when heterogeneity is moderate (large), power is only about 60% (55%), making the likelihood of replication little better than a coin toss.

In Figure 2, we present the sample size per condition required to achieve 80% power as a function of the effect size μ and heterogeneity τ^2. As can be seen, when heterogeneity is zero, the requisite sample size matches that calculated by standard formulae and indicated by the dashed horizontal lines. However, as heterogeneity increases, the requisite sample size increases rather dramatically, and it can be many multiples of that suggested by standard formulae; this increase is particularly prominent when the effect size is small and moderate.

Many Labs and choice overload data

To present the effect of heterogeneity on power and the sample size requisite for adequate power in the context of actual psychological research, we use experimental data from both close and general replications. In particular, we examine how heterogeneity affects power and requisite sample sizes for 36 studies of the 16 effects examined by the Many Labs authors (close replications) and for 17 studies of the choice overload effect (general replications). All studies were unmoderated (i.e., they were two-condition, single-effect studies).

Effect size and heterogeneity estimates for the difference in the means of the two conditions based on random effects meta-analyses of each effect appear in Table 1; all estimates are presented on the standardized Cohen's *d* scale to facilitate comparison. Using these estimates, we can calculate power and the sample size requisite for adequate power for future replication studies of these effects. These calculations can provide guidance for, in the case of the 16 effects studies by the Many Labs authors, future replication studies in different social contexts and with new subject pools in which their materials are used (i.e., close replications) and, in the case of the choice overload effect, future replication studies in which different operationalizations of the dependent variable and treatment manipulation (i.e., general replications) are used.

The results in Table 1 are divided into three sections: null effects (currency priming and flag priming), normal-sized effects (imagined contact through low vs. high category scales), and very large effects (anchoring and the allowed/forbidden effect). Heterogeneity is most relevant for normal-sized effects, though we discuss each of the three in turn. The Many Labs authors found that two effects (i.e., currency priming and flag priming) did not replicate. For these two effects (and only these two effects), heterogeneity was estimated at zero (it was very small but nonzero for gains vs. loss framing). Consequently, n_0, the sample size per condition requisite for adequate

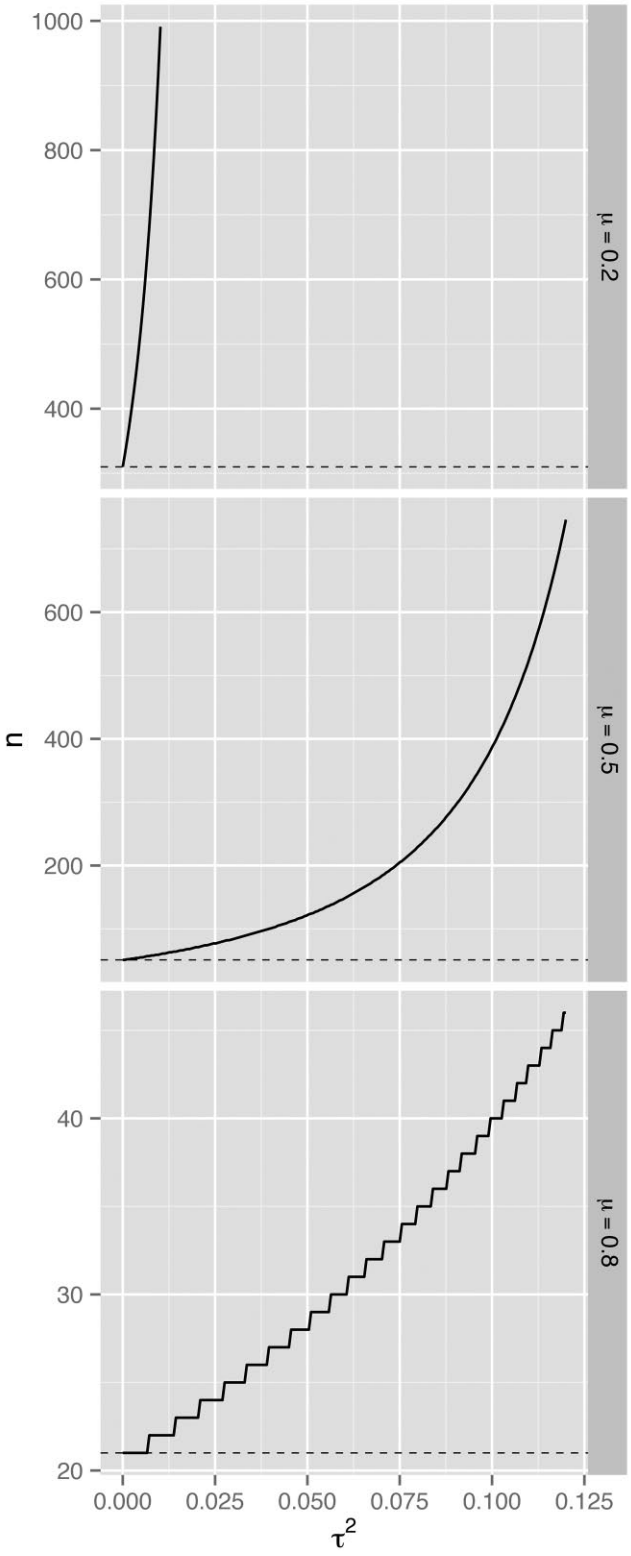

Fig. 2. Sample size per condition requisite for 80% power. The dashed line indicates the sample size indicated by standard formulae. 80% power is achieved at this sample size when heterogeneity τ^2 is zero, but a considerably larger sample size is required as heterogeneity increases. The *y*-axis is capped at 1,000, and, in some cases, 80% cannot be achieved even with infinite sample sizes.

power computed with standard formulae, is equal to n_τ, the sample size computed per condition requisite for adequate power computed with our formulae, and it provides adequate power; that said, n_0 is unrealistically large because the effects are miniscule.

For the five very large effects (i.e., the four anchoring effects plus the allowed/forbidden effect), the effect sizes are so large that adequate power is obtained with very small sample sizes. Consequently, heterogeneity does not have much of an impact (i.e., n_0 and n_τ are similar or identical).

Turning to the studies with effect sizes more typical in psychology (i.e., imagined contact through low vs. high category scales), the effect of heterogeneity can be quite dramatic. For example, standard sample size formulae suggest that a sample size of $n_0 = 885$ subjects per condition (i.e., 1,770 in total) is adequate to obtain 80% power for the imagined contact effect; however, because of heterogeneity, power is only 66.7% with 885 subjects per condition, and our formulae require a sample size of n_τ = 2,434 subjects per condition (i.e., 4,868 in total) for adequate power. The results are even more striking for the choice overload effect in which heterogeneity is much larger (i.e., because the studies are general rather than close replications). Standard sample size formulae suggest that a sample size of $n_0 = 71$ subjects per condition is adequate to obtain 80% power. However, because of heterogeneity, power is only 64.3% with 71 subjects per condition, and our formulae require a sample size of n_τ = 389 subjects per condition for adequate power.

As can be seen in Table 1, the impact of heterogeneity is relatively modest when either the effect size is large or heterogeneity is small (noting, as per Figures 1–2, that even small heterogeneity τ^2 can have a large impact if the effect size μ is also small). What can be done when effect size heterogeneity is not so modest (e.g., imagined contact, quote attribution, choice overload)? In the case of choice overload, the 17 studies were general replications. Therefore, one could search for study-level design factors that moderated the effects observed across 17 studies; such moderators might explain some of the variation in the effects thereby reducing τ^2. However, in the case of the imagined contact and quote attribution effects, the 36 studies were close replications with identical materials. Therefore, one would have to search for moderators pertaining to the social context or subject pool; this is more difficult (and possibly intractable), and realistically one might simply need larger sample sizes to replicate these effects.

Analyses of Multiple Effects

Up to the present, the discussion has centered on a single effect of interest. In practice, however, researchers are

Table 1. Effect Size (μ), Heterogeneity (τ), Standard Sample Size per Condition Requisite for 80% Power (n_0), Our Sample Size per Condition Requisite for 80% Power (n_τ), and Power at the Standard Sample Size Requisite for 80% Power (Power) for 17 Effects

Effect	μ	τ	n_0	n_τ	Power (%)
Currency priming	−0.02	.00	40,377	40,377	80.0
Flag priming	0.02	.00	36,924	36,924	80.0
Imagined contact	0.12	.08	885	2,434	66.7
Sunk costs	0.29	.05	145	155	78.1
Quote attribution	0.31	.16	131	239	69.6
Norm of reciprocity	0.36	.09	95	108	76.4
Choice overload	0.42	.35	71	389	64.3
Gender differences in implicit math attitudes	0.57	.11	39	42	78.0
Retrospective gambler fallacy	0.61	.09	34	36	78.8
Gains versus loss framing	0.66	.00	30	30	81.0
Correlation between implicit and explicit math attitudes	0.82	.10	20	20	80.5
Low versus high category scales	0.88	.16	17	19	78.0
Anchoring—Babies born	1.21	.15	10	10	81.5
Allowed/forbidden	1.93	.50	5	5	81.3
Anchoring—Mount Everest	2.00	.36	5	5	85.4
Anchoring—Chicago	2.41	.69	4	4	83.3
Anchoring—Distance to New York City	2.53	.30	3	4	79.2

Note: Effect sizes are presented on the standardized Cohen's *d* scale, and τ denotes heterogeneity of a mean difference on the standardized Cohen's *d* scale reported as a standard deviation (rather than as a variance τ^2). Effect size and heterogeneity estimates are based on random effects meta-analyses conducted by the Many Labs authors for 15 of the 17 effects. We conducted our own meta-analysis of the choice overload data presented in Appendix A and the correlation between implicit and explicit math attitudes data after converting the raw correlations to the Cohen's *d* scale. The effect of heterogeneity on the sample size requisite for 80% power and on power at the standard sample size is most notable for effect sizes typical in psychology (i.e., imagined contact through low vs. high category scales).

often interested in multiple effects (e.g., a simple effect and an interaction effect in a 2 × 2 study). Standard power formulae tend not to deal with this situation—whether there is heterogeneity or not—and researchers often simply determine the sample size on the basis of the smallest of the multiple effects. This approach can be optimistic because multiple effects are typically correlated with one another; when they are negatively correlated, as frequently occurs in practice (e.g., a simple effect and an interaction effect in a 2 × 2 study), the sample size required to achieve a given level of power for the multiple effects jointly can be dramatically higher than that required for the smallest of the effects. In this section, we extend the analysis in the prior section to show the impact of heterogeneity on the sample size required for adequate power when a researcher is interested in multiple effects.

Consider a 2 × 2 between-subjects experiment with equal sample size *n* in each condition in which the simple effect of Experimental Factor A is 0.5, the simple effect of Experimental Factor B is 0.8, the interaction effect is 0.8, and the variance of the observations in each condition is assumed, without loss of generality, to be 1 so that the simple effects are on the standardized Cohen's *d* scale (i.e., it can be assumed without loss of generality that the observations in each condition have condition-specific mean *ab* = 0.0, *Ab* = 0.5, *aB* = 0.8, and *AB* = 2.1, and variance = 1).[2] Further, suppose there is independent condition-specific heterogeneity with variance τ^2.

Imagine there are two researchers with different theories about the effects under study. The first researcher develops a hypothesis about both simple effects, whereas the second researcher develops a hypothesis about the simple effect of Experimental Factor A and the interaction; both researchers wish to achieve 80% power jointly for one-sided tests of both effects of interest with α = .05. Note that both of the researchers are interested in two effects that are the same size (i.e., one effect of size 0.5 and another of size 0.8).

The sample size required so that the first researcher achieves 80% power for both effects of interest jointly is shown for various values of τ^2 by the solid curve in the left panel of Figure 3. The dashed and dotted curves show the sample size required to achieve 80% power for each of the simple effects separately; consequently, these are identical to corresponding curves in Figure 2. For relatively low values of τ^2, the sample size required to achieve 80% power for both effects jointly is negligibly larger than that required for the smaller of the two simple effects. However, as τ^2 gets larger, more subjects are required beyond those required for the smaller simple effect.

The first researcher's sample size calculations (see the left panel of Figure 3) present an optimistic portrait of the

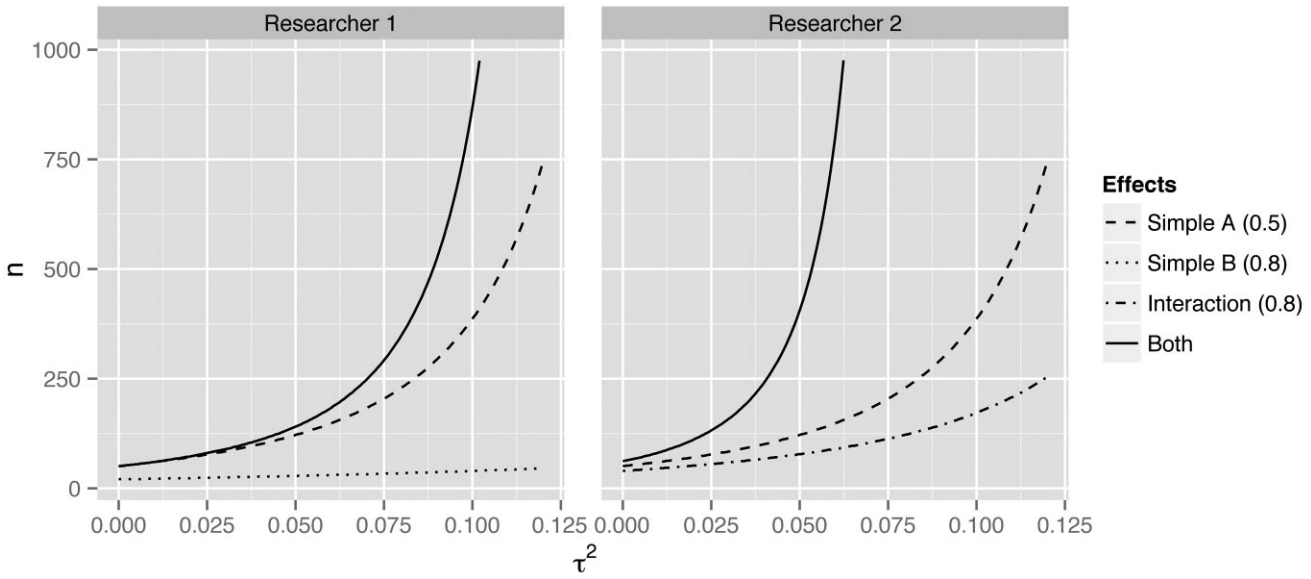

Fig. 3. Sample size per condition requisite for 80% power. The solid curve indicates the sample size requisite for the two effects of interest jointly, whereas the dashed curve indicates the sample size requisite for the simple effect of Experimental Factor A alone, the dotted curve indicates the sample size requisite for the simple effect of Experimental Factor B alone, and the dot-dashed curve indicates the sample size requisite for the interaction alone. The *y*-axis is capped at 1,000, and, in some cases, 80% cannot be achieved even with infinite sample sizes. The sample size requisite for the two simple effects jointly is not much greater than that requisite for the smaller of the two simple effects when heterogeneity is small to moderate, but it is when heterogeneity is large. However, the sample size requisite for the simple effect of Experimental Factor A and the interaction effect jointly is much greater than that requisite for either effect separately regardless of the degree of heterogeneity τ^2.

sample size required to achieve adequate power for multiple effects: The sample size required is only modestly larger than that required for the smaller of the two effects (except when heterogeneity is large). Unfortunately, this is optimistic because, in 2 × 2 experiments, the estimates of the two simple effects are positively correlated. If instead a researcher were interested in, say, a single simple effect as well as the interaction effect (as the second researcher is), the divergence between the sample size required to achieve adequate power for both effects is substantially larger than that required for the smaller of the two. This occurs for two reasons. First, in a 2 × 2 experiment, whereas the estimates of each simple effect are positively correlated with one another, they are each negatively correlated with the estimate of the interaction effect. Second, interactions are in general estimated with greater error.

The sample size required so that the second researcher achieves 80% power for both the simple effect of size 0.5 and the interaction effect of size 0.8 jointly is shown for various values of τ^2 by the solid curve in the right panel of Figure 3. The dashed and dot-dashed curves show the sample size required to achieve 80% power for each effect separately. As can be seen, even for small values of τ^2, there is a considerable divergence among the solid curve and the dashed and dot-dashed curves, and this divergence only increases with τ^2. Much larger sample sizes are required to achieve 80% power for both effects jointly as compared with each separately. In sum, there is a substantial difference in the sample sizes required by the two researchers (and depicted in the two panels of Figure 3), even though they are both interested in two effects jointly and even though the two effect sizes of interest are identical.

Discussion

In this discussion, we briefly recapitulate our findings and recommendations. We then provide an in-depth discussion of how to quantify heterogeneity to facilitate the implementation of our sample size formulae in practice. Next, we discuss how the Many Labs approach could serve as a model for improving the replicability of research findings. Finally, in our conclusion, we outline a proposal for advancing the standards of what constitutes a successful replication.

Recapitulation and recommendations

In this article, we have argued for and presented evidence that between-study variation in excess of sampling variation is present in many psychological settings: "no replications in psychology can be absolutely 'direct' or 'exact,'" (Brandt et al., 2014, p. 218) or, more poetically,

"you cannot step into the same river twice for other waters are continually flowing in" (Heraclitus quoted in Plato's *Cratylus*, Section 402a). We have shown that between-study variation causes power to be lower than expected on the basis of standard formulae and that this is particularly pronounced for small to moderate effect sizes or when there is interest in multiple simultaneous effects. Consequently, if researchers are setting their sample sizes on the basis of standard formulae, it is unsurprising that they are finding replication difficult. These difficulties are only exacerbated in an environment in which researchers may be applying the statistical significance filter (Gelman & Weakliem, 2009) and engaging in questionable research practices (Fanelli, 2010; John, Loewenstein, & Prelec, 2012; Simmons et al., 2011; Simonsohn, Nelson, & Simmons, 2013), both of which tend to upwardly bias effect size estimates.

We have proposed a remedy for this situation, namely, setting sample sizes on the basis of our power formulae that explicitly account for between-study variation.[3] Like standard formulae, our approach requires an estimate of the population average effect size μ. It also requires an estimate of heterogeneity τ^2. An estimate of the former may be obtained from prior research, as is currently done by researchers before using standard formulae. This strategy is more difficult to use for the latter because it is uncommon for researchers to report heterogeneity in prior research except in meta-analytic studies; consequently, we discuss it in greater depth later.

As mentioned throughout the article, we have created an easy-to-use website that implements our formulae for a wide variety of cases most common in psychological research so that researchers may immediately begin accounting for heterogeneity in sample size calculations. The website is available at http://spark.rstudio.com/blakemcshane/hetsampsize/, and it contains a tutorial that explains how to reproduce the results contained in this article. In particular, the tutorial demonstrates how to reproduce the calculations of n_τ presented in Table 1 as well as the plots of Figures 2–3. It also demonstrates how to extend them for different values of μ and τ^2. By following the tutorial as well as the additional instructional material on the website, researchers should easily be able to account for heterogeneity in their own sample size determinations.

Quantifying heterogeneity

For our proposal to be most effective, it is important that researchers have a reliable estimate of heterogeneity. This estimate can easily be obtained when a large number of prior studies are available via a random effects meta-analysis. However, this is the situation in which researchers may be least interested in conducting a single replication study; instead, they are typically more interested in evaluating the evidence in a more cumulative fashion.

Consequently, an important consideration is how to estimate heterogeneity when there are either no or few prior studies available, for example, in new research areas. Cumming (2014) has noted that meta-analysis is possible even with as few as two prior studies, and therefore meta-analysis can be used to estimate heterogeneity. However, with few prior studies, a meta-analysis is less reliable—particularly for estimating heterogeneity (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013). Instead, we discuss three alternative approaches and advocate for a sensitivity analysis approach.

First, a "best guess" for heterogeneity could be used, as is currently done for effect sizes when there are no or few prior studies available (i.e., researchers could do for τ^2 what they currently do for μ when there are no or few prior studies available). Although this approach is not completely satisfactory because any best guess is unlikely to be precisely correct, it is an improvement on the current approach, which amounts to assuming a best guess of zero and thus ignoring heterogeneity.

Second, an improvement on the first approach involves examining how power or the sample size requisite for adequate power varies when heterogeneity is below or above the best guess (i.e., conducting a sensitivity analysis). Sensitivity analyses to determine how power or the sample size requisite for adequate power varies as a function of the inputs (e.g., the effect size μ) are standard practice in power and sample size analyses, and thus it is natural to also conduct a sensitivity analysis with respect to heterogeneity.

The hypothetical data and actual data results presented in this article (see Figures 2–3 and Table 1) portend the results of such a sensitivity analysis in practice. For small effect sizes, the sample size requisite for adequate power is quite sensitive to the best guess of heterogeneity, and potentially unrealistic sample sizes would be required even for low heterogeneity. For moderate effect sizes, the sample size is reasonably insensitive provided heterogeneity is low to moderate. Finally, for large effect sizes, the sample size is comparably insensitive to heterogeneity. Sensitivity would in any case be exacerbated when there is interest in multiple simultaneous effects.

An important feature of our website is that it automatically conducts this sensitivity analysis for heterogeneity. In particular, it returns a plot analogous to Figures 2–3 for each effect of interest as well as all effects simultaneously. These plots can be used to assess the implications for the sample size requisite for adequate power when heterogeneity is below or above the specified value (i.e., the best guess).

Finally, a third possibility is to use data from other more established research areas to estimate heterogeneity

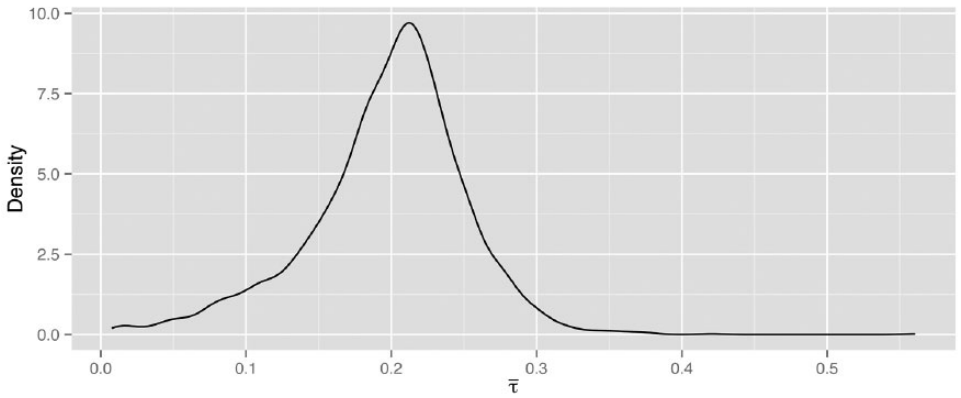

Fig. 4. Posterior distribution of average heterogeneity. The average heterogeneity $\bar{\tau}$ for a difference in means on the standardized Cohen's *d* scale reported as a standard deviation (rather than as a variance) for the Many Labs series of studies could be as high as 0.30, suggesting that heterogeneity in typical replication settings is likely to be higher than that.

and then to apply this estimate to the area that has no or few prior studies. This seems like a reasonable approach because studies in research areas with no or few prior studies are likely to be more heterogeneous than those in more established research areas (because, e.g., the operationalization of the dependent variable or the treatment manipulation are likely to be less precisely calibrated). Consequently, one could estimate heterogeneity in research areas in which there are a large number of studies and use this estimate as a lower bound on heterogeneity in research areas with no or few prior studies. One could then conduct a sensitivity analysis with respect to this lower bound.

One version of this third approach would be to use the Many Labs data to establish such a lower bound. Because most replications in psychology are not as close as those from the Many Labs series of studies, these studies seem particularly well-suited to establish a practical lower bound on heterogeneity. How can one obtain a lower bound with these data? One possibility is to simply use the median value reported across the 16 effects. This would imply heterogeneity of about $\tau = .10$ (i.e., $\tau^2 = .10^2 = .01$) for a difference in means on the standardized Cohen's *d* scale; this equates to condition-specific heterogeneity of $\tau = .10/\sqrt{2} = .07$ or $\tau^2 = .10^2/2 = .005$. Researchers attempting replication when there are no or few prior studies available could potentially use this lower bound as the estimate of heterogeneity. As illustrated previously, even this seemingly low number can have a dramatic effect on sample size calculations.

An approach that is more sophisticated than simply taking the median would be to build a Bayesian hierarchical model for heterogeneity across the full set of Many Labs studies. We implemented such a model (see Appendix C for details) and found that the posterior distribution for $\bar{\tau}$, the average heterogeneity for a difference in means on the standardized Cohen's *d* scale, favors values substantially larger than $\tau = .10$; this posterior is plotted in Figure 4, and it is consistent with heterogeneity as large as $\tau = .30$ (i.e., $\tau^2 = .09$), suggesting that heterogeneity in typical settings, in which replications are less close than the Many Labs replications, is likely to be even larger than that. Consequently, researchers may prefer to use this larger value as a more conservative estimate.

We emphasize that we do not wish to enshrine these particular heterogeneity estimates in the literature as certain other statistical rules (e.g., $p < .05$) have been. Instead, we note that because they are based on replications that are closer than those typical in psychology, they provide reasonable benchmarks and lower bounds. Of course, the effects studied by and design choices made by the Many Labs authors are not a random sample of effects and designs across psychology more broadly: Heterogeneity in a given research area or for a given design choice could thus differ nontrivially from these numbers, and this difference could be in either direction. The sensitivity analysis approach that we advocate as the best approach allows researchers to examine the potential consequences of this difference. As an additional matter, we advocate research into quantifying heterogeneity across a variety of experimental settings (one possible way to achieve this is discussed in the following subsection); such research would provide other researchers with heterogeneity estimates and lower bounds most relevant to their setting.

The many benefits of Many Labs

The recommendations presented in this article are beneficial for single-study replication. However, we also wish to emphasize a number of benefits of the Many Labs approach to science. As these two approaches seek

different aims, we view them and their benefits as complementary rather than competitive.

The core of the Many Labs approach lies in running multiple smaller studies of a given phenomenon distributed across multiple labs rather than one large study of the phenomenon in one lab. Because heterogeneity is the norm in psychological research, this approach has three direct benefits: (a) the explicit quantification of heterogeneity in a particular research area, (b) more efficient estimation of the population average effect size, and (c) better calibration of the Type I error rate. We discuss each in turn.

First, multiple studies allow researchers to directly quantify heterogeneity via meta-analysis of their own data. This is beneficial because the estimate of heterogeneity is tied specifically to the phenomenon and operationalizations under study; alternative estimates of heterogeneity typically require recourse to studies that use conceptually similar but still distinct phenomena or that use somewhat different operationalizations.

Such estimates of heterogeneity, particularly when gathered across a variety of domains, provide a number of benefits. They provide knowledge of how heterogeneity varies as a function of research domain and subdomain. They also provide direct inputs to our power and sample size formulae. Finally, they suggest areas that are likely ripe for theory enrichment (i.e., areas in which heterogeneity is large), which could come via study of, for example, new moderators.

Second, multiple studies provide a more efficient estimate of the overall population average effect size. For instance, consider a researcher interested in a two-condition, between-subjects study with $d = 0.223$ and heterogeneity $\tau^2 = .01$ (in which τ^2 denotes heterogeneity of a mean difference on the standardized Cohen's d scale and thus corresponds to condition-specific heterogeneity of $\tau^2 = .10^2/2 = .005$). Standard power calculations require 250 subjects per condition (i.e., a total of 500 subjects) for 80% power (this is of course optimistic because heterogeneity is ignored); suppose the researcher considers either running one large study with 250 subjects per condition or, as a parallel to the Many Labs series of studies, running one smaller study with 50 subjects per condition and sending the study materials to four colleagues and asking them to also run the study with 50 subjects per condition (i.e., so there are a total of 500 subjects in both scenarios). In the second scenario, not only can the researcher pool across the five smaller studies to obtain an estimate of heterogeneity (not possible in the first scenario) but the overall estimate of the effect is 44% more efficient in the second as compared with the first scenario (i.e., the variance of the estimate in the second scenario is 44% lower than in the first).

Third, multiple studies allow for better calibration of the realized Type I error under the null hypothesis that the overall population average effect size μ is zero. In the presence of heterogeneity, standard null hypothesis significance tests will reject the null more frequently than the size α of the test because they test whether the study-specific (as opposed to overall average) effect is zero. Consequently, researchers may believe they have found something that is generalizable but that, in reality, can only be attributed to heterogeneity (i.e., study-specific method factors). However, by pooling across multiple studies and accounting for heterogeneity, researchers can test whether the overall population average effect size μ is zero in a manner that preserves the stated size α of the test.

The Many Labs approach to science is of course more costly: It requires the coordination of a large number of labs across the world as well as a potentially larger number of subjects. These costs are real, but they must be assessed against the benefits listed earlier as well as costs of alternative approaches (e.g., the cost of failed single-study replications, particularly those that are not properly powered). It is our hope that improvements in technology will ease the burden of coordinating studies across multiple labs and will allow this approach to become more common in psychological research.

Conclusion

The approach outlined in this article constitutes a principled strategy for dealing with heterogeneity—a fact that has largely been ignored until the present—in the context of single-study replications. Though no panacea for all ills that ail replication, the likelihood of future studies replicating prior ones will increase when heterogeneity is explicitly accounted for in sample size determinations, thus mitigating, at least to some extent, the current difficulties in replicating psychological research.

Nonetheless, current difficulties in replicating psychological research may stem directly from the notion of replication used: that estimates of one or more effects of interest from a subsequent study match the direction of those from one or more prior studies and attain statistical significance. This definition fails to reflect many important features of such estimates (e.g., magnitude, variability), and, thus, in closing, we would like to raise the possibility of altering the standards for what constitutes a successful replication.

An ample literature has decried the null hypothesis significance testing paradigm on which the current standards for replication rely (Bakan, 1966; Cohen, 1994; Cumming, 2014; Gigerenzer, 2004; Gill, 1999; Hunter, 1997; Meehl, 1978; Rozenboom, 1960; Schmidt, 1996; Schwab, Abrahamson, Starbuck, & Fidler, 2011; Serlin & Lapsley, 1993) and has instead noted that "the primary product of a research inquiry is one or more measures of effect size, not *p*-values" (Cohen, 1990, p. 1310). This is particularly relevant as the difference between one estimate that attains statistical significance and another estimate that

fails to attain statistical significance is not in general statistically significant itself[4]—an issue that is distinct from the arbitrariness of the conventional $\alpha = .05$ threshold (Cochran, 1974; Cowles & Davis, 1982; Cramer, 1955; Fisher, 1926; Yule & Kendall, 1950) but that is often misunderstood in practice (Gelman & Stern, 2006).

Consequently, one might adopt an alternative notion of replication (Asendorpf et al., 2013; Brandt et al., 2014; Gelman, 2014) that involves comparing estimates of effect sizes and their variability from subsequent studies with those from prior studies for consistency in a more holistic sense. Relatedly, one might seek to "power" future studies so that effect size estimates from them are likely to be consistent with those from prior literature (e.g., whereas researchers are currently advised to choose sample sizes so that future studies have adequate power or probability [typically 80%] of rejecting a null hypothesis at size α [typically 5%], researchers could instead "power" studies so that they have adequate probability of lying "near" the effect size in which nearness would depend on both α and the degree of heterogeneity). It is possible that—under such alternative notions of replication that are based on "one or more measures of effect size, not *p*-values" (Cohen, 1990, p. 1310) and the consistency among them—the replicability crisis may even turn out to be no such thing at all.

Appendix A

Choice Overload Studies

Article	ID	μ	σ	n_{total}	Product category	Detail
Iyengar and Lepper (2000)	1	0.82	.13	249	Jam	Study 1
	2	0.30	.15	193	Essays	Study 2
	3	0.44	.15	193	Essays	Study 2
	4	0.82	.24	67	Chocolates	Study 3
	5	1.15	.26	67	Chocolates	Study 3
Shah and Wolford (2007)	6	0.77	.22	80	Pens	Study 1
Scheibehenne, Greifeneder, and Todd (2009)	7	−0.11	.22	80	Restaurant coupons	Study 1
	8	−0.18	.23	75	Charities	Study 2b
	9	−0.25	.16	80	Music	Study 3a
	10	−0.05	.15	87	Music	Study 3b
Sela, Berger, and Liu (2009)	11	0.38	.18	121	Ice cream	Study 1a
	12	0.45	.23	75	Food	Study 1b
	13	0.89	.28	51	Printers and MP3 players	Study 2
	14	0.35	.16	156	Printers and MP3 players	Study 3
Diehl and Poynor (2010)	15	0.32	.16	165	Camcorders	Study 2
	16	0.53	.25	65	Computer wallpaper	Study 3
Inbar, Botti, and Hanko (2011)	17	1.14	.42	27	DVDs	Study 1

Note: Choice overload studies. The effect size estimate for each study is given on the standardized Cohen's *d* scale by μ, the standard error of this estimate is given by σ, and the total sample size of the study is given by n_{total}. Iyengar and Lepper (2000), in both Studies 2 and 3, measured two dependent variables; in each case, both measurements are included and, for simplicity, are considered as independent. A positive μ is associated with a negative impact of larger assortments (i.e., the choice overload effect), whereas a negative μ is associated with a positive impact of larger assortments. For simplicity, we considered only studies with no moderators (i.e., two-condition, single-effect studies).

Appendix B

Calculations

In this appendix, we derive the sampling variance of effect size estimates in the presence of heterogeneity and compare it with the sampling variance assumed by standard null hypothesis significance tests. These derivations are general in that they account for multiple effects of interest and relatively unrestricted forms of heterogeneity. They follow directly from multilevel models, in particular the random intercepts and slopes model with no group-level predictors (Gelman & Hill, 2006).

In calculating the power for effect sizes, we assume that $\bar{\boldsymbol{y}}$, the vector of condition-specific measurements of interest (e.g., means, proportions) for each of the conditions in a given study, has overall population mean $\boldsymbol{\alpha}$. We let $\mathbf{C}$ be a matrix that yields the contrasts of interest such that $\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\alpha}$ denotes the overall population mean of the contrasts of interest. For example, in a two-condition experiment, the contrast matrix may be $\mathbf{C} = (-1\ 1)$, and,

consequently, $\boldsymbol{\mu}$ will have one element that is the difference between the measurements in the two conditions. Similarly, in a 2 × 2 experiment, the contrast matrix may be $\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & -1 \end{pmatrix}$, and, consequently, the first two elements of $\boldsymbol{\mu}$ will represent simple effects, whereas the third will represent the interaction effect.

If we believe that (a) study-level, condition-specific effect size heterogeneity is quantified by $\boldsymbol{\Sigma}_{\boldsymbol{\tau}}$, (b) the sampling variation of $\bar{\boldsymbol{y}}$ is quantified by $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}$, and (c) these sources of variance are independent of one another, then $\hat{\boldsymbol{\mu}}$, our estimate of $\boldsymbol{\mu}$, will have expectation $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{s}} = \mathbf{C}\boldsymbol{\Sigma}_{\boldsymbol{\tau}}\mathbf{C}' + \mathbf{C}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}\mathbf{C}'$. However, standard null hypothesis significance tests account for sampling variation alone, and, thus, assume that the variance is $\boldsymbol{\Sigma}_{\mathbf{0}} = \mathbf{C}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}\mathbf{C}'$. Because of the overdispersion of $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ relative to $\boldsymbol{\Sigma}_{\mathbf{0}}$, such tests will typically have lower power than expected if it is assumed, as is standard, that $\boldsymbol{\Sigma}_{\boldsymbol{s}} = \boldsymbol{\Sigma}_{\mathbf{0}}$.

Before proceeding, we note that $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}$ should be written more formally as $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}(\boldsymbol{n})$ as it is a function of $\boldsymbol{n}$, the vector of sample sizes in each condition. Consequently, $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ and $\boldsymbol{\Sigma}_{\mathbf{0}}$ should be written more formally as $\boldsymbol{\Sigma}_{\boldsymbol{s}}(\boldsymbol{n})$ and $\boldsymbol{\Sigma}_{\mathbf{0}}(\boldsymbol{n})$, respectively. This detail is suppressed but is clearly relevant for calculating the sample size requisite for adequate power.

If we assume that a normal distribution aptly models heterogeneity and sampling variation and that the variance of these distributions are known, then the sampling distribution of $\hat{\boldsymbol{\mu}}$ is $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{s}})$, whereas the null distribution under the sharp point null hypothesis of zero effect is $\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{0}})$. Power can be calculated by integrating the sampling distribution of $\hat{\boldsymbol{\mu}}$ over the critical region determined by the null distribution. Typically, we are interested in testing the null hypothesis for only one or several elements of $\boldsymbol{\mu}$, thereby simplifying the integration considerably. Given this form for power, we can easily numerically solve for $\boldsymbol{n}$ that give adequate power; typically, to simplify matters, we assume an equal sample size n per condition (and thus a total sample size of n times the number of conditions), thereby allowing for a numerical solution in one dimension.

When (a) the condition-specific heterogeneity is independent with variance τ^2 and (b) the study is between-subjects with sampling variance σ^2 and equal sample size n in each condition, $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ simplifies considerably reducing to $\tau^2\mathbf{C}\mathbf{C}' + \frac{\sigma^2}{n}\mathbf{C}\mathbf{C}' = (\tau^2 + \frac{\sigma^2}{n})\mathbf{C}\mathbf{C}'$. Similarly, $\boldsymbol{\Sigma}_{\mathbf{0}}$ reduces to $\frac{\sigma^2}{n}\mathbf{C}\mathbf{C}'$. If interest in this setting centers on the treatment effect in a two-condition experiment or either of the simple effects in a 2 × 2 experiment, the distributions listed earlier simplify to $\mathbf{N}(\mu, 2(\tau^2 + \frac{\sigma^2}{n}))$ and $\mathbf{N}(0, 2\frac{\sigma^2}{n})$, respectively, where μ is the single effect of interest. Letting $s_S = \sqrt{2(\tau^2 + \frac{\sigma^2}{n})}$ and $s_0 = \sqrt{2\frac{\sigma^2}{n}}$, the power of the one-tailed test of $H_0 : \mu = 0$ is $1 - \Phi(\frac{z_\alpha s_0 - |\mu|}{s_S})$ while the power for the two-tailed test is $1 - \Phi(\frac{z_{\alpha/2} s_0 - |\mu|}{s_S}) + \Phi(\frac{-z_{\alpha/2} s_0 - |\mu|}{s_S})$ where z_α is the $100(1-\alpha)$ percentile of the standard normal distribution and $\Phi(x)$ is the standard normal cumulative distribution function. These formulae can be easily solved numerically for the smallest n such that adequate power is achieved. Further, these formulae hold for the interaction effect in a two by two experiment replacing the twos in s_S and s_0 by fours.

In this discussion, we assume that $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}_{\boldsymbol{\tau}}$, and $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}$ are known. In practice, they are often not. If we believe that our uncertainty in $\boldsymbol{\alpha}$ can be quantified by Σ_α and that our uncertainty is independent of all other sources of variation, then we can instead set $\boldsymbol{\Sigma}_{\boldsymbol{s}} = \mathbf{C}\Sigma_{\boldsymbol{\alpha}}\mathbf{C}' + \mathbf{C}\boldsymbol{\Sigma}_{\boldsymbol{\tau}}\mathbf{C}' + \mathbf{C}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}\mathbf{C}'$ to account for this uncertainty. Alternatively, if prior study-level data are available, one can bootstrap (Efron & Tibshirani, 1994) the studies to derive sampling distributions for $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\tau}}$; furthermore, if subject-level data are available, one can bootstrap that data to obtain sampling distributions for $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}$. One can then use these values in combination with our formulae to compute an approximate sampling distribution for, for example, the sample size required for adequate power; then, a value from the upper part of that distribution can be specified as the sample size for a future replication. Another method for accounting for uncertainty in $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}_{\boldsymbol{\tau}}$, and $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}$ is to conduct a sensitivity analysis. Finally, adjustments can be made directly to the power formulae presented earlier to make them hold exactly under unknown parameters (e.g., moving from a normal distribution to a t distribution when the sampling variance $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\varepsilon}}}$ is unknown); in practice, adjustments for unknown sampling variance have a negligible impact on the sample size requisite for adequate power when uncertainty in $\boldsymbol{\alpha}$ or heterogeneity is nonzero because, in such an environment, the sample size requisite is generally sufficiently large that the normal approximation holds reasonably well.

Appendix C

Bayesian hierarchical model

Our Bayesian hierarchical model for estimating heterogeneity in the Many Labs data is

$$d_{s,l} \sim \mathrm{N}\left(\mu_s + \beta_{s,l}, \sigma_s^2 / n_{s,l}\right),$$

where $d_{s,l}$ denotes the estimated effect size for study s from lab l (so that s ranges from 1 to 16, and l ranges from 1 to 36), and $n_{s,l}$ denotes the sample size for study s from lab l.

The hierarchical model for the lab-specific deviation from the overall study mean can be written as

$$\beta_{s,l} \sim \mathrm{N}\left(0, {\tau_s}^2\right).$$

Primary interest in this case centers not on the hierarchical model for the means $\beta_{s,l}$ but rather on the hierarchical model for the variances ${\tau_s}^2$, which is log-normal

$$\log\left(\tau_s\right) \sim \mathrm{N}\left(\bar{\lambda}, \eta^2\right).$$

We are most interested in the posterior distribution of $\bar{\tau} = e^{\bar{\lambda}}$.

All that remains for us to specify the full model is to give our priors for the μ_s, the σ_s^2, $\bar{\lambda}$, and η. In all cases, we use relatively diffuse and, thus, noninformative priors:

$$\mu_s \sim \mathrm{N}\left(0, 100^2\right), \qquad {\sigma_s}^2 \sim \mathrm{IG}\left(0.001, 0.001\right)$$

$$\bar{\lambda} \sim \mathrm{U}\left(-10, 10\right), \qquad \eta \sim \mathrm{U}\left(0, 5\right).$$

All models were estimated in WinBUGS (Spiegelhalter, Thomas, & Best, 1999).

Acknowledgment

We thank Alison Ledgerwood for the helpful comments and suggestions she made throughout the review process.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. We follow Brandt et al. (2014, p. 218) and "use the term close replications because it highlights that no replications in psychology can be absolutely 'direct' or 'exact' recreations of the original study (for the basis of this claim see Rosenthal, 1991; Tsang & Kwan, 1999)."
2. Here and throughout, by the "simple effect of Experimental Factor A (B)," we mean the "the simple effect of Experimental Factor A (B) in the low condition of Factor B (A)." We use the former for simplicity.
3. We note that our approach may be somewhat optimistic in that we assume that the population average effect size μ and heterogeneity τ^2 are known—analogous to standard formulae assuming μ is known. Though both are never truly known, our approach can be generalized to accommodate uncertainty in them; because these extensions are technical in nature, we present them in Appendix B.
4. For an example, consider Studies 10 and 12 of the choice overload effect listed in Appendix A.

References

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trotschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, *81*, 1014–1027.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Brodeur, A., Le, M., Sangnier, M., & Zylberberg, Y. (2012). *Star wars: The empirics strike back*. Paris, France: Paris School of Economics.

Brown, S. D., Furrow, D., Hill, D. F., Gable, J. C., Porter, L. P., & Jacobs, W. J. (2014). A duty to describe: Better the devil you know than the devil you don't. *Perspectives on Psychological Science, 9*, 626–640.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A non-degenerate estimator for hierarchical variance parameters via penalized likelihood estimation. *Psychometrika*, *78*, 685–709.

Cochran, W. G. (1974). *Early development of techniques in comparative experimentation*. Cambridge, MA: Harvard University.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of significance. *American Psychologist*, *44*, 1276–1284.

Cramer, H. (1955). *The elements of probability theory*. New York, NY: Wiley.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Diehl, K., & Poynor, C. (2010). Great expectations?! Assortment size, expectations and satisfaction. *Journal of Marketing Research*, *47*, 312–322.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap*. Boca Raton, FL: Chapman and Hall/CRC.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4), e10068.

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.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, *33*, 503–513.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.

Gelman, A. (2014). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*. Advance online publication. doi:10.1177/0149206314525208

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*, 328–331.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power. *American Scientist*, *97*, 310–316.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*, 647–674.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3–7.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects versus random effects meta-analysis models: Implications for cumulative knowledge in psychology. *International Journal of Selection and Assessment*, *8*, 275–292.

Inbar, Y., Botti, S., & Hanko, K. (2011). Decision speed and choice regret: When haste feels like waste. *Journal of Experimental Social Psychology*, *47*, 533–540.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*, 996–1006.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532.

Klein, R. A., Ratliff, K., Nosek, B. A., Vianello, M., Pilati, R., Devos, T., . . . Kappes, H. (2014). *Investigating variation in replicability: A "many labs" replication project*. Open Science Framework. Retrieved from https://osf.io/wx7ck/

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Counseling and Clinical Psychology*, *46*, 806–834.

Nosek, B. A., & Lakens, D. (2014). Editorial: Registered reports. *Social Psychology*, *45*(3), 137–141.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

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

Rosenthal, R. (1991). Replication in behavioral sciences. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–30). Newbury Park, CA: Sage.

Rozenboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2009). What moderates the too-much-choice effect? *Psychology & Marketing*, *26*, 229–253.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115–129.

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, *22*, 1105–1120.

Sela, A., Berger, J., & Liu, W. (2009). Variety, vice, and virtue: How assortment size influences option choice. *Journal of Consumer Research*, *35*, 941–951.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal psychological research and the good enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shah, A. M., & Wolford, G. (2007). Buying behavior as a function of parametric variation of number of choices. *Psychological Science*, *18*, 369–370.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. doi:10.1037/a0033242

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). *WinBUGS Version 1.2 user manual*. Cambridge, United Kingdom: MRC Biostatistics Unit.

Tsang, E. W., & Kwan, K.-M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review*, *24*, 759–780.

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*, 298–300.

Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London, England: Griffin.