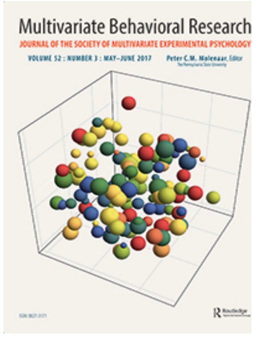

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: www.tandfonline.com/journals/hmbr20

Addressing the “Replication Crisis”: Using Original Studies to Design Replication Studies with Appropriate Statistical Power

Samantha F. Anderson & Scott E. Maxwell

To cite this article: Samantha F. Anderson & Scott E. Maxwell (2017) Addressing the “Replication Crisis”: Using Original Studies to Design Replication Studies with Appropriate Statistical Power, Multivariate Behavioral Research, 52:3, 305-324, DOI: 10.1080/00273171.2017.1289361

To link to this article: https://doi.org/10.1080/00273171.2017.1289361

Published online: 07 Mar 2017.

MULTIVARIATE BEHAVIORAL RESEARCH
2017, VOL. 52, NO. 3, 305–324
http://dx.doi.org/10.1080/00273171.2017.1289361

Routledge
Taylor & Francis Group

Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power

Samantha F. Anderson and Scott E. Maxwell

Department of Psychology, University of Notre Dame

ABSTRACT

Psychology is undergoing a replication crisis. The discussion surrounding this crisis has centered on mistrust of previous findings. Researchers planning replication studies often use the original study sample effect size as the basis for sample size planning. However, this strategy ignores uncertainty and publication bias in estimated effect sizes, resulting in overly optimistic calculations. A psychologist who *intends* to obtain power of .80 in the replication study, and performs calculations accordingly, may have an *actual* power lower than .80. We performed simulations to reveal the magnitude of the difference between actual and intended power based on common sample size planning strategies and assessed the performance of methods that aim to correct for effect size uncertainty and/or bias. Our results imply that even if original studies reflect actual phenomena and were conducted in the absence of questionable research practices, popular approaches to designing replication studies may result in a low success rate, especially if the original study is underpowered. Methods correcting for bias and/or uncertainty generally had higher actual power, but were not a panacea for an underpowered original study. Thus, it becomes imperative that 1) original studies are adequately powered and 2) replication studies are designed with methods that are more likely to yield the intended level of power.

KEYWORDS

Effect size; power; replication; sample size planning

A decade ago, scientists received a rude awakening: the bold claim that "most published research findings are false" (Ioannidis, 2005). Over the years that followed, an increasing appreciation of the benefits of replication and reproducibility has ensued. Now, the status of replication as the "gold standard" of science (Jasny, Chin, Chong, & Vignieri, 2011, p. 1225) is hard to ignore. In 2014, Simons noted, "if an effect is real and robust, any competent researcher should be able to obtain it when using the same procedures with adequate statistical power" (p. 76). Along similar lines, the Open Science Collaboration (OSC, 2015) recently wrote that findings "should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence" (p. 943). Yet, the recent data coming from replications have only reinforced the notion that published findings are not what they claim to be. For example, a recent review of 100 psychology replication attempts in the Reproducibility Project-Psychology (RPP) found that only 36% had statistically significant results (OSC, 2015). Along similar lines, the pharmaceutical company Bayer was able to replicate less than 25% of a sample of 67 cancer research studies in another large-scale replication attempt (Prinz, Schlange, & Asadullah, 2011). Many have taken these findings to mean that numerous reported psychological effects simply are not real. Although these data are clearly sobering, we argue that current methodological practices make it difficult to fairly evaluate replication studies, especially when they appear to contradict previous findings. Replication is integral to science, but providing an original study with a fair chance of replication does not depend only on the existence of the effect in question. In recent reports on replication, the statistical power (henceforth, simply power) of the *original* study has received relatively little attention. Yet, the power of the original study influences the extent to which a replication study can be adequately powered, regardless of whether formal sample size planning is performed for the replication study. The goal of this article is to explore the magnitude by which the actual proportion of studies likely to replicate differs from the power that researchers believe they have achieved. Further, we determine how this difference is affected by two factors: 1) the selected sample size planning approach for the replication study, and 2) the power of the original study, as determined by both sample size and true effect size.

CONTACT Scott E. Maxwell smaxwell@nd.edu Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556, USA.

Assumptions and definitions

First, it is important to note that throughout this article, we distinguish between what we call "intended power" and "actual power." Intended power is the idealized benchmark set by the researcher, often .80 or .90. Researchers often incorrectly assume that the power they have achieved will equal their intended power as long as they collect data from the planned number of participants. In contrast, actual power is the probability that the true effect in question will be detected, given the population effect size parameter. However, the magnitude of the true effect to be detected in replication studies is unknown and is typically estimated from the sample effect size in the original study. If there is a difference between the estimated effect size of the original study and the population effect size of the phenomenon of interest, then there will be a difference between the actual and intended power in the replication study. Neither the extent of the difference between intended and actual power in replication studies nor the effectiveness of possible strategies to reduce this gap has been systematically studied.

Second, this article defines a successful replication as one with a statistically significant result. Thus, sample size planning has the goal of determining the sample size necessary to detect a significant effect, rather than to achieve a certain degree of precision. Other definitions of replication are certainly possible (see Anderson & Maxwell, 2016; Cook, Shadish, & Wong, 2008, for criteria for comparing results across studies), and accuracy in parameter estimation (AIPE; Maxwell, Kelley, & Rausch, 2008) approaches are valuable in many situations. However, we adopt the basic replication definition for three reasons. First, the discussion surrounding replication has typically focused on this definition, and our arguments will be most transparent by remaining grounded in this definition. As "some fuzziness in determining how close ... study results should be is inevitable" (Cook et al., 2008, p. 729), even authors who discuss alternate definitions still often compare studies via statistical significance. Second, the concept of power, heavily dependent on statistical significance, figures heavily in this article. Third, directional hypotheses have an important place in psychology. Most researchers are more familiar with the Campbell perspective, which has "historically focused more on determining the direction of a causal effect" (West & Thoemmes, 2010, p. 18), as opposed to the Rubin perspective, which places greater emphasis on magnitude. Similarly, Dawes (2004) stressed the value of directional hypotheses for theory development and testing in psychology. These directional predictions may be most effectively tested with a significance testing approach (e.g. Anderson & Maxwell, 2016; Fraley & Vazire, 2014).

Finally, we limit our discussion to direct or exact replications (sometimes called close replications, given that "no replications in psychology can be absolutely 'direct' or 'exact'"; Brandt et al., 2014, p. 218). Until recently, conceptual replications were more common, because "all causal relationships are context dependent" (Shadish, Cook, & Campbell, 2002, p. 5). The emphasis was on extending prior research, generalizing to new populations or conditions, or otherwise including novel components. However, conceptual replications are most relevant when the original effect has been already corroborated by direct replications. Following the lack of successful replications in the field, there has been a call for more direct replications (Simons, 2014). In order to isolate the effects of power, we assume that the replication researcher has conducted the replication study in accordance with the procedures and measures of the original study. Although this assumption ignores the between-study heterogeneity (e.g., differences in "time, space, human populations"; Shadish et al., 2002, p. 5) that is inevitable even with direct replications, the performance of current approaches to sample size planning would be even worse with this additional wrinkle (McShane & Bockenholt, 2014).

A hypothetical example

Let us begin with a hypothetical scenario that mimics what replication researchers often face in the real world and introduces the motivation for the current study. Suppose that a researcher is interested in attempting to replicate a recent study published in a major journal. The original study reported its findings in terms of a *t* test for two independent groups and followed the minimum recommendation of Simmons, Nelson, and Simonsohn (2011) using 20 participants per group ($n = 20$).[1,2] Let us assume that the hypothetical original study reported an observed *t*-value of 2.21, which corresponds to a *p*-value of .0332, with a Cohen's *d* (henceforth, simply "*d*") of 0.70. Based on this reported *d*, the prospective replicator determines that a sample size of 44 participants per group would be necessary to have .90 power and thus designs the replication study accordingly. What is likely to be the actual power of the replication study?

[1] Throughout the article, *n* refers to the sample size per group and *N* refers to total sample size.

[2] Sample sizes can vary quite widely by research area (e.g., median $N = 175$ for regression studies, Jaccard & Wan, 1995; modal $n = 10$ for behavioral neuroscience, Talboom, West, & Bimonte-Nelson, 2015). However, Simonsohn and colleagues' recommendation was geared toward experimental studies, which have been the focus of recent high-profile replication attempts. Further, in Simonsohn's (2015) own review of *Psychological Science* articles, $n = 20$ was the median sample size reported during the years 2003–2010. Finally, the April and June 2016 issues have a median per-group sample size of 25, indicating that Simonsohn's analysis is still appropriate today.

An intuitive answer might seem to be .90. However, this assumes that the reported value of d is in fact the true population value (δ). In reality, the observed d of 0.70 will probably be different from δ due to sampling error. Even so, it might seem that .90 should be a reasonable replication study power estimate, because sampling error is often assumed to be random. However, it is quite sensible to assume that the journal in question only publishes studies with statistically significant results (studies with $p > .05$ are censored). In this case (observed d of 0.70 and $n = 20$), the most likely δ is 0.19, not 0.70. We will discuss the theory and calculations behind this corrected estimate in the "Proposed Solutions in the Literature" section. If the actual δ is in fact 0.19, the resultant actual power of the replication study will be only .14, much smaller than the intended power of .90. We will see that, especially for smaller original study sample sizes, actual power values are much lower than the power intended by the researcher, unless the true effect size is quite large.

Why is the most likely power of the replication study so much less than the intended power? One potential explanation is that the effect size reported in the original study is the result of p-hacking, the "garden of forking paths" (Gelman & Loken, 2014), or hypothesizing after results are known (HARKing; Kerr, 1998). If the significant result reported in the original study is an artifact of these post hoc approaches, it naturally follows that the power to detect the effect in a replication study is likely to be even less than it would otherwise appear to be. In fact, in cases of p-hacking, the Type I error rate increases, as it is no longer a managed component of the significance testing process. We agree that this explanation is plausible, but we feel it is crucial to first consider the ideal situation where the authors of the original study planned to conduct only a single t test and that is exactly what they have reported. Although it may seem that this ideal situation would produce actual replication study power values very close to intended power, we will see that this assumption can be quite inaccurate. To the extent that actual research involves the additional practices specified above, the power of the replication study will generally be even lower because these practices generally exacerbate the extent to which the original study sample effect size is a biased estimate of the population effect size. Further, if p-hacking in the original study is suspected following a failed replication, it is important to first determine that the failure is unlikely due to problems with power. It is not enough that the intended power of the replication study be sufficient; instead, the replication study needs to have been designed with sufficient actual power, which is why we focus on the likely gap given current practices and also evaluate possible methods to eliminate the gap.

Literature review: Power and sample size planning for replication studies

Alongside the growth of replication research more generally, psychologists have begun to appreciate the importance of designing a replication study with adequate power specific to the domain (e.g. Brandt et al., 2014; Simonsohn, 2015). Yet, there is still variability in how researchers approach power analysis and sample size planning, if they do so at all. In fact, much of Sedlmeier and Gigerenzer's (1989) somewhat ironic report remains true today. They warned that the power of psychological studies had surprisingly not benefited from articles emphasizing the importance of power. Although some areas of psychology may approach desirable levels of power (e.g., health psychology; Maddock & Rossi, 2001), typical power reported in recent reviews is .35 (Bakker, van Dijk, & Wicherts, 2012) and even as low as .21 in neuroscience (Button et al., 2013). These estimates are even lower than those observed by Cohen in the 1960s.

Commonly used approaches

In order to determine how replication researchers generally conduct sample size planning today, we performed a focused literature review encompassing direct replications from 2013 to 2016.[3] The replications spanned several areas of psychology, though the modal subject areas were social and cognitive domains. Of those replication studies that reported a power analysis at all, the most common approach was to base sample size on the sample effect size estimate from the original study (employed in 9/15 replications in our review), as did our replicator in our hypothetical scenario. Other approaches were varied, but most were based generally on the sample size of the original study. The most common replication study sample size planning strategies will be discussed in more detail subsequently.

However, these replications found scattered throughout the general psychological literature may not be representative of the sample size planning approaches used by replication experts. Thus, we supplemented our general review with both the RPP and the 2014 special issue of *Social Psychology* dedicated to replication. As an example of good science practices, the RPP guidelines specified that researchers determine the "samples needed for 80%, 90%, and 95% power" and to "plan to collect at least enough data for 80% power, but to collect more data "if

[3] Specifically, we searched PsycINFO for articles containing "replicat*" in the title and either "direct replicat*" or "exact replicat*" or "close replicat*" in the abstract, for the years 2013 to March 2016. We excluded articles that were about the topic of replication, rather than replications themselves. We also excluded articles from the 2014 *Social Psychology* replication issue, as we reviewed those articles separately.

it is feasible to reach for greater power" (Open Science Framework, 2015). These guidelines go far beyond that of most journals, which still do not require a power analysis. However, even in this replication-focused project, most researchers (76/108) followed the standard procedure of using the original study's published effect size as the basis for their power calculations, as the replicator did in our introductory scenario. The findings from reviewing *Social Psychology* echoed this pattern (9/14 replications used this approach). Thus, most researchers aware of the importance of sample size planning have followed the standard procedure of using the original study's published effect size as the basis for their replication study power calculations.

Nonetheless, using the original sample effect size as the basis for replication study sample size planning can be problematic, as we saw in our hypothetical scenario. Sample effect sizes can be taken at face value as purely descriptive measures in a sample. However, sample effect sizes typically differ from unknown population effect sizes for two reasons. First, as with other statistics, the sample effect size is only an estimate of the true, unknown population value. Thus, it carries with it a distribution that represents that uncertainty around this value (Maxwell, Lau, & Howard, 2015). In other words, using the sample effect size in power calculations assumes that the reported value is fixed, when it should be treated as random (Taylor & Muller, 1996). In line with this, Batterham and Hopkins (2013) emphasized that "an observed large treatment effect in a small study would almost invariably be only *possibly* large" (p. 768, italics in original). Consequently, Pereira, Horwitz, and Ioannidis (2012) criticized the overreliance on point estimates of effect sizes, without considering the confidence interval surrounding them. For example, for a study finding an effect size of $d = 0.8$ with 20 participants per group, a 95% confidence interval around the effect size extends from 0.15 to 1.44. The confidence interval width improves with sample size (e.g., the confidence interval for the same effect size with $n = 250$ extends from 0.62 to 0.98), but still reflects nonnegligible uncertainty. Power calculations ignoring uncertainty often lead to underpowered studies (Dallow & Fina, 2011), but researchers seem generally unaware of the need to consider uncertainty: in a sample of 50 psychology articles published in 2013, only 4 articles of the 23 that provided an effect size estimate reported a confidence interval around it (Anderson & Maxwell, 2016). Further, only two of the 147 replications across all years and journals from our current review provided any mention of uncertainty in sample effect sizes.

Second, sample effect sizes tend to be upwardly biased, a result of strong journal preference for significant findings and selective reporting of multiple tests (Lane & Dunlap, 1978; Maxwell et al., 2015). Bias shifts the entire confidence interval surrounding the sample effect size upward (Ioannidis, Pereira, & Horwitz, 2013). Simulation studies have shown that this bias can be large, especially in underpowered studies (Brand, Bradley, Best, & Stoica, 2008). To illustrate, Figure 1 presents funnel plots showing the impact of publication bias for studies of varying sample sizes ($n = 10$–80, 5–10 randomly generated studies for each sample size) with a δ of 0.5 (vertical line). The left panel assumes no publication bias. The right panel takes the same set of studies, but assumes that only the subset of studies with $p < .05$ are published.[4] Figure 1 emphasizes two points. First, when publication bias is present, the points on the graph are not symmetric around δ: studies observing smaller values of d are not published. For example, if $n = 20$, sample effect size values lower than $d = 0.64$ will never appear in a journal that requires $p < .05$ for publication, even if δ is less than 0.64.[5] Second, the

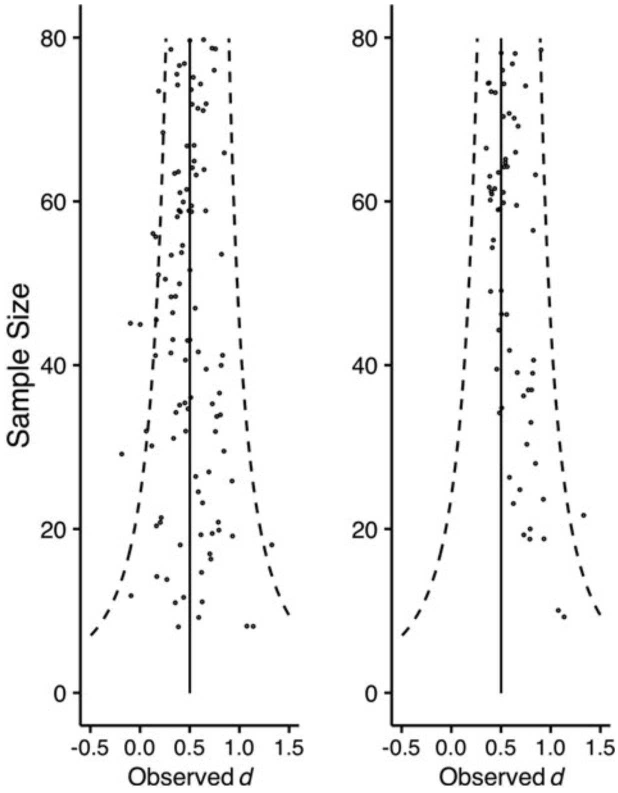

Figure 1. Funnel plot showing the impact of publication bias on a body of literature. The left panel shows a literature with no publication bias. The right panel shows a literature where studies reporting $p > .05$ are not published. The vertical line represents the true δ of 0.5. The dashed lines represent a 95% confidence interval around the meta-analytic estimate, d. Points are vertically jittered for visual ease. Sample sizes are per group.

[4] The precise shape of the funnel plot is specific to our choice of the magnitude of δ and the distribution of the sample sizes, but the overall pattern would be similar for other population values of δ and sample sizes.

[5] Assuming a sample size of 20 per group and an α level of .05, the critical t-value (38 degrees of freedom) = 2.02. We can calculate the minimum d

asymmetry decreases with larger sample sizes. For example, if $n = 80$, $d \geq 0.31$ can be published.

Moreover, the consequences of effect size variability and bias are often made worse by the nonlinear association between effect size and power. The power loss from using an overestimate of effect size is generally larger than the power gain from using an underestimate that differs from the population value by the same magnitude as the overestimate (Maxwell et al., 2015). The practical result of these issues is replication study power calculations that are too optimistic. Well-intentioned replicators who believe they are achieving .80 or .90 power may have actual power at values much lower than their goal, and their required sample sizes may be underestimates. However, the extent to which this optimism is influencing the field is not yet known. In other words, we do not yet know the magnitude of the difference between a replication study's intended power and the actual power achieved after correcting for these sources of error. The potential for bias has been noted occasionally in more recent replications, although sample size planning approaches that directly correct for bias were not used in any of our reviewed articles.

Beyond approaches capitalizing on the original sample effect size, other, more straightforward, strategies are also common in replication sample size planning. One intuitively appealing suggestion is to use the sample size of the original study for the replication. In all likelihood, the original study was able to find a statistically significant effect with the chosen sample size (otherwise it probably would not have been published), so it might seem reasonable to expect that a replication study should be able to do the same. In our review of the RPP replications, 19 of 108 studies ultimately employed this technique, despite some authors providing direct power calculations based on effect size. A related strategy is to use a sample size rule such as a 25% larger sample size than the original or include some additional participants as an intended cushion (employed in 6/108 studies in the RPP). Although they seem sensible, these simpler methods may be unsuccessful in either of two ways. First, they may augment the magnitude of difference between intended and actual power of the replication study, given the preponderance of underpowered studies in the literature (Button et al., 2013). If the power of the original study was typical of psychology, and thus had a power around .35 (Bakker et al., 2012), the power of the replication study will also be only .35 with the same sample size and may not be much larger even if the sample size is increased by 25%. Second, if the original study was truly high powered, these methods may also provide unnecessarily large suggested sample sizes, resulting in potentially added cost in terms of time and resources. Although larger samples have many benefits, and some areas of psychology may have a virtually unlimited number of participants (e.g., recruiting participants from a subject pool), not all areas are able to achieve such large samples. For example, in some subfields, it is difficult to identify and study special types of participants (e.g., autistic children, minorities), and in other areas, obtaining data from participants can be very costly (e.g., fMRI). For either or both of these reasons, it may be important not to overshoot the required sample size.

Our literature review identified a few additional ad-hoc replication study sample size planning methods for which we reserve discussion at the end of the article. However, critically, any approach that determines the sample size of the replication study based solely on the sample size of the original study does not use any information regarding the power or effect size of the original study, so resulting actual replication power (and whether it is an over- or underestimate of the intended power) may as well be a shot in the dark.

Proposed solutions in the literature

Several strategies have been developed to specifically take the distribution surrounding the sample effect size and/or publication bias into account when planning sample size for a future study. Perugini, Gallucci, and Constantini's (2014) safeguard power method "uses the uncertainty" in the reported effect size to "protect [the original study] from being underpowered" (p. 319). However, this protection comes at a potential cost. The method uses the lower limit of the 90% confidence interval surrounding the sample effect size, which may often result in an overly conservative estimate and larger replication sample sizes than necessary. Despite this potential disadvantage, its ease of implementation may make the lower bound method an appealing choice among researchers in planning replication studies. We abbreviate this method as "lower bound" throughout the remainder of the paper.

Hedges (1984) proposed a method to correct for publication bias in the sample effect size estimate. He first derived the distribution of d under the influence of publication bias (denoted h^*), when values of d associated with nonsignificant p-values are unobserved and thus censored. This distribution is as follows.

$$h^*(t|\delta, n) = \frac{h(t|\delta, n)}{A(\delta, n, \alpha)} \quad (1)$$

The numerator is a probability density function of a noncentral t, with degrees of freedom (df) $= 2n\text{-}2$

that would be observed in a published equal-n study with the following: $d\sqrt{\frac{n}{2}} \geq 2.02$. Rearranging the terms, and plugging in $n = 20$, we obtain: $d \geq 2.02\sqrt{\frac{2}{20}}$. Thus, $d \geq 0.64$. The calculations proceed similarly for $n = 250$.

and noncentrality parameter $= \delta\sqrt{\frac{n}{2}}$. The noncentrality parameter shifts the distribution of t away from the null distribution. The denominator is the probability that a noncentral F distribution ($df = 1$, $2n$-2, noncentrality parameter $= n\delta^2/2$) is above the relevant critical value (i.e., the power of a test) with the given noncentrality parameter. This density function provides the basis for a maximum likelihood-based estimate of δ that corrects for censoring. The new, bias-corrected estimate of δ is equal to the value of δ that maximizes the logarithm of Equation (1). Equation (2) is the resulting log-likelihood function of the sample effect size (t is replaced by d),

$$\log[h^*(d|\delta, n)] = \log[h(d|\delta, n)] - \log[A(\delta, n, \alpha)] \quad (2)$$

This method corrects the sample effect size estimator for upward bias, but does not consider the uncertainty around the sample estimate, as the purpose of Hedges' study was not to design replication studies. We henceforth denote this simply as "Hedges' method."

Recall the hypothetical example presented earlier in the article. The corrected effect size ($d = 0.19$) that we presented (based on an original study with an observed d of 0.70 and a sample size of 20 per group) was calculated using Hedges' (1984) method. Table 1 presents calculations of most likely replication study actual power for several common sample sizes and sample effect sizes if standard power calculations are used, in addition to the corrected effect sizes via Hedges' method, assuming .90 intended replication study power. We say "most likely" here to emphasize that, even when using Hedges' method, the population effect size is still unknown and thus the actual power is still unknown. If publication bias is operating, Hedges' method will arrive at estimated power values closer to the unknown actual power by selecting an effect size estimate closer to the true value, but given that estimates are still involved, the true replication study power is still unknown.

Table 1. Replication study power estimates for three observed *d*s, corrected for publication bias using Hedges' (1984) method.

	Sample *t*	Replication *n*	Corrected *d*	Actual power
		Observed $d = 0.3$		
$n = 20$	0.93	235	.06	.10
$n = 40$	1.33	235	.06	.10
		Observed $d = 0.5$		
$n = 20$	1.54	86	.11	.11
$n = 40$	2.21	86	.14	.10
		Observed $d = 0.7$		
$n = 20$	2.21	44	.19	.14
$n = 40$	3.09	44	.59	.78

Note. n = per-group sample size; d = Cohen's d effect size.

A third approach, proposed by Taylor and Muller (1996), aims to correct the sample d for both publication bias and uncertainty (essentially a combination of the lower bound (Perugini et al., 2014) and Hedges' (1984) methods). The following likelihood function is used.

$$f_L[f_S; \nu_1, \nu_2, \omega] = \frac{f_F[f_S|\nu_1, \nu_2, \omega]}{1 - F_F[f_{crit}(1 - \alpha_S)|\nu_1, \nu_2, \omega]} \quad (3)$$

The formula represents the density of a truncated noncentral F distribution, where f_F is the probability density function of the noncentral F distribution from the original study, f_s is the observed F-value for the original study, F_F is the cumulative distribution function (CDF) indicating the probability that a noncentral F distribution exceeds the critical value, f_{crit} is the critical F-value for α_s, the specified α-level of the original study, ν_1 and ν_2 are the *df* of the original study, and ω is the noncentrality parameter we are trying to estimate for our sample size planning. The numerator can be thought of as the likelihood of obtaining the observed original study F-value given the noncentrality parameter (in the absence of censoring), and the denominator can be thought of as the power of the original study test given the specified noncentrality parameter. In the two-group case, this formula reduces to a truncated noncentral t distribution. Equation (3) corresponds to the case of left censoring, where original studies are only published if they achieved statistical significance. Because researchers often plan replication studies using sample effect sizes from prior significant studies, the left censoring formula is most relevant for our purposes. A parallel formula is also available for right censoring. In the case of a t test, the method works by finding the value of δ associated with specific CDF probabilities. Choosing the δ corresponding to probability .5 results in a median-unbiased estimator of δ. Selecting a more conservative value of δ (from confidence bounds) allows the researcher to incorporate the uncertainty in the sample d. Taylor and Muller recommended using the lower limit of a one-sided 95% confidence interval (5th percentile) as a more conservative alternative to using the median-unbiased estimator. We will often abbreviate these methods as "Taylor and Muller (.5)" and "Taylor and Muller (.05)."

A potential caveat regarding both Taylor and Muller (1996) methods is the possibility that the δ value selected by the likelihood function is estimated to be zero. In other words, the original study was significant, but after correcting for publication bias and uncertainty, the corrected sample d is consistent in this case with no true population effect. In these cases, the sample size for a replication study cannot be estimated, and considering power is irrelevant. Selecting a δ value of zero would be especially appropriate if the original published study reflected a Type I error. This occurrence will be tracked in our simulations.

Critically, these four strategies that directly attempt to ameliorate the issues of publication bias and/or uncertainty in the sample effect size estimate are simply not currently being used to plan psychology replication studies. A review of the articles citing the lower bound method and Taylor and Muller's (1996) methods, for example, yielded no psychology replications using these methods *a priori* to plan replication sample sizes.[6] Thus, it is important to critically evaluate these proposed solutions to determine whether they should be recommended to substantive researchers.

The current study

Our rationale for the current study is as follows. A naïve view of replication might be that if all reported findings in original studies are real, then 100% of replication studies would be successful. Given the prominent recent lack of replicability, it seems natural to assume that many effects simply do not exist. Of course, this view would only be accurate if replication studies could be designed to have 100% power, which is unrealistic. A seemingly logical compromise is to design replication studies with power of .80 or even .90 as the goal, with an expectation that updates the naïve view: 80% or 90% of effects will replicate, if original effects are truly all nonnull in the population. Even with this updated view, the low rates of successful replications reported appear to cast doubt on the validity of many original studies. The purpose of this article is to explore whether other explanations for this low rate of replication success may be plausible, particularly with regard to power.

More specifically, it is clear that despite increased recognition of the importance of power analysis in both original studies and replication studies, commonly used sample size planning approaches may still be inadequate to provide an appropriate sample size for a replication study. We conduct a simulation study in line with the aim of determining the magnitude by which actual power to detect an effect in a replication study differs from intended power. In other words, we investigate whether, based on commonly used sample size planning conventions, the low replication rates reported in psychology would be surprising, or if they are rather a likely consequence of optimistic power calculations. We also assess the performance of "proposed solutions" and other sample size planning approaches that aim to improve upon the common strategies.

It is important to note that most of our demonstration is based on a best-case scenario. Unless otherwise specified, we assume that all published effects are nonnull, defined as having a nonzero population effect size. Although this is clearly not realistic, it enables us to answer the question in which we are most interested: Namely, if an original study reports an effect that is in fact real, what is the probability that a subsequent replication study will be a success or a failure? Our simulation results for replication success assuming 100% of effects are real will indicate power values that are larger (and thus differences between actual and intended power that are smaller) than a scenario that included original studies representing Type I errors. Further, the population proportion of null effects can never be known. We aim to show that even in a world where all psychological phenomena are real, actual power for a replication study may be lower than intended. However, we also believe that it is important to assess how well the proposed solutions of Perugini and colleagues (2014), Hedges (1984) and Taylor and Muller (1996) perform in the case where the null hypothesis is true. In reality, a replication researcher who reads a published study will not know whether or not the significant result is a Type I error. It would be especially useful if these methods could reliably signal a potential Type I error.

To summarize, we compare various sample size planning approaches, based on original studies with varying levels of actual power, with regard to the proportion of replications that would be successful. To manipulate the power of the original study, we vary both sample size and the population effect size. We reference the effect size designations of Cohen (1988) to provide the basis for the population effect sizes typically found in the psychological literature. We first assess the power separately for each of Cohen's small, medium, and large δ values (Simulation 1). We then mimic what may be seen across multiple studies within a discipline such as psychology by assuming that various effects have different true δs, which we will reflect with random δ draws from a normal distribution (Simulation 2). Finally, we assess the performance of the proposed solution methods in the case where the true δ is zero (Simulation 3).

Method

To assess the actual proportion of studies that we would expect to replicate based on various sample size planning approaches, we conducted analytical work and a simulation study using R Statistical Software. The analytical work assessed two variations of sample size planning methods that base replication sample sizes only on the sample size of the original study. The simulation study assessed six methods basing replication sample sizes on some form of the original study effect size estimate,

[6] Hedges' method did not have the goal of instructing sample size planning, so it is not expected that researchers would be citing Hedges (1984) for replication sample size planning.

including two methods that take this sample estimate at face value and the four approaches described in the "Proposed Solutions in the Literature" section, which correct this sample estimate for uncertainty and/or publication bias. Our work assumed an independent samples t test as the analysis method for the effect of interest, with d as the specified sample effect size. The specific values we obtain would differ for other types of statistical tests, but the overall pattern of results would remain largely the same. All tests were two-tailed for the following reasons. Regarding the original study: 1) two-tailed tests are more common, and 2) authors of original studies may not have yet hypothesized a direction of effect. Regarding the replication study: 1) authors of replication studies are often interested in results in both the expected and unexpected directions, 2) even if the original study reported a significant effect in the positive direction, it is possible that the reported effect was incorrect in sign, so it may be important for the replication to detect an effect in the opposite tail (Type-S error, Gelman & Carlin, 2014), and 3) additional simulations showed that the results were nearly identical when the replication study was one-tailed. Finally, we assumed that only original studies reporting a p-value $< .05$ would be published and available for subsequent replication.

Analytical work

The simplest methods of sample size planning in our investigation are 1) using the original study sample size and 2) using a sample size 25% larger than the original study in the replication study. Determining the proportion of studies that will successfully replicate, under different original study sample sizes and population effect sizes, does not require simulation. Thus, we analytically determined the actual power for replication studies based on these two methods, under three δs (0.2, 0.5, and 0.8) and four original study sample sizes (20, 40, 80, and 250 per group). These parameters are identical to those implemented in the simulation study, which we discuss subsequently. The actual replication power values were calculated with the pwr package in R for an independent sample t test. Results are presented in the first two columns of Table 2.

Table 2. Actual replication study power based on four basic sample size planning approaches.

	Same n	25% Larger n	Intended power of .80	Intended power of .90
		$\delta = 0.2$		
$n = 20$	**.095**	**.107**	**.108**	**.148**
$n = 40$	**.143**	**.168**	**.173**	**.231**
$n = 80$	**.242**	**.291**	**.305**	**.401**
$n = 250$	**.607**	**.705**	**.626**	**.726**
		$\delta = 0.5$		
$n = 20$	**.338**	**.410**	**.406**	**.497**
$n = 40$	**.598**	**.697**	**.609**	**.726**
$n = 80$	.882	.940	**.740**	**.835**
$n = 250$	1	1	**.787**	**.882**
		$\delta = 0.8$		
$n = 20$	**.693**	**.791**	**.661**	**.761**
$n = 40$	.942	.977	**.771**	**.850**
$n = 80$	.999	1	**.794**	**.881**
$n = 250$	1	1	.806	**.894**
		$\delta \sim N(0.5, 0.15^2)$		
$n = 20$	**.421**	**.507**	**.455**	**.547**
$n = 40$	**.661**	**.744**	**.635**	**.720**
$n = 80$	.863	.905	**.727**	**.819**
$n = 250$	.982	.989	**.781**	**.872**

Note. Actual power values below their intended power are bolded. Intended power methods indicate using the uncorrected original study sample d to obtain the specified level of intended power. n = per-group sample size; δ = population value of Cohen's d effect size.

Simulation 1

In the first simulation, the following parameters were varied in order to manipulate the, typically unknown, actual power of the original study: 1) δ fixed across repetitions at 0.2, 0.5, and 0.8, and 2) original study sample size (20, 40, 80, and 250 per group) for the effect of interest. Sample sizes were chosen to mirror those found in typical psychological studies. Specifically, our "small" sample size was based on the minimum per-cell sample size suggestion provided by Simmons et al. (2011). Based loosely on the distribution of sample sizes for experimental studies we have seen in our literature review, we also assessed a "medium" sample size of 40 per group and a "large" sample size of 80 per group. Finally, we selected a fourth more extreme sample size, $n = 250$, to showcase what happens in the case of a very high-powered original study. In terms of the population effect size, we selected the values of δ per Cohen's (1988) benchmarks.

Procedure

The simulation can be conceptualized in terms of two phases. In the original study phase, we drew random samples of data from a standard normal distribution ($\mu = 0$, $\sigma^2 = 1$). The N observations of each sample were split into two groups of size n, coded 0 and 1. The observations were then transformed by adding the product of δ and the group code (0 or 1) to the original observation value, resulting in population group differences equal to the specified δ. Independent samples t tests were performed on each sample, without assuming homogeneity of variance, and the resulting p-values and sample ds were recorded. Only those samples that achieved a significant p-value ($<.05$) were then passed to the replication study phase.

In the replication study phase, simulations of replication studies were conducted using each of the following six replication study sample size planning methods: 1) using the original study sample d as the assumed δ to obtain an intended power of .80, 2) using the original study sample d as the assumed δ to obtain an intended power of .90, 3) using the lower bound of the confidence interval surrounding the sample d to obtain an intended power of .80 (Perugini et al., 2014), 4) using Hedges' (1984) method's maximum likelihood estimate of the sample effect size, corrected for publication bias, to obtain an intended power of .80, 5) using Taylor and Muller's (1996) median-unbiased estimator of the sample effect size to obtain an intended power of .80 (Taylor and Muller (.5)), and 6) using Taylor and Muller's lower limit of a 95% confidence bound estimator of the sample effect size to obtain an intended power of .80 (Taylor and Muller (.05)). We opted to assess these six sample size planning strategies because they represent approaches commonly seen in the literature or proposed as potential solutions to correct for publication bias and uncertainty. The number of repetitions (10,000; we refer to simulation replications as "repetitions," to distinguish this term from replication studies) was chosen to ensure a small standard error surrounding each reported power value, even when a proportion of simulated original studies would not be passed along to the second phase. The minimum number of repetitions remaining in the second phase is 934, leading to a maximum standard error for intended power of .016.[7]

Quantities assessed

As aforementioned, we operationalized actual replication study power as the proportion of replication studies achieving statistical significance at $p < .05$. Thus, the proportion of "successful" replication studies was recorded for each of these six approaches to choosing the sample size of the replication study, in addition to the mean and median sample size that each method would calculate as necessary to achieve the specified power (regardless of whether that sample size really was sufficient to obtain the intended power).

The calculation of the proportion of successful studies was slightly more complicated for methods 4–6.[8] As we discussed earlier, the likelihood estimates for d in these methods could sometimes equal zero. In these cases, one would not be able to conduct sample size planning for a replication study, so these studies never really make it to the replication phase of the simulation. For these methods, we calculated the average replication power in two ways. First, we divided the number of significant replication studies by the number of studies that made it to the replication phase (which excludes nonsignificant original studies and significant original studies that had a corrected d of zero). This calculation may be an accurate estimate of actual power, but it does not reveal all the cases in which a replication study cannot even be conducted with the selected approach. Therefore, we also used a denominator that included the studies that had a corrected d of zero. These success-rate values will be lower, as they include the situation of not even being able to conduct a replication study in the category of nonsuccess.

For each method, we also calculated assurance, which we would like to distinguish from the more common concept of power. Assurance is the proportion of times the power would be at or above the intended level, if the experiment were reproduced many times. For example, suppose an effect exists with a δ of 0.5. An original study is conducted, which means (in the absence of publication bias) there is approximately an equal probability that the observed effect size will be either less than or greater than the population value of 0.5. Suppose a replication study is then planned to obtain a power of .80 based on the observed effect size of the original study. If the original study reported a d less than 0.5, the replication study will have a power above .80. On the other hand, if the original study reported a d greater than 0.5, the replication study will have a power less than .80. The practical implication is that even in the absence of publication bias, a replication study that uses an original observed effect size to determine sample size is likely to have appropriate intended power only about 50% of the time (an assurance of 50%). However, a replication researcher should arguably be dissatisfied with achieving his/her intended power only 50% of the time, so methodologists have developed methods that provide the intended power a higher percentage of time. Assurance quantifies this percentage. Methods producing higher assurance will also have higher average actual power because the goal is to have power above the intended level more of the time.

Simulation 2

The second simulation was similar to the first simulation, with the exception of how the population effect size was simulated. In this simulation, δ was randomly drawn from a normal distribution with $\mu = 0.5$ and $\sigma = 0.15$.

[7] This estimate is conservative, as the calculation assumes that $p = .5$. Because we know that p is much less than .5 in the scenario that results in the lowest number of usable repetitions, we can expect a maximum standard error that is more precise than this estimate. However, this estimate does not take into account the repetitions in which Taylor and Muller's corrected d is estimated to be zero.

[8] Although Hedges (1984) did not mention the possibility of zero estimates using his method, we assessed the power using both denominators for this method as well. However, Hedges' method produced no zero estimates in any condition.

These values were chosen to mirror the distribution of δs commonly seen in the published literature, representing a medium δ as average, with small to large δs falling within ±2 *SD* from the mean, while allowing for the possibility of more extreme δs on either end of the distribution.[9] The other simulation parameters were identical to the first simulation, and again, 10,000 repetitions were used. Again, the proportion of successful replication studies, assurance, and replication sample sizes based on the aforementioned sample size planning methods were recorded. This simulation included the two approaches basing the replication sample size on the sample size of the original study (same n and 25% larger n), as the outcome quantities depend on the particular random draw of the population δ.

Simulation 3

In our third simulation, we again ran 10,000 repetitions on original studies, with δ set to zero, varying the sample size with the levels used in the previous simulations. We assessed the following four methods: the lower bound method (Perugini et al., 2014), Hedges' (1984) method, Taylor and Muller (1996) (.5), and Taylor and Muller (.05). This simulation only involved a single phase, the original study phase, as our interest involved the distribution of corrected sample ds estimated by each method. We wanted to know how reliably these methods signaled a possible effect size of zero, when in reality the effect was null. Therefore, we calculated the proportion of zeroes estimated by each method.

Results

Results for power are displayed in Tables 2 and 3. Table 2 shows the actual power obtained (proportion of studies that replicate) for replication researchers employing each of the four basic sample size planning strategies: original sample size, a sample size of 25% larger than the original study, .80 power based on the original published effect size, and .90 power based on this same effect size. Table 3 displays actual power results for .80 intended power based on the four proposed solutions: the lower bound method (Perugini et al., 2014), Hedges' (1984) method, Taylor and Muller (1996) (.5), and Taylor and Muller (.05). The power resulting from these strategies was evaluated when the original study δ was 0.2, 0.5, and 0.8, and drawn from $N(0.5, 0.15^2)$, respectively.

Table 3. Actual replication study power based on four proposed solutions (all with intended power of .80).

	LB	Hedges	TM .5	TM .05
		$\delta = 0.2$		
$n = 20$	**.672**	**.544**	**.283** (.160)	**.587** (.049)
$n = 40$	.803	**.670**	**.419** (.259)	**.764** (.082)
$n = 80$	.900	**.731**	**.516** (.347)	.839 (.122)
$n = 250$	.956	**.779**	**.692** (.563)	.936 (.328)
		$\delta = 0.5$		
$n = 20$	.928	**.745**	**.577** (.407)	.857 (.153)
$n = 40$	.960	**.770**	**.678** (.550)	.932 (.319)
$n = 80$	.965	**.792**	**.764** (.703)	.953 (.581)
$n = 250$	.949	.804	**.796**	.947 (.942)
		$\delta = 0.8$		
$n = 20$	.962	**.775**	**.708** (.591)	.934 (.350)
$n = 40$	.964	**.793**	**.777** (.741)	.958 (.677)
$n = 80$	.956	**.794**	**.796** (.795)	.955 (.934)
$n = 250$	.923	.805	.800	.924
		$\delta \sim N(0.5, 0.15^2)$		
$n = 20$	.929	**.742**	**.603** (.443)	.879 (.208)
$n = 40$	.949	**.783**	**.706** (.590)	.941 (.377)
$n = 80$	.960	**.786**	**.764** (.704)	.955 (.622)
$n = 250$	.948	**.789**	**.782** (.774)	.947 (.886)

Note. Actual power values below their intended power are bolded. Parenthetical values refer to the proportion of successful replications when studies with a corrected d of zero are included in the denominator. δ = population value of Cohen's d effect size; n = per-group sample size; LB = lower bound; TM.5 = Taylor and Muller 50th percentile; TM.05 = Taylor and Muller 5th percentile.

Actual power: Basic methods

We first discuss the results from Table 2. Of most interest to the current study is the general trend toward much lower replication study power values than the nominal values that researchers expect. The differential between actual and intended power is especially pronounced in the small δ condition, which is especially important given that many effects that psychologists study are small in size (Prentice & Miller, 1992). For example, when the original sample size is 20 per group, a researcher who has selected a replication sample size based on a power analysis that claims a power of .80 has less than a 12% chance of a statistically significant replication. Moreover, this is true for a psychological effect that really does exist. In this condition, even a researcher who goes above and beyond the typical .80 power benchmark and believes that the replication sample size will guarantee a power of .90 has less than a 15% chance of replicating. When the true δ is small, even a replication attempt of an original study with a sample size of 80 per group that intends a power of .90 will only have less than a 40% chance of success. The actual power values for researchers who duplicate the sample size of the original study are even lower, failing to reach .30 even when the original study used a sample size of 80 per group. Thus, the simple methods of using the original study sample size or adding a 25% cushion were generally

[9] Richard, Bond, and Stokes-Zoota (2003) found the mean reported effect size in social psychology to be $r = .21$, which translates to $d = 0.43$, just under Cohen's medium effect size.

among the poorest of choices with regard to power. The proportion of replicated studies vastly improves when the per-group sample size is 250, but even in this situation, the proportions do not reach their intended values.

When the δ of the phenomenon under study is truly medium, the actual power of replication studies is somewhat improved but still often vastly different from the .80 minimum benchmark recommended to psychology researchers. The situation only becomes somewhat acceptable when the original study has a large sample size of 80 per group. In this case, using the same sample size as the original study (and the 25% cushion method) actually outperformed both methods using the sample effect size estimate as a guide. However, as we previously noted, these methods based solely on the sample size of the original study are quite disconnected from the size of the effect in question, and this overestimation has its own consequences in terms of feasibility and wasted resources.

Not surprisingly, the best scenario occurs when a researcher is attempting to replicate a truly large δ. In this scenario, the expected proportion of replicating studies is very close to the intended level for each of the methods when the original study had a sample size of 80 per group or larger. However, even under a large δ, the actual replication power values for the small sample size condition are lower than ideal. Again, using the same sample size as the original study, or adding a 25% cushion, tended to overshoot the intended power.

In a scenario set to more broadly mirror psychology as a discipline, where δ was drawn from a normal distribution centered at a medium effect size with a standard deviation of 0.15, the results indicated actual replication power values similar to those when a medium δ was fixed. Here, an intended power value of .90 for a replication with a small original sample size ($n = 20$) corresponds to an actual chance for replication success of only about 55%.

Actual power: Proposed solutions

We now discuss results from the four proposed solutions in Table 3. Overall, the proportion of studies expected to replicate based on these methods is larger than the basic methods. In almost all conditions, the lower bound (Perugini et al., 2014) was conservative, providing replication study power values often well above the intended level. For example, with a large δ, the lower bound method reached a power of .95 in all sample size conditions (with the exception of $n = 250$). This result is due to the lower bound's achievement of high assurance (see "Assurance" section). Hedges' (1984) method and Taylor and Muller (1996) (.5) performed relatively similarly throughout, which is unsurprising given the similarity in their formulas. Both methods aim to correct publication bias, but do not consider uncertainty in the original published effect size. However, especially when the original study was more underpowered, Hedges method tended to have higher power than Taylor and Muller (.5). With a δ of medium or larger and a larger original study sample size, both these methods tended to be fairly close to the intended replication study power benchmark of .80. However, in the small δ condition, these methods still produced gross underestimates of power, with the exception of the $n = 250$ condition.

Taylor and Muller (1996) (.05) produced a large proportion of replicated studies in almost all conditions, with power values exceeding the intended level in the majority of cases, again due to its goal of high assurance. However, even this dual solution—dealing with both publication bias and uncertainty—only produces a replication study power of about .60 when the true effect size and sample size are small, as is true of many psychology studies. Note that the actual replication study power values for the Taylor and Muller methods are optimistic, given that they do not account for the situations in which the corrected d is zero. The parenthetical values present the proportion of replicated studies, using a denominator that includes cases in which the study never makes it to the replication phase. These alternate "replication success" values are much lower, even as low as 5% in the small δ small sample size condition. We emphasize that this does not signal a problem with the Taylor and Muller methods. In fact, these methods are drawing attention to the fact that effect sizes with high levels of uncertainty may not only be incorrect in magnitude but also in whether or not the claimed effect is even real.

Suggested replication sample size

Table 4 shows the mean and median "required" replication sample sizes based on each method, for the same δ and original study sample size specifications as Tables 2 and 3. It is important to note that these sample sizes only result in the power values reported in Tables 2 and 3, *not necessarily* the intended power values. Required replication sample sizes generally increase for original studies with smaller δs. Somewhat nonintuitively, the required sample size generally increases as the original study sample size increases. We elaborate on this result in the "Discussion" section. The most noteworthy piece of Table 4 is the extremely large mean sample sizes required by some of the methods, which can be even upward of several thousand participants per group in the small δ condition. These extremely large values are influenced by outlier corrected ds that are close to zero in some repetitions, thus requiring extreme sample sizes. The median values provide a better estimate of the typical required sample size in these cases. However, again, the protection offered by methods such as Taylor and Muller (1996) (.05) and the

Table 4. Mean required replication study sample sizes based on eight sample size planning approaches.

	Same n	25% Lgr	.80 Pwr	.90 Pwr	LB	Hedges	TM .5	TM .05
				$\delta = 0.2$				
$n = 20$	20	25	28 (28)	37 (38)	439 (328)	266 (253)	359 (50)	708 (274)
$n = 40$	40	50	55 (56)	73 (74)	880 (604)	507 (436)	575 (83)	3769 (545)
$n = 80$	80	100	107 (109)	143 (145)	1632 (1097)	928 (699)	1162 (155)	6078 (930)
$n = 250$	250	313	275 (260)	368 (347)	3301 (1614)	1880 (545)	1899 (298)	9563 (1571)
				$\delta = 0.5$				
$n = 20$	20	25	25 (25)	33 (33)	341 (210)	197 (116)	319 (34)	1214 (165)
$n = 40$	40	50	45 (42)	59 (56)	528 (272)	290 (79)	533 (47)	1802 (217)
$n = 80$	80	100	67 (58)	89 (77)	567 (232)	303 (67)	418 (60)	2026 (217)
$n = 250$	250	313	72 (64)	95 (85)	170 (127)	76 (64)	77 (64)	291 (130)
				$\delta = 0.8$				
$n = 20$	20	25	21 (19)	28 (26)	225 (110)	123 (31)	200 (21)	949 (86)
$n = 40$	40	50	29 (25)	39 (33)	199 (85)	101 (27)	117 (25)	881 (86)
$n = 80$	80	100	30 (26)	39 (34)	86 (57)	37 (26)	50 (26)	279 (57)
$n = 250$	250	313	27 (26)	36 (34)	42 (39)	27 (26)	27 (26)	42 (39)
				$\delta \sim N(0.5, 0.15^2)$				
$n = 20$	20	25	25 (24)	32 (31)	333 (186)	177 (79)	390 (28)	1935 (130)
$n = 40$	40	50	42 (38)	55 (51)	458 (214)	258 (64)	246 (41)	1495 (146)
$n = 80$	80	100	62 (53)	83 (70)	524 (190)	290 (57)	339 (52)	1514 (146)
$n = 250$	250	313	89 (62)	118 (83)	409 (124)	200 (62)	205 (62)	913 (116)

Note. Median sample sizes shown in parentheses. .80 and .90 pwr methods indicate using the original study uncorrected sample d to obtain .80 and .90 intended power. δ = population value of Cohen's d effect size; n = per-group sample size; pwr = power; 25% Lgr = using a replication sample size 25% larger than the original study; LB = lower bound; TM.5 = Taylor and Muller 50th percentile; TM.05 = Taylor and Muller 5th percentile.

lower bound method can come at a cost especially when the true δ is small.

Assurance

Table 5 displays the results for the assurance of each method under the different effect size conditions. Again, assurance represents the proportion of replications in which the actual power value is as large as the intended target. Note that the assurance of the two methods based solely on the original study sample size is either zero or one for the conditions in which δ was fixed, as these methods either do or do not reach the intended power target. Taylor and Muller (1996) (.5) and (.05) aim to have assurance of .5 and .95, respectively, representing the proportion of replications in which the power would be higher than the specified value. It is evident that in the more ideal conditions (δ medium or larger, original study sample size not extremely small), the observed assurance values were consistent with the expectations. Although not specifically mentioned, Hedges' (1984) method seems to provide 50% assurance and the lower bound method reaches 95% assurance in these situations as well. However, especially in the small δ condition, the assurance for all three methods did not meet the target. In fact, when $\delta = 0.2$ and the original study $n = 20$, none of the methods produced an assurance as large as 45%. Even so, these values alone give a somewhat unclear picture of the performance of Taylor and Muller's methods with regard to assurance because the assurance estimates do not include studies where the Taylor and Muller corrected d is zero. If instead, corrected ds of zero are followed with a replication sample size large enough to reach the intended power, the assurance values for Taylor and Muller's methods are consistent with the expectations, even when the original study is underpowered.

This exclusion of zero estimates from assurance also tempers the comparison of Taylor and Muller (1996) (.05) and the lower bound method, both of which achieve close to 95% assurance in some conditions. It appears that the assurance of the lower bound method slightly exceeds Taylor and Muller (.05). However, again these numbers are derived from only the studies that were able to be replicated, which for Taylor and Muller's (.05) method, do not include the studies with a corrected d of zero. When these same studies are removed from the lower bound method as well, Taylor and Muller (.05) has a higher assurance than the lower bound method. This is consistent with the expectations, given that Taylor and Muller (.05) corrects for both uncertainty and bias, leading to larger suggested replication sample sizes and more replications with power as large as intended.

Zero effect size condition

All our previous results assumed the hypothetical case where the effect of interest was real. However, published effects may represent a Type I error. How well did the proposed solutions do at properly estimating the null effect size? Taylor and Muller (1996) (.05) reliably signaled that the true δ could be zero. The proportion of zero estimates ranged from 93% to 96%. Taylor and Muller (.5) correctly

Table 5. Assurance for eight sample size planning approaches.

	Same *n*	25% Lgr	.80 Pwr	.90 Pwr	LB	Hedges	TM .5	TM .05
				$\delta = 0.2$				
$n = 20$	0	0	0	0	.418	.344	.126	.387
$n = 40$	0	0	0	0	.658	.542	.202	.580
$n = 80$	0	0	0	0	.803	.583	.263	.710
$n = 250$	0	0	.186	.185	.919	.577	.405	.884
				$\delta = 0.5$				
$n = 20$	0	0	0	0	.857	.589	.298	.742
$n = 40$	0	0	.168	.160	.924	.554	.378	.858
$n = 80$	1	1	.436	.427	.945	.534	.460	.919
$n = 250$	1	1	.503	.490	.953	.521	.522	.954
				$\delta = 0.8$				
$n = 20$	0	0	.291	.294	.933	.553	.408	.861
$n = 40$	1	1	.485	.488	.952	.522	.486	.937
$n = 80$	1	1	.515	.520	.955	.503	.507	.953
$n = 250$	1	1	.532	.538	.956	.517	.521	.953
				$\delta \sim N(0.5, 0.15^2)$				
$n = 20$	.007	.039	.048	.047	.861	.566	.319	.779
$n = 40$	.284	.466	.238	.237	.911	.564	.418	.876
$n = 80$	.736	.844	.404	.401	.940	.527	.462	.928
$n = 250$	.975	.986	.498	.495	.955	.500	.490	.948

Note. .80 and .90 pwr methods indicate using the original study uncorrected sample *d* to obtain .80 and .90 intended power. δ = population value of Cohen's *d* effect size; *n* = per-group sample size; pwr = power; 25% Lgr = using a replication sample size 25% larger than the original study; LB = lower bound; TM.5 = Taylor and Muller 50th percentile; TM.05 = Taylor and Muller 5th percentile.

detected a null effect about half of the time (46%–50%). However, neither Hedges' (1984) method nor the lower bound method (Perugini et al., 2014) produced zero estimates. The δ estimates (absolute value of the sample *d*) for all four proposed solutions are graphed in Figure 2. We can see from the plot that the Taylor and Muller (.05) was the only method that reliably signaled a null effect to the researcher.

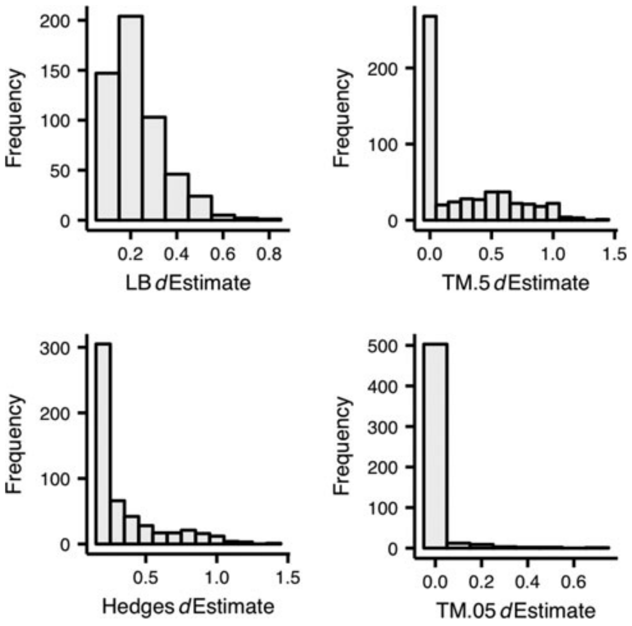

Figure 2. Plot showing the distribution of corrected sample *d* estimates when δ is zero, based on an original study sample size of 20 per group, for the four proposed solutions. LB = lower bound; TM.5 = Taylor and Muller 50th percentile; TM.05 = Taylor and Muller 5th percentile.

The role of original study power

We have seen that the four proposed solutions work better than the more commonly used methods of sample size planning for replication studies. However, actual power values were still much lower than intended when both original study sample size and δ are small. Figure 3 shows the likelihood function underlying Taylor and Muller's (1996) method. Taylor and Muller (.05) and the lower bound method (Perugini et al., 2014) work with the lower tail of a likelihood function. When the original study is underpowered, these methods select corrected *d*s near zero, where even small differences in these estimates can imply large differences in suggested replication study sample sizes. For example, decreasing the lower bound method's corrected *d* from 0.13 to 0.12 results in the suggested sample size changing from 938 to 1168, whereas a similar difference in the corrected *d* in the upper tail of the likelihood function results in a much smaller change in suggested sample size: decreasing from 0.87 to 0.86 changes the suggested replication sample size by only one participant. The large fluctuations in suggested sample sizes can have implications for how successful the replication will be in achieving the intended power. Hedges' method (Hedges, 1984) and Taylor and Muller (.5) work more with the middle portion of the likelihood,

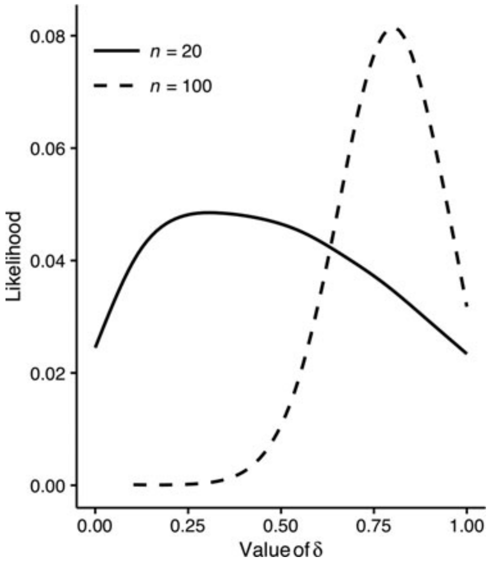

Figure 3. Likelihood function for Taylor and Muller's method for an original study sample *d* of 0.70, with curves for smaller ($n = 20$ per group) and larger ($n = 100$ per group) original study sample sizes.

where these issues are somewhat ameliorated. However, they achieve around 50% assurance, which means that unlike the other two methods, only half of the time will actual replication study power reaches intended power even when the methods work well.

Discussion

The results of this study indicate that common methods of planning the sample size for replication studies provide inadequate estimates of the actual power achieved. This general pattern is unsurprising. However, more noteworthy is that the magnitude of these overestimates can be quite substantial. For example, a replication researcher intending to achieve .80 power based on the observed effect size of the original study may actually have only around a 10% chance of detecting a real effect. The common approaches that performed inadequately in our simulation would presumably be less popular among replicators if researchers were aware of their limitations, which cloud attempts to accurately interpret the widespread tendency for so many replication studies to fail to achieve statistically significant results. In fact, Stanley and Spence (2014) showed that measurement error alone can make replication studies difficult to interpret, without adding in the issue of power. With this knowledge, it is not so surprising that fewer than 40% of effects replicated successfully in the RPP, and that the general replication success rate in psychology is so low. Recall that in our simulations, the difference between intended and actual power was larger in the small sample size conditions than the larger sample size conditions. When reminded of the small sample sizes found in many psychological studies, our results suggest that we might expect the proportion of replicated studies to be even lower than it has been in recent reports simply as a function of the likely difference between intended and actual power.

What do our results mean, then, for the replication crisis? The discussion surrounding the issue of replication has focused mainly on facets of the replication study and on the suddenly untrustworthy nature of many claimed effects. In other words, psychologists are assuming that when replications are planned as carefully as they have been, failures to replicate signal suspicion toward the existence of original effects. Yet, our simulation shows that other explanations may be valid alternatives. The issue with the original study may have nothing to do with the credibility of the original claim, but rather with the actual power of that study. Ironically, in order to begin to emerge from the *replication* crisis, the burden falls on researchers of *original* studies to design their studies with larger sample sizes and greater power in mind.

Probability of successful replication: The original study matters

From an alternate perspective, the results of this study indicate that the probability of a successful replication depends heavily on the sample size of the original study, regardless of the method used to plan the sample size of the replication study. Researchers of original studies rarely consider the implications of sample size and power after finding and publishing a significant effect. This is because power is often taught in relation to avoiding Type II errors, which by definition did not occur for an effect reaching statistical significance. Yet, we have shown that the sample size of the original study does play a pivotal role in the likelihood of successful replication, and thus, the likelihood of an effect being accepted as real by the scientific community. The sample effect size from larger original studies is more trustworthy for two reasons. First, the amount of uncertainty surrounding the effect size is likely to be smaller due to the increase in precision. Second, the likelihood and amount of publication bias that needs to be accounted for are smaller. Thus, when the original study is sufficiently powered, basic methods that do not correct for these two issues will produce less overestimation. Further, methods that do account for these concerns will need to make less of a correction to the sample effect size used in the calculations. These advantages to a high-powered original study will lead to a smaller discrepancy between actual and intended power, and thus maximize the ideal conditions for successful replication.

A second caveat regarding the original study sample size relates to the following general trend: increases in original study sample size are associated with increases in suggested replication study sample sizes. This result may seem somewhat counterintuitive, given our emphasis on the advantage of a larger original sample size in the previous paragraph. However, original studies with larger sample sizes also are able to detect smaller effects, which in turn require a larger sample size to replicate. Again, as discussed in the "Literature Review: Power and Sample Size Planning for Replication Studies" section, this pattern emphasizes the shrinking confidence interval surrounding the sample effect size when the original study has larger power, a reflection of decreasing uncertainty.

To finish our discussion on original study sample size, we note that small original studies are not always inherently flawed, as it is the small sample size *in conjunction with* a nonlarge true effect size that results in uncertainty and bias. To briefly illustrate, we ran a simulation where the original study used only 20 participants per group, but the true δ was 1.6 (twice the size of Cohen's large effect size). In this case, all the replication sample size planning methods achieved at least .80 power. Effect sizes of this size are not necessarily nonexistent in psychology (see Fraley & Vazire, 2014).[10] However, we still caution that many psychological phenomena are small in nature, even if the effect sizes present in the literature seem larger due to publication bias.

Unpacking the performance of the potential solutions

We have explored several solutions to the low power inherent in current sample size planning conventions: Perugini and colleagues' (2014) lower bound method, Hedges' (1984) method, and Taylor and Muller's (1996) methods. These methods generally outperformed more basic methods of sample size planning, with regard to power and assurance. For example, Hedges' method and Taylor and Muller (.5) can provide 50% assurance in ideal conditions, and the lower bound method and Taylor and Muller (.05) increases this value to 95%. However, our simulations showed that even accounting for both publication bias and uncertainty is not a panacea for a highly underpowered original study. This power issue is smartly signaled when a zero estimate arises for the corrected effect size. In the real world, replication researchers are unaware of whether a claimed original effect is real. The zero effect size estimate says that the power is such that we cannot safely assume that the effect in question is real. As we discuss later, this does not mean that a replication of an underpowered study is unmerited but rather that several replications may be warranted.

We saw that this "safety mechanism" can be especially valuable when the published effect is a Type I error. Taylor and Muller (.05) arrived at a corrected *d* of zero almost all the time in this condition, alerting the researcher to the possibility of a Type I error. One might wonder why this method does not arrive at zero 100% of the time in this case. Recall that even when the population null hypothesis is true, the sample effect size estimate will not likely be zero. Furthermore, somewhat nonintuitively, the sample effect size estimates can be quite high, especially with smaller sample sizes. In fact, as discussed previously, when a study is published, the reported effect size has to be larger than a certain critical value—meaning that published Type I errors often report large effect sizes. It is thus even more impressive that Taylor and Muller's methods selected $d =$ zero as often as they did in our simulation. The lower bound method never arrived at an estimate of zero because it is based on the confidence interval for the observed effect size, which will not include zero in a published study. Hedges' (1984) method also did not estimate any effect size as zero, which agrees with Hedges' original calculations (see Hedges, 1984, Figure 2).

Another potential solution that deserves increased attention is the Registered Replication Report, which offers a different way to assess replication. Rather than replicating many studies one time each, as in the RPP, Registered Replication Reports allow a single study to be replicated many times, by teams of independent researchers. Critics of the RPP design have noted that combining the results of multiple replication attempts, as in the Registered Replication Report (and the similar Many Labs Project), results in larger power in aggregate (Gilbert, King, Pettigrew, & Wilson, 2016). This method allows the sample size to increase iteratively with each new replication. For example, a recent replication (Alogna et al., 2014) of the verbal overshadowing effect by 31 laboratories found a meta-analytic effect size smaller than the original study, but with a much smaller confidence interval due to the benefit of a larger total sample size and increased precision. More nuanced determinants of a "successful replication" beyond the simple significant-nonsignificant distinction could also provide a more detailed picture of an effect's replicability (Anderson & Maxwell, 2016; Cook et al., 2008).

The general lesson to be learned here is that the proposed solutions can only do so much when the true issues lie with the power of the original studies themselves. The issue at hand is simple: underpowered original studies

[10] A population effect size as large as 1.6 may be quite unusual in between-subject designs, but may be more prevalent in within-subject designs. For example, with all other things being equal, the noncentrality parameter of a dependent *t* will be larger than the corresponding noncentrality parameter of an independent *t* by a factor of $\frac{1}{\sqrt{1-\rho}}$ (Liu, 2014, p. 45). Consequently, if ρ equals .75, the noncentrality parameter for the dependent *t* will be twice as large as for the independent *t*. For example, a large effect size of 0.8 for *d* in a between-subject design would correspond to an effect size of 1.6 in a within-subject design if ρ equals .75.

often do not provide enough information to make an educated guess at the underlying effect size—when there is a large possibility that the true effect size is zero, computing an appropriate replication sample size becomes mathematically impossible, and the concept of power no longer applies. Again, this does not mean that it is useless to attempt to replicate an underpowered study. Understandably, underpowered original studies are often the very studies that investigators most desire to replicate. Instead, we argue, it is especially important in such cases not to put too much weight on a single replication study and to consider the benefits of an approach such as the Registered Replication Report mentioned previously or meta-analysis and the related idea of a continuously cumulative science.

We encourage researchers to attend to the power of the original study and its implications for future replicability. In circumstances under which large sample sizes are difficult to achieve, we remind researchers that there are other ways to increase power. Researchers can consider stronger manipulations (increase effect size; e.g., increasing the length of treatment exposure), more precise measurement instruments (decrease variance; e.g., time participants in seconds, not minutes), and alternative experimental designs (increase power via the design; e.g. within-subjects) to achieve higher statistical power in original studies (Shadish et al., 2002).

Additional sample size planning strategies in the literature

Although we assessed the strategies most common in our literature review, as well as strategies specifically designed to ameliorate the issues of effect size bias and uncertainty, we would like to highlight a few other strategies that have recently been recently recommended or were identified in our more general replication literature review. We identify four variants of a general goal of aiming for "conservatively higher levels of power" (Brandt et al., 2014 p. 220). First, most literally following Brandt and colleagues' (2014) advice, a replicator could aim for. 90 or .95 power as opposed to .80.[11] Second, along similar lines, Simonsohn's 2.5*n rule multiplies the per-cell study sample size by 2.5 (Simonsohn, 2015). Third, an even simpler strategy that has been suggested is to use a total sample size of 250 (Fraley & Vazire, 2014). Finally, yet another approach is to use an arbitrarily smaller effect size than that reported in the original study. However, these approaches will be quite ineffective in some situations. Considering the hypothetical example presented in the introduction to the paper (original study $n = 20$, $d = 0.70$) may shed light on this. Following the first variant, a replicator aiming for .95 power would collect data on 64 participants and have an actual power of .39. A replicator using the 2.5*n rule would collect data on 50 participants per group, resulting in power of .32. A replicator using the $N = 250$ rule would have a power of .66. Finally, a replicator using an arbitrarily smaller effect size (we chose $d = 0.50$, an effect size .20 smaller than the sample effect size, a difference in effect sizes much larger than we saw in our literature review) would collect data on 55 participants per group, resulting in a power of .34. Thus, despite their appealing simplicity, these approaches may result in replication sample suggestions that are too small.

On the other side of the spectrum, these rules may end up requiring arbitrarily large samples, which, as mentioned previously, may be unfeasible in certain subfields. Consistent with this possibility of an unnecessarily large sample size, we note that the 2.5*n rule was actually developed for the task of having sufficient power to find a "just-detectable" effect (i.e. to find an effect so small, one could justify the null hypothesis as true; Simonsohn, 2015), rather than simply to find a significant effect. Overall, these additional strategies suffer from many of the same issues as the methods we tested using the sample size as the original study or increasing that sample size by 25%. We argue that it is most efficient to tackle the issues of uncertainty and bias directly, rather than arbitrarily selecting a higher intended power, larger sample size, or smaller effect size, which may not be an appropriate choice.

Limitations

The magnitude of difference between intended and actual power is not ignorable, even in the ideal situation we have analyzed in this study. Notably, the reported proportions of replicated studies in our simulation are likely on the optimistic side, given that we have not included the possibility that the eventual sample size may be smaller than planned. *A priori* power calculations often result in the sample that the researcher intends to collect, but missing data can occur for a variety of reasons. Further, some argue that it is not possible for a replication to be direct or exact (McShane & Bockenholt, 2014). These authors suggest that between-study variation can lead power analysis methods based on an original study effect size to be too optimistic. Moreover, as we mentioned early in the paper, our simulation did not take p-hacking or "the garden of forking paths" (Gelman & Loken, 2014) into account, factors, which could make conventional replication study sample size planning methods even less reliable. These topics are beyond the scope of the current

[11] "Aim" is the critical word, as this method only changes the intended power to a higher level. We have already seen that actual power may often differ from intended power.

study, but would jointly serve to make the discrepancy between intended and actual power even more extreme than seen in this study. Most importantly, we reiterate that in order to determine whether an original study may reflect p-hacking hidden from readers, it is important to design a replication study so that its actual power can be very close to its intended power. In other words, it is important to first eliminate the possibility that a failure to replicate was due to insufficient power if there is suspicion that statistical significance in the original study occurred only because of p-hacking.

Recommendations for replication researchers

Let us provide some guidance on a question that remains: knowing that power may be more nuanced of an issue than is often assumed, what can be done to improve the state of replication in psychology? First, we advise researchers of replication studies to attend to the discrepancy between actual and intended power, and to design replication studies so as to minimize this discrepancy. All the proposed solutions described in this article (Table 3) generally provide larger power than the more basic methods commonly employed (Table 2), which will aid in more effective interpretation of the success or failure of replication studies.

Because the lower bound method (Perugini et al., 2014) and Taylor and Muller's (1996) (.05) approach provided the highest assurance, we provide a closer comparison of these methods to aid in our recommendations (Table 6 and Figure 4). Specifically, Table 6 displays the percentage of times both methods, neither method, and only one of the methods is successful in reaching the intended power, for both small and very large original study sample sizes and small and large δs. From Table 6, it is clear that the lower bound method adjusts less than Taylor and Muller (.05), leading to more corrected ds just above zero when the Taylor and Muller method produces estimates equal to zero. This results in the lower bound method having both more nonsuccesses and more successes than Taylor and Muller (.05). When both methods provide a nonzero corrected d, Taylor and Muller (.05) tends to be successful more often. When Taylor and Muller (.05) produces corrected d of zero and either δ is small and n is very large or δ is large and n is small, the lower bound method tends to be successful. However, when δ and sample size are small, the lower bound method is unsuccessful in reaching the intended replication study power more often than it is successful in this situation.

Table 6. Percentages of success and nonsuccess in reaching intended power for the lower bound method and Taylor and Muller (.05) for small and large original study sample sizes and population δs.

	$n = 20$			
	$\delta = 0.2$		$\delta = 0.8$	
	LB unsuccessful	LB successful	LB unsuccessful	LB successful
TM unsuccessful	7.1	0	4.9	0
TM successful	2.4	0	2.1	30.7
TM = 0	50.0	40.5	0	62.3
	$n = 250$			
TM unsuccessful	5.0	0	5.2	0.4
TM successful	3.7	28.9	0	94.4
TM = 0	0	62.4	0	0

Note. δ = population value of Cohen's d effect size; n = per-group sample size; LB = lower bound; TM = Taylor and Muller; TM = 0 indicates a corrected Cohen's d of zero.

Figure 4 shows whether each method is successful in reaching or exceeding replication study intended power for a range of δ values based on 350 random draws from a uniform (0.1, 1.0) distribution. We first focus on the y-axis, d. When $n = 250$ (panels b and d), for original study sample d values above 0.3, the two methods perform quite similarly in terms of whether they reach .80 power, regardless of δ (x-axis). For sample d values below 0.3 in the $n = 250$ condition, and for all sample d values in the $n = 20$ condition (panels a and c), the methods sometimes differ in their success rates, depending on the value of δ. When $n = 20$ and δ is below 0.3, the lower bound method more often results in power below the intended level (panel a), whereas Taylor and Muller (.05) often leads to a corrected d of zero (panel c). For δs above 0.3, the lower bound method more consistently reaches the intended power. Although Taylor and Muller (.05) continues to produce some corrected ds of zero, it is still unsuccessful less often than the lower bound method. Turning to the $n = 250$ condition, when δ is between 0.1 and 0.4, the lower bound method is again more consistently successful (panel b), and Taylor and Muller (.05) produces some corrected ds of zero (panel d). For δs above 0.4, the methods perform similarly. Overall, the difference between these two methods is most prominent for original studies whose p-values are just under the threshold for statistical significance. Taylor and Muller (.05) notes the high level of uncertainty and bias, resulting more often in corrected d of zero. The lower bound method instead results in a positive corrected d, which can lead to large actual power if δ is truly nonzero.

Especially when the original study δ is small, any of the proposed solutions are likely to offer an advantage as compared to the more common methods. However, as Figure 4 and other results show, no approach is uniformly best. Furthermore, researchers will never know δ, the piece of information that can distinguish among the effectiveness of the proposed solutions (and which would

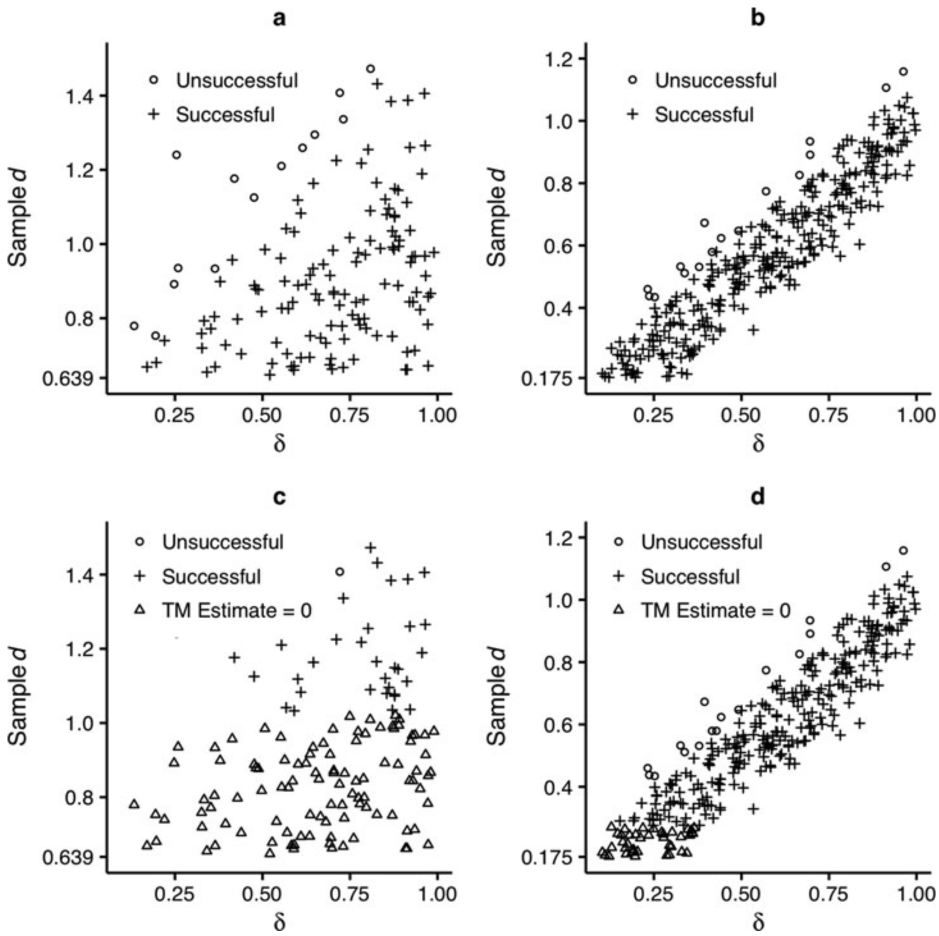

Figure 4. Comparing lower bound method (a and b) and Taylor and Muller (.05) (c and d) on whether or not the replications were successful in reaching the intended power, for small (a and c; $n = 20$) and very large (b and d; $n = 250$) original study per-group sample sizes, with population effect sizes δ drawn from uniform distribution (0.1, 1.0). TM = Taylor and Muller.

eliminate the need for these strategies). Researchers may need to take their specific circumstances into consideration when selecting a method. For example, the lower bound method and Taylor and Muller (.05) often result in the largest actual power, but in situations where collecting additional participants is costly, researchers may elect to use a method with a lower assurance. Overall, Taylor and Muller's (.05) method is safer: when it does provide a positive corrected *d*, it will suggest larger replication sample sizes, which will result in larger replication study actual power.

One option we suggest is to perform both the lower bound method and Taylor and Muller's (.05) method. When both methods suggest similar replication sample sizes, both are likely to work well. However, when the methods provide drastically different sample sizes, two explanations are possible. First, publication bias may be a factor, and only Taylor and Muller (.05) corrects for bias. Second, the amount of uncertainty could be so large that the corrected effect size estimates for both methods hover close to zero, where, as we previously mentioned, small differences in the effect size estimate could lead to drastic differences in the suggested sample sizes. In this situation, the suggested sample sizes may be quite large: again, favoring several replications and meta-analysis.

Conclusion

It is reassuring that all the proposed solutions are an improvement over the more commonly used methods in terms of power and assurance. We have seen that understanding that sample effect sizes are hindered in their accuracy by both publication bias and uncertainty, and correcting for those issues appropriately when conducting sample size planning for a replication study, can yield replication studies that have actual power much closer to the intended level. The proposed solutions presented in this article emphasize thinking beyond only power: via assurance, researchers can determine how consistently they would like their sample size planning approach to meet or exceed intended power. Yet, we have seen that none of these proposed solutions are a panacea, especially when the original study is underpowered. Thus, we urge researchers of original studies to heed to one of the many benefits of power: a better chance at later replication.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Lijuan Wang, Joe Rodgers, and Steve West for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred. This article is based in part on the second author's presidential address to SMEP in Vancouver, BC in 2012.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*, 556–578. doi:10.1177/1745691614545653

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*, 1–12. doi:10.1037/met0000051

Bakker, M., vanDijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

Batterham, A. M, & Hopkins, W. G. (2013). Emergence of large treatment effects from small trials. *JAMA*, *309*, 768–769. doi:10.1001/jama.2012.208828

Brand, A., Bradley, M. T., Best, L. A., & Stoica, G. (2008). Accuracy of effect size estimates from published psychological research. *Perceptual and Motor Skills*, *106*, 645–649. doi:10.2466/pms.106.2.645-649

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R … Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. doi:10.1016/j.esp.2013.10.005

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munaf, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 1–12. doi:10.1038/nrn3475

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750. doi:10.1002/pam.20375

Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, *10*, 311–317. doi:10.1002/pst.467

Dawes, R. M. (2004). Commentary on Meehl's theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Applied and Preventive Psychology*, *11*, 23–25. doi:10.1016/j.appsy.2004.02.002

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS One*, *9*, e109019. doi:10.1371/journal.pone.0109019

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, *351*, 1037–b. doi:10.1126/science.aad7243

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651. doi:10.10177/1745691614551642

Gelman, A., & Loken, E. (2014) The statistical crisis in science. *American Scientist*, *102*, 460–465. doi:10.1511/2014.111.460

Hedges, L. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, *9*, 61–85.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124. doi/10.1371/journal.pmed.0020124

Ioannidis, J. A, Pereira, T. V, & Horwitz, R. I. (2013). Emergence of large treatment effects from small Trials—Reply. *JAMA*, *309*, 768–769. doi:10.1001/jama.2012.208831

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*, 348–357.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again… *Science*, *334*, 1225. doi:10.1126/science.334.6060.1225

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217. doi:10.1207/s15327957pspr0203_4

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112.

Liu, X. S. (2014). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. New York: Taylor and Francis.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology*, *20*, 76–78. doi:10.1037/0278-6133.20.1.76

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. doi:10.1146/annurev.psych.59.103006.093735

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does failure to replicate really mean? *American Psychologist*, *70*, 487–498. doi:10.1037/a0039400

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

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

Open Science Framework. (2015). *Reproducibility project: Researcher guide*. Retrieved from https://osf.io/ru689/

Pereira, T. V., Horwitz, R. I., & Ioannidis, J. P. A. (2012). Empirical evaluation of very large treatment effects of medical interventions. *JAMA*, *308*, 1676–1684. doi:10.1001/jama.2012.13444

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

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160–164. doi:10.1037/0033-2909.112.1.160

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.

Richard, F., Bond, C., & Stokes-Zoota, J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363 doi:10.1037/1089-2680.7.4.331

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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. doi:10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. doi:10.1177/1745691613514755

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. doi:10.1177/0956797614567341

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318. doi:10.1177/1745691614528518

Talboom, J. S., West, S. G., & Bimonte-Nelson, H. A. (2015). A primer of methods in biobehavioral research: Improving a study's design, analysis, and write up. In H. A. Bimonte-Nelson (Ed.), *The maze book: Theories, practice, and protocols for testing rodent cognition*. New York, NY: Springer.

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics: Theory and Methods*, *25*, 1595–1610. doi:10.1080/03610929608831787

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, *15*, 18–37. doi:10.1037/a0015917