General Article

Psychological Science
2017, Vol. 28(11) 1547–1562
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797617723724
www.psychologicalscience.org/PS
SAGE

Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty

Samantha F. Anderson, Ken Kelley, and Scott E. Maxwell
University of Notre Dame

Abstract

The sample size necessary to obtain a desired level of statistical power depends in part on the population value of the effect size, which is, by definition, unknown. A common approach to sample-size planning uses the sample effect size from a prior study as an estimate of the population value of the effect to be detected in the future study. Although this strategy is intuitively appealing, effect-size estimates, taken at face value, are typically not accurate estimates of the population effect size because of publication bias and uncertainty. We show that the use of this approach often results in underpowered studies, sometimes to an alarming degree. We present an alternative approach that adjusts sample effect sizes for bias and uncertainty, and we demonstrate its effectiveness for several experimental designs. Furthermore, we discuss an open-source R package, BUCSS, and user-friendly Web applications that we have made available to researchers so that they can easily implement our suggested methods.

Keywords

statistical power, sample size, effect size, publication bias, methodology

Received 12/14/16; Revision accepted 7/11/17

Readers of the Association for Psychological Science's (APS's) journals have seen serious criticisms leveled at the field of psychological science lately. For example, authors have stated that "the typical study in psychology appears to be underpowered" (Bakker, van Dijk, & Wicherts, 2012, p. 544), that "the average power of typical psychological research is estimated to be embarrassingly low" (Perugini, Gallucci, & Constantini, 2014, p. 319), and, perhaps worst of all, that "psychology is suffering from a 'crisis of confidence'" (Pashler & Wagenmakers, 2012, p. 528; this latter comment referred to the low rate of successful replication, for which the prevalence of underpowered studies appears to be a contributing factor). Given all of this, there has never been a more important time to carefully consider how studies can be properly planned so as to have appropriate statistical power (the probability of rejecting the null hypothesis of no effect when the true effect is nonnull in the population—which we henceforth refer to simply as *power*).[1] The goal of the present article is to provide guidance on effective strategies for sample-size planning for sufficient power in common experimental designs used in psychology and related fields.

When the goal of a study is to determine the existence or direction of an effect, the appropriate sample size for a given level of power depends on the size of the effect. Given that the true (population) effect size is unknown, there are two general types of approaches to using effect size as a basis for planning sample size. In one type of approach, the researcher determines a sample size that will provide desired power for detecting a minimally important effect,[2] regardless of the true value of the effect size. In the other type of approach, the researcher attempts to estimate how large the true effect is likely to be and to calculate a corresponding

Corresponding Author:
Samantha F. Anderson, Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556
E-mail: sander10@nd.edu

sample size based on that estimate. The estimated effect size can be found in a few ways, but one option is to use a sample effect size from a prior published or pilot study. We focus on this specific approach for three reasons. First, it is a popular method for theory testing. In fact, this was the most commonly used approach to sample-size planning reported in two recent issues of *Psychological Science*. Second, we show that although this method is intuitively appealing, it can be severely flawed if sample effect sizes are taken at face value. The desired level of power will be achieved only if the sample effect size is an accurate estimate of the population value. Later in this article, we illustrate that intuitions regarding the relationship between an effect-size estimate and the true value may sometimes be surprisingly inaccurate, especially when the estimate comes from an underpowered study. In some situations, even small differences between the true and estimated values can have a dramatic impact on power when these estimates are used in sample-size planning. Third, we have developed an adjustment that yields a more appropriate value of the effect-size estimate from a prior study, which can lead to more accurate sample-size planning. We evaluate this proposed method using Monte Carlo simulations, show that it overcomes the shortcomings of the widely used approach, and discuss an R package (BUCSS; Anderson & Kelley, 2017) and user-friendly Web applications that we have made freely available to allow researchers to use our suggestions in their own sample-size planning.

Statistical Power in Psychological Studies

The topic of power in psychology is not new, but has become a mainstream point of discussion in recent years. In fact, a PsycINFO search of titles that included "power analysis" or "sample size" in the decade from 1990 through 1999 yielded 239 results, whereas the same search criteria yielded 583 results from 2006 through 2015. However, it is not obvious that this increase in writing on the topic has led to a corresponding increase in power in the published literature. Consider that in 1962, Cohen reported that across the experiments reported in 70 articles from the *Journal of Abnormal and Social Psychology*, the average power to detect a medium-size effect was .48 (although the criteria for what constitutes a medium-effect size were lowered following this work). This eye-opening review catalyzed an entire subdiscipline devoted to formal power analysis. Twenty-seven years later, Sedlmeier and Gigerenzer (1989) published a report on the change in power following Cohen's and other researchers' writings on this subject. The findings were paradoxical: Twenty-five years of articles emphasizing the importance of power had not resulted in an improvement of the power of experiments conducted in the field of psychology.

Unfortunately, research indicates that the situation may not be much better today. Despite the abundance of articles on the topic, navigating the complex literature to determine the appropriate sample size for sufficient power in a study can be daunting and confusing. Further, journals are just beginning to require formal justification for the sample sizes of experiments reported in submitted manuscripts.

In a recent survey, Bakker et al. (2012) posited that average power in psychological experiments is .35, assuming typical effect sizes and sample sizes. Other surveys indicate that the average may vary by subject area (see, e.g., Fraley & Vazire, 2014, and Button et al., 2013, for literature reviews of social psychology and neuroscience, respectively). Further, in a comprehensive review of studies assessing average power in psychology, Smaldino and McElreath (2016) demonstrated that power to detect a small-size effect has shown "no sign of increase over six decades" (p. 5; $R^2 = .00097$). Correspondingly, sample sizes for experimental research are often small. For example, Marszalek, Barber, Kohlhart, and Holmes (2011) found a mean total sample size, N, of 40 in a representative survey of four top-tier psychology journals, and our analysis of data[3] from a review published in the same year (Wetzels et al., 2011) revealed that the median per-group sample size, n, was 24 for independent-samples t tests and 23 for dependent-samples t tests. Moreover, when we surveyed four *Psychological Science* issues from 2016 (March–June), we found a median n of 26 for in-person studies (and 250 for online studies).

Consequences of Low Power

Low power has some nonintuitive consequences, which go beyond low probability of detecting the effect of interest when it exists. Because published articles are almost exclusively restricted to studies with statistically significant results, high power can certainly benefit the individual researcher. However, power also has implications for the field at large and the trust that can be placed in published studies.

First, when studies are underpowered, the false-discovery rate (i.e., the proportion of studies falsely finding an effect) is increased. Ioannidis (2005) showed that, all other things being equal, there is a higher probability that a published research claim is false in a literature with lower-powered studies. Second, even underpowered studies that do find real effects tend to yield inflated effect-size estimates (Lane & Dunlap, 1978; Maxwell, Lau, & Howard, 2015), a problem that we describe in more detail later. Third, underpowered studies decrease the likelihood that a researcher will be able to successfully replicate the original results. Reflected by the Open Science Collaboration's (2015) article on the low rate of reproducibility in psychology, mistrust in published findings has become an increasingly serious concern. The replication crisis in psychology is multiply determined, and factors such as multiple testing and *p*-hacking likely play a role (see Simmons, Nelson, & Simonsohn, 2011, and John, Loewenstein, & Prelec, 2012, for descriptions of questionable research practices, or QRPs). However, even when researchers conduct a well-intentioned power analysis to determine the sample size of a replication study, the power of the original study limits their ability to determine an accurate sample size, sometimes severely (Anderson & Maxwell, 2016, 2017; Button et al., 2013).

Current Practices

Despite the growing emphasis on achieving adequate power, formal power analysis is still not typical in many areas within psychology. In fact, Bakker et al. (2012) noted that power was referred to "as a rationale for the choice of sample size or design" (p. 544) in only 11% of a recent sample of psychological publications reporting results of null-hypothesis significance testing. A more recent survey of psychology researchers found that, although almost half mentioned power analysis when they explained how they decided on sample sizes, practical considerations were sometimes valued more than a rigorous approach to sample-size planning even within this group. Moreover, the half who did not mention power analysis used strategies involving "intuition, rules of thumb, and prior practice" (Bakker, Hartgerink, Wicherts, & van der Maas, 2016, p. 1069), in addition to accommodating practical constraints.

Our own focused analysis of common strategies for sample-size planning reported in two recent issues of *Psychological Science* (April and June 2016) also revealed a variety of justifications for sample size. Details about sample-size planning were provided in 16 of the 29 articles, but in almost half of these cases (7 out of 16), there was no mention of effect size. Six articles reported that the sample sizes used were comparable to those found in published studies on a similar topic, and the authors of one article relied on practical considerations. Unfortunately, when researchers use these ad hoc strategies, they may be unsuccessful in reaching the level of power that they believe they are achieving (*intended* power; e.g., .80), as the chosen sample sizes are not based on the size of the effects in question. Strategies based on prior sample size alone ("rules of thumb") can result in selecting sample sizes that are either too large or too small, depending on the power of the prior studies involved (Green, 1991; see Anderson & Maxwell, 2017, for a simulation). An overestimate of the required sample size may not always be an advantage either, especially in fields where participants are costly or belong to specialized subpopulations that are difficult to recruit, and where oversampling can be considered an ethical issue (Maxwell & Kelley, 2011).

Effect size was taken into account by the authors of 10 of the 16 articles that reported details about sample-size planning.[4] In 2 of these cases, the targeted sample size was based on an assumption of a medium-size effect in the population. In the other 8 cases, the authors calculated the suggested sample size for an intended level of power (e.g., with G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) using a sample effect size from a prior published study (5 studies) or a pilot study (3 studies) as a substitute for the population effect size. This was the most common strategy we found in our review. Despite *Psychological Science*'s assertion that "it is typically not appropriate to base sample size solely on the sample sizes and/or effect sizes reported in prior research or on the results of small pilot studies" (APS, 2016), "there has been a venerable tradition of using pilot studies to estimate . . . effect sizes that, in turn, are used to inform the design of subsequent larger scale hypothesis testing studies" (Leon, Davis, & Kraemer, 2011, p. 628). Although using an estimated effect size to plan the sample size of a future study seems logical because the sample effect size does estimate the population value, researchers who use this approach typically design studies with less power than intended, and thus reduce their probability of obtaining statistically significant results.

Specifically, there are two main problems with this "effect size at face value" approach as currently applied. First, sample effect sizes reported in published research are upwardly biased, because of journals' preference for statistically significant findings[5] (Brand, Bradley, Best, & Stoica, 2008; Lane & Dunlap, 1978; Maxwell et al., 2015) and the prevalence of multiple testing. Moreover, the nonlinear relationship between effect size and power means that an overestimated effect size will decrease power more than an equivalent underestimate increases power (Maxwell et al., 2015).

To understand the impact of publication bias, suppose that it is Tuesday morning and a researcher eagerly opens a new article via the weekly "This Week in *Psychological Science*" e-mail. The article reports an interesting result with an effect size of 0.6, as measured by Cohen's *d*, based on an independent-samples *t* test with 25 participants per group. The astute researcher may wonder what population effect size (δ) is most

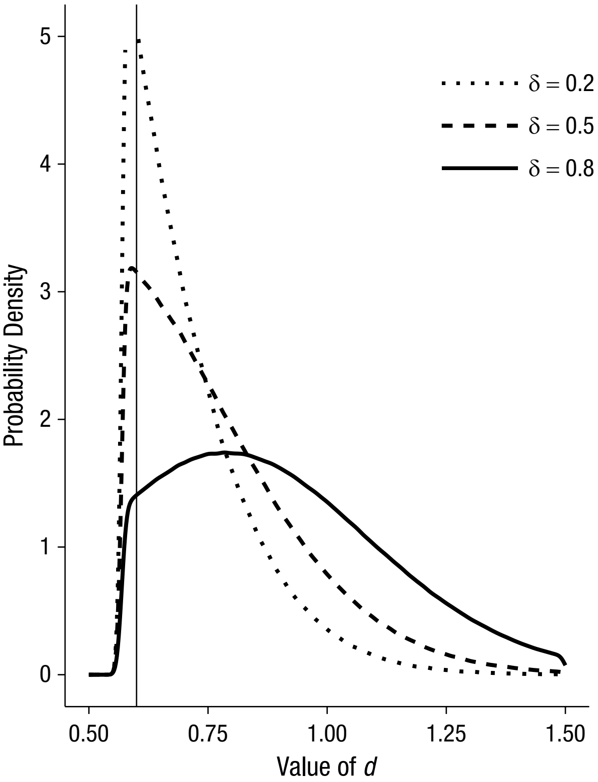

Fig. 1. Distribution of Cohen's *d* for experiments with 25 participants per group. Separate distributions are shown for three values of the population effect size, δ: 0.2, 0.5, and 0.8. Values of *d* less than 0.50 are not shown because with 25 participants per group, *d* values below 0.57 will not be seen in a journal that publishes only experiments resulting in *p* values less than .05, regardless of the value of δ. The vertical line designates the hypothetical sample *d* of 0.6.

consistent with this sample *d* of 0.6.[6] Consider three alternatives: δ = 0.2, δ = 0.5, and δ = 0.8, which correspond to Cohen's small, medium, and large effects, respectively. Intuition would probably suggest that the observed sample *d* should be a good estimate (or at a minimum, the only available estimate) of the unknown population δ, so that a reasonable estimate for δ would be 0.6 or, among the three alternatives, 0.5. However, publication bias makes this intuition misguided. Figure 1 depicts the distribution of *d* for these three δ values, given 25 participants per group, in the presence of publication bias.[7] We note that the *x*-axis does not go below 0.50. In a journal that publishes only experiments resulting in *p* values less than .05, the minimum sample *d* reported among studies with *n* of 25 is 0.57, regardless of how large or small the effect is in the population. Publication bias will thus prevent any published study with this sample size from producing an accurate estimate of the population effect unless δ is 0.57 or larger. Moreover, unless δ is much larger than 0.57, publication bias will still cause the sample *d* to systematically overestimate δ. A published sample *d* of 0.6 (vertical line in Fig. 1) is not most consistent with a δ of 0.5, but is most consistent with a δ of 0.2. In fact, the most likely value of δ for this sample *d* is 0.16 (Hedges, 1984).

The second main problem with the current "effect size at face value" approach is that even if publication bias is not a concern,[8] sample effect sizes are only estimates of their population counterparts. Thus, there is uncertainty in these estimates. This uncertainty is reflected in the width of a confidence interval for the population value, but is not reflected when a researcher uses only the point estimate for sample-size planning (e.g., Dallow & Fina, 2011; Taylor & Muller, 1996).[9] Consequently, to make a sample effect size more effective as the basis for planning a future study for a desired power, the effect-size estimate should be adjusted for uncertainty in addition to publication bias.

To understand the joint impact of publication bias and uncertainty, suppose a second researcher reads a different article, which reports a *d* of 0.8, also for an independent-samples *t* test with 25 participants per group. What is the most likely value of δ in this case? Figure 2 illustrates the possibilities from a different perspective than Figure 1, with values of δ rather than *d* on the *x*-axis. The graph shows the likelihood of various values of δ between zero and 1.0 for *n*s of 25 and 100, given a sample *d* of 0.8 and the presence of publication bias. When *n* is 25, the most likely δ for this study is not 0.8, but rather 0.56, much lower than the reported sample value. What is even more interesting is the relative flatness of the curve for this *n*: A large effect size reported in the published literature when *n* is 25 tells readers relatively little about the true effect size. In fact, δ is almost as likely to be zero as 0.8. The curve for an *n* of 100 highlights the fact that larger sample sizes are necessary to obtain a more precise estimate of δ.[10]

However, going beyond the two fundamental problems of publication bias and uncertainty, the process of basing the appropriate sample size on a sample effect size is not straightforward for two additional reasons. First, effect-size measures commonly used in analysis of variance (ANOVA) are inconsistently defined. For example, even in a case as simple as a dependent-samples *t* test, *d* can be defined in at least three ways (see Lakens, 2013, for a primer on these various representations of *d*), and the effect-size measure f^2 for ANOVA can also be defined in several distinct ways. However, published studies rarely specify the particular definition or formula used for the effect size presented (e.g., Kelley & Preacher, 2012), so researchers must hope that the software they use for their power analysis uses a definition of effect size that matches the published study's formulation. Second, the effect-size measures f^2 and f, which are needed for the popular

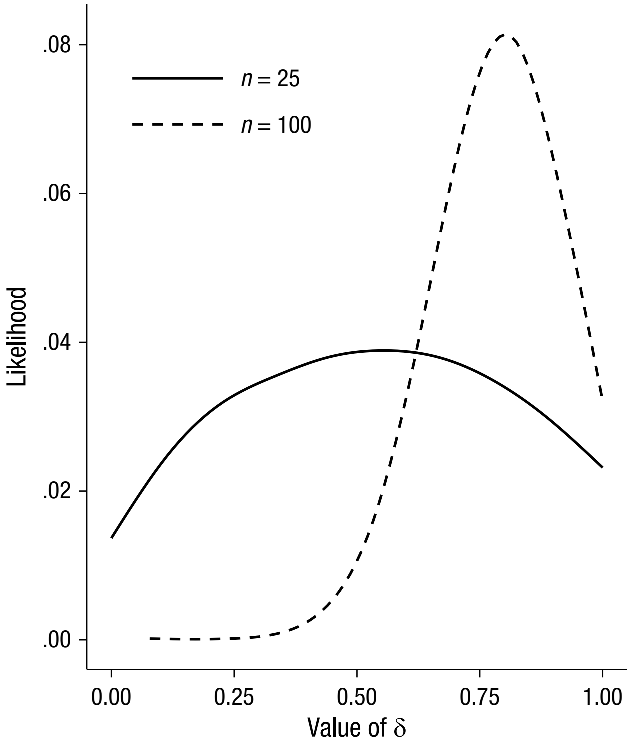

Fig. 2. Relative likelihood of various values of the population effect size, δ, given a sample Cohen's *d* of 0.8 and the presence of publication bias. Separate functions are shown for studies with 25 and 100 participants per group.

power-analysis software G*Power (Faul et al., 2007), and the related effect-size measure $\hat{\eta}^2$ can be particularly problematic to use for sample-size planning, as the sample estimators result in upwardly biased estimates of the true population values (Olejnik & Algina, 2000; Skidmore & Thompson, 2013). This bias is due to the formulas for the estimators themselves and is distinct from publication bias.

These factors have caused methodologists to bemoan the difficulties in accurately estimating effect sizes and thus planning appropriate sample sizes. Lipsey (1990) aptly referred to effect size as the "problematic parameter" (p. 47) in sample-size planning, most importantly because it is "unknown and difficult to guess" (p. 47). Similarly, Gelman and Carlin (2014) stated that "the largest challenge in a design calculation [is] coming up with reasonable estimates of plausible effect sizes based on external information" (p. 641). Thankfully, sample effect-size measures are only an intermediate step in sample-size planning, and are not necessary for determining power, as we show in the next section.

What Determines Power

To calculate power in ANOVA, one has to consider the noncentral *F* distribution.[11] This distribution is defined by appropriate numerator and denominator degrees of freedom and the *noncentrality parameter*, which describes how far the center of the (alternative) distribution shifts away from the distribution under the null hypothesis.[12] A general expression for the noncentrality parameter (λ) in an ANOVA with equal *n*s in all cells is

$$\lambda = n\xi, \tag{1}$$

where the general population effect size ξ is given (in matrix form) by

$$\xi = \frac{(\mathbf{C}\boldsymbol{\mu})'\,(\mathbf{CC}')^{-1}(\mathbf{C}\boldsymbol{\mu})}{\sigma^2}, \tag{2}$$

where $\boldsymbol{\mu}$ is a vector of population means, $\mathbf{C}$ is a contrast matrix (or vector) reflecting the type of effect of interest (e.g., main effect, interaction, contrast), and σ^2 is the population within-group variance (Muller & Fetterman, 2002, pp. 450–451). The parameter ξ reflects the magnitude of the population mean differences to be tested relative to the population variance. For example, in an independent-samples *t* test, ξ reduces to $\delta^2/2$. Notably, Equation 1 shows that the noncentrality parameter combines effect size and sample size.

Central to our approach to sample-size planning is the utility of separating sample size from effect size, which is evident in Equation 1. One can isolate *n*, holding ξ constant at any value. We emphasize a simple relationship between sample size and the noncentrality parameter: Doubling *n* doubles λ, halving *n* halves λ, and so forth.[13] More generally, one can choose a multiplicative factor for *n* that produces a λ that corresponds to one's intended power. An immediate complication is that power depends on λ, whose value one cannot know in practice because it is a population parameter. We now turn our attention to how the results from a published study can be used to more accurately estimate λ.

Proposed Method: A Better Approach to Sample-Size Planning

As discussed, when researchers conduct a power analysis, they often base their calculations on the sample effect size obtained in a study on their topic of interest. The method we propose here allows researchers to use such an effect-size estimate in a way that allows their power analysis to more accurately reflect the population effect size. In this section, we describe this method. The underlying logic, and more methodological detail, is provided in the Supplemental Material available online (see Conceptual Logic of Taylor and Muller's Method).

Taylor and Muller (1996) proposed a likelihood-based procedure to adjust sample effect-size estimates for varying degrees of publication bias and uncertainty.

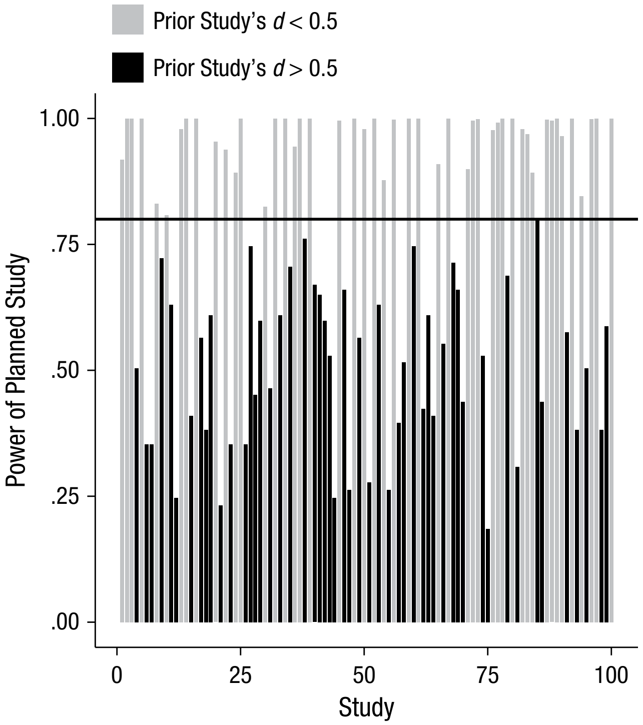

Fig. 3. Plot illustrating 50% assurance with .80 power. The plot shows the actual power achieved in 100 future studies planned from 100 hypothetical prior studies, given an underlying population effect size, δ, of 0.5. Because no publication bias is assumed, half of the sample *d*s from the hypothetical prior studies fall below the population δ, and the corresponding future studies exceed .80 power (the horizontal bar). The other half of the sample *d*s fall above δ and lead to future studies with less than .80 power.

Specifically, a new, adjusted estimate of the noncentrality parameter, $\hat{\lambda}_A$,[14] which will be used in determining sample size, is derived from a truncated noncentral *F* distribution, in which the truncation reflects publication bias. The likelihood function, L_T, of this truncated distribution is given by

$$L_T[F_P; \nu_{num}, \nu_{den}, \lambda] = \frac{L_F[F_P \mid \nu_{num}, \nu_{den}, \lambda]}{1 - CDF_F[F_{crit}(1-\alpha_P) \mid \nu_{num}, \nu_{den}, \lambda]}. \quad (3)$$

The numerator is the likelihood of obtaining the *F* value observed in the prior study, F_P, given the noncentrality parameter, λ, and the degrees of freedom in the numerator and denominator, ν_{num} and ν_{den}, respectively, in the absence of publication bias. The denominator can be thought of as the power of the test given the specified λ, ν_{num}, and ν_{den}; α_P is the α level that was required in order for the published study to be eligible for publication, F_{crit} is the critical value at the (1 − α)100th percentile of the distribution, and CDF_F is the cumulative distribution function of the noncentral *F* distribution. The denominator serves to truncate what would otherwise be a standard *F* likelihood, as, given our model of publication bias, nonsignificant results are censored from the distribution.

The method produces a likelihood distribution for the value of $\hat{\lambda}_A$, given the observed *F* value from a prior study, under an assumption that publication bias may be operative. Note that the researcher does not have to make any assumptions regarding the number of unpublished studies (the "file drawer") in order to adequately adjust for publication bias. The only information needed is the α level required in order for the study to be accessible to the reader (usually .05). The value of α_P can be specified as greater than .05 when the degree of publication bias is thought to be lower. If publication bias is thought not to apply, α_P can be set equal to 1.

A benefit of obtaining a distribution of values is the option of deciding how much (if at all) to accommodate uncertainty in the estimate of the noncentrality parameter. For example, choosing the value of $\hat{\lambda}_A$ associated with the median (50th percentile) of the likelihood distribution results in a median-unbiased estimator of the population noncentrality parameter, that is, an estimator adjusted for publication bias only. Researchers can also adjust for both uncertainty and publication bias by selecting the $\hat{\lambda}_A$ associated with a more conservative (lower) quantile of the likelihood distribution. Taylor and Muller (1996) recommended using the lower limit of a one-sided 95% confidence interval of the noncentrality parameter, or the 5th percentile of the distribution.

The researcher's choice of quantile for the uncertainty adjustment is directly related to the *assurance* the researcher desires to reach: quantile = 1 − desired assurance. Assurance is the percentage of times power would reach or exceed the intended level if the sample-size planning process was to be reproduced many times. Taylor and Muller's (1996) method is designed to reach 50% and 95% assurance when one uses the 50th and 5th percentiles, respectively.

Figure 3 makes the concept of assurance more concrete. It shows the power of a future study based on the *d* observed in each of 100 hypothetical prior studies (n = 25) and a population effect size, δ, of 0.5, in a world where all studies get published, so as to isolate the role of uncertainty. The unadjusted effect size from each of these individual studies forms the basis for the sample size of a future study. Because we know the value of δ, we know that 64 participants per group are needed for the future study to have .80 power (the horizontal line in Fig. 3). Because of uncertainty alone, approximately one half of the 100 sample estimates will happen to fall above the population value (too large), and the other half will fall below (too small). When the studies obtaining a *d* value less than 0.5 are used in sample-size planning, they will suggest sample sizes larger than the necessary 64 per group. The resulting future studies will have power at or above .80. However, when the studies obtaining sample *d* values greater than 0.5 are used to plan future sample sizes, the suggested sample sizes will be smaller than 64 per group. These future studies will have power lower than .80. Because half of the future studies reach or exceed the intended power, Figure 3 represents a situation with 50% assurance.

Now that we have described the properties of our proposed method, we provide an example of an investigator using the approach in practice. Recall our hypothetical researcher who read an article in which an independent-samples *t* test (with 25 participants per group) indicated that the sample *d* was 0.8. Using Taylor and Muller's (1996) 50th-percentile method to adjust the sample estimate of 0.8 for publication bias results in a new (adjusted) δ estimate, $\hat{\delta}_A$, of 0.64. This value of $\hat{\delta}_A$ can then be used for subsequent sample-size planning, in place of the face value of *d* (i.e., 0.8).

Alternatively, a lower percentile could be chosen so as to adjust for uncertainty in addition to publication bias (lower percentiles lead to more adjustment and thus higher assurance). For the example in the previous paragraph, using the more conservative 5th percentile, recommended by Taylor and Muller (1996), results in an uncertainty-adjusted $\hat{\delta}_A$ of zero. This highlights an important consideration with this method: When a researcher takes a conservative approach to adjusting a published effect-size estimate for uncertainty and bias, the resulting estimate may suggest that the true population effect could be zero. This does not indicate a problem with the method, but rather is a signal that the prior study's effect size is so uncertain and biased that the population effect size cannot be accurately estimated at the desired level of assurance (e.g., Gelman & Carlin, 2014; Yuan & Maxwell, 2005).

Practically speaking, if a zero estimate arises for a specified level of assurance, a researcher may choose to decrease the intended assurance (generally a minimum of 50%), to increase the α level required for the prior study to be published (α_P), or both. The benefit of doing either or both is that sample-size planning can proceed with a nonzero estimate with levels of assurance and adjustment for publication bias that are still known and planned for ahead of time. Nevertheless, a zero estimate is a signal that there is not as much information in the observed estimate as would be desirable. In practice, we encourage researchers to adjust assurance downward or α_P upward (or both) as minimally as possible, so as to achieve a nonzero estimate with assurance and bias adjustment as close to the desired levels as possible.

Comparison of Approaches to Sample-Size Planning: A Simulation Study

To thoroughly evaluate the performance of the proposed method and compare it with a typically used procedure for sample-size planning, we conducted a Monte Carlo simulation study. Readers can likely recall many strategies for determining sample size. However, several reasons led us to compare our suggested method with using a sample effect size at face value in this simulation study. First, using a sample effect size directly is common among researchers reporting sample-size justifications, even when their studies make novel contributions and are not simply replications. Second, this approach gives the impression of being effective, objective, and informed by the literature, as it requires formal power-analysis software and an empirical basis for estimating the magnitude of the effect. Third, as we mentioned earlier, other practices that were common in the *Psychological Science* articles we reviewed have been shown to provide overestimates or underestimates of the necessary sample sizes for future studies (e.g., rules of thumb; Anderson & Maxwell, 2017). Consequently, it is most efficient to compare approaches aimed at providing an intended level of power. Nevertheless, we discuss several other approaches to sample-size planning later (see Alternative Approaches to Sample-Size Planning).

Our simulation tested the effectiveness of four strategies (all with intended power of .80): (a) using the face value of the sample effect size obtained in a previously published study, (b) using Taylor and Muller's (1996) method at the 50th percentile, (c) using Taylor and Muller's method at the 20th percentile,[15] and (d) using Taylor and Muller's method at the 5th percentile. We generated data consistent with a variety of effects in experimental designs: the mean difference in a two-independent-group design (independent-samples *t* test), the paired mean difference in a two-level repeated measures design (dependent-samples *t* test), the three-level main effect in a 3 × 2 between-subjects ANOVA design, and the interaction effect in a 3 × 4 split-plot design (mixed-model ANOVA). We assumed that the researcher was basing sample-size calculations for a future study on the effect size reported for a previously published study with 25 participants per group, and that only studies yielding results with *p* values less than .05 were published. For each design, we assessed the performance of the four methods for three different magnitudes of the population noncentrality parameter,

Table 1. Results of the Monte Carlo Simulations for a 3 × 4 Split-Plot Analysis of Variance

	Approach to sample-size planning			
		Taylor and Muller's (1996) method		
Population ES	Taking the sample ES at face value	50th-percentile version	20th-percentile version	5th-percentile version
		Actual power		
Small	.139	.616	.771	**.825**
Medium	.529	.792	**.911**	**.972**
Large	.641	**.818**	**.902**	**.958**
		Median suggested sample size		
Small	17	84	126	147
Medium	11	20	36	76
Large	6	8	11	14
		Achieved assurance		
Small	0.0%	38.3%	54.1%	61.3%
Medium	4.0%	54.3%	82.6%	95.7%
Large	15.2%	66.1%	89.2%	98.0%

Note: The simulations were based on a prior study with 25 participants per group. The intended power was .80, and values that exceeded this level are highlighted in boldface. The values used for small, medium, and large effect sizes were based on Cohen's (1988) guidelines. ES = effect size.

based on Cohen's small, medium, and large effect sizes.[16] Specifically, we calculated the mean actual power, the median suggested sample size, and the assurance achieved in the future study. (Details regarding the simulation procedure are presented in the Supplemental Material—see Full Simulation Procedure and Results—and simulation code is available from the first author upon request.)

For the sake of brevity, we present results only for the interaction in a 3 × 4 split-plot ANOVA (Table 1). Results for the other designs showed similar patterns and can be found in the Supplemental Material (see Full Simulation Procedure and Results). Table 1 shows that the average power was lowest when the face value of the sample effect size was used to determine sample size for a new study. In fact, average power using this technique was less than .15 for a small population effect size, a value distressingly below the intended .80. Taylor and Muller's (1996) method was more powerful, with the 5th-percentile version providing the highest power, as is to be expected because it provides maximum allowance for uncertainty by choosing a smaller estimate of the noncentrality parameter than either the 20th- or the 50th-percentile versions. With larger population effect sizes, the methods performed more similarly to one another, though Taylor and Muller's method consistently outperformed using the sample effect size at face value.

Table 1 also shows the median sample sizes suggested by each method. At first, it may seem preferable to opt for using the sample effect size at face value in determining one's sample size, given that the sample sizes for this method seem to be the most reasonable and obtainable. However, recall that these sample sizes are adequate to detect only an effect of the published effect size, which we have emphasized can be quite inaccurate (e.g., for a small population f^2 of 0.01, the median "published" f^2 in our simulation was 0.07, larger than Cohen's criterion for a medium effect size). Consequently, these sample sizes provide the power that was actually obtained with each method (i.e., as low as .139 in our simulations; see Table 1), rather than the .80 benchmark. The suggested sample sizes from Taylor and Muller's (1996) method are larger, increasing with the conservativeness of the chosen percentile. Critically, however, the more effective the method is at ensuring .80 power, the larger the sample size required. Such larger sample sizes may sometimes necessitate collaboration among multiple investigators.

Finally, Table 1 also presents the assurance achieved by each method. For a small population effect size, the assurance from using the sample effect size at face value was zero. That is, in 10,000 replications, this method never resulted in the intended .80 power. Taylor and Muller's (1996) method is designed to reach 50%, 80%, and 95% assurance when the median, 20th percentile, and 5th percentile, respectively, are used. When the population effect size was medium or large, these goals were met or exceeded. In fact, the method worked as intended in all cases, despite the fact that the desired assurances (and adjustments for publication bias) were lowered when $\hat{\lambda}_A$ was zero.

It is clear from the simulations that actual power achieved using the common strategy of taking the sample effect size at face value can be far removed from the intended power. To be clear, this poor performance was due to the uncertainty and bias in sample effect sizes; given this uncertainty and bias, the face value of a reported sample effect size is not generally an accurate representation of the population effect size. If uncertainty and publication bias were not present, this method would be effective. However, given the ubiquity of bias and uncertainty in estimates of effect size, researchers who conscientiously plan their sample sizes using published effect sizes from prior studies can have actual power that is abysmal, especially when the population effect size is small. This is true when the prior study has an *n* of 25, which, as we have noted, is quite typical for experimental studies. Further, this finding is consistent across several common research designs (see Full Simulation Procedure and Results in the Supplemental Material). Taylor and Muller's (1996) method performs rather well in comparison, particularly in the case of the 20th- and 5th-percentile versions, which adjust for both bias and uncertainty. Of course, Taylor and Muller's method is not a panacea when a prior study both assessed a small effect size and used a small sample size.

However, we note that studies with small sample sizes are not necessarily uninformative: Factors such as the population effect size, the correlation between repeated measures, and the study design contribute to power as well. To illustrate this point, we performed another simulation comparing the four strategies for sample-size planning when the prior study involved a dependent-samples *t* test with 15 participants. We assumed a very large population effect size of 0.8 and a large correlation between pairs, $\rho = .75$. An effect size this large is not atypical in some areas of psychology, especially in repeated measures designs that may have high correlations between experimental conditions. In this simulation, all three variations of Taylor and Muller's (1996) method achieved an average power of just below .80 or higher, and all assurances met the target.

Three Illustrative Examples

Now that we have shown the general effectiveness of our proposed method, to place our simulations into a more practical context, we describe its performance in the specific cases of three hypothetical researchers, each planning a sample size for .80 intended power on the basis of a similar previously published study.[17] In addition to showing how various methods of sample-size planning work for various designs, we show how close each method would come to achieving the intended power. Although true population power would not generally be known, we adopt an omniscient perspective.

Underpowered

Suppose Researcher 1 is interested in studying the effect of working memory load (high vs. low) on associative activation. We assume that the true but unknown population effect size, δ, is 0.2. The researcher reads a previously published article about a study on this topic. This article indicates that an independent-samples *t* test with 25 participants per group yielded a *d* of 0.68.[18] The authors correctly rejected the null hypothesis despite the study's extremely low power (.107) to detect the effect of interest, but did so only because they had badly overestimated the population effect size. Although Researcher 1 would not know this additional information, we can use it to give an indication of the likely power of her new study, given various methods of sample-size planning.

It turns out that if Researcher 1 uses the sample effect size directly in power calculations, she will recruit 35 participants per group, achieving an abysmally low power of .13. Taylor and Muller's (1996) 20th-percentile method returns a $\hat{\lambda}_A$ of zero in this case, which means that the method is working exactly as designed: The prior study was severely underpowered, and the sample *d* reported is quite inaccurate. Given the results of the 20th-percentile method, Researcher 1 may decide to decrease her target assurance. Taylor and Muller's 50th-percentile method suggests 332 participants per group, but results in a power of .73, much higher than the power that would be achieved if Researcher 1 plans her sample size by taking the sample effect size at face value.

Adequately powered

Researcher 2 is studying the effects of sex ratio of the participants in the room (majority men, even split, majority women) and participant's sex (male, female) on diversification of financial resources in a 3 × 2 between-subjects ANOVA. His primary interest is the main effect of sex ratio. He reads that a similar study obtained a sample f^2 of 0.09 for this effect, with 25 participants per group ($N = 150$). Unbeknownst to Researcher 2, the population f^2 is 0.0625 (Cohen's medium effect size), and the true power to detect the effect of interest in the published study was .78.[19]

If Researcher 2 conducts his power analysis taking this sample f^2 at face value, he will need 19 participants per group and achieve a power of .65. Using Taylor and Muller's (1996) 50th-percentile method suggests 24 participants per group, which will result in a power of .76. The 20th- and 5th-percentile methods suggest 64 and 1,418 participants per group, respectively, and both lead to a power of almost 1. The latter suggested *n* is almost certainly prohibitive unless data can be collected at little or no cost online. Even so, we believe that one of the benefits of our suggested approach is that it can specify the degree of assurance for a broad range of options instead of pretending that there is only one "correct" sample size for a future study given a desired power, which would be true only if one could know the population effect size with certainty. Again, using the sample effect size at face value results in lower power than using Taylor and Muller's method, though the magnitude of the difference is smaller than in the first scenario. As Researcher 2 will not know the actual power of each method, he might select the 20th-percentile method, aware of the fact that, on average, it will reach .80 or higher power 80% of the time (80% assurance).

Overpowered

Finally, Researcher 3 is interested in studying the effect of performance pressure on attention using event-related potentials. She reads a prior study comparing participants' attention in high-pressure versus low-pressure trials using a dependent-samples *t*-test with 100 participants that is, unbeknownst to her, overpowered (power > .80 to detect the focal effect, for which δ is 0.5). The article reports a Cohen's d_z of 0.5.[20] Note that, in this case, the sample effect size is consistent with the population δ, and the true power is almost 1.

Researcher 3's power will be .81 if she uses the sample effect size at face value to plan her sample size. If she uses Taylor and Muller's (1996) method, her power will be .81, .93, and .99 for the 50th-, 20th-, and 5th-percentile versions, respectively. These power values are achieved with sample sizes of 34, 34, 49, and 79, in the four methods, respectively. What is most important about this third scenario is that any of the methods of sample-size planning is adequate, and that even the most conservative method suggests a new sample size smaller than that of the published study. Thus, Taylor and Muller's method does not always suggest a sample size larger than in the prior study. In this case, the prior study in the literature is overpowered, which is a benefit resulting from its large sample size. Of course, studies with smaller sample sizes can also be overpowered, but the population effect size must be quite large for this to be true. Although Researcher 3 has the most flexibility in terms of which method to use, overpowered studies are likely quite rare, as indicated by several literature reviews assessing power in psychology. The *n* of 25 that we used in our simulations is much more consistent with the sample sizes used in many experimental areas.

R Package and Web Applications: Bias- and Uncertainty-Corrected Sample Size (BUCSS)

We have developed an R package, BUCSS (Anderson & Kelley, 2017; freely available on CRAN), to allow researchers to use our suggested method to plan appropriate sample sizes that can provide power closer to the intended level. For researchers who do not use R, we have also provided Shiny Web applications (available at www.DesigningExperiments.com, the site that accompanies Maxwell, Delaney, & Kelley, 2018) to run the various functions from BUCSS using a user-friendly interface (see the Supplemental Material for a screenshot taken from these apps). Our package and apps are simple to use. Users only need to input their desired levels of assurance, power, and publication-bias adjustment, as well as readily available information regarding the effect of interest: the prior study's observed *F* or *t* value, the total sample size, and (for ANOVA designs) the number of levels of each factor. By directly inputting the observed *F* or *t*, users circumvent other software's potential proneness to error due to inconsistently defined effect-size measures. Some designs necessitate providing the type of effect of interest (e.g., interaction, main effect) because of varying formulas for degrees of freedom. Our package and apps will compute an adjusted effect size for any fixed effect in an ANOVA analysis and output the suggested sample size for the future study in a single step. We have provided separate functions for the independent-samples *t* test, dependent-samples *t* test, between-subjects ANOVA, within-subjects ANOVA, and split-plot ANOVA. The functions support omnibus effects for one- and two-way designs in which each factor can have any number of levels. We have also provided more general functions that can support effects beyond omnibus tests (e.g., contrasts), as well as designs beyond two-way designs.

Recommendations and Limitations

We have shown that Taylor and Muller's (1996) method can allow researchers to use information from a prior study in planning the sample size for a future study. However, researchers may wonder what level of assurance is appropriate, given that there are many options available. We emphasize that selecting the desired level of assurance is akin to choosing the desired power: There is no single correct value for power or assurance. An important advantage of Taylor and Muller's method is that researchers can report both the target power and the target assurance when describing their sample-size planning, so that readers and reviewers can know how often the chosen approach will be successful in achieving the desired level of power.

Although our proposed method has many advantages over several currently used practices, we recognize that it has limitations. It is designed to work with a single prior study, but there may be situations in which researchers have access to more than one prior study. In such situations, researchers can use a meta-analytic approach or perform sample-size planning with each available prior study, selecting the median sample size, the largest sample size, or the sample size associated with the most similar study. Conceptually, we assume a stringent *p*-value cutoff for publication: Studies in which the *p* values are above a certain threshold are not published, although what that threshold is can vary per user specification; in fact, the user can specify no publication bias whatsoever. Although publication bias is likely more complex than our model assumes, this assumption is generally consistent with the existing state of the literature in many areas of psychology. First, although published "marginally significant" results (e.g., *p* values between .05 and .10) are not unheard of, there has been recent speculation that "the tolerance for marginally significant" *p* values has decreased (Lakens, 2015a, p. 4). Second, simulations support the existence of a strict cutoff for significance, which mirrors our own assumption of a stringent cutoff at *p* values less than .05. For example, in a distributional analysis of *p* values from the published literature, Lakens (2015b) recently found "a clear drop" in *p* values greater than .05, a finding that supports a "strong effect of publication bias" (p. 832).

The method proposed here does not deal with additional issues that affect power, such as *p*-hacking. However, no method of sample-size planning that we are aware of corrects for *p*-hacking, as, by definition, researchers do not often reveal in their published articles how many tests they ran and what liberties they took. Statistically, we assume that all assumptions have been met (e.g., normality, homogeneity of variance). In particular, for tests of within-subjects effects with three or more levels, we assume sphericity.

Finally, in our simulations, we assumed equal *n*s across groups for the prior study as well as the proposed study. However, this supposition is fairly reasonable in experimental studies, in which researchers have more control over equating sample sizes across groups than they do when conducting field studies. Moreover, using equal *n*s usually maximizes power and enhances robustness to violation of assumptions, and power is determined most strongly by the group with the smallest *n* (Maxwell & Delaney, 2004). Consequently, our proposed method can be thought of as appropriate for planning the smallest *n*.

Alternative Approaches to Sample-Size Planning

The method we have presented is not the only appropriate approach for planning sample sizes and certainly has its own limitations. However, it does allow researchers to continue to use the sample effect-size estimate from a prior study in sample-size planning, after appropriate adjustments for publication bias and uncertainty. Yet other methods should not go unmentioned. We note that the choice of method often depends on one's goal, the experimental design, and the information available. We encourage readers interested in learning more about these approaches to consult the sources we cite or a review (e.g., Maxwell, Kelley, & Rausch, 2008).

First, an often-recommended strategy is to calculate the sample size for a theoretical minimum effect size of interest, which may be especially useful when a future study is the first on a topic. This approach is well suited for applied research and interventions, in which it may be clear how large an effect size needs to be to make a practical difference. For example, suppose a researcher wishes to detect an effect only if it is 0.8 or larger in an intervention study because only a large effect will convince policymakers that the intervention is worthy of implementation. Setting 0.8 as the minimum effect size of interest and calculating a sample size based on this value is appropriate for this goal regardless of the population effect size. Although power to detect a true effect will be less than the desired level if the effect size is less than 0.8, the researcher will presumably not care about this because if the effect size is less than 0.8, there is no practical value in implementation and thus the failure to find a significant result will be of no practical consequence even though the null hypothesis is false. However, "psychological theories are almost exclusively qualitative rather than quantitative," so they may be "not well equipped to help us identify when an effect is too small to be of theoretical interest" (Simonsohn, 2015, p. 560). Thus, sample-size planning using a minimally important effect size may be of limited value in theory testing.

Second, uncertainty can be handled from alternative perspectives. In one approach, researchers simply calculate the necessary sample sizes for a range of possible effect-size values (e.g., 0.2, 0.4, 0.6; O'Brien & Muller, 1993). However, this still requires the researcher to choose the most likely values for the population effect size. In another approach, the confidence interval surrounding the sample effect size is used to adjust for uncertainty (Perugini et al., 2014). However, this method does not adjust for publication bias and was shown to be more conservative than Taylor and Muller's (1996) method when equivalent percentiles were selected (Anderson & Maxwell, 2017).

Third, in lieu of using a sample effect-size estimate from a published study, researchers conducting ANOVAs may instead have an idea of what the individual group means are likely to be; such expectations may be based either on prior studies or on theoretical expectations. Raw means from a published study may not always be as susceptible to publication bias as the sample effect size, as group means may come from nonsignificant comparisons that are also reported in the article.

Fourth, researchers may consider group-sequential methods of sample-size planning, which involve intermediate analyses on the data. These methods can increase efficiency, because significant results can be obtained in interim analyses based on a fraction of the planned sample size at which data collection would be terminated. Thus, these methods potentially require a much smaller final sample size than originally planned (e.g., Lai, Lavori, & Shih, 2012; Lakens, 2014). Sequential methods require researchers to specify in advance the number of interim analyses as well as the number of participants for each stage (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015). Peace and Chen (2011) explained that the sample size at each stage is simply the total sample size of a nonsequential study divided by the number of planned stages.[21] As they emphasized, sequential methods still require specification of a total (i.e., maximum) sample size. The sample size derived from our suggested approach can provide this total sample size for sequential methods.

All of these approaches to sample-size planning are generally aimed at declaring an effect to be statistically significant. Other approaches, in contrast, plan sample sizes for alternative purposes. Simonsohn (2015) recently proposed using the sample size that would result in .80 power to detect an effect of the size that the original study would have had .33 power to detect, an approach geared toward situations in which the researcher aims to show that the original study was not adequately powered to detect the effect that was reported. An alternative approach when the goal is to show the absence of an effect would entail sample-size planning for equivalence (see Anderson & Maxwell, 2016; Chow, Shao, & Wang, 2003; and Lakens, in press, for more on equivalence-testing approaches). Finally, another option is to forgo the concept of power, which is based on statistical significance, and instead plan sample sizes for a certain degree of precision (accuracy in parameter estimation, or AIPE). AIPE approaches allow the researcher to specify the desired width (or half-width) of a confidence interval that brackets the population effect size of interest (e.g., Kelley & Maxwell, 2003; Kelley & Rausch, 2006; Maxwell et al., 2008). We note that methods for achieving accurate estimates can often yield required sample sizes that are very different from those obtained using methods designed to achieve intended power. AIPE is an effective approach when the goal is to achieve a certain degree of accuracy, to avoid "embarrassingly wide confidence intervals" (Cohen, 1994, p. 1002), but may be less useful when the goal is to show the existence and direction of an effect, as we have assumed here.[22]

Conclusion

Current commonly used strategies for sample-size planning have resulted in individual studies that have much lower power than intended, and this has resulted in a psychological literature that is, on average, quite underpowered. Multiple problems have ensued: a high rate of replication failures (e.g., Asendorpf et al., 2013), inconsistencies in the literature (e.g., Maxwell, 2004), inflated and inaccurate effect-size estimates (e.g., Kelley & Rausch, 2006; Maxwell et al., 2008; Maxwell et al., 2015), and the use of QRPs to obtain significant results despite low power (e.g., John et al., 2012).

Although our proposed method does not specifically correct for *p*-hacking, it does represent a methodologically sound alternative that can reduce the motivation for *p*-hacking and other QRPs by providing researchers with higher odds of detecting the effect of interest via a more appropriate sample size. The field often seems to be trapped in a vicious cycle: Researchers run underpowered studies, and, in turn, some may engage in QRPs in order to publish, despite low power, in a system that heavily values statistical significance (e.g., John et al., 2012). This results in upwardly biased effect sizes, which in turn lead other researchers to underestimate the sample sizes needed for future studies. Finally, these underestimates lead to more underpowered studies and, correspondingly, more QRPs. Our proposed method can help to break this cycle by providing a way to determine a sample size that will yield appropriate power, and thus an honest pathway to publication.

The logical strategy of basing power analysis on the effect size observed in a prior study does not often result in an accurate estimate of the true effect size, because of publication bias, uncertainty, or both. We have provided an approach to sample-size planning that allows users to adjust for both of these sources of error, and we have shown that using this method provides power that is much closer to the intended level for several experimental designs. Our suggested method has gone largely unnoticed, but our evaluation suggests that it is a hidden gem in the methods literature. We have provided an R package and Web apps that researchers can use immediately to implement the method we recommend. We hope that more accurate estimates of effect size will result in new psychological studies that are more adequately powered and will lead to a replicable literature that inspires more confidence and is less in crisis.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

S. F. Anderson and S. E. Maxwell developed the study concept. S. F. Anderson performed the simulations and interpretation of the results under the supervision of S. E. Maxwell. S. F. Anderson drafted the manuscript, and K. Kelley and S. E. Maxwell provided critical revisions. All the authors approved the final version of the manuscript for submission.

Acknowledgments

The authors would like to thank Lijuan Wang, Bradley Gibson, G. A. Radvansky, Harold Delaney, and Kerry Marsh for their insightful comments on earlier versions of this manuscript, as well as Eric-Jan Wagenmakers for providing access to his *t*-test data.

Declaration of Conflicting Interests

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

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617723724

Notes

1. A given study may include multiple statistical tests. Throughout this article, when we refer to the power or statistical significance of a study, we are referring to the power to detect the focal effect, or effect of interest, and the statistical significance of the specific test for that effect.

2. *Minimally important* implies that the effect size to be tested is chosen on theoretical or practical grounds, rather than estimated. The researcher is essentially saying that he or she is not interested in the effect unless it is above a minimum threshold. We discuss this approach further in the Alternative Approaches to Sample-Size Planning section.

3. These data can be downloaded at http://www.ejwagenmakers.com/papers.html, by clicking on the data link in the reference entry for Wetzels et al. (2011).

4. One of our reviewed articles fell into two categories. The authors reported basing their sample size on the sample size of a prior study, but also noted that the chosen sample size provided .80 power to detect an effect of the size obtained in a prior study. We included this article in both categories.

5. Although we use the term *publication bias* throughout this article, readers should be aware that effect sizes in some unpublished pilot studies might show a similar type of upward bias. Just as a journal editor may be more likely to publish an article if it reports a *p* value less than .05 (which results in bias in the subset of studies that are selected for publication), a researcher may be more inclined to conduct a future study if a pilot study's results are encouraging (e.g., a *p* value reaches or approaches statistical significance), rather than to simply discard the results. This latter phenomenon could result in upwardly biased effect sizes in the subset of pilot studies that form the basis for future studies.

6. In the case of the independent-samples *t* test, the population effect size, δ, is defined as $\frac{\mu_2 - \mu_1}{\sigma}$, where μ_1 and μ_2 are the two population group means, and σ is the pooled population standard deviation. With sample data, δ is estimated by d as $\frac{\bar{Y}_2 - \bar{Y}_1}{s}$, where $\bar{Y}_1$ and $\bar{Y}_2$ are the two sample group means, and s is the pooled sample standard deviation.

7. R code to reproduce all the figures in this article is available in the Supplemental Material.

8. Publication bias may not apply if the journal in question publishes all findings, regardless of statistical significance.

9. Leon et al. (2011) asserted that pilot studies, even if unpublished, may better serve as a check on the experimental protocol than as the basis for estimating the magnitude of the effect. The sample size may be so small that the underlying confidence interval is quite wide.

10. The distributions shown in Figures 1 and 2 are based on the assumption that journals never publish reports of experiments that yielded *p* values above .05 for the effect of interest, but our general point holds with less extreme publication bias (e.g., if some studies with *p* values of .10 are accepted for publication).

11. We conceptualize power in terms of the *F* distribution because it is more general, has simpler formulas, and has a more developed literature than the *t* distribution. However *t* is a special case of *F*, so our approach can be conceptualized in terms of the *t* distribution when a *t* test is appropriate. Readers interested in a more general overview of power can consult Cohen (1988) and Kraemer and Blasey (2016).

12. The following applies for testing a fixed effect, but different methods are applicable for testing a random effect. Similarly, different methods are relevant for multilevel models or nonnormal dependent variables.

13. If the *t* distribution is used, *n* must be quadrupled to double λ, as sample size falls under a square root.

14. For a fixed *n*, the distribution of the noncentrality parameter mirrors that of the effect size. Thus, adjusting the noncentrality parameter for uncertainty and bias is akin to adjusting the effect size.

15. Taylor and Muller did not report results for or recommend this percentile. However, we chose to test this percentile as it is aimed at 80% assurance and is somewhat less conservative than the 5th-percentile estimator.

16. When $\hat{\lambda}_A$ was zero, we used the following approach. For Taylor and Muller's 50th percentile, we increased α_P from .05 to .10. For Taylor and Muller's 20th percentile, we first decreased assurance to 50%. If $\hat{\lambda}_A$ was still 0, we then increased α_P from .05 to .10. For Taylor and Muller's 5th percentile, we first decreased assurance to 80%, and if $\hat{\lambda}_A$ was still 0, we then

decreased assurance to 50%. If the estimate was still zero, we increased α_p from .05 to .10. This simplified, practical solution resulted in no zero estimates.

17. The research interests of the hypothetical researchers were inspired from *Psychological Science* articles (Ackerman, Maner, & Carpenter, 2016; Baror & Bar, 2016; and Reinhart, McClenahan, & Woodman, 2016).

18. This d value is simply a hypothetical sample value a researcher might see in the literature. However, we selected this value because our simulations showed that the median d resulting from all published studies with 25 participants per group was 0.68 when the population δ was 0.2. Thus, 0.68 is a realistic value for a reported sample d in this situation, for a publication in a journal that functionally requires statistical significance as a prerequisite for publication.

19. As in the previous example, the reported sample f^2 is typical for the true but unknown population f^2; 0.09 is the median published f^2 resulting from a study with 25 participants per group and a population f^2 of 0.0625.

20. Cohen's d_z is a within-subjects version of the traditional Cohen's d. This d_z value is not directly from our simulations, as they all assumed an n of 25. This is instead the median d_z resulting from all published dependent-samples t tests in a simulation with 100 participants and a δ of 0.5.

21. Cook and DeMets (2008) pointed out that with sequential methods, power is reduced relative to fixed-sample tests with the same total sample size. They described how to determine the increase in total sample size that will maintain power comparable to that of a fixed-sample test with the desired level of power.

22. One might also consider sequential methods for estimation, as opposed to sequential methods for finding statistical significance. These methods can be used either to obtain a compromise between accuracy of the estimate and study cost (Chattopadhyay & Kelley, 2017) or to obtain a narrow confidence interval around the sample effect size (sequential AIPE; Kelley, Darku, & Chattopadhyay, 2017).

References

Ackerman, J. M., Maner, J. K., & Carpenter, S. K. (2016). Going all in: Unfavorable sex ratios attenuate choice diversification. *Psychological Science*, *27*, 799–809. doi:10.1177/0956797616636631

Anderson, S. F., & Kelley, K. (2017). BUCSS: Bias and Uncertainty Corrected Sample Size (Version 0.0.2) [R package]. Retrieved from https://cran.r-project.org/web/packages/BUCSS/index.html

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

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*, 305–324. doi:10.1080/00273171.2017.1289361

Asendorpf, J. B., Conner, M., de Fruyt, F., de Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–199. doi:10.1002/per1919

Association for Psychological Science. (2016). *Submission guidelines*. Retrieved from http://www.psychologicalscience.org/publications/psychological_science/ps-submissions#PG

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*, 1069–1077. doi:10.1177/0956797616647519

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

Baror, S., & Bar, M. (2016). Associative activation and its relation to exploration and exploitation in the brain. *Psychological Science*, *27*, 775–789. doi:10.1177/0956797616634487

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

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

Chattopadhyay, B., & Kelley, K. (2017). Estimating the standardized mean difference with minimum risk: Maximizing accuracy and minimizing cost with sequential estimation. *Psychological Methods*, *22*, 94–113.

Chow, S.-C., Shao, J., & Wang, H. (2003). *Sample size calculations in clinical research*. Boca Raton, FL: Taylor & Francis.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

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

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

Cook, T. D., & DeMets, D. L. (2008). *Introduction to statistical methods for clinical trials*. Boca Raton, FL: Chapman & Hall/CRC.

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

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

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*(10), Article e109019. doi:10.1371/journal.pone.0109019

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:10.1177/1745691614551642

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*, 499–510.

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*(8), Article e124. doi:10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2017). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*. Advance online publication. doi:10.1037/met0000127

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*, 305–321. doi:10.1037/1082-989X.8.3.305

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152. doi:10.1037/a0028086

Kelley, K., & Rausch, J. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*, 363–385. doi:10.1037/1082-989X.11.4.363

Kraemer, H. C., & Blasey, C. M. (2016). *How many subjects? Statistical power analysis in research* (2nd ed.). Thousand Oaks, CA: Sage.

Lai, T. L., Lavori, P. W., & Shih, M. C. (2012). Adaptive trial designs. *Annual Review of Pharmacology and Toxicology*, *52*, 101–110. doi:10.1146/annurev-pharmtox-010611-134504

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. doi:10.3389/fpsyg.2013.00863

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701–710. doi:10.1002/ejsp.2023

Lakens, D. (2015a). On the challenges of drawing conclusions from *p*-values just below 0.05. *PeerJ*, *3*, e1142. doi:10.7717/peerj.1142

Lakens, D. (2015b). What *p*-hacking really looks like: A comment on Masicampo and Lalande (2012). *The Quarterly Journal of Experimental Psychology*, *68*, 829–832. doi:10.1080/17470218.2014.982664

Lakens, D. (in press). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological & Personality Science*.

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.

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research*, *45*, 626–629. doi:10.1016/j.jpsychires.2010.10.008

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*, 331–348. doi:10.2466/03.11.pms112.2.331-348

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163. doi:10.1037/1082-989X.9.2.147

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). New York, NY: Taylor & Francis.

Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. In A. Panter & S. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 159–184). New York, NY: Taylor & Francis.

Maxwell, S. E., Kelley, K., & Rausch, J. (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

Muller, K. E., & Fetterman, B. A. (2002). *Regression and ANOVA: An integrated approach using SAS software*. Cary, NC: SAS Institute.

O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for *t*-tests through multivariate hypotheses. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 297–344). New York, NY: Marcel Dekker.

Olejnik, S., & Algina, A. (2000). Measures of effect sizes for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286. doi:10.1006/ceps.2000.1040

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943. doi:10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

Peace, K. E., & Chen, D.-G. (2011). *Clinical trial methodology*. Boca Raton, FL: Taylor & Francis.

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

Reinhart, R. M. G., McClenahan, L. J., & Woodman, G. F. (2016). Attention's accelerator. *Psychological Science*, *27*, 790–798. doi:10.1177/0956797616636416

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*. Advance online publication. doi:10.1037/met0000061

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

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

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

Skidmore, S. T., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods*, *45*, 536–546.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*, 160384. doi:10.1098/rsos.160384

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculations due to estimating noncentrality. *Communications in Statistics: Theory and Methods*, *25*, 1–14. doi:10.1080/03610929608831787

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298. doi:10.1177/1745691611406923

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