AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2024 American Psychological Association
ISSN: 0096-3445

2024, Vol. 153, No. 4, 1139–1151
https://doi.org/10.1037/xge0001273

COMMENTARY

Uncertainty Limits the Use of Power Analysis

Jolynn Pek, Mark A. Pitt, and Duane T. Wegener
Department of Psychology, The Ohio State University

The calculation of statistical power has been taken up as a simple yet informative tool to assist in designing an experiment, particularly in justifying sample size. A difficulty with using power for this purpose is that the classical power formula does not incorporate sources of uncertainty (e.g., sampling variability) that can impact the computed power value, leading to a false sense of precision and confidence in design choices. We use simulations to demonstrate the consequences of adding two common sources of uncertainty to the calculation of power. Sampling variability in the estimated effect size (Cohen's *d*) can introduce a large amount of uncertainty (e.g., sometimes producing rather flat distributions) in power and sample-size determination. The addition of random fluctuations in the population effect size can cause values of its estimates to take on a sign opposite the population value, making calculated power values meaningless. These results suggest that calculated power values or use of such values to justify sample size add little to planning a study. As a result, researchers should put little confidence in power-based choices when planning future studies.

Keywords: statistical power, experimental design, uncertainty, sampling variability, random effects

Supplemental materials: https://doi.org/10.1037/xge0001273.supp

Concerns about the (in)ability to replicate empirical results in psychology have led some researchers to push for stricter standards and greater scrutiny in the conduct of research. These include greater attention to effect sizes and statistical power calculations in designing studies, preregistering experimental designs and analyses, publishing replication reports, and championing open science (e.g., sharing data and code). The hope is that, together, these actions will improve replicability and the reputation of the science. Some of these good intentions carry with them challenges that are not widely known. The current paper is a synthesis of old and newer methodological developments, including recommendations about using power analysis. We posit that complexities and uncertainties in power analysis limit its informativeness and usefulness for justifying sample size. Our take-home message is that researchers should be particularly careful in trusting power calculations for determining sample size, and instead treat such calculations as only one of many factors when planning a study.

It is widely believed that using calculated statistical power values to determine study features, especially sample size, can improve the dependability and replicability of findings (Appelbaum et al., 2018; Asendorpf et al., 2013; Funder et al., 2014). Leading publishers (e.g., American Psychological Association, American Psychological Society) and journals of professional societies (e.g., Psychonomic Society, Society for Personality and Social Psychology) recommend, and sometimes require, that researchers perform a power analysis as a quality-control measure to justify the sample size of an experiment.[1] The outcome of the power analysis serves to inform experimental design and thus is taken to reflect the integrity of the experiment. For example, studies using designs higher in power have been linked to lower rates of false findings (i.e., lower proportions of significant effects that represent Type I errors—rejections of true null hypotheses, e.g., Ioannidis, 2005; Pashler & Harris, 2012; but see Wegener et al., 2022 for qualifications on such links).

The purpose of this review is to caution the research community about putting too much faith in (or emphasis on) the power values they calculate when designing an experiment. Specifically, the paper presents simulations to demonstrate the uncertainty inherent in

Jolynn Pek https://orcid.org/0000-0002-9694-4967

Code for the simulations and all data are available on the Open Science Framework (https://osf.io/cpsrj/). An accompanying RShiny application for the simulations is available at https://seeing-statistics.shinyapps.io/UncertainPower/.

Correspondence concerning this article should be addressed to Jolynn Pek, Department of Psychology, The Ohio State University, 228 Lazenby Hall, 1835 Neil Avenue, Columbus, OH 43210, United States. Email: pek.5@osu.edu

[1] Although psychologists tend to use power analysis to justify sample sizes, sample sizes can also be determined by the precision of the estimated effect size approach (e.g., see Maxwell et al., 2008; Rothman & Greenland, 2018).

most common power calculations, which are often used to determine an appropriate sample size for future studies. Standard practice is to plug in an effect size (e.g., one's best guess or an estimated effect from previous research) and treat the calculated value of power (or number of participants) as a meaningful value from which to design an experiment. However, the opposite might be true: Substantial uncertainty underlying these calculations can make it unwise to rely heavily on them when making design decisions. To complicate matters, the sources of such uncertainty are highly varied, which makes it difficult to trace their origins and formally incorporate them in power calculations to quantify the precision of these calculations.

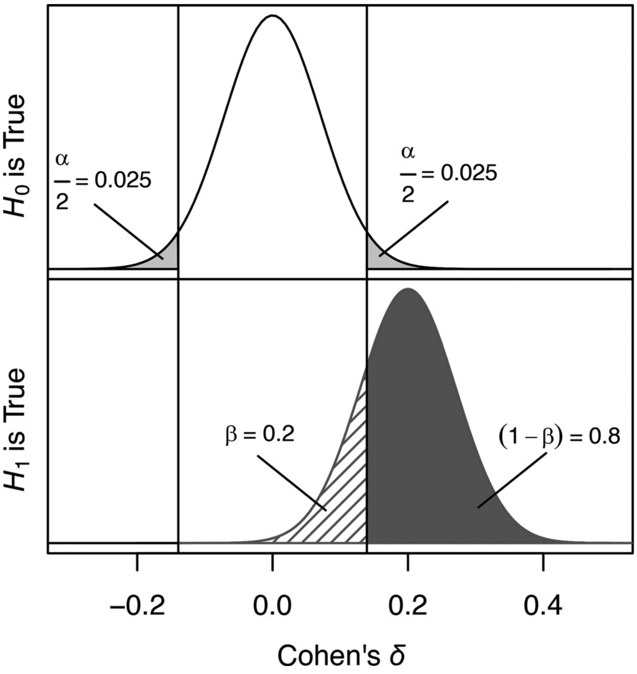
We use a simple two condition, within-subjects design throughout our examples, though the principles involved would readily extend to other designs. Consider a research scenario that typifies how the power of an experimental design is assessed. Imagine that a researcher (R) is interested in the link between emotion and memory, hypothesizing that high-valence words (positive or negative) will be remembered better in a recall test than low-valence (neutral) words. As a first step, R refers to recently published studies for guidance on design decisions. To calculate the power of the design for planned studies with similar effects and sample size, R might survey the existing literature for the typical effect size in similar studies and the approximate sample sizes used. Alternatively, R might take the effect size from previous work in their lab and a desired level of power and calculate how large the sample should be for planned studies. In the following sections, we illustrate two sources of uncertainty that underlie the calculation of power and sample size, both of which raise questions about how they should be used in experimental design. Before doing so, we provide a brief conceptual refresher on power and sample size determination.

Power: Definitions and Review

Calculating Power

Power is a concept within the null hypothesis significance testing (NHST) framework, in which a null hypothesis, H_0, is evaluated against an alternative hypothesis, H_1. Power is defined as the probability of rejecting H_0 (e.g., no difference between conditions) over repeated sampling assuming that a particular value of H_1 (e.g., a particular hypothesized value for the difference between conditions) is true. Figure 1 depicts the power calculation graphically for a two-sided dependent samples *t*-test with the Type I error rate of α = .05. The design consists of two conditions (within-subjects) with sample size N = 198, chosen for illustrative purposes. The effect size of interest is Cohen's *d* in the population (denoted by δ). Throughout the article, we denote *estimates* of δ, those based on data, with the symbol *d*. The top panel contains a distribution centered on 0 and reflects the state of the world assuming H_0 is true (no difference between conditions). The lower panel contains a corresponding distribution, assuming that a particular value of H_1 is true—that is $\delta = .2$, believed or hypothesized by the researcher (representing a best guess, sometimes based on data from a similar past study or studies). Vertical lines denote the critical value that guides NHST decision making. Power is the dark gray area of the lower distribution that exceeds the critical value. In this case, the dark gray area depicts a power of .8, a commonly used value.

Figure 1
Illustration of Decisional Error Probabilities and Power



Note. Power = .8 as area under the sampling distribution of $H_1 : \delta = 0.2$, with $H_0 : \delta = 0$, when $N = 198$ and α = .05 for a two-sided dependent samples *t*-test.

Calculating Sample Size

The formula used to calculate power can be rearranged to obtain a value of the necessary sample size (Cohen, 1988) to have a desired level of power. If using the same δ of .2 and specifying a desired power level of .8, the necessary sample size would be 198, as shown in the preceding example. By assuming a smaller or larger δ, the relation between power and sample size will change. We illustrate these changes below.

Challenges Due to Uncertainty

The classical formulas used to calculate power and calculate sample size can give the impression of certainty because they take fixed inputs and output a fixed value. The single power value or single sample size one obtains makes them seem certain and precise. However, uncertainty is ever-present because there are many unknowns in a power calculation. Although researchers are generally aware that specification of the effect size in the power calculation can be uncertain in reality (e.g., Du & Wang, 2016; Pek & Park, 2019; Perugini et al., 2014), presentations of power (e.g., in textbooks) tend to omit sources of uncertainty in the calculated power (or sample size). This idealized presentation of power calculations can make it difficult for researchers to appreciate how much such sources of uncertainty affect the calculated values of power or sample size. The classical power formula oversimplifies the relations among power, sample size, and α by excluding multiple sources of uncertainty when estimating power. We illustrate two of these sources of uncertainty (sampling variability in estimated effect sizes and random population effect sizes) on the calculation of power and sample size. Awareness of the impact they alone can have on calculated power and calculated sample size highlights the limitations of power calculations in informing experimental design.

A Common Misunderstanding of Power

The discipline-wide orientation to derive and justify sample size from the power formula can lead to the false belief that power is a property of a single experiment (e.g., Lakens, 2022). In the present scenario, R wishes to design an appropriately powered experiment and sample size is a key ingredient under their control. Because sample size is readily determined from a power analysis, R concludes the experiment will have the specified level of power by testing the prescribed number of participants. However, this well-intentioned reasoning extends the concept of power beyond its formal definition. As illustrated in Figure 1, power is defined by sampling distributions made from an infinite number of experiments using a hypothesized effect size, sample size, and alpha level. Power is thus a "what if" abstraction of R's experiment (Judd et al., 2017, p. 62). It refers only to the context of an infinite number of repetitions of an experiment. A single experiment itself is not "underpowered," "overpowered," or "correctly powered." Rather, these terms refer to the statistical test applied to repeated samples of data. In short, power is a performance measure of the statistical test that has consequences for experimental design in the long run. In what follows, the applicability of power is called into question when realistic sources of variability are included in its calculation.

Sampling Variability of an Estimated Effect Size

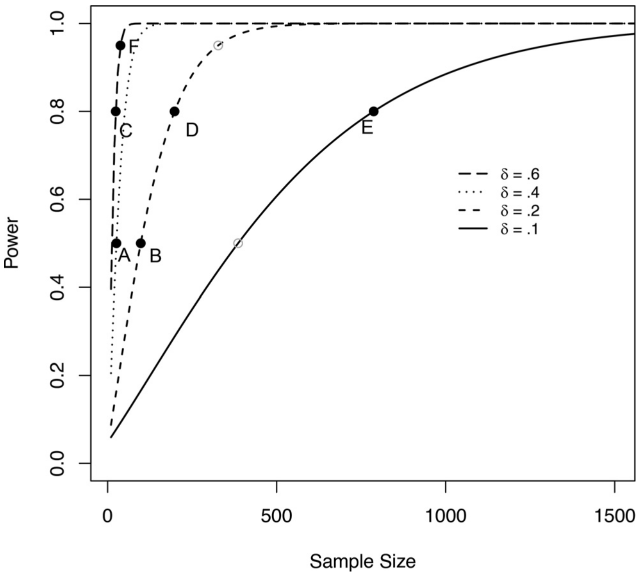
The first approach R might naturally take to design future studies is to identify an effect size equal to the typical effect in the literature that is accompanied by a typical sample size. R might identify a single study that seems representative of studies in the literature (e.g., one that seems to fall in the middle of the distribution of effect sizes and sample sizes), or R might take an estimated effect size from an existing meta-analysis of a related literature (and similarly attempt to calculate a reasonable sample size from that literature). R could plug these effect size and sample size values into a classical power formula and get a sense of whether the design would have "sufficient" power. Alternatively, R could plug in the desired level of power and the same effect size to calculate what sample size would be "sufficient" (to obtain that desired level of power—often .8 by convention).

Fixed Effect Size

We begin with the most common approach to power where R treats the effect size in the power calculation as the believed (hypothesized) effect size in the population (usually an unrealistic assumption) and calculated power (or calculated sample size) is treated as an exact value from the power/sample-size formula. Consider power of the two-condition within-subjects design described above. Figure 2 is a standard power-by-sample-size plot, with the lines denoting four potential effect sizes. If R's notion is that the effect of interest has an effect size of $\delta = 0.2$, R would consider a relation between sample size and power that is depicted by the line with short dashes (second from the right on which the points B and D are labeled). As the sample size plugged into the calculation increases, power increases, but at a different rate depending on the assumed effect size. For a small effect size (e.g., $\delta = 0.1$; solid line on which point E is labeled), it takes large changes in sample size to change the level of power. As the effect size increases (moving from lines on the right to lines on the left), larger increases in power occur with smaller changes in sample size, with power plateauing at a value of 1.0 at all assumed effect sizes once sufficient sample size (N) is reached.

Figure 2

The Nonlinear Relation Between Sample Size (N), Effect Size (δ), and Power for a Two-Condition Within-Subjects Design for a Two-Sided t-Test Where α = .05



Note. Points labeled A to F relate to the six conditions (unique combinations of N, δ, and power) in the simulation.

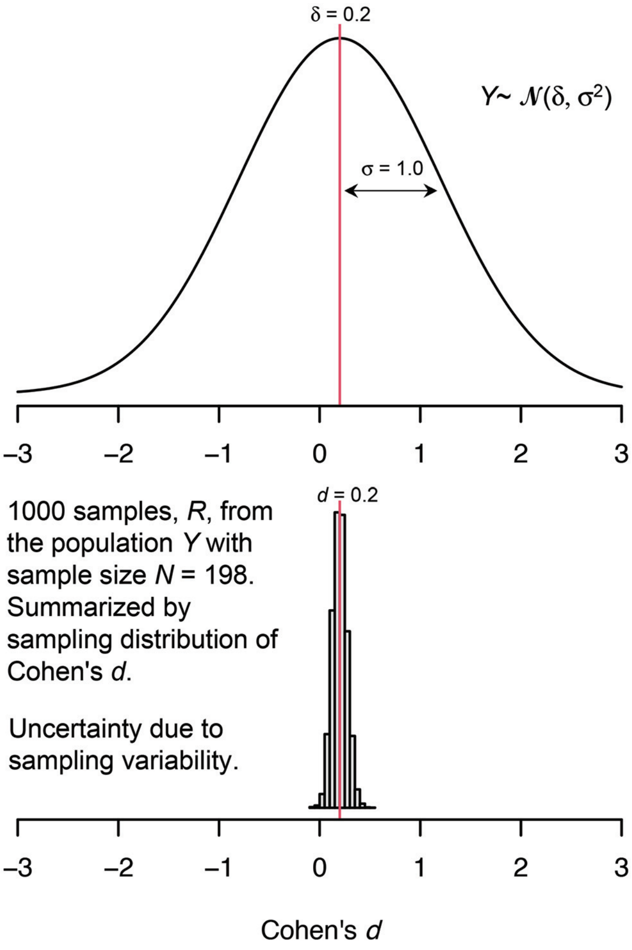
Estimated Effect Size

Recall that when applying power calculations to study design, it is easy to overlook the fact that each of the inputs in the calculations of power and sample size involves uncertainty. In other words, the inputs themselves are not error-free values but possess particular forms of uncertainty that can be formally expressed with distributions. To illustrate the impact of one such type of uncertainty on power estimation, we present simulations assuming only one source of uncertainty in the inputs: sampling variability of the estimated effect size (see Yuan & Maxwell, 2005). How much does uncertainty in the precision of an estimated effect size (i.e., a summary statistic from a study sampled from a population),[2] based on some fixed N, impact calculated power and sample size?

[2] Estimates from pilot studies tend to be accompanied by high sampling variability because of their small N. This has led to warnings against use of pilot studies to establish effect size estimates for power calculations (see Albers & Lakens, 2018; Kraemer et al., 2006; Leon et al., 2011; see also Gelman, 2019a, 2019b).

Simulations. To demonstrate the impact of sampling variability of effect size estimates on calculated power values, we simulated six combinations of believed effect size and N from the power curves in Figure 2 (labeled A through F). There were three target levels of power (.5, .8, and .95), four different population effect sizes ($\delta = .1, .2, .4$, and .6) and six different sample sizes ($N = 24, 26, 38, 98, 198$, and 787). These six conditions were chosen to span a range of combinations of effect sizes, sample sizes, and power that might be found in studies ranging from sensory processing to social cognition. The mean of each population distribution was made equal to δ, and the standard deviation of the outcome Y (e.g., scores) in the population was $\sigma = 1$ (i.e., the effect sizes were on a standardized scale; top panel of Figure 3). More formally, observed data (y) of size N are drawn from a normally distributed population $Y \sim \mathcal{N}(\delta, 1)$ with mean δ and standard deviation 1. For each of the six combinations, we drew $R = 1000$ samples (replicates) from the populations of standardized outcomes (e.g., scores) from a distribution with mean δ, with sample size N associated with the target level of power. Cohen's d was estimated based on each sample, thereby creating a sampling distribution of Cohen's d values from the 1000 samples, illustrated in the bottom panel of Figure 3.

Figure 3
Illustration of the Simulation Including Sampling Variability in the Estimated Effect Size



Note. Population distribution with mean δ = 0.2 and standard deviation σ = 1 and sampling distribution of Cohen's d for 1,000 samples. See the online article for the color version of this figure.

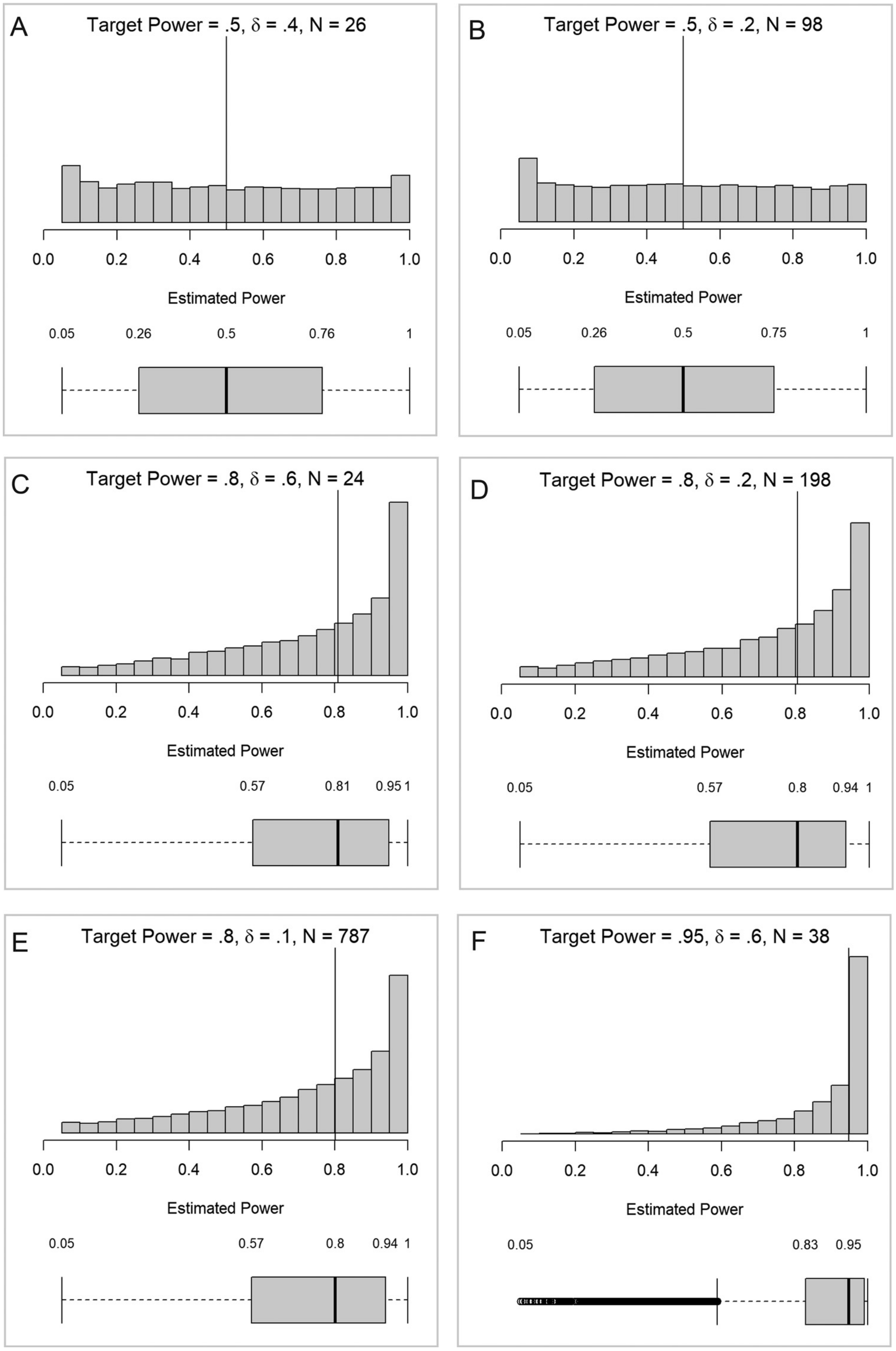
Calculating Power. Power was calculated for each of the 1,000 estimated Cohen's d values. The distribution of power estimates (in the six histograms and boxplots in Figure 4) communicates the degree of uncertainty due to the sampling variability in the estimate of d alone across the 1,000 samples. There are a few key points to note about these power distributions. First, the median of each power distribution comes very close to power calculated in the classical way (i.e., ignoring sampling variability). That is, the median comes very close to what we call *target power* (i.e., the value of power calculated in the classical way based on the population mean effect size and sample size). The distributions in Panels C through F are skewed, an indication that use of the mean of the distribution of power estimates (reported in the online supplemental materials) is a poor estimate of target power in these conditions. The most salient property of all six distributions, however, is that they are very broad, with all spanning the entire range of possible values for power (i.e., .05 −1). With a simulated population effect size associated with target power of .5 for a specific N (reflecting the points A and B in Figure 2), the shape of the distribution of power estimates is uniform (i.e., more, less, or equally likely to obtain a power value anywhere across the full range). As simulated population effect sizes or sample sizes are changed to increase the target power of the design, the distribution of power spans the same range of .05 to 1 but becomes more negatively skewed at higher target power values of .8 or .95 (reflecting the points C through F in Figure 2). Note also that the shape of the distributions of power estimates is determined by the target level of power, agnostic to different combinations of effect size and N that obtain the same target level of power (for analytic work that demonstrates the generality of these results, see Korn, 1990a, 1990b; Yuan & Maxwell, 2005; see also Pek et al., in press). Similar observations about the uncertainty (noisiness) of calculated power based on estimated effect sizes have been made in the literature on power calculated using the collected data from a single study (e.g., see Gelman, 2019a, 2019b).

A further challenge created by the presence of so much sampling variability is determining which distribution an estimated effect size, d, came from. For example, if R uses a Cohen's d of .4 from a previous study, R does not know whether that d came from a distribution in which δ = .2, .4, or .6, among other possible values. This situation further obscures the link between any obtained power estimate and the target power value that corresponds to the population.

Can Sampling Variability in Power Calculations Be Reduced by Increasing *N*? Suppose R is uncomfortable with the idea of the power estimate representing a single draw from a distribution that is as broad as those in Figure 4. R might reasonably wonder whether an alternate approach might help, such as dramatically increasing the sample size of the study being planned, even if the means for doing so might undermine the effectiveness of the manipulations or measures. For example, R could conduct the experiment online. This would give R access to much larger samples than could be brought into the lab, but the manipulations might be less effective in a less controlled environment or the previous experiences of at least some of the participants with similar

Figure 4

Variation in Calculated Power When Incorporating Sampling Variability



Note. Histograms and boxplots of power estimates for a two-sided *t*-test from a two-condition within-subjects design that reflect uncertainty due to the sampling variability in the effect size estimate (d) for six simulated conditions. Within each panel, the title denotes the population effect size value, δ, and sample size (N) that obtains target power calculated via the classical approach.

manipulations might undermine their effectiveness (cf. Anderson et al., 2019). Even so, would this approach help reduce variability and thereby obtain a more accurate estimate of power if sample size of the planned study is substantially increased? Unfortunately, no, unless the target level of power is dramatically increased, and even then, substantial variability would remain. As one can readily see in Figure 4, the shapes of the power distributions are largely insensitive to changes in sample size when the classically calculated target value of power (ignoring uncertainty) is the same across scenarios. For target power = .5, for example, the distribution in Panels A and B are very similar in central tendency, spread, and shape despite sample size being more than three times greater when δ = .2 compared to δ = .4. Perhaps this outcome would be somewhat expected when target power is relatively low, but the same insensitivity to *N* is also observed when population effect size places target power at the more conventional level of .8 (Panels C, D, and E). Whether δ = .6 or δ = .1, the power distributions are almost indistinguishable despite the sample sizes differing dramatically (26 vs. 787). Only for a fixed δ with increasing sample size can the variability in estimated power be reduced (Panel A vs. Panel E), though the long tail in the distribution remains until power estimates approach 1.0.[3]

Determining Sample Size. Having seen Figure 4, R's attention might shift to determining sample size for a desired level of power. We refer to calculated sample size based on the distribution of power estimates as *determined N*. Perhaps determined *N* is not cursed by such wide variability as the power estimates. Furthermore, during study planning, determined *N* is often of most interest, especially once key variables have been operationalized. To the dismay of R, this approach does not skirt the problem of sampling variability. To demonstrate the extent and nature of uncertainty in determined *N*, we next depict variability in sample size determination for the estimated effect sizes, *d*, that correspond to the four δ values corresponding to the lines labeled in Figure 2 We used the *d* value for each of the 1,000 samples (bottom panel of Figure 3) of size *N* for the corresponding point in Figure 2 as input to calculate the "required sample size for power of .8" as the output (see Figure 5).

Unlike the distributions of power in Figure 4, the shape of these determined *N* distributions differs at the same target level of power as effect size (δ) and *N* change (e.g., panels A vs. B, or C vs. D. vs. E). The distributions are all positively skewed as there is a long tail including extremely large sample sizes. Because determined *N* is not bounded on the upper end, that long tail extends so far in some cases that the histograms had to be truncated for presentation purposes (see Panel E; full ranges and specific quantiles are available in the online supplemental materials). In general, to design future studies with power = .8, the determined *N* distribution increases in variability as δ drops. With δ = .1, the determined *N* distribution is so flat that the median (784) is difficult to justify as any more appropriate a sample size than values that are considerably higher or lower than that value (see Panel E). With δ = .2, a common assumption in some research areas, the distribution of determined *N* is also sufficiently wide to caution placing undue importance on any one exact value (such as 198 and 196, respectively, the mean and median of the distributions; see Panels B and D). Thus, when the population effect size, δ, is small, even when attempting to develop a design with power of .8, uncertainty in determined *N* will be high and the median of that distribution becomes an arbitrary, minimally informative choice in determining the sample size needed to reach the desired level of power for that design. R would have to be dealing with a large effect size (δ = .6) to obtain a determined *N* that seems reasonably informative (see Panels C and F). However, with a population effect size this large, the experimental outcome is likely to be so certain that there is little need to fret about sample size.

Increasing Estimated Effect Size Precision

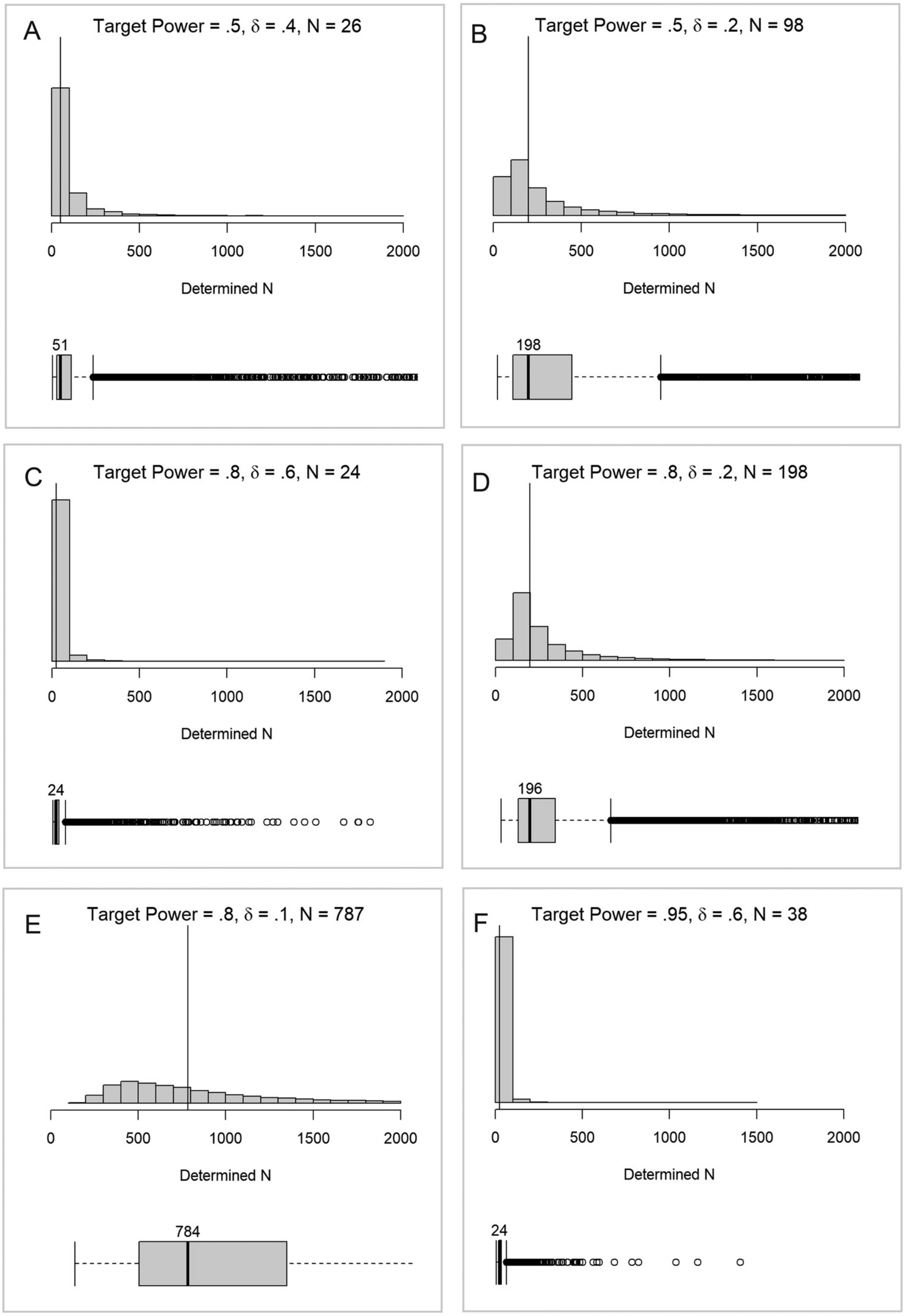
A related question R might ask is, "does having a better estimate of effect size, *d*, reduce uncertainty in power or sample size estimation?" Considering the uncertainties inherent in attempting to use the estimated effect size from a single study to estimate power or determined *N* for future studies, R might seek out a relevant meta-analysis for a better estimate of effect size. After all, a meta-analytic effect size is based on a combination of data from a larger number of studies and thus should have less sampling error compared to an effect size from a single study. McShane et al. (2020) examined the variability of power estimates based on a meta-analytic effect size (an estimate they called average power). From analytical and simulation evidence, average power has less variability than estimated power in Figure 4. However, McShane et al. (2020) "found that point estimates of average power are too variable and inaccurate for use in application" and that "our assessments are optimistic in that point estimates of average power will be more variable, and thus less accurate, in more realistic scenarios" (p. 195). Results in Figures 4 and 5 echo these conclusions. As discussed in more detail later, multiple studies using the same design might produce a more informative summary effect size, but meta-analyses generally include studies that use varied methods that are tested across varied samples and settings. Thus, in addition to potential sampling variability, effect sizes from literature-wide meta-analyses often involve sources of uncertainty that go beyond sampling variability per se—sources that cannot be reduced by increases in sample size (captured, in part, using random effect models in meta-analysis, Hedges & Vevea, 1998). Uncertainty that cannot be reduced with more information (e.g., larger sample sizes) is called aleatory uncertainty (O'Hagan, 2004).

Methodologists have recognized the importance of sampling variability inherent in effect size estimates used in power computations. New approaches to calculate power and determine *N* incorporate sampling variability (e.g., safeguard power by Perugini et al., 2014; power-calibrated effect size by McShane & Böckenholt, 2016; bias and uncertainty corrected sample size determination by Anderson et al., 2017). These approaches vary in terms of complexity in their application, development for specific procedures, sample size recommendations, and how uncertainty in estimated power or determined *N* is summarized (e.g., by a lower quantile or mean of the power distribution). In general, these methods point to requiring much larger sample sizes relative to the classical approach of power analysis that does not consider

[3] This outcome might be surprising because it seemingly violates the well-known observation that the standard error (e.g., of means) is reduced by increasing sample size. However, the sampling variability in calculated power comes from the estimated Cohen's *d* (see Figure 3) of the completed study, which cannot be reduced by increasing *N* of the planned (future) studies.

Figure 5

Variation in Determined N When Incorporating Sampling Variability



Note. Histograms and boxplots of determined *N* for a two-sided *t*-test with 80% power from the example design. Within each panel, the title denotes the effect size value, δ, and sample size (*N*) that obtains the target power calculated via the classical approach. The boxplots display nontruncated data for the sample sizes, but the histograms are truncated at *N* = 2,000.

sampling variability. Larger sample sizes, however, do not necessarily reduce uncertainty (i.e., shrink the variance of the distributions of power), especially if how sample size is increased negatively impacts qualities of the design, such as reducing effect size by reducing the effectiveness of manipulations or reducing the reliability of outcome measures. All else remaining equal, these new approaches to power calculation increase the probability that the design might have at least the intended level of power over repeated sampling (i.e., not the single experiment that R conducts next). Yet, as discussed in the following section, substantial uncertainty would nonetheless remain.

Sampling Variability and Random Population Effect Size

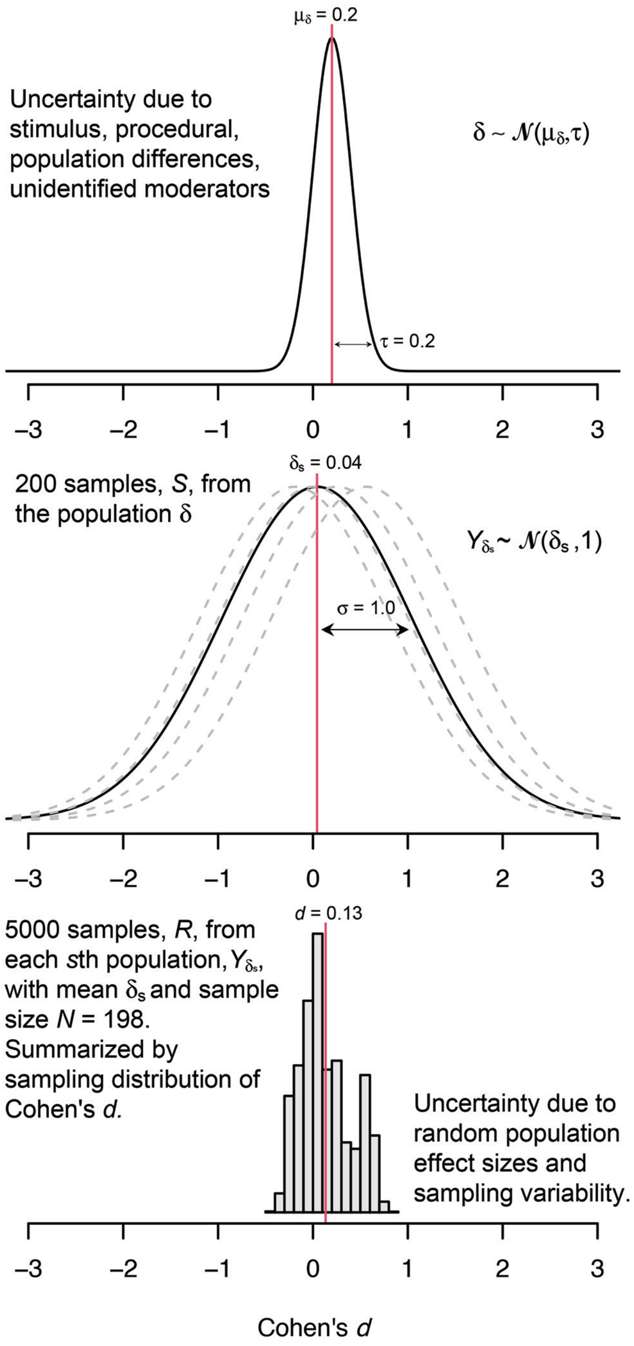
The distributions in Panels A to F of Figures 4 and 5 are contrived in that the simulations assumed a known and fixed population effect size, δ, a virtual impossibility in everyday research. Even in this purposefully constrained and simplified situation, sampling variability in estimated *d* creates substantial variability in estimated power and determined *N*. Yet, there is often variability in effects based on factors beyond sampling variability, such as those due to unknown moderators (Shrout & Rodgers, 2018), measurement error (Stanley & Spence, 2014), imperfect validity (Hunter & Schmidt, 2004), and changes across time, contexts (task, stimuli, instructions), and participants, which result from factors such as history or culture (e.g., Markus & Kitayama, 1991; Schwartz & Rubel, 2005). Together these are all classified as random effects (aleatory uncertainty) in the population, and thus impact the size of the population effect, δ, independently of sampling variability.[4]

To account for random effects, rather than assuming a single fixed population effect size, researchers must consider a range of plausible effect sizes for the phenomenon of interest. This goal is challenging to accomplish without knowledge of one or more of the various potential sources of variability. Indeed, if there is one proposition on which all researchers might agree, it is that no research program identifies all possible moderators of an effect (or even all possible operationalizations of the relevant variables). These persistent gaps in knowledge lead to uncertainty about the appropriate effect size to use in power estimates (see McShane & Böckenholt, 2014; O'Hagan et al., 2005; Pek & Park, 2019). Our next simulation includes aleatory uncertainty and sampling variability in the effect size. This approach parallels the random effects meta-analytic framework by allowing the effect size to have variability in the population (Hedges & Schauer, 2019; Kenny & Judd, 2019). These simulations show that the challenges R faces in the form of sampling-based uncertainty in power (see Figure 4) and determined *N* (see Figure 5) when the population effect size, δ, is fixed compound when effect sizes are more realistically assumed to be random.

Recall the top panel of Figure 3, where the population has a mean = δ and standard deviation σ = 1. Suppose, now, that δ is random. That is, δ follows a normal distribution with mean μ_δ and standard deviation τ: denoted as $\delta \sim \mathcal{N}(\mu_\delta, \tau)$. The uncertainty in population values of δ (quantified by τ) represents sources of aleatory uncertainty. There are thus two levels of sampling in this more comprehensive simulation, depicted in Figure 6. Shown in the top panel is uncertainty in the population effect size, δ. In the middle panel, we might sample *S* = 200 "populations" from which to draw sample data to quantify this aleatory uncertainty. Next, we draw

Figure 6

Illustration of the Simulation of Random Population Effect Sizes and Sampling Variability



Note. Uncertainty in the population effect size (τ) is combined with sampling variability (samples drawn from *S* populations) and summarized in the sampling distribution of Cohen's *d*. See the online article for the color version of this figure.

[4] Aleatory uncertainty represented as random effects is also termed "heterogenous effects" in the literature.

R = 5,000 samples of size N from each population distribution in the middle panel and estimate d within each of the S populations as we did in the preceding simulations: $Y_{\delta_s} \sim \mathcal{N}(\delta_s, 1)$. The bottom panel illustrates the distribution of d for (S = 5) × (R = 5,000) samples (25,000 total). Compared to the fixed effect situation, each distribution of Y is subscripted with s, denoting each sth draw from the distribution of random effect sizes, which reflects the uncertainty about the population effect size specified by τ. Thus, the total number of samples of size N would be $S \times R = 200 \times 5000 = 1,000,000$. One can increase N to reduce sampling variability in estimated d (bottom panel of Figure 6) but the between-population variability captured by τ (top panel of Figure 6) remains regardless of the number of S populations that are considered. We illustrate the effect of τ in downstream estimates of power using two relatively small values (.1 and .2) equal to the estimated 25th and 50th quantiles of between-study variability, respectively (see van Erp et al., 2017, who estimated τ from a survey of meta-analyses).[5]

The leftmost panel of Figure 7 shows a sampling distribution of estimated d values calculated from the $S \times R$ samples. The sampling distribution is centered on the value of d = .2 (because μ_δ = .2). However, even when the variability in δ is relatively small (standard deviation of τ=0.1) there are a substantial number of estimated effect sizes that fall in the opposite direction (i.e., a negative d) compared with the population mean value (i.e., μ_δ = .2). When the variability in δ is larger (standard deviation of τ = .2) there are even more estimated effect sizes that fall in the opposite direction to the population value. Thus, the sampling distribution of d increases in variance as the uncertainty (τ) in the effect size δ increases. The consequences of this increase in variance on power estimation are both expected and unexpected.

Calculating Power

When these estimated d values are used to compute power (middle panel in Figure 7), we see a pattern like the usual power function when the estimated Cohen's $d > 0$ (i.e., a sigmoid shape increasing and flattening as the maximum value of power [1.0] is approached). That shape is clear when the variation in δ is relatively small ($\tau = 0.1$), except that some negative d estimates produce power values for those negative values of d. Because the negative d values are generally small, these calculated power values are also relatively small, $<$ .25. These small power values for negative estimated d produce a small "hump" on the left side of the density plot of power estimates in the right panel of Figure 7. The hump is created because the negative d values in the middle plots fold onto their positive counterparts to create the density plots on the right (rotate the middle graphs on their side to see this). When there is greater variability in the d estimates (i.e., $\tau = 0.2$), however, the negative d estimates become extreme enough (≤.2) that some estimates of power become large for d values opposite in sign to the population δ value! Note that in the top middle panel of Figure 7, d estimates approaching $-.5$ have an estimated power of 1. The result of this extreme variability is that the density plot of estimated power (right panel of Figure 7) produces a more pronounced peak near power of 1 when $\tau = 0.2$ than when $\tau = 0.1$. Yet, because a sizable portion of estimated d get the direction of the effect wrong, the power calculation is obviously not conveying meaningful information related to the population of interest. With such a large number of power estimates based on directionally incorrect effects (negative instead of positive) when variation in the estimated effect size is substantial, the distribution is misleading and cannot be reasonably summarized by a single value. More broadly, the power estimates based on an estimated effect size from any one sample are called into question. For example, the central tendency of the distribution of calculated power values (right panel of Figure 7), which one might use as a best estimate of power, is meaningless. Indeed, "power" associated with negative effects should instead be considered Type III errors (i.e., a statistically significant effect in the direction opposite to the population effect; Kaiser, 1960). As suggested by a reviewer, one might alternatively consider power as zero for effect size estimates that get the direction in the population wrong. For an alternative plot of the resulting power calculations, see the online supplemental materials accompanying this article. Because of the complication of Type III errors in relation to estimates of power, we do not elaborate on determining N in the presence of sampling variability and random population effects (but see the online supplemental materials for results on determined N, ignoring Type III errors).

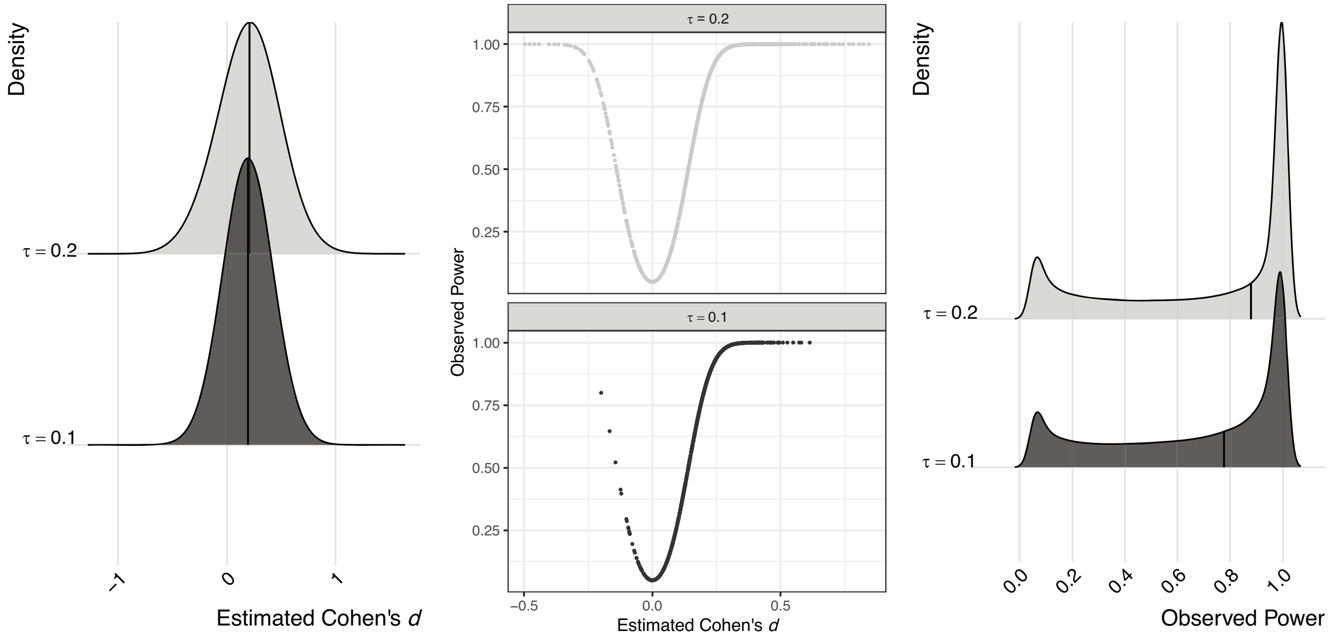
Consequences of Uncertainty in Calculating Power

Though well-intentioned, the use of power calculations to "establish credibility" of a design can be grossly misleading. The results of the preceding simulations reveal the extent of the uncertainty underlying power calculations. Any precision implied by getting a specific number out of a power calculation or a determined sample size is quite illusory. Precision cannot be achieved when the population effect size is unknown except for situations with very low variability (e.g., maybe a psychophysics study) or high information (e.g., collecting the whole population). Sampling-related sources of uncertainty are ubiquitous but are only one of many sources of uncertainty that must be considered in calculating power. As we illustrate, considering additional sources of uncertainty compound to the point of sometimes generating large "power" estimates for effect sizes that run counter to the direction of the population effect. Our simulations were incomplete in that they did not address additional potential sources of uncertainty, such as incorrect statistical models. Our simulations assumed that the analysis model matched the generating model. Stated differently, the generating mechanism that gives rise to collected data is seldom the generating model used in simulations or assumed in power calculations (MacCallum, 2003). When enough data are available to reduce epistemic sources of uncertainty (i.e., uncertainty that can be reduced with more information, e.g., sampling variability) to a negligible level, power analysis would be largely unnecessary. The researcher would understand the phenomenon of interest well enough to predict with confidence the likely outcome, making the goal of the researcher one of engineering a system rather than discovering the properties of that system (see Gigerenzer, 2004; for a perspective on applying power in a quality control context). Even in such settings, however, there is likely to be uncertainty about the population effect size (as in Hedges & Schauer, 2019; Kenny & Judd, 2019), many of which are aleatory

[5] One can model between-study variability in a meta-analytic context by analyzing study variables (e.g., properties of the manipulations, measures, settings, or participants) that can "explain" that variance, but the variance remains largely unexplained in most settings. Note that the estimates of τ in van Erp et al. (2017) likely contain two sources of uncertainty (i.e., sampling variability and between-study variability).

Figure 7

Joint Effects of Random Population Effect Size and Sampling Variability on Calculated Power



Note. The left panel illustrates the sampling variability in estimated d when $\mu_\delta = 0.2$, $N = 198$ and $\tau = 0.1$ or 0.2. The middle panel presents calculated power values in relation to the estimated effect size d. The right panel presents the density plot of power values, ignoring the direction (sign) of the estimated effect d (visible in the middle panels).

sources of variability, such as use of an incorrect statistical model or effect size uncertainty due to incomplete knowledge of moderators, that cannot be reduced by simply adding more data.

All this uncertainty has consequences for researchers when thinking about the relation between the level of power that is actually embodied in the procedures they are using in their studies (i.e., target power in our simulations) and the power values researchers calculate. The uncertainty due to sampling variability and random population effect sizes means that one cannot directly take the calculated power value as applying to the study one designs (even if repeated across a number of studies). That is, there will be necessary gaps between calculated power values and the results observed in completed studies. It is a misconception to assume that a design with classically-calculated .8 power would produce 80% significant results across a set of studies because of two main reasons. First, power calculations tend to be imprecise because of various sources of uncertainty (see Figures 4, 5, and 7). Our illustrations show that the values one obtains using estimated effect size from a previous study can be quite different from the long-run value of power associated with that design. Unless one's effect size is very large (e.g., $\delta = 0.6$), the distributions for calculated power and determined N are so wide (see Figures 4 and 5) that any such values should be taken with considerable caution. Even when assuming a large effect size, that effect could be from a random population (uncertain because of unknown moderators, varied instantiations, etc.), adding much uncertainty to calculated power and determined N that is difficult to eliminate. Second, power is a probability from a sampling distribution representing an infinite number of samples (see Figure 1). Because an infinite number of studies cannot be conducted, the probability of observing statistically significant results in a given set of studies is better modeled by a binomial distribution (see also Wegener et al., 2022).

Power tends to be misunderstood and misapplied (e.g., Bakker et al., 2016) despite insights and complexities highlighted by methodologists and statisticians (e.g., on the problem of single-study post hoc power [see Hoenig & Heisey, 2001; Lenth, 2007; Senn, 2002]; on the inconsistency between power and statistical nonsignificance [see Greenland, 2012]). Our simulations add to the literature on using power to determine N while acknowledging a high level of uncertainty.

Uses of Power and Alternative Paths Forward

Statistical power is a pure abstraction. It is a property of a test (e.g., a t test) nested within a statistical model (e.g., linear regression) whose purpose is to assess the presence of "a signal" (vs. noise) in the observed data (see also Pek et al., 2020). Such statistical models are imprecise and incomplete because they cannot incorporate all potential influences or sources of uncertainty, as many of these are unknown when a statistical model is specified. In this context, it is advisable to recall that all models are wrong at some level (Box & Draper, 1987). At best they provide a simplified mathematical approximation of the psychological effects and mechanisms under study (MacCallum, 2003). This lack of specificity leads to a high level of uncertainty that is ever-present in the calculation of power. Put another way, power has such a tenuous link to the phenomenon of interest, through no more than a (noisy) estimate the researcher provides, that there has to be a high degree of uncertainty in its estimation. So little is known about most psychological phenomena early in the research process that the situation could hardly be otherwise. Psychological phenomena are routinely moderated and sometimes completely reversed across different conditions or perceivers. There is considerable distance between the physical properties of stimuli and the corresponding psychological constructs of interest. For these and other reasons, throughout much of the development and testing of a psychological theory, there must be uncertainty about the random nature of effect sizes for particular operational manipulations or measures of the phenomena of interest. This uncertainty should be neither surprising nor necessarily dismaying. It does pose challenges, among them being the blunt nature of any tool involving calculations of statistical power or sample size determination (and probably other tools that involve calculations derived *not directly from data* but from use of other statistics based in an abstract model that have their own—compounding—uncertainties). Yet, by acknowledging the uncertainties involved in such calculations, we can focus on those aspects of the uncertainty that can be reduced and on understanding and explaining those aspects of the uncertainty that require theory to disentangle.

Although one might reasonably wonder, “Isn’t use of power better than not using power at all?,” the answer from our perspective would not be a simple “yes” or “no.” Values of power from a calculation cannot be treated as precise. In the same vein, determined *N* from such power analyses cannot remain as a universal gold standard for justifying sample size. It is simply folly to believe that a point estimate (effect size) from a noisy system (human participants) can tell us anything precise about the design of the next experiment (sample size). Thus, basing study design decisions on imprecise and uncertain estimates of power or *N* and treating them as precise or certain is likely to do more harm than good. For example, such decisions could lead to wasted resources or to faulty dismissals of reasonable evidence based on assumptions of “low power” (cf. Wegener et al., 2022). They could also lead to overreliance on results that are accompanied by assumptions of “high power” (cf. Greenland, 2012).

Consideration of power and determined *N* are potentially important, however, because such considerations motivate “statistical thinking” when planning an experiment (Chance, 2002). Partly, such statistical thinking is aided by helping the researcher to develop a data analysis plan early in the research process. Statistical thinking requires a mindset in which design and analysis decisions are thought through and justified. R should be able to explain the reasoning of the experiment from start to finish, articulating the logic of predictions and how they will be evaluated statistically. The size of one’s sample requires considering properties of the design, accompanying analyses, and the population being tested. Such decisions should also crucially depend on whether the current experiment is designed as a single one-off study that is meant to produce compelling data on its own or whether it is intended to become part of a set of studies all examining the same phenomenon. Through experience, R learns how to weigh these factors to arrive at what they believe is an appropriate sample size for any given study. There is not a single ideal sample size. Many are appropriate, and the choice is subject to updating as additional information is considered (cf. notions of optimal experimental design for model discrimination and parameter estimation, Myung & Pitt, 2009). As we have seen in the current simulations, merely scratching the surface of the uncertainty underlying power calculations had the effect of substantially increasing uncertainty about the level of power that might be associated with a particular research design (see also Anderson et al., 2017; McShane & Böckenholt, 2016; Pek & Park, 2019; Perugini et al., 2014).

Our current treatment of power has yet to bridge other conceptual distances, which highlight the peril of applying power analysis in a mechanistic manner (see Stark & Saltelli, 2018; who lament cargo-cult statistics or the ritualistic mining of statistics rather than conscientious practice). First, there is an inescapable conundrum. Power analysis is encouraged when little is known about the phenomena under study to aid study design, but the numbers obtained from a power analysis are so highly imprecise to be of little practical use. Only when much is known about the effect size, in which case power analysis is unnecessary because the test is largely confirmatory, do we obtain values that can be practicably applied. Second, power is based on a simplified (statistical) model that purportedly approximates the mechanisms that produce the phenomenon or phenomena that appear in the expected data. As noted above, this statistical model is an abstraction without any direct connection to data or the mechanisms generating the data (e.g., an explanatory mental model). Third, it should be recognized that the collected data include not only any signals of the mechanisms at work, but also any mechanisms involved in the translation of participants’ experiences, intentions, design attributes (e.g., manipulations, participant task, length of the experiment), and the like into responses that can be measured in the study. Thus, to unquestioningly accept a power value as a quantification of one’s expectation of observing a significant finding in completed studies disregards the gap between hypothetical versus collected data, the gap between the simplified statistical model versus the complex mechanistic model, and the uncertainties that underlie the inputs to power calculations.

We recommend that researchers place limited confidence on design decisions when using power to design experiments, or not use power at all as a direct justification for determining *N*. Alternative justifications exist that are probably much more compelling, such as using distributional properties of data to determine sample sizes at which statistics using normal approximations make sense (e.g., using the Central Limit Theorem or precision-based justifications in Footnote 1). The large amount of uncertainty in power calculations makes it most appropriate to use power to generate “what if” scenarios that can be useful in thinking through a host of design issues (e.g., appropriate statistical models, how the statistical model connects to data, what kind of operationalizations of manipulations or measures aptly instantiate variables of interest). R’s overarching goal is to ensure that the choices made in designing an experiment yield results that will be useful and believable however they turn out. How does R ensure that an experiment can result in high quality data that can answer the research question? Reliance on past work is a traditional and obvious place to start. Ideas about sample size, number of stimuli, and number of observations can be used to plan an experiment, but because R is testing a novel hypothesis, educated guesses are required; indeed, they are integral to scientific discovery. Pilot experiments can be carried out to calibrate some aspects of the design (e.g., stimulus difficulty, determine number of conditions) and further inform these decisions, but ultimately R must evaluate the quality of those choices by conducting the experiment (and, in most cases, probably a series of experiments). There is no single best design, but through careful thought, reliance on experience and intuitions, and advice from others, a sound and justifiable design can be created and implemented. The quality of the data depends on the quality of the design, not only in statistical terms, but regarding issues such as construct validity, internal validity, and external validity as well (e.g., Fabrigar et al., 2020; Flake et al., 2017).

Regardless of whether the data fall as R predicts, confidence in the design leads to confidence in the data. R's first experiment is an initial step in developing such confidence. Follow-up experiments, whether direct replications or variants that test related hypotheses, will modify the believability of the outcome and the theoretical explanation that is being tested. Clear statistical evidence might be present in a single study or might require multiple studies. Evaluation of the theoretical explanation must involve the clarity of the design (e.g., internal and construct validity) as well as the strength of the statistical signal in the data. Sufficient signal in the data themselves (rather than in the assumptions about power) make the case compelling when the other design elements are also clear. This is true regardless of whether one takes a frequentist approach (as we have in this paper) or a Bayesian approach. The amount of noise in most data ensures that effect sizes vary across experiments (or samples of participants) due to sampling variability, differences in operational manipulations or measures, and the like. But the overall trend across studies should ultimately lead R to a conclusion. It could be that the effect is so unpredictable in magnitude and direction (positive or negative) that confidence in an initial encouraging outcome evaporates. Similarly, a string of small effects in the same direction across a number of studies might solidify one's belief in the outcome; it has shown reliability across samples. Or a smaller number of large effects in a predicted direction might strengthen one's belief based on relatively few studies. Ultimately, one's confidence (or trust) in a finding should be rooted in the empirical data, not originate from reliance on the output of a ritualized power calculation with tenuous links to the data themselves.

Context

Motivated by improving the quality of scientific results, justification of sample size in study design using power analysis has come to be considered best practice. Methodological developments in power analysis encourage incorporating sources of uncertainty to yield more realistic calculations of power (e.g., Anderson et al., 2017; McShane & Böckenholt, 2016; Pek & Park, 2019; Perugini et al., 2014). This paper demonstrates that when two sources of uncertainty (sampling variability and random effects) are incorporated into power calculations, sample size determination for future experiments becomes highly imprecise. Thus, power-based decisions should acknowledge this uncertainty and power calculations should play less of a role in sample size justifications (see also Pek et al., in press; Wegener et al., 2022).

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*(1), 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850. https://doi.org/10.1177/0146167218798821

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology (with discussion). *European Journal of Personality*, *27*(2), 108–144. https://doi.org/10.1002/per.1919

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education: An International Journal on the Teaching and Learning of Statistics*, *10*(3), 11910677. https://doi.org/10.1080/10691898.2002.11910677

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic Press.

Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research*, *51*(5), 589–605. https://doi.org/10.1080/00273171.2016.1191324

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological & Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*(1), 3–12. https://doi.org/10.1177/1088868313507536

Gelman, A. (2019a). Comment on "post-hoc power using observed estimate of effect size is too noisy to be useful." *Annals of Surgery*, *270*(2), e64. https://doi.org/10.1097/SLA.0000000000003089

Gelman, A. (2019b). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, *269*(1), e9–e10. https://doi.org/10.1097/SLA.0000000000002908

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Greenland, S. (2012). Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology*, *22*(5), 364–368. https://doi.org/10.1016/j.annepidem.2012.02.007

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557–570. https://doi.org/10.1037/met0000189

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (3rd ed.). Routledge. https://doi.org/10.4324/9781315744131

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*(3), 160–167. https://doi.org/10.1037/h0047595

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Korn, E. L. (1990a). Projecting power from a previous study: Maximum likelihood estimation. *The American Statistician*, *44*(4), 290–292. https://doi.org/10.2307/2684350

Korn, E. L. (1990b). Projection from previous studies. A caution. *Controlled Clinical Trials*, *11*(1), 67–69. https://doi.org/10.1016/0197-2456(90)90033-X

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, *63*(5), 484–489. https://doi.org/10.1001/archpsyc.63.5.484

Lakens, D. (2022). Sample size justification. *Collabra Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lenth, R. V. (2007). *Post hoc power: Tables and commentary (TR378)*. University of Iowa. https://stat.uiowa.edu/sites/stat.uiowa.edu/files/techrep/tr378.pdf

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*(5), 626–629. https://doi.org/10.1016/j.jpsychires.2010.10.008

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

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*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625. https://doi.org/10.1177/1745691614548513

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods*, *21*(1), 47–60. https://doi.org/10.1037/met0000036

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2020). Average power: A cautionary note. *Advances in Methods and Practices in Psychological Science*, *3*(2), 185–199. https://doi.org/10.1177/2515245920902370

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518. https://doi.org/10.1037/a0016104

O'Hagan, A. (2004). Dicing with the unknown. *Significance*, *1*(3), 132–133. https://doi.org/10.1111/j.1740-9713.2004.00050.x

O'Hagan, A., Stevens, J. W., & Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, *4*(3), 187–201. https://doi.org/10.1002/pst.175

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Pek, J., Hoisington-Shaw, K. J., & Wegener, D. T. (in press). Avoiding questionable research practices surrounding statistical power analysis. In W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding questionable practices in applied psychology*. Springer. https://doi.org/10.1007/978-3-031-04968-2_11

Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods*, *24*(5), 590–605. https://doi.org/10.1037/met0000208

Pek, J., Wegener, D. T., & McClelland, G. H. (2020). Signal detection continues to be part of science. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(24), 13199–13200. https://doi.org/10.1073/pnas.2005860117

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519

Rothman, K. J., & Greenland, S. (2018). Planning study size based on precision rather than power. *Epidemiology*, *29*(5), 599–603. https://doi.org/10.1097/EDE.0000000000000876

Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: Cross-cultural and multimethod studies. *Journal of Personality and Social Psychology*, *89*(6), 1010–1028. https://doi.org/10.1037/0022-3514.89.6.1010

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *BMJ*, *325*(7375), 1304. https://doi.org/10.1136/bmj.325.7375.1304

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. https://doi.org/10.1177/1745691614528518

Stark, P. B., & Saltelli, A. (2018). Cargo-cult statistics and scientific crisis. *Significance*, *15*(4), 40–43. https://doi.org/10.1111/j.1740-9713.2018.01174.x

van Erp, S., Verhagen, J., Grasman, R. P., & Wagenmakers, E. J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data*, *5*(1), 4. https://doi.org/10.5334/jopd.33

Wegener, D. T., Fabrigar, L. R., Pek, J., & Hoisington-Shaw, K. (2022). Evaluating research in personality and social psychology: Considerations of statistical power and concerns about false findings. *Personality and Social Psychology Bulletin*, *48*(7), 1105–1117. https://doi.org/10.1177/01461672211030811

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*(2), 141–167. https://doi.org/10.3102/10769986030002141

Received November 5, 2021
Revision received June 2, 2022
Accepted June 7, 2022 ■