aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–11
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614551642
pps.sagepub.com
SAGE

Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors

Andrew Gelman[1] and John Carlin[2,3]
[1]Department of Statistics and Department of Political Science, Columbia University; [2]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria, Australia; and [3]Department of Paediatrics and School of Population and Global Health, University of Melbourne

Abstract
Statistical power analysis provides the conventional approach to assess error rates when designing a research study. However, power analysis is flawed in that a narrow emphasis on statistical significance is placed as the primary focus of study design. In noisy, small-sample settings, statistically significant results can often be misleading. To help researchers address this problem in the context of their own studies, we recommend design calculations in which (a) the probability of an estimate being in the wrong direction (*Type S [sign] error*) and (b) the factor by which the magnitude of an effect might be overestimated (*Type M [magnitude] error* or *exaggeration ratio*) are estimated. We illustrate with examples from recent published research and discuss the largest challenge in a design calculation: coming up with reasonable estimates of plausible effect sizes based on external information.

Keywords
design calculation, exaggeration ratio, power analysis, replication crisis, statistical significance, Type M error, Type S error

You have just finished running an experiment. You analyze the results, and you find a significant effect. Success! But wait—how much information does your study really give you? How much should you trust your results? In this article, we show that when researchers use small samples and noisy measurements to study small effects—as they often do in psychology as well as other disciplines—a significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an effect.

In this article, we examine some critical issues related to power analysis and the interpretation of findings arising from studies with small sample sizes. We highlight the use of external information to inform estimates of true effect size and propose what we call a *design analysis*—a set of statistical calculations about what could happen under hypothetical replications of a study—that focuses on estimates and uncertainties rather than on statistical significance.

As a reminder, the power of a statistical test is the probability that it correctly rejects the null hypothesis. For any experimental design, the power of a study depends on sample size, measurement variance, the number of comparisons being performed, and the size of the effects being studied. In general, the larger the effect, the higher the power; thus, power calculations are performed conditionally on some assumption of the size of the effect. Power calculations also depend on other assumptions, most notably the size of measurement error, but these are typically not so difficult to assess with available data.

It is of course not at all new to recommend the use of statistical calculations on the basis of prior guesses of effect sizes to inform the design of studies. What is new about the present article is as follows:

1. We suggest that design calculations be performed after as well as before data collection and analysis.

Corresponding Author:
Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, New York, NY 10027
E-mail: gelman@stat.columbia.edu

2. We frame our calculations not in terms of Type 1 and Type 2 errors but rather Type S (sign) and Type M (magnitude) errors, which relate to the probability that claims with confidence have the wrong sign or are far in magnitude from underlying effect sizes (Gelman & Tuerlinckx, 2000).

Design calculations, whether prospective or retrospective, should be based on realistic external estimates of effect sizes. This is not widely understood because it is common practice to use estimates from the current study's data or from isolated reports in the literature, both of which can overestimate the magnitude of effects.

The idea that published effect-size estimates tend to be too large, essentially because of publication bias, is not new (Hedges, 1984; Lane & Dunlap, 1978; for a more recent example, also see Button et al., 2013). Here, we provide a method to apply to particular studies, making use of information specific to the problem at hand. We illustrate with recent published studies in biology and psychology and conclude with a discussion of the broader implications of these ideas.

One practical implication of realistic design analysis is to suggest larger sample sizes than are commonly used in psychology. We believe that researchers typically think of statistical power as a trade-off between the cost of performing a study (acutely felt in a medical context in which lives can be at stake) and the potential benefit of making a scientific discovery (operationalized as a statistically significant finding, ideally in the direction posited). The problem, though, is that if sample size is too small, in relation to the true effect size, then what appears to be a win (statistical significance) may really be a loss (in the form of a claim that does not replicate).

Conventional Design or Power Calculations and the Effect-Size Assumption

The starting point of any design calculation is the postulated effect size because, of course, the true effect size is not known. We recommend thinking of the true effect as that which would be observed in a hypothetical infinitely large sample. This framing emphasizes that the researcher needs to have a clear idea of the population of interest: The hypothetical study of very large (effectively infinite) sample size should be imaginable in some sense.

How do researchers generally specify effect sizes for power calculations? As detailed in numerous texts and articles, there are two standard approaches:

1. *Empirical*: assuming an effect size equal to the estimate from a previous study (if performed prospectively, in which case the target sample size is generally specified such that a desirable level of power is achieved) or from the data at hand (if performed retrospectively).
2. *On the basis of goals*: assuming an effect size deemed to be substantively important or more specifically the minimum effect that would be substantively important.

We suggest that both of these conventional approaches are likely to lead either to performing studies that are too small or to misinterpreting study findings after completion. Effect-size estimates based on preliminary data (either within the study or elsewhere) are likely to be misleading because they are generally based on small samples, and when the preliminary results appear interesting, they are most likely biased toward unrealistically large effects (by a combination of selection biases and the play of chance; Vul, Harris, Winkelman, & Pashler, 2009). Determining power under an effect size considered to be of "minimal substantive importance" can also lead to specifying effect sizes that are larger than what is likely to be the true effect.

After data have been collected, and a result is in hand, statistical authorities commonly recommend against performing power calculations (see, e.g., Goodman & Berlin, 1994; Lenth, 2007; Senn, 2002). Hoenig and Heisey (2001) wrote, "Dismayingly, there is a large, current literature that advocates the inappropriate use of post-experiment power calculations as a guide to interpreting tests with statistically nonsignificant results" (p. 19). As these authors have noted, there are two problems with retrospective power analysis as it is commonly done:

1. Effect size—and thus power—is generally overestimated, sometimes drastically so, when computed on the basis of statistically significant results.
2. From the other direction, post hoc power analysis often seems to be used as an alibi to explain away nonsignificant findings.

Although we agree with these critiques, we find retrospective design analysis to be useful, and we recommend it in particular when apparently strong (statistically significant) evidence for nonnull effects has been found. The key differences between our proposal and the usual retrospective power calculations that are deplored in the statistical literature are (a) that we are focused on the sign and direction of effects rather than on statistical significance and, most important, (b) that we base our design analysis (whether prospective or retrospective) on an effect size that is determined from literature review or other information external to the data at hand.

Our Recommended Approach to Design Analysis

Suppose you perform a study that yields an estimate d with standard error s. For concreteness, you may think of d as the estimate of the mean difference in a continuous outcome measure between two treatment conditions, but the discussion applies to any estimate of a well-defined population parameter. The standard procedure is to report the result as statistically significant if $p < .05$ (which in many situations would correspond approximately to finding that $|d/s| > 2$) and inconclusive (or as evidence in favor of the null hypothesis) otherwise.[1]

The next step is to consider a true effect-size D (the value that d would take if observed in a very large sample), hypothesized on the basis of *external* information (other available data, literature review, and modeling as appropriate to apply to the problem at hand). We then define the random variable d^{rep} to be the estimate that would be observed in a hypothetical replication study with a design identical to that used in the original study.

Our analysis does not involve elaborate mathematical derivations, but it does represent a conceptual leap by introducing the hypothetical d^{rep}. This step is required so that general statements about the design of a study—the relation between the true effect size and what can be learned from the data—can be made without relying on a particular, possibly highly noisy, point estimate. Calculations in which d^{rep} is used will reveal what information can be learned from a study with a given design and sample size and will help to interpret the results, statistically significant or otherwise.

We consider three key summaries based on the probability model for d^{rep}:

1. The *power*: the probability that the replication d^{rep} is larger (in absolute value) than the critical value that is considered to define "statistical significance" in this analysis.
2. The *Type S error rate*: the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero.
3. The *exaggeration ratio* (expected Type M error): the expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero.

We have implemented these calculations in an R function, retrodesign(). The inputs to the function are D (the hypothesized true effect size), s (the standard error of the estimate), α (the statistical significance threshold; e.g., .05), and df (the degrees of freedom). The function returns three outputs: the power, the Type S error rate, and the exaggeration ratio, all computed under the assumption that the sampling distribution of the estimate is t with center D, scale s, and dfs.[2]

We sketch the elements of our approach in Figure 1. The design analysis can be performed before or after data collection and analysis. Given that the calculations require external information about effect size, one might wonder why a researcher would ever do them after conducting a study, when it is too late to do anything about potential problems. Our response is twofold. First, it is indeed preferable to do a design analysis ahead of time, but a researcher can analyze data in many different ways—indeed, an important part of data analysis is the discovery of unanticipated patterns (Tukey, 1977) so that it is unreasonable to suppose that all potential analyses could have been determined ahead of time. The second reason for performing postdata design calculations is that they can be a useful way to interpret the results from a data analysis, as we next demonstrate in two examples.

What is the relation among power, Type S error rate, and exaggeration ratio? We can work this out for estimates that are unbiased and normally distributed, which can be a reasonable approximation in many settings, including averages, differences, and linear regression.

It is standard in prospective studies in public health to require a power of 80%, that is, a probability of 0.8 that the estimate will be statistically significant at the 95% level, under some prior assumption about the effect size. Under the normal distribution, the power will be 80% if the true effect is 2.8 standard errors away from zero. Running retrodesign() with $D = 2.8$, $s = 1$, $\alpha = .05$, and df = infinity, we get power = 0.80, a Type S error rate of 1.2×10^{-6}, and an expected exaggeration factor of 1.12. Thus, if the power is this high, we have nothing to worry about regarding the direction of any statistically significant estimate, and the overestimation of the magnitude of the effect will be small.

However, studies in psychology typically do not have 80% power for two reasons. First, experiments in psychology are relatively inexpensive and are subject to fewer restrictions compared with medical experiments in which funders typically require a minimum level of power before approving a study. Second, formal power calculations are often optimistic, partly in reflection of researchers' positive feelings about their own research hypotheses and partly because, when a power analysis is required, there is a strong motivation to assume a large effect size, as this results in a higher value for the power that is computed.

Figure 2 shows the Type S error rate and exaggeration ratio for unbiased estimates that are normally distributed for studies with power ranging from 0 to 1. Problems with the exaggeration ratio start to arise when power is

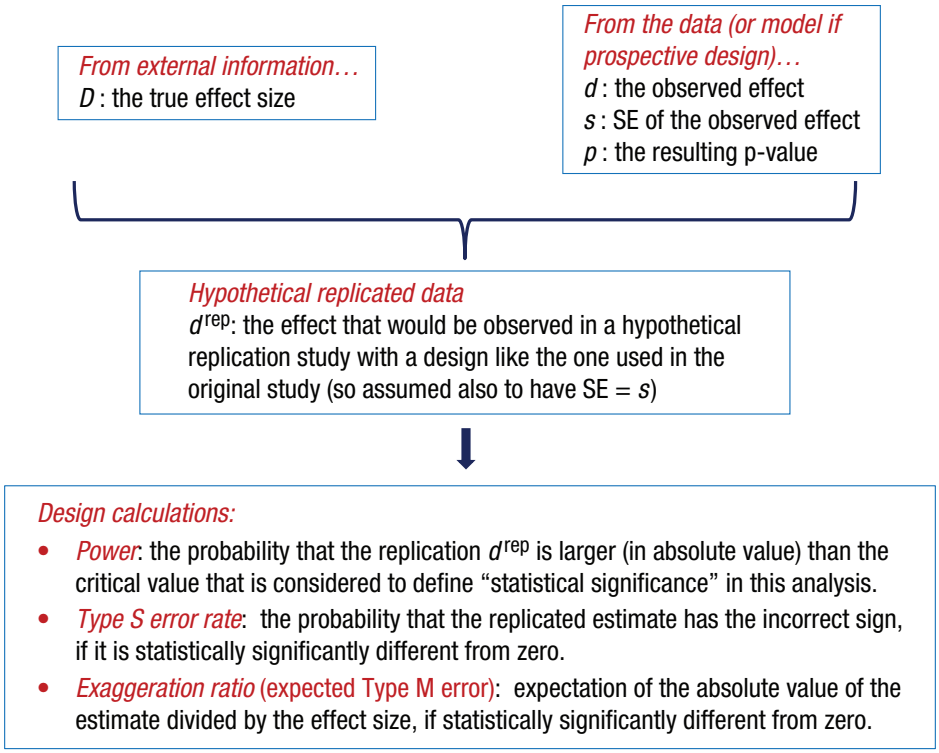

Figure 1. Diagram of our recommended approach to design analysis. It will typically make sense to consider different plausible values of D, the assumed true effect size.

less than 0.5, and problems with the Type S error rate start to arise when power is less than 0.1. For reference, an unbiased estimate will have 50% power if the true effect is 2 standard errors away from zero, it will have 17% power if the true effect is 1 standard error away from 0, and it will have 10% power if the true effect is 0.65 standard errors away from 0. All these are possible in psychology experiments with small samples, high

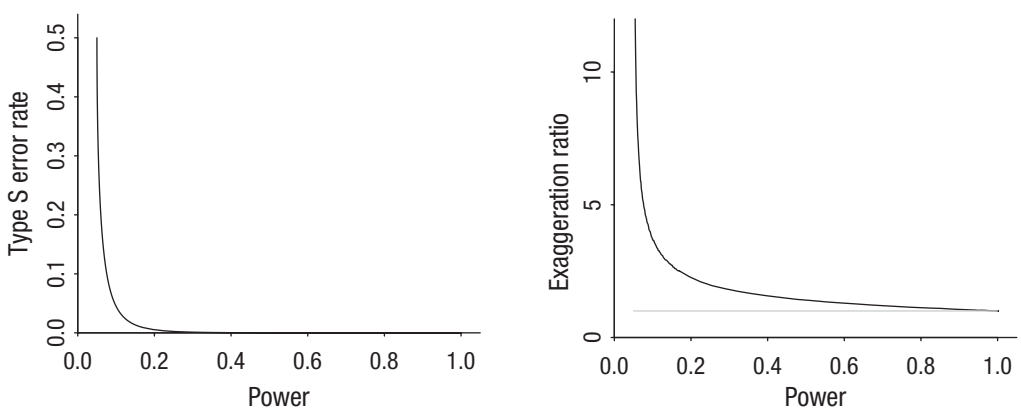

Figure 2. Type S error rate and exaggeration ratio as a function of statistical power for unbiased estimates that are normally distributed. If the estimate is unbiased, the power must be between 0.05 and 1.0, the Type S error rate must be less than 0.5, and the exaggeration ratio must be greater than 1. For studies with high power, the Type S error rate and the exaggeration ratio are low. But when power gets much below 0.5, the exaggeration ratio becomes high (that is, statistically significant estimates tend to be much larger in magnitude than true effect sizes). And when power goes below 0.1, the Type S error rate becomes high (that is, statistically significant estimates are likely to be the wrong sign).

variation (such as arises naturally in between-subjects designs), and small effects.

Example: Beauty and Sex Ratios

We first developed the ideas in this article in the context of a finding by Kanazawa (2007) from a sample of 2,972 respondents from the National Longitudinal Study of Adolescent Health.

This is not a small sample size by the standards of psychology; however, in this case, the sizes of any true effects are so small (as we discuss later) that a much larger sample would be required to gather any useful information.

Each of the people surveyed had been assigned an "attractiveness" rating on a 1–5 scale and then, years later, had at least one child. Of the first-born children of the parents in the most attractive category, 56% were girls, compared with 48% in the other groups. The author's focus on this particular comparison among the many others that might have been made may be questioned (Gelman, 2007). For the purpose of illustration, however, we stay with the estimated difference of 8 percentage points with a p value .015—hence, statistically significant at the conventional 5% level. Kanazawa (2007) followed the usual practice and just stopped right here, claiming a novel finding.

We go one step further, though, and perform a design analysis. We need to postulate an effect size, which will not be 8 percentage points. Instead, we hypothesize a range of true effect sizes using the scientific literature:

> There is a large literature on variation in the sex ratio of human births, and the effects that have been found have been on the order of 1 percentage point (for example, the probability of a girl birth shifting from 48.5 percent to 49.5 percent). Variation attributable to factors such as race, parental age, birth order, maternal weight, partnership status and season of birth is estimated at from less than 0.3 percentage points to about 2 percentage points, with larger changes (as high as 3 percentage points) arising under economic conditions of poverty and famine. That extreme deprivation increases the proportion of girl births is no surprise, given reliable findings that male fetuses (and also male babies and adults) are more likely than females to die under adverse conditions. (Gelman & Weakliem, 2009, p. 312)

Given the generally small observed differences in sex ratios as well as the noisiness of the subjective attractiveness rating used in this particular study, we expect any true differences in the probability of girl birth to be well under 1 percentage point. It is standard for prospective design analyses to be performed under a range of assumptions, and we do the same here, hypothesizing effect sizes of 0.1, 0.3, and 1.0 percentage points. Under each hypothesis, we consider what might happen in a study with sample size equal to that of Kanazawa (2007).

Again, we ignore multiple comparisons issues and take the published claim of statistical significance at face value: From the reported estimate of 8% and p value of .015, we can deduce that the standard error of the difference was 3.3%. Such a result is statistically significant only if the estimate is at least 1.96 standard errors from zero; that is, the estimated difference in proportion of girls, comparing beautiful parents with others, would have to be more than 6.5 percentage points or less than −6.5 percentage points.

The results of our proposed design calculations for this example are displayed in the Appendix for three hypothesized true effect sizes. If the true difference is 0.1% or −0.1% (probability of girl births differing by 0.1 percentage points, comparing attractive with unattractive parents), the data will have only a slightly greater chance of showing statistical significance in the correct direction (2.7%) than in the wrong direction (2.3%). Conditional on the estimate being statistically significant, there is a 46% chance it will have the wrong sign (the Type S error rate), and in expectation the estimated effect will be 77 times too high (the exaggeration ratio). If the result is not statistically significant, the chance of the estimate having the wrong sign is 49% (not shown in the Appendix; this is the probability of a Type S error conditional on nonsignificance)—so that the direction of the estimate gives almost no information on the sign of the true effect. Even with a true difference of 0.3%, a statistically significant result has roughly a 40% chance of being in the wrong direction, and in any statistically significant finding, the magnitude of the true effect is overestimated by an expected factor of 25. Under a true difference of 1.0%, there would be a 4.9% chance of the result being statistically significantly positive and a 1.1% chance of a statistically significantly negative result. A statistically significant finding in this case has a 19% chance of appearing with the wrong sign, and the magnitude of the true effect would be overestimated by an expected factor of 8.

Our design analysis has shown that, even if the true difference was as large as 1 percentage point (which we are sure is much larger than any true population difference, given the literature on sex ratios as well as the evident noisiness of any measure of attractiveness), and even if there were no multiple comparison problems, the sample size of this study is such that a statistically significant result has a one-in-five chance of having the wrong sign, and the magnitude of the effect would be overestimated by nearly an order of magnitude.

Our retrospective analysis provided useful insight, beyond what was revealed by the estimate, confidence interval, and p value that came from the original data summary. In particular, we have learned that, under reasonable assumptions about the size of the underlying effect, this study was too small to be informative: From this design, any statistically significant finding is very likely to be in the wrong direction and is almost certain to be a huge overestimate. Indeed, we hope that if such calculations had been performed after data analysis but before publication, they would have motivated the author of the study and the reviewers at the journal to recognize how little information was provided by the data in this case.

One way to get a sense of required sample size here is to consider a simple comparison with n attractive parents and n unattractive parents, in which the proportion of girls for the two groups is compared. We can compute the approximate standard error of this comparison using the properties of the binomial distribution, in particular the fact that the standard deviation of a sample proportion is $\sqrt{(p \times (1 - p)/n)}$, and for probabilities p near 0.5, this standard deviation is approximately $0.5/\sqrt{n}$. The standard deviation of the difference between the two proportions is then $0.5 \times \sqrt{(2/n)}$. Now suppose we are studying a true effect of 0.001 (i.e., 0.1 percentage points), then we would certainly want the measurement of this difference to have a standard error of less than 0.0005 (so that the true effect is 2 standard errors away from zero). This would imply $0.5 \times \sqrt{(2/n)} < 0.0005$, or $n > 500{,}000$, which would require that the total sample size $2n$ would have to be at least a million. This number might seem at first to be so large as to be ridiculous, but recall that public opinion polls with 1,000 or 1,500 respondents are reported as having margins of error of around 3 percentage points.

It is essentially impossible for researchers to study effects of less than 1 percentage point using surveys of this sort. Paradoxically, though, the very weakness of the study design makes it difficult to diagnose this problem with conventional methods. Given the small sample size, any statistically significant estimate will be large, and if the resulting large estimate is used in a power analysis, the study will retrospectively seem reasonable. In our recommended approach, we escape from this vicious circle by using external information about the effect size.

Example: Menstrual Cycle and Political Attitudes

For our second example, we consider a recent article from *Psychological Science*. Durante, Arsena, and Griskevicius (2013) reported differences of 17 percentage points in voting preferences in a 2012 preelection study, comparing women in different parts of their menstrual cycle. However, this estimate is highly noisy for several reasons: The design is between- rather than within-persons, measurements were imprecise (on the basis of recalling the time since last menstrual period), and sample size was small. As a result, there is a high level of uncertainty in the inference provided by the data. The reported (two-sided) p value was .035, which from the tabulated normal distribution corresponds to a z statistic of $d/s = 2.1$, so the standard error is $17/2.1 = 8.1$ percentage points.

We perform a design analysis to get a sense of the information actually provided by the published estimate, taking the published comparison and p value at face value and setting aside issues such as measurement and selection bias that are not central to our current discussion. It is well known in political science that vote swings in presidential general election campaigns are small (e.g., Finkel, 1993), and swings have been particularly small during the past few election campaigns. For example, polling showed President Obama's support varying by only 7 percentage points in total during the 2012 general election campaign (Gallup Poll, 2012), and this is consistent with earlier literature on campaigns (Hillygus & Jackman, 2003). Given the lack of evidence for large swings among any groups during the campaign, one can reasonably conclude that any average differences among women at different parts of their menstrual cycle would be small. Large differences are theoretically possible, as any changes during different stages of the cycle would cancel out in the general population, but are highly implausible given the literature on stable political preferences. Furthermore, the menstrual cycle data at hand are self-reported and thus subject to error. Putting all this together, if this study was to be repeated in the general population, we would consider an effect size of 2 percentage points to be on the upper end of plausible differences in voting preferences.

Running this through our retrodesign() function, setting the true effect size to 2% and the standard error of measurement to 8.1%, the power comes out to 0.06, the Type S error probability is 24%, and the expected exaggeration factor is 9.7. Thus, it is quite likely that a study designed in this way would lead to an estimate that is in the wrong direction, and if "significant," it is likely to be a huge overestimate of the pattern in the population. Even after the data have been gathered, such an analysis can and should be informative to a researcher and, in this case, should suggest that, even aside from other issues (see Gelman, 2014), this statistically significant result provides only very weak evidence about the pattern of interest in the larger population.

As this example illustrates, a design analysis can require a substantial effort and an understanding of the relevant literature or, in other settings, some formal or informal meta-analysis of data on related studies. We return to this challenge later.

When "Statistical Significance" Does Not Mean Much

As the earlier examples illustrate, design calculations can reveal three problems:

1. Most obvious, a study with low power is unlikely to "succeed" in the sense of yielding a statistically significant result.
2. It is quite possible for a result to be significant at the 5% level—with a 95% confidence interval that entirely excludes zero—and for there to be a high chance, sometimes 40% or more, that this interval is on the wrong side of zero. Even sophisticated users of statistics can be unaware of this point—that the probability of a Type S error is not the same as the p value or significance level.[3]
3. Using statistical significance as a screener can lead researchers to drastically overestimate the magnitude of an effect (Button et al., 2013). We suspect that this filtering effect of statistical significance plays a large part in the decreasing trends that have been observed in reported effects in medical research (as popularized by Lehrer, 2010).

Design analysis can provide a clue about the importance of these problems in any particular case.[4] These calculations must be performed with a realistic hypothesized effect size that is based on prior information external to the current study. Compare this with the sometimes-recommended strategy of considering a minimal effect size deemed to be substantively important. Both these approaches use substantive knowledge but in different ways. For example, in the beauty-and-sex-ratio example, our best estimate from the literature is that any true differences are less than 0.3 percentage points in absolute value. Whether this is a substantively important difference is another question entirely. Conversely, suppose that a difference in this context was judged to be substantively important if it was at least 5 percentage points. We have no interest in computing power or Type S and Type M error estimates under this assumption because our literature review suggests that it is extremely implausible, so any calculations based on it will be unrealistic.

Statistics textbooks commonly give the advice that statistical significance is not the same as practical significance, often with examples in which an effect is clearly demonstrated but is very small (e.g., a risk ratio estimate between two groups of 1.003 with a standard error of 0.001). In many studies in psychology and medicine, however, the problem is the opposite: an estimate that is statistically significant but with such a large uncertainty that it provides essentially no information about the phenomenon of interest. For example, if the estimate is 3 with a standard error of 1, but the true effect is on the order of 0.3, we are learning very little. Calculations such as the positive predictive value (see Button et al., 2013) showing the posterior probability that an effect that has been claimed on the basis of statistical significance is true (i.e., in this case, a positive rather than a zero or negative effect) address a different though related set of concerns.

Again, we are primarily concerned with the sizes of effects rather than the accept/reject decisions that are central to traditional power calculations. It is sometimes argued that, for the purpose of basic (as opposed to applied) research, what is important is whether an effect is there, not its sign or how large it is. However, in the human sciences, real effects vary, and a small effect could well be positive for one scenario and one population and negative in another, so focusing on "present versus absent" is usually artificial. Moreover, the sign of an effect is often crucially relevant for theory testing, so the possibility of Type S errors should be particularly troubling to basic researchers interested in development and evaluation of scientific theories.

Hypothesizing an Effect Size

Whether considering study design and (potential) results prospectively or retrospectively, it is vitally important to synthesize all available external evidence about the true effect size. In the present article, we have focused on design analyses with assumptions derived from systematic literature review. In other settings, postulated effect sizes could be informed by auxiliary data, meta-analysis, or a hierarchical model. It should also be possible to perform retrospective design calculations for secondary data analyses. In many settings it may be challenging for investigators to come up with realistic effect-size estimates, and further work is needed on strategies to manage this as an alternative to the traditional "sample size samba" (Schulz & Grimes, 2005) in which effect size estimates are more or less arbitrarily adjusted to defend the value of a particular sample size.

Like power analysis, the design calculations we recommend require external estimates of effect sizes or population differences. Ranges of plausible effect sizes can be determined on the basis of the phenomenon being studied and the measurements being used. One concern here is that such estimates may not exist when one is conducting basic research on a novel effect.

When it is difficult to find any direct literature, a broader range of potential effect sizes can be considered. For example, heavy cigarette smoking is estimated to reduce life span by about 8 years (see, e.g., Streppel, Boshuizen, Ocke, Kok, & Kromhout, 2007). Therefore, if the effect of some other exposure is being studied, it would make sense to consider much lower potential

effects in the design calculation. For example, Chen, Ebenstein, Greenstone, and Li (2013) reported the results of a recent observational study in which they estimated that a policy in part of China has resulted in a loss of life expectancy of 5.5 years with a 95% confidence interval of [0.8, 10.2]. Most of this interval—certainly the high end—is implausible and is more easily explained as an artifact of correlations in their data having nothing to do with air pollution. If a future study in this area is designed, we think it would be a serious mistake to treat 5.5 years as a plausible effect size. Rather, we would recommend treating this current study as only one contribution to the literature and instead choosing a much lower, more plausible estimate. A similar process can be undertaken to consider possible effect sizes in psychology experiments by comparing with demonstrated effects on the same sorts of outcome measurements from other treatments.

Psychology research involves particular challenges because it is common to study effects whose magnitudes are unclear, indeed heavily debated (e.g., consider the literature on priming and stereotype threat as reviewed by Ganley et al., 2013), in a context of large uncontrolled variation (especially in between-subjects designs) and small sample sizes. The combination of high variation and small sample sizes in the literature imply that published effect-size estimates may often be overestimated to the point of providing no guidance to true effect size. However, Button et al. (2013) have provided a recent example of how systematic review and meta-analysis can provide guidance on typical effect sizes. They focused on neuroscience and summarized 49 meta-analyses, each of which provides substantial information on effect sizes across a range of research questions. To take just one example, Veehof, Oskam, Schreurs, and Bohlmeijer (2011) identified 22 studies providing evidence on the effectiveness of acceptance-based interventions for the treatment of chronic pain, among which 10 controlled studies could be used to estimate an effect size (standardized mean difference) of 0.37 on pain, with estimates also available for a range of other outcomes.

As stated previously, the true effect size required for a design analysis is never known, so we recommend considering a range of plausible effects. One challenge for researchers using historical data to guess effect sizes is that these past estimates will themselves tend to be overestimates (as also noted by Button et al., 2013), to the extent that the published literature selects on statistical significance. Researchers should be aware of this and make sure that hypothesized effect sizes are substantively plausible—using a published point estimate is not enough. If little is known about a potential effect size, then it would be appropriate to consider a broad range of scenarios, and that range will inform the reader of the article, so that a particular claim, even if statistically significant, only gets a strong interpretation conditional on the existence of large potential effects. This is, in many ways, the opposite of the standard approach in which statistical significance is used as a screener, and in which point estimates are taken at face value if that threshold is attained.

We recognize that any assumption of effect sizes is just that, an assumption. Nonetheless, we consider design analysis to be valuable even when good prior information is hard to find for three reasons. First, even a rough prior guess can provide guidance. Second, the requirement of design analysis can stimulate engagement with the existing literature in the subject-matter field. Third, the process forces the researcher to come up with a quantitative statement on effect size, which can be a valuable step forward in specifying the problem. Consider the example discussed earlier of beauty and sex ratio. Had the author of this study been required to perform a design analysis, one of two things would have happened: Either a small effect size consistent with the literature would have been proposed, in which case the result presumably would not have been published (or would have been presented as speculation rather than as a finding demonstrated by data), or a very large effect size would have been proposed, in which case the implausibility of the claimed finding might have been noticed earlier (as it would have been difficult to justify an effect size of, say, 3 percentage points given the literature on sex ratio variation).

Finally, we consider the question of data arising from small existing samples. A researcher using a prospective design analysis might recommend performing an $n = 100$ study of some phenomenon, but what if the study has already been performed (or what if the data are publicly available at no cost)? Here, we recommend either performing a preregistered replication (as in Nosek, Spies, & Motyl, 2013) or else reporting design calculations that clarify the limitations of the data.

Discussion

Design calculations surrounding null hypothesis test statistics are among the few contexts in which there is a formal role for the incorporation of external quantitative information in classical statistical inference. Any statistical method is sensitive to its assumptions, and so one must carefully examine the prior information that goes into a design calculation, just as one must scrutinize the assumptions that go into any method of statistical estimation.

We have provided a tool for performing design analysis given information about a study and a hypothesized population difference or effect size. Our goal in developing this software is not so much to provide a tool for routine use but rather to demonstrate that such

calculations are possible and to allow researchers to play around and get a sense of the sizes of Type S errors and Type M errors in realistic data settings.

Our recommended approach can be contrasted to existing practice in which *p* values are taken as data summaries without reference to plausible effect sizes. In this article, we have focused attention on the dangers arising from not using realistic, externally based estimates of true effect size in power/design calculations. In prospective power calculations, many investigators use effect-size estimates based on unreliable early data, which often suggest larger-than-realistic effects, or on the minimal substantively important concept, which also may lead to unrealistically large effect-size estimates, especially in an environment in which multiple comparisons or researcher *df*s (Simmons, Nelson, & Simonsohn, 2011) make it easy for researchers to find large and statistically significant effects that could arise from noise alone.

A design calculation requires an assumed effect size and adds nothing to an existing data analysis if the postulated effect size is estimated from the very same data. However, when design analysis is seen as a way to use prior information, and it is extended beyond the simple traditional power calculation to include quantities related to likely direction and size of estimate, we believe that it can clarify the true value of a study's data. The relevant question is not "What is the power of a test?" but rather is "What might be expected to happen in studies of this size?" Also, contrary to the common impression, retrospective design calculation may be more relevant for statistically significant findings than for nonsignificant findings: The interpretation of a statistically significant result can change drastically depending on the plausible size of the underlying effect.

The design calculations that we recommend provide a clearer perspective on the dangers of erroneous findings in small studies, in which "small" must be defined relative to the true effect size (and variability of estimation, which is particularly important in between-subjects designs). It is not sufficiently well understood that "significant" findings from studies that are underpowered (with respect to the true effect size) are likely to produce wrong answers, both in terms of the direction and magnitude of the effect. Critics have bemoaned the lack of attention to statistical power in the behavioral sciences for a long time: Notably, for example, Cohen (1988) reviewed a number of surveys of sample size and power over the preceding 25 years and found little evidence of improvements in the apparent power of published studies, foreshadowing the generally similar findings reported recently by Button et al. (2013). There is a range of evidence to demonstrate that it remains the case that too many small studies are done and preferentially published when "significant." We suggest that one reason for the continuing lack of real movement on this problem is the historic focus on power as a lever for ensuring statistical significance, with inadequate attention being paid to the difficulties of interpreting statistical significance in underpowered studies.

Because insufficient attention has been paid to these issues, we believe that too many small studies are done and preferentially published when "significant." There is a common misconception that if you happen to obtain statistical significance with low power, then you have achieved a particularly impressive feat, obtaining scientific success under difficult conditions.

However, that is incorrect if the goal is scientific understanding rather than (say) publication in a top journal. In fact, statistically significant results in a noisy setting are highly likely to be in the wrong direction and invariably overestimate the absolute values of any actual effect sizes, often by a substantial factor. We believe that there continues to be widespread confusion regarding statistical power (in particular, there is an idea that statistical significance is a goal in itself) that contributes to the current crisis of criticism and replication in social science and public health research, and we suggest that the use of the broader design calculations proposed here could address some of these problems.

Appendix

```
retrodesign <- function(A, s, alpha=.05, df=Inf, n.sims=10000){
  z <- qt(1-alpha/2, df)
  p.hi <- 1 - pt(z-A/s, df)
  p.lo <- pt(-z-A/s, df)
  power <- p.hi + p.lo
  typeS <- p.lo/power
  estimate <- A + s*rt(n.sims,df)
  significant <- abs(estimate) > s*z
  exaggeration <- mean(abs(estimate)[significant])/A
  return(list(power=power, typeS=typeS, exaggeration=exaggeration))
}
```

```
# Example: true effect size of 0.1, standard error 3.28, alpha=0.05
retrodesign(.1, 3.28)
# Example: true effect size of 2, standard error 8.1, alpha=0.05
retrodesign(2, 8.1)

# Producing Figures 2a and 2b for the Gelman and Carlin paper

D_range <- c(seq(0,1,.01),seq(1,10,.1),100)
n <- length(D_range)
power <- rep(NA, n)
typeS <- rep(NA, n)
exaggeration <- rep(NA, n)
for (i in 1:n){
  a <- retrodesign(D_range[i], 1)
  power[i] <- a$power
  typeS[i] <- a$typeS
  exaggeration[i] <- a$exaggeration
}

pdf("pow1.pdf", height=2.5, width=3)
par(mar=c(3,3,0,0), mgp=c(1.7,.5,0), tck=-.01)
plot(power, typeS, type="l", xlim=c(0,1.05), ylim=c(0,0.54), xaxs="i", yaxs="i",
  xlab="Power", ylab="Type S error rate", bty="l", cex.axis=.9, cex.lab=.9)
dev.off()

pdf("pow2.pdf", height=2.5, width=3)
par(mar=c(3,3,0,0), mgp=c(1.7,.5,0), tck=-.01)
plot(power, exaggeration, type="l", xlim=c(0,1.05), ylim=c(0,12), xaxs="i", yaxs="i",
  xlab="Power", ylab="Exaggeration ratio", bty="l", yaxt="n", cex.axis=.9, cex.lab=.9)
axis(2, c(0,5,10))
segments(.05, 1, 1, 1, col="gray")
dev.off()
```

Acknowledgments

We thank Eric Loken, Deborah Mayo, and Alison Ledgerwood for helpful comments; we also thank the Institute of Education Sciences, Department of Energy, National Science Foundation, and National Security Agency for partial support of this work.

Declaration of Conflicting Interests

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

Notes

1. See Broer et al. (2013) for a recent empirical examination of the need for context-specific significance thresholds to deal with the problem of multiple comparisons.
2. If the estimate has a normal distribution, then the power is $Pr(|d^{rep}/s| > 1.96) = Pr(d^{rep}/s > 1.96) + Pr(d^{rep}/s < -1.96) = 1 - \Phi(1.96 - D/s) + \Phi(-1.96 - D/s)$, where Φ is the normal cumulative distribution function. The Type S error rate is the ratio of the second term in this expression for power, divided by the sum of the two terms; for the normal distribution, this becomes the following probability ratio (assuming D is positive): $\Phi(-1.96 - D/s)/\{[1 - \Phi(1.96 - D/s)] + \Phi(-1.96 - D/s)\}$. The exaggeration ratio can be computed via simulation of the hypothesized sampling distribution, truncated to have absolute value greater than the specified statistical significance threshold.
3. For example, Froehlich (1999), who attempted to clarify p values for a clinical audience, described a problem in which the data have a one-sided tail probability of .46 (compared with a specified threshold for a minimum worthwhile effect) and incorrectly wrote, "In other words, there is a 46% chance that the true effect" exceeds the threshold (p. 236). The mistake here is to treat a sampling distribution as a Bayesian posterior distribution—and this is particularly likely to cause a problem in settings with small effects and small sample sizes (see also Gelman, 2013).
4. A more direct probability calculation can be performed with a Bayesian approach; however, in the present article, we are emphasizing the gains that are possible using prior information without necessarily using Bayesian inference.

References

Broer, L., Lill, C. M., Schuur, M., Amin, N., Roehr, J. T., Bertram, L., . . . van Duijn, C. M. (2013). Distinguishing true from false positives in genomic studies: *P* values. *European Journal of Epidemiology*, *28*, 131–138.

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

Chen, Y., Ebenstein, A., Greenstone, M., & Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences, USA*, *110*, 12936–12941.

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

Durante, K., Arsena, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, *24*, 1007–1016.

Finkel, S. E. (1993). Reexamining the "minimal effects" model in recent presidential campaign. *Journal of Politics*, *55*, 1–21.

Froehlich, G. W. (1999). What is the chance that this study is clinically significant? A proposal for Q values. *Effective Clinical Practice*, *2*, 234–239.

Gallup Poll (2012). *U.S. Presidential Election Center*. Retrieved from http://www.gallup.com/poll/154559/US-Presidential-Election-Center.aspx

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology*, *49*, 1886–1897.

Gelman, A. (2007). Letter to the editor regarding some papers of Dr. Satoshi Kanazawa. *Journal of Theoretical Biology*, *245*, 597–599.

Gelman, A. (2013). P values and statistical practice. *Epidemiology*, *24*, 69–72.

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

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*, 373–390.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power: Statistical challenges in the estimation of small effects. *American Scientist*, *97*, 310–316.

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, *121*, 200–206.

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

Hillygus, D. S., & Jackman, S. (2003). Voter decision making in election 2000: Campaign effects, partisan activation, and the Clinton legacy. *American Journal of Political Science*, *47*, 583–596.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*, 1–6.

Kanazawa, S. (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology*, *244*, 133–140.

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.

Lehrer, J. (2010, December 13). The truth wears off. *New Yorker*, pp. 52–57.

Lenth, R. V. (2007). Statistical power calculations. *Journal of Animal Science*, *85*, E24–E29.

Nosek, B. A., Spies, J. R., & Motyl, M. (2013). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *Lancet*, *365*, 1348–1353.

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *British Medical Journal*, *325*, Article 1304. doi:10.1136/bmj.325.7375.1304

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

Sterne, J. A., & Smith, G. D. (2001). Sifting the evidence—What's wrong with significance tests? *British Medical Journal*, *322*, 226–231.

Streppel, M. T., Boshuizen, H. C., Ocke, M. C., Kok, F. J., & Kromhout, D. (2007). Mortality and life expectancy in relation to long-term cigarette, cigar and pipe smoking: The Zutphen Study. *Tobacco Control*, *16*, 107–113.

Tukey, J. W. (1977). *Exploratory data analysis*. New York, NY: Addison-Wesley.

Veehof, M. M., Oskam, M.-J., Schreurs, K. M. G., & Bohlmeijer, E. T. (2011). Acceptance-based interventions for the treatment of chronic pain: A systematic review and meta-analysis. *Pain*, *152*, 533–542.

Vul, E., Harris, C., Winkelman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition (with discussion). *Perspectives on Psychological Science*, *4*, 274–290.