Beyond power calculations to a broader design analysis, prospective or retrospective, using external information[1]

Andrew Gelman[2] and John Carlin[3]

3 December 2013

Abstract. As classically defined, statistical power is the probability, conditional on some hypothesized effect size, of observing a statistically significant result in a particular sample. Statisticians often emphasize that power calculations should be performed only before, not after, data have been collected and analyzed. But, in noisy, small-sample settings, statistically significant results can often be misleading. To help researchers understand this problem in the context of their own studies, we recommend *design calculations* of the probability of an estimate being in the wrong direction ("Type S error") and the estimated amount by which the magnitude of an effect is overestimated ("exaggeration factor").

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

Introduction

In this paper we examine some critical issues in statistical power calculations. We highlight the importance of using external information to inform estimates of true effect size and propose a form of "design analysis" that focuses on estimates and uncertainties rather than on "statistical significance." Our approach is ultimately Bayesian in that we are using external information and we are interested in quantifying uncertainty. But formally it can be done in a Bayes or non-Bayesian framework, in the latter conditioning on assumed effect sizes rather than averaging over a distribution. In this paper we present the non-Bayesian version.

It is of course not at all new to recommend the use of statistical calculations based on prior guesses of effect sizes, to inform the design of studies. What is new about the present paper is, first, that we suggest such studies be performed after as well as before data collection and analysis; and, second, that we frame our calculations not in terms of type 1 and type 2 errors but rather type S and type M errors, which relate to the probability that claims with confidence have the wrong sign or are far in magnitude from underlying effect sizes. We also emphasize that retrospective design calculations be performed using external estimates of effect sizes rather than using point estimates from available data, which typically overestimate the magnitude of effects.

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

[2] Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, http://www.stat.columbia.edu/~gelman/

[3] Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute and School of Population & Global Health, University of Melbourne, Parkville, Victoria, Australia.

The practical implications of realistic design analysis are to require larger sample sizes than are commonly used, at least in many studies in the social and health sciences. We believe that researchers typically think of statistical power as a tradeoff between the cost of performing a study (particularly 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 an unreplicable claim that does not generalize to the larger population).

Power is often thought of in a cost-benefit setting: the desire to obtain a statistically significant result is balanced against the cost of gathering a larger sample. But there is another potential cost. A low-power study is not merely wasteful of resources; it can also mislead. With a low-power study, in the improbable event that the data attain statistical significance, it is likely that the result is just noise (as recently emphasized by Button et al., 2013).

In this paper we first examine the conventional approach to power calculations, highlighting the popularity of retrospective calculations, especially in the presence of "non-significant" study findings. The critical issue in all power calculations is the specification of the effect size under the alternative hypothesis. We suggest that failure to think clearly about this is the main cause of confusion with respect not only to power itself but also to the interpretation of study findings "after the event." With this motivation, we propose that in general it may be useful to plan a study or reflect on the results of a statistical analysis by performing what we call a *design calculation* which extends the conventional power analysis to consider further aspects of the sampling distribution of estimates under an assumed effect size. We illustrate the approach with an example based on a recent published study. We conclude with a discussion of the broader implications of these ideas.

Conventional design or power calculations and the effect size assumption

The starting point of any design calculation is the postulated effect size. As detailed in numerous texts and articles, power calculations commonly take one of the following approaches:

- Assume an effect size equal to the estimate from the data (if performed retrospectively) or from someone else's data (if performed prospectively, in which case the target sample size is generally specified such that a desirable level of power is achieved);

- Determine power under an effect size deemed to be substantively important or more specifically the minimum effect that would be substantively important (without a good argument that such effects are likely to be true); or

- Compute the effect size required to reach a specified power (given an available sample size).

The essence of a power calculation is to obtain the probability that a certain null-hypothesis-significance-testing result (traditionally a result for which p<0.05; see Broer et al., 2013, for a recent empirical examination of the need for context-specific thresholds) *would* arise *given* an assumption about the underlying true effect. Although the centrality of the assumption of the true effect is clear, the range of recommendations for specifying this effect, and the continuing appeal of retrospective power calculations, indicate that some researchers may not fully grasp its importance. We recommend that the best way to conceptualize the true effect is to think of the effect that would be observed in a hypothetical infinitely large sample. This framing also emphasizes that the researcher needs to have a clear idea of the population to which they seek to generalize: the hypothetical study of very large (effectively infinite) sample size should be imaginable in some sense.

We suggest that each of the approaches above is likely to lead either to performing studies that are too small or to misinterpreting study findings after completion, if these are couched in terms of "significance" or p-values more generally. Estimates of effect size obtained from 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 are "interesting" they are most likely biased towards unrealistically large effects (by the play of chance). Determining power under an effect size considered to be of "minimal substantive importance" also often leads to specifying effect sizes that are larger than what is likely to be the true effect, because researchers wrongly perceive that as long as the minimal effect "of interest" is "detectable" then it is not necessary to further consider whether the actual effect might be smaller than this. The third approach is invoked when the researcher has no control over the sample size, so could be considered relatively harmless, except that it may lead the researcher to overlook critical issues in data interpretation associated with low power.

Once data have been collected and a result is in hand, the relevance of a power calculation—a hypothetical probability calculation based on an assumed effect size that may or may not be somehow informed by the partial data already obtained—is unclear. Despite this, however, the appeal of retrospective power calculations does not seem to have waned since 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."

We have seen numerous requests for such calculations from various sources including the editors of major medical journals. The central flaw in the thinking behind such calculations is a failure to understand the concept of the true effect size and a consequent conflation of observed effect (from interim data) with true effect. The main motivation for performing such calculations appears to be to provide an alibi for non-significant results, by reassuring oneself that sample size was too small to discover what was being looked for, although another more perverse reason may be to reassure that a "negative" finding is well supported. We agree with many other authors that a simple confidence interval does a better job at the task of expressing the uncertainty of findings when considered on their own (e.g. Goodman and Berlin, 1994; Senn, 2002; Lenth, 2007).

However, we suggest that it is important to consider findings in context, to interpret the results of an experiment in light of prior information. In the next section we describe our proposal for performing a design analysis that is *based on an effect size that is determined from literature review or other information external to the data at hand.* We believe that our design analysis calculation is relevant both prospectively and retrospectively, and in the latter case is most relevant in the context of "statistically significant" results rather than less apparently conclusive findings.

Our recommended approach to design analysis

Suppose you perform a study that yields an estimate d with standard error s. For concreteness the reader 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 traditional standard procedure based on a hypothesis testing framework is to compute the ratio $z=d/s$ and report the result as statistically significant if $|z|>2$, and inconclusive (or as evidence in favor of the null hypothesis) otherwise.[4]

Our recommended design analysis (prospective or retrospective) requires that you use *external* information (other available data, literature review, and modeling as appropriate to extrapolate this additional information to apply to the problem at hand) to hypothesize a true effect size, D (the value that d would take if observed in a very large sample). In many settings this information will already have been gathered before the data have been collected and should be used in sample size planning where that is possible. In practice, though, hypothesized effect sizes are often set using the "minimal substantively important" criterion or indeed just as a matter of convenience (for example, an investigator who has a budget for a sample size of 150 and expects a 20% attrition rate will figure out the effect size needed to have 80% power with a sample size of 120, and then the power calculation is reported accordingly); hence if these calculations are being performed retrospectively we recommend that the researcher think anew about plausible effect sizes in this phase of the analysis.

As with traditional power analysis, it can make sense to consider several plausible effect sizes. From a Bayesian perspective, it would be natural to express the estimated effect and its uncertainty as a prior distribution, but in the present article we take a classical approach and consider a range of point assumptions.

Our design analysis is based on the distribution of the estimate d^{rep} from a hypothetical replicated study with effect size D and standard error s. There are three key summaries, all conditional on the hypothesized effect size:

- 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.

[4] The threshold for the z-ratio is not necessarily 2. It could be higher because of a t-distribution correction for degrees of freedom, a more stringent significance level, or a multiple comparisons correction. For our present purposes, however, the exact level of the statistical-significance cutoff is not important; all that matters is that some threshold is being used.

- The *Type S error rate*: the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero.

- The *exaggeration factor*: the expected (absolute) value of the estimate divided by the effect size, if it is statistically significantly different from zero.

If the estimate has a normal distribution, then as per standard calculations the power is the value $\Pr(|d^{\text{rep}}/s|>1.96) = \Pr(d^{\text{rep}}>1.96) + \Pr(d^{\text{rep}} < -1.96) = 1 - \Phi\,(1.96 - D/s)) + \Phi\,(-1.96 - D/s)$, where Φ is the normal cumulative distribution function.

The Type S error rate is the ratio of the second term in the above expression for power, divided by the sum of the two terms; for the normal distribution, this becomes the following probability ratio: $\Phi\,(-1.96 - D/\text{s}))/[(1 - \Phi\,(1.96 - D/\text{s})) + \Phi\,(-1.96 - D/\text{s})]$.

The exaggeration factor can be computed via simulation of the hypothesized sampling distribution, truncated to have absolute value greater than the specified statistical significance threshold.

We have implemented these calculations in an R function, `retrodesign()`. The arguments to the function are D (the hypothesized effect size), s (the standard error of the estimate), α (the statistical significance threshold, assumed above to be 0.05), and df (the degrees of freedom). The function returns a list with three items: the power, the type S error rate, and the exaggeration factor, all computed under the assumption that the sampling distribution of the estimate is t with center D, scale s, and df degrees of freedom. 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.

The above calculations can be performed before or after data collection and analysis. Given that the calculations require external information about effect size, one might ask why would a researcher ever do them after conducting a study, when it is too late to do anything about potential problems, rather than before? Our response is twofold. First, it indeed is 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 post-data design calculations is that they can be a useful way to interpret the results from a data analysis, as we demonstrate with an example.

Example: Beauty and sex ratios

We demonstrate our recommended approach with an example from Gelman and Weakliem (2009). The story begins with a finding by Kanazawa (2007) from a sample of 2972 respondents from the National Longitudinal Study of Adolescent Health. Each of these people had been assigned an “attractiveness” rating on a 1-5 scale and then, years later, had at least one child.

56% of the first-born children of the parents in the highest attractiveness category were girls, compared to 48% in the other groups. This observed difference of 8 percentage points has a standard error of 3.5 percentage points (based on the binomial model, which we agree is appropriate for these data) and is statistically significant at the conventional 5% level, with a p-value of 0.015.

This p-value has been questioned for reasons of multiple comparisons (Gelman, 2007). Instead of comparing attractiveness category 5 to categories 1,2,3,4, the researcher also had the equally reasonable options of comparing 4,5 to 1,2,3, or comparing 3,4,5 to 1,2, or comparing 2,3,4,5 to 1, or comparing 5 to 1, or comparing 4 and 5 to 1 and 2, or simply fitting a linear regression. It turns out that none of these other potential analyses produces a statistically significant p-value: thus, the p-value of 0.015 represents the winner among at least seven possible comparisons. The most reasonable summary of the attractiveness/sex-ratio pattern in these data is probably the linear regression, which estimates the difference in probability of a girl birth between parents who are one standard deviation more attractive than the mean and those who are one standard deviation less attractive at 0.047 (that is, 4.7 percentage points) with standard error 0.043.

For the purpose of illustration, however, we shall stick with the original estimate of 8 percentage points with p-value 0.015. Kanazawa (2007) followed the usual practice and just stopped right here. We will go one step further, though, and perform a design analysis.

We need to postulate an effect size, which will *not* be 8 percentage points, or even the 4.7 percentage points estimated from the regression. Instead, we form our assumptions in the same way that we recommend in a prospective design calculation, using the scientific literature. From Gelman and Weakliem (2009):

> 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.

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 population 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 shall 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 t-statistic of 2.44, we can

True underlying population difference, D	Power: Pr (p<0.05 | D)	Type S error rate: Pr (Type S error | D, p<.0.05)	Exaggeration factor: E (|Exaggeration| | D, p<0.05)
0.1% or -0.1%	5.0%	46%	77
0.3% or -0.3%	5.1%	40%	25
1.0% or -1.0%	6.1%	19%	8

Table 1: *Results of design calculations for the example on beauty and sex ratios. We consider three different magnitudes for the underlying population difference D. The largest possibility, ±1.0%, is highly implausible given the background literature on sex ratios and is included only as a comparison point.*

deduce that the standard error of the difference was 3.28%. Such a result will be statistically significant only if the estimate is at least 1.96 standard errors from zero; that is, the estimated difference in proportion girls, comparing beautiful parents to others, would have to be more than 6.43 percentage points or less than -6.43.

The results of our proposed design calculations for this example are displayed in Table 1, for three hypothesized true effect sizes. For a true difference of 0.1% or -0.1% (probability of girl births in the population is 0.1 percentage points different, comparing attractive to among unattractive parents), an unbiased estimate will have only a slightly greater chance of being statistically significantly in the correct direction (2.7%) than of being statistically significant and in the wrong direction (2.3%). Conditional on the estimate being statistically significant, there is a 46% chance it will be in the wrong direction (the Type S error rate) and, in expectation the estimated effect will be 77 times higher than the true effect (the exaggeration factor). If the result is not statistically significant, the chance of the estimate being the wrong sign is 49% percent (not shown in the table; this is the probability of a Type S error conditional on non-significance), so that the direction of the estimate would provide 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 any statistically significant finding overestimates the magnitude of the true effect 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 would overestimate the magnitude of the true effect by an expected factor of 8.

Our design analysis has shown that, *even if* the true difference were 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 comparisons 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 would overestimate the magnitude of the effect 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 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 reviewers at the journal to recognize how little information was provided by the data in this case.

The statistical significance filter: overestimating the magnitude of small effects

As the above example illustrates, underpowered studies suffer from three deficiencies:

1. Most obviously, a study with low power is unlikely to "succeed" in the sense of yielding a statistically significant result.

2. Any statistically significant findings are likely to be in the wrong direction. 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 chance of 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.[5]

3. The final problem with underpowered studies is that when they do happen to result in statistical significance, they often drastically overestimate the magnitude of any effect (Button et al., 2013). This filtering effect of statistical significance likely plays a large part in the decreasing trends that have been observed in the magnitudes of reported effects in medical research (as popularized by Lehrer, 2010).

Design analysis can give a clue as to how important these problems will be in any particular case. A more direct probability calculation can be obtained using the posterior distribution, but in the present article we have emphasized the gains that are possible using prior information without necessarily using Bayesian inference.

We emphasize that these calculations must be performed with a realistic hypothesized effect size that is based on prior information external to the current study. Compare this to 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

[5] For example, in a paper attempting to clarify p-values for a clinical audience, Froehlich (1999) describes a problem in which the data have a one-sided tail probability of 0.46 (compared to a specified threshold for a minimum worthwhile effect) and incorrectly writes: "In other words, there is a 46% chance that the true effect" exceeds the threshold. The mistake here is to treat a sampling distribution as a Bayesian posterior distribution—and is particularly likely to cause a problem in settings with small effects and small sample sizes (see also Gelman, 2013).

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 were judged to be substantively important if it were at least 5 percentage points. We have no interest in computing power under this assumption, since our literature review suggests 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 where an effect is clearly demonstrated but is very small (for example, an estimate of 0.003 with standard error 0.001). In many studies in psychology and medicine, however, the problem is the opposite: an estimate may be statistically significant but the uncertainty is so large 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 (PPV; see Button et al., 2013) showing the posterior probability of an effect being true (i.e., in this case, a positive rather than a zero or negative effect) address a different set of concerns.

Again, we are primarily concerned with the sizes of effects, rather than the accept/reject decisions that are central to power traditional power calculations (whether classical or Bayesian). It is sometimes argued that, for the purpose of research (as opposed to policy analysis), what is important is whether an effect is there, not its sign or how large it is. But 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.

Discussion

Design calculations surrounding null hypothesis test statistics are among the few places in which the incorporation of numerical prior incorporation has a place in classical 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. In this paper 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 degrees of freedom (Simmons et al., 2011) make it easy for researchers to find large and statistically significant effects that could arise from noise alone (see Bem, 2011, for a notorious recent example).

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. But when design analysis is seen as a way to use prior information, and is extended beyond the simple traditional power calculation to include quantities related to likely direction and size of estimate, we believe 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, "What might be expected to happen in studies of this size?" Also, contrary to the common impression, retrospective design calculation is not only appropriate but possibly more relevant for statistically-significant as for non-significant findings—as long as the analysis uses real external information.

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. The present article has focused on design analyses with assumptions derived from systematic literature review. In other settings, postulated effect sizes could be set using auxiliary data, meta-analysis, or a hierarchical model. It should also be possible to perform retrospective design calculations for secondary data analyses: for example, if subgroup analysis is performed in a randomized trial, one can apply a design analysis based on an external estimate of the magnitude of interactions. 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 and Grimes, 2005) in which effect size estimates are more or less arbitrarily adjusted in order 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. One concern here is that no such estimates may exist when one is conducting basic research on a novel effect. Our response is that, yes, such estimates are necessarily speculative—after all, if the true effect size were known, there would be no need to conduct an experiment in the first place—but that, in the examples we have seen, some sense of plausible magnitudes can be obtained based on some combination of theory and the literature on related studies.

To put it another way: In research studies where statistical significance is used to demonstrate the existence of a new effect, the case is never made from the p-value alone. The data are typically framed as being novel, but as being predicted by theory and consistent with an earlier literature. Ranges of plausible effect sizes can be determined based on the phenomenon being studied and the measurements being used. For example, Durante et al. (2013) reported differences of over 20 percentage points in vote preferences in a 2012 pre-election study, comparing women in different parts of their menstrual cycle. 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 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 and Jackman, 2003). Given the lack of evidence for large swings among any groups during the campaign, one can only suppose that any average differences between 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. Indeed, one could just as easily explain this as a differential nonresponse pattern: maybe liberal or conservative women in different parts of their cycle are more or less likely to participate in a survey. Beyond all this, the menstrual cycle data at hand are self-reported and thus subject to error. Putting all this together, we would consider an effect size of 2 percentage points to be on the upper end of plausible differences in vote preferences in that study.

As this example illustrates, a design analysis can require a bit of effort and an understanding of the relevant literature or, in other settings, some formal or informal meta-analysis of data on related studies. We believe this can often be done, if necessary in consultation with outside experts (for example, political scientists or survey researchers in the menstrual-cycle-and-voting example). 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 lifespan by about 8 years (see, e.g., Streppel et al., 2007). So if the effect of some existing environmental exposure is being studied, it would make sense to consider much lower potential effects in the design calculation. A similar process can be undertaken to consider possible effect sizes in psychology experiments, by comparing to demonstrated effects on the same sorts of outcome measurements from other treatments.

It is not necessary in a design calculation for the hypothesized effect size to be correct or even unbiased; indeed, we recommend performing the design analysis under several different plausible underlying effects. One challenge in using historical data to guess effect sizes is that these past estimates will themselves tend to overestimates, 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 very 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 paper, 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, especially in between-subject designs with high variability and small sample sizes. Second, the requirement of a design calculation can stimulate a more careful interaction with 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 performed a design calculation, 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 problem 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 prospective design calculation 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 recognizing the uncertainty and either performing a pure replication (as in Nosek, Spies, and Motyl, 2013) or else reporting design calculations that clarify the limitations of the data.

The design calculations that we recommend provide a clearer perspective on the dangers of erroneous findings in small studies, where “small” must be defined relative to the true effect size (and variability of estimation, which is particularly important in between-subject 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.

Because insufficient attention has been paid to these issues, we believe too many small studies are done and preferentially published when “significant.” The usual thinking is 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. But 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 can grossly overestimate the absolute values of any actual effect sizes. There is 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 replication in social science and public health research, and we believe that formal design calculations could address some of these problems.

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* 100, 407-25.

Broer, L., Lill, C. M., Schuur, M., Amin, N., Roehr, J. T., Bertram, L., Ioannidis, J. P. A., and 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., and Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews: Neuroscience* 14 (May), 1-12.

Durante, K., Arsena, A., and 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.

Francis, G. (2013). Replication, statistical consistency, and publication bias (with discussion). *Journal of Mathematical Psychology*.

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. http://www.gallup.com/poll/154559/US-Presidential-Election-Center.aspx

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

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

Goodman, S. N., and 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.

Hillygus, D. S., and 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., and 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.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association* 44, 1372-1381.

Lehrer, J. (2010). The truth wears off. *New Yorker,* 13 Dec, 52-57.

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

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

Schulz, K. F., and 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, 1304.

Simmons J., Nelson L., and 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., and 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., and 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: Addison-Wesley.

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