Published in final edited form as:
*Curr Psychol*. 2020 June ; 39(3): 870–877. doi:10.1007/s12144-018-0018-1.

# The Interpretation of Statistical Power after the Data have been Gathered

**John J. Dziak**[1], **Lisa C. Dierker**[1,2], **Beau Abar**[3]

[1]The Methodology Center, The Pennsylvania State University, University Park, PA, USA

[2]Department of Psychology, Wesleyan University, Middletown, CT, USA

[3]Department of Emergency Medicine, School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, NY

## Abstract

Post-hoc power estimates (power calculated for hypothesis tests after performing them) are sometimes requested by reviewers in an attempt to promote more rigorous designs. However, they should never be requested or reported because they have been shown to be logically invalid and practically misleading. We review the problems associated with post-hoc power, particularly the fact that the resulting calculated power is a monotone function of the *p*-value and therefore contains no additional helpful information. We then discuss some situations that seem at first to call for post-hoc power analysis, such as attempts to decide on the practical implications of a null finding, or attempts to determine whether the sample size of a secondary data analysis is adequate for a proposed analysis, and consider possible approaches to achieving these goals. We make recommendations for practice in situations in which clear recommendations can be made, and point out other situations where further methodological research and discussion are required.

### Keywords

power; post-hoc; equivalence; replicability; null hypothesis; exploratory data analysis

The importance of *prospectively* considering statistical power when choosing the sample size for a planned future study is well known (see, e.g. Cohen 1988, 1990; Button et al. 2013). However, whether and how to assess power *post-hoc*, when the data have already been gathered, is more controversial. This article will review why it is invalid to try to determine the power of a test after having performed the test. We begin with a brief review of prospective power, the form of power analysis that is generally agreed to be valid and highly recommended. We then consider post-hoc power analysis and explain why it should

Correspondence concerning this article can be addressed to John Joseph Dziak, The Methodology Center, 408 Health and Human Development, The Pennsylvania State University, University Park, PA, 16801, USA: jjd264@psu.edu; 1-814-863-9806.

Conflict of Interest Statement
On behalf of all authors, the corresponding author states that there are no conflicts of interest.

Compliance with Ethical Standards

**Ethical approval**: This article does not contain any studies with human participants or animals performed by any of the authors.

**Informed Consent**: Informed consent is not required on this study.

never be used in practice, following explanations given by Hoenig and Heisey (2001), and discuss better alternatives such as confidence intervals. We follow by describing some situations that fall in between these extremes, in which power analysis is potentially meaningful but requires special care and consideration, and we point out areas where future methodological research is required. Thus, the paper is intended to serve both as a tutorial overview of existing work on post-hoc power analysis, and as a call for further work and discussion.

## A Valid Use of Statistical Power: Prospective Power Analysis

Statistical power is defined as the probability of rejecting a null hypothesis ($H_0$), assuming that it is false, and given additional assumptions about the true values of population parameters (see, e.g., Cohen 1988, 1992; Norcross, Hogan, Koocher & Maggio, 2017). It differs for different study designs and different statistical tests; for example, it can sometimes be improved by using pretests or repeated measures (see Guo et al., 2013; Vickers and Altman, 2001). However, for a given study design and a given analysis plan, power depends mainly on effect size and sample size, so we focus on these two factors for simplicity.

To review the ideas behind formulas for calculating power and sample size, we work through a simple case, that of a *z*-test for comparing the means of two independent samples. Readers who are primarily interested in heuristic ideas rather than mathematical details, or who are already familiar with power formulas, may skim over the following subsection.

### Power Formulas

Consider a comparison of two sample means $\overline{Y}_1$ and $\overline{Y}_2$ , both samples having size *n*, for a total size of 2*n*. Suppose the investigator intends to do a one-sided test at $\alpha = .05$ of $H_0$: $\mu_1 \le \mu_2$, versus the alternative $H_1$: $\mu_1 > \mu_2$, where $\mu_1$ and $\mu_2$ are the population means of the variable of interest in the respective populations. Suppose that in fact, $\mu_1 - \mu_2$ equals a positive constant Δ. Finally, for simplicity, suppose that both populations are approximately normally distributed with the same error variance $\sigma^2$, and that the sample size will be large enough to ignore uncertainty in estimating $\sigma^2$ . The two-sample *z*-test statistic, which is a simpler approximate form of the well-known two-sample *t*-test, can then be used:

$$Z = \frac{\overline{Y}_1 - \overline{Y}_2}{\frac{\sqrt{2\sigma^2}}{\sqrt{n}}} \tag{1}$$

where the numerator is the difference in sample means and the denominator is the standard error of the difference. Since, by assumption, $\Delta > 0$, the $H_0$ is in fact false. The critical *z*-score is $\Phi^{-1}(1 - \alpha)$, where $\Phi(z)$ is the area to the left of *z* on a standard normal curve. In particular, $z_{\text{crit}} = \Phi^{-1}(0.95)$ is the number such that 95% of the area of the standard normal curve is to its left, and this is approximately 1.645. Power is the probability that *Z*, as defined in (1), exceeds $_{\text{crit}}$. From basic statistical theory, $Z - \Delta$ is normally distributed with mean 0 and standard error $\frac{\sqrt{n}\,\Delta}{\sqrt{2\sigma^2}}$ Thus,

$$\text{Power} = \Pr(|Z| > \ \Phi^{-1}\ (1-\alpha)) = 1 - \ \Phi\left(Z_{\text{crit}} - \frac{\sqrt{n}\ \Delta}{\sqrt{2\sigma^2}}\right). \tag{2}$$

Power will be greater than the false positive rate $\alpha$, to the extent that $\Delta > 0$. If a two-sided (two-tailed) test of $H_0$: $\mu_1 = \mu_2$ versus $H_1$: $\mu_1 \neq \mu_2$ is desired instead of a one-sided (one-tailed) test, then Expression (2) can still be used, with $z_{\text{crit}} = \Phi^{-1}(1 - \alpha)$ replaced by $z_{\text{crit}} = \Phi^{-1}(1 - 2\alpha)$. The one-sided test would reject $H_0$ only for differences in the expected direction. (In theory, the two-sided test would reject $H_0$ for either direction, but here we are counting only the probability of a $H_0$ rejection in the correct direction when calculating power. This is partly for simplicity, in that the rejection probability in the wrong direction is tiny when the rejection probability in the right direction is adequate, and also for practicality, in that rejecting $H_0$ is not a desirable outcome if it is rejected in the wrong direction.) Power is reviewed further in Cohen (1988, 1990, 1992) and Norcross, Hogan, Koocher and Maggio (2017).

Instead of asking how much power is obtained by a specified available sample size, another approach is to ask how what sample size is needed in order to obtain a desired amount of power. Some algebra can be used to re-express (2) as the usual sample size expression for obtaining the needed sample size for a balanced independent-sample *z*-test:

$$n = \left(\frac{\sqrt{2}\left(z_{\text{pow}} + z_{\text{crit}}\right)\sigma}{\Delta}\right)^2 \tag{3}$$

Where $z_{\text{pow} =}$ $\Phi^{-1}$(Power), e.g., $z_{\text{pow}} = 0.84$ for a desired power of 0.80.

## Choosing Values to Use in the Formula

To calculate power or required sample size using Expression (2) or (3), one must specify hypothetical values of the effect size $\Delta$ and the standard deviation $\sigma$. One approach is to use estimates of $\Delta$ and $\sigma$ obtained from an analogous study in the previously published literature. Although using these estimates is better than doing no power analysis at all, they are subject to sampling error and likely publication bias (see Anderson et al., 2017) and furthermore the implementation details of the previous studies are likely to differ somewhat from the current study. A similar approach is to conduct a small pilot study to obtain crude estimates of $\Delta$ and $\sigma$. This approach is also somewhat problematic, because a study that is too small to allow a statistical hypothesis to be tested with adequate power is probably also too small to provide reliable estimates of $\Delta$ and $\sigma$; that is, the standard errors of the estimates will be very high.

Expressions (2) and (3) depend only on the ratio $\Delta/\sigma$ (i.e., the standardized effect size), rather than or themselves. Thus, one could specify $\Delta/\sigma$ instead of trying to specify the parameters separately. The standardized effect size $\Delta/\sigma$ can be useful as a heuristic approach for describing the factors that contribute to power. Standard benchmark values for "low," "medium" and "high" effect sizes were reluctantly provided by Cohen (1988, 1992). For comparisons of means, a value of .2 for $d = \Delta/\sigma$ is commonly considered "small," .5 is "medium" and .8 is "large." If these are specified, the only remaining unknown in (2) is $n$. In particular, for a desired power of .80 and for an α level of .05, a sample size of about 310, 50 or 20 per group is necessary to detect a small, medium or large difference in means in a one-sided *z*-test, and a sample size of about *n*=393, 63, or 25 per group is necessary to detect a small, medium or large difference in means in a two-sided *z*-test. Almost the same sample sizes are required for a two-sided *t*-test. However, as Cohen himself pointed out, the meanings of "small," "medium" and "large" here are arbitrary, and may not apply well to all fields (see, e.g., Szucs and Ioannidis, 2017a). A difference of 0.20 standard deviation units in symptom-free days for patients with a serious illness such as sickle-cell anemia may be considered quite a large effect. Conceptualizing differences in standard deviation units may tend to obscure consideration of practical implications of an effect, relative to using interpretable real-world units. Furthermore, the standardized benchmarks arguably make power analysis too simplistic; for example, for 80% power and α =.05, the needed sample size per group for either a *z*-test or *t*-test will always be given as either 393, 63, or 25, depending only on whether the effect is assumed to be "small," "medium," or "large." Thus, a skeptical view of these standard benchmarks for small, medium or large *effect sizes* is that they are only disguised conventions for what social scientists consider as large, medium or small *sample sizes* (Lenth 2001).

It may be better to specify Δ by deciding on the minimum value $\Delta_1$ considered clinically or theoretically significant (Eng 2004; Lenth, 2001, Schulz & Grimes 2005), in the units of practical interest rather than in standard deviations. Although choosing a minimum practically significant effect seems subjective, it avoids the temptation to rely on optimistic guesses about Δ. That is, suppose the investigator hopes that Δ will be 40 raw units, but would consider the treatment successful if Δ is at least 20 raw units. Although assuming 20 units would lead to a larger recommended sample size, it would be a safer choice than gambling the outcome of the study on having an effect size of 40 units. An investigator can also use sensitivity analysis, that is, calculate power or sample size given a number of plausible values of Δ and $\sigma$, and then take multiple scenarios into consideration when making a final sample size recommendation.

## An Invalid Use of Statistical Power: Post-Hoc Power Calculation

Power calculations are very important in planning an empirical study to be done in the future. However, if the study has already been done then the meaning and utility of power becomes much less clear. A post-hoc power estimate calculated for a test that has already been performed is no longer easily interpreted as the probability of a desired future event. In the *z*-test example, suppose a researcher performs a study and obtains estimates for $\widehat{\Delta} = \bar{Y}_1 - \bar{Y}_2$ and for $\hat{\sigma}$ , and plugs these numbers into (2) in an attempt to estimate the power that the test must have had. This is called "post-hoc power" (PHP) or "observed (*sic*) power." This is likely to be done for one of two reasons: either to try to demonstrate to reviewers that the study was appropriately designed with respect to statistical power and replicability, or else to defend one's research hypothesis by explaining away a lack of observed statistically significant findings as having been due merely to inadequate power. Unfortunately, PHP is not valid for either of these purposes because of two problems which we review below.

The first problem is that an equation like (2) treats $\bar{Y}_1$ and $\bar{Y}_2$ as random variables, but after the study is done they are instead realized constants. Thus, power as a probability of rejecting the $H_0$ has its usual meaning only before $\bar{Y}_1$ and $\bar{Y}_2$ are observed. After the test, we still are not sure whether $H_0$ is true or false, but we do know whether $H_0$ was rejected or not in our sample: "If the ax chopped down the tree, it was sharp enough" (Goldstein 1964, p. 62; Kraemer and Thiemann 1987, p. 25). Presumably then, the answer that potential users of PHP really want is not the probability that they rejected $H_0$ (which they already know to be either zero or one), but the probability that they *should have* rejected $H_0$. This could be interpreted to mean one of the following: either an estimate of the probability that $H_0$ was actually false, or an estimate of the probability that a future sample from the same population would also reject $H_0$. Unfortunately, the probability that $H_0$ is false is undefined in classical statistics, because population parameters are assumed to be unknown constants rather than random variables. Such a probability can be explored using Bayesian statistics, but these are beyond the scope of the current discussion. In contrast, the question of whether a future sample of the same size from the same population would also reject $H_0$ is meaningful, and superficially seems to be a good way to address the vital concern of replicability of findings – yet it still cannot be usefully estimated via PHP. The second problem with PHP, and the reason why PHP does not actually measure replicability, can be presented either conceptually or algebraically.

## Heuristic Explanation of the Interpretational Problem

Heuristically, power is a characteristic of the sampling distribution (i.e., distribution of possible samples), rather than of a particular sample. One could think of each sample of size $n$ as a marble in a jar, marked with the $p$-value that would be calculated from this particular sample. The difficulty is that, regardless of the number $n$ of participants inside the sample, the number of samples an investigator has is generally 1. Thus, the user of PHP is in the unenviable position of having to make an inference about the jar of marbles, using information from only a single marble. In such a situation, the best the investigator could do would be to assume that whatever kind of marble was drawn must have been from the majority. Thus, if the observed $p$-value is less than $\alpha$, then a good provisional guess is that most of the $p$-values are less than $\alpha$. If the observed $p$-value is greater than $\alpha$, then a good provisional guess is that most of the $p$-values are greater. Following this logic, if the observed $p$-value is equal to $\alpha$, the best guess would be that half of the $p$-values are larger than $\alpha$ and half are smaller than $\alpha$. Thus, a $p$-value equal to $\alpha$ (e.g., .05) should lead to PHP of exactly 50%. Because the PHP is really the $p$-value in disguise, measuring the importance of a finding via the PHP is equivalent to measuring its importance via the $p$-value, a practice which is widely criticized (e.g., Cohen 1990, Demidenko, 2016; Longford, 2016; Wasserstein & Lazar, 2016). In this paper we do not address the controversy over whether or not it is meaningful for a point $H_0$ to be precisely true (e.g., for an effect to be exactly zero), especially for two-sided tests (see, e.g., Cohen 1994; Baril and Cannon 1995; Jones and Tukey 2000; Kirk 2007; Amrhein, et al., 2017; Norcross, Hogan, Koocher and Maggio, 2017); but either way the PHP would not indicate whether $H_0$ were true or not.

Perhaps in one somewhat bizarre situation PHP might be a meaningful quantity: specifically a situation in which an investigator wishes to predict the power of a future study with exactly the same sample size and design as the current study (Lenth 2007). However, even then, PHP would only be a very imprecise estimate of the future power (see Korn 1990. Yuan & Maxwell, 2005). This should not be surprising since it is an estimate based only on a single sample (marble from the jar of possible studies), regardless of the number of participants within this sample. Therefore, PHP is never a good measure of replicability.

### Algebraic Explanation of the Interpretational Problem

Algebra bears out the intuition that PHP is a function of the *p*-value. If the observed estimates of Δ and σ are substituted into expression (2), then power simplifies to $1 - \Phi(z_{\text{crit}} - Z)$ where $Z$ is the observed sample *z*-test statistic in (1). Recall that the *p*-value for this *z*-statistic is defined as $1 - \Phi(Z)$. Thus, after some algebra, the post-hoc power simplifies to $1 - \phi\left(z_{\text{crit}} - \Phi^{-1}(1 - p)\right)$ which is a somewhat complicated but still deterministic and monotone function of the *p*-value. Furthermore, if = .05, then $\Phi^{-1}(1 - p) = z_{\text{crit}}$, so the post-hoc power is $1 - \Phi(0) = 1/2$. Thus, calculating power for observed data provides no information beyond what the ordinary *p*-value already provides. A policy of requiring a PHP of .80 in order to publish a result, therefore, would simply be a disguised way of requiring a *p*-value such that $1 - \Phi\left(1.645 - \Phi^{-1}(1 - p)\right) > .8$ Such a *p*-value would have to be less than roughly 0.0065 instead of .05, leading to a much more conservative and less powerful test. That could seriously impede scientific progress (see Esarey, 2017). The algebra shown for the z-test above does not apply exactly to more complicated tests, but the general intuition that the PHP is a decreasing function of the *p*-value is valuable (see Hoenig and Heisey, 2001; Lenth, 2007). Thus, it is misleading to report both a *p*-value and a power estimate for the same test from the same data.

## Partially Post Hoc Power and Interpretation of Null Findings

Sometimes an investigator who has not been able to reject $H_0$ (i.e., who has obtained $p$>.05) may wish to calculate PHP as a way of interpreting this failure to reject. The investigator wants to decide whether this failure took place because there was not enough power to detect an effect and more study (and funding) is needed, or whether instead no meaningful effect ever existed and science should move on. Unfortunately PHP cannot meaningfully answer this question. Because a PHP is really only another way of expressing a low *p*-value and vice versa, it provides no information about whether the sample being evaluated is representative of the population, let alone whether the research topic was worthwhile. In particular, that the less significant a *p*-value is, the larger PHP analysis will consequently indicate that the sample should have been (Hoenig & Heisey, 2001; Schulz and Grimes, 2005).Thus, one cannot meaningfully report a *p*-value and a power estimate for the same test from the same data. That is, uncritically plugging sample estimates into an expression like (2) or (3) leads to a meaningless result. However, while PHP calculations should never be performed, reported, or interpreted, there are still situations in which the question of whether an existing intervention had adequate power may still be important.

Fortunately, there are more principled ways to use estimates from an existing study to plan a new study. One reasonable approach would be as follows: use the sample estimates for nuisance parameters such as of $\sigma$ in (2) or (3), but then plug in a different value of raw effect size parameters like $\Delta$. For example, Δ could be set to the minimum value Δ1 considered clinically or theoretically significant (Eng 2004; Schulz and Grimes 2005). Alternatively, one could use a benchmark relative effect size and ask, "What would the power have been if the true population effect size was 3 units (or 5%, or .2σ, etc.)?" Even though some information from the current study is used, this is still not exactly PHP, because the researcher is not trying to find the power required to detect a difference of the observed size, but instead some fixed size argued on theoretical grounds to be meaningful (Fagley 1985; Hoenig and Heisey 2001; Lenth 2007). One could similarly report what the power would have been under the given sample size for some benchmark standardized effect size. In this paper, we informally call these approaches "partially post-hoc power" (PPHP), because they seem to lie in between prospective and post-hoc power analysis. This terminology is admittedly somewhat awkward, but this approach of mixing post-hoc and non-post-hoc information does not seem to have a formal name in the literature, although it was mentioned by Thomas (1997), as well as by Yuan and Maxwell (2005, p. 142) who called them the "third possibility" between completely prospective power and PHP.

PPHP is one possible way to address the dilemma mentioned earlier: namely, deciding how to interpret a failure to reject $H_0$. Although students are often told never to "accept the null," it is common in practice to see nonsignificant *p*-values interpreted as demonstrating the absence of an effect or relationship (see Amrhein et al., 2017). It is very easy for investigators, let alone citizens or policymakers, to misinterpret "the intervention was not found to have an effect" as the seemingly almost identical "the intervention was found not to have an effect." However, accepting $H_0$ in this way can be dangerously misleading, especially if based on an underpowered test. Far from being mere non-statements, in practice Type II errors can be very serious: ignoring real medical or environmental risks, or discarding a helpful treatment (Lipsey et al. 1985, Peterman 1990). On the other hand, if an intervention shows very little evidence of efficacy then it is important to look for a better intervention, rather than simply arguing that the existing intervention might conceivably still have some small undetected effect. For these reasons, neither scientists nor funding agencies are likely to cast aside an expensive study as inconclusive without making some attempt to interpret why $H$0 was not rejected: whether it was merely an underpowered study or also represents an unpromising intervention. Thus, even though PHP is never advisable, PPHP or something like it may be necessary.

PPHP is more interpretable than PHP because it uses a theoretical judgment of what population-level effect would be scientifically meaningful, instead of just setting the population Δ to whatever was observed in the study. Peterman (1990) recommended that journal editors require the reporting of power (probably PPHP rather than PHP in his context) whenever $H$0 is not rejected, in order to alert readers of grossly underpowered studies. Reporting PPHP in this way is better than ignoring power altogether and can help to encourage better planning and prevent the publication of uninterpretable studies (Thomas, 1997; Szucs and Ioannidis, 2017b; Button et al., 2013). Thus, power considerations are not necessarily post-hoc simply because the study has already been done. Unfortunately, PPHP may still pose some risk of misinterpretation due to its superficial similarity to PHP. Also, its use as a journal policy might translate in practice to a required minimum sample size for studies, because of the previously mentioned equivalence between standardized benchmarks and sample sizes, and this could be harmful to researchers with small budgets or rare populations of interest. Thus, while PPHP is better than nothing, alternatives should be considered. We focus on confidence intervals because readers are likely to be most familiar with them, but briefly discuss Bayesian approaches.

## Better Alternatives: Confidence Intervals or Bayesian Analyses

Several authors have pointed out that the question that PPHP is supposed to answer may be better addressed by simply reporting the relevant confidence interval (Detsky and Sackett 1985; Hoenig and Heisey 2001; Schulz & Grimes 2005; Vandenbroucke et al. 2007), or alternatively a one-tailed confidence bound in the case of a one-sided hypothesis. For instance, suppose we are testing $H_0 : \mu_1 - \mu_2 = 0$, and suppose that $\Delta_1$ is 5 units on some scale of measurement. If the 95% confidence interval for $\mu_1 - \mu_2$ is (−.2, +.1) on this scale, a researcher may provisionally conclude that the available data support viewing $H_0$ as at least approximately true. On the other hand, if the confidence interval is (−10, +20) in the same units of measurement, it clearly is quite unsafe to assume that the true effect size is zero simply because it is statistically nonsignificant; it might be zero, but it might be highly positive or highly negative, and recognition of this uncertainty is necessary. Although both (−.2, +.1) and (−10, +20) correspond to “nonsignificant” hypothesis tests in that they include $H_0$, they do not have equivalent practical implications. The careful use of a confidence interval to interpret a failure to reject $H_0$ is conceptually related to an approach called “equivalence” or “non-inferiority” testing in medical studies, discussed further in Durrleman & Simon (1990), Hoenig and Heisey (2001), Lakens (2017), and Lindley (1998).

Bayesian model selection provides a different way to think about whether findings are conclusive or not. In a Bayesian analysis, a researcher can specify prior probabilities for different population hypotheses, and/or prior distributions over possible values of population parameters. Based on these priors, which must often be chosen subjectively, Bayes’ theorem can then be used to calculate posterior probabilities (see Kruschke, 2015, for an introduction). This approach makes it possible to avoid the $p$-value altogether, and actually calculate the posterior probability for $H_0$ given the data. To implement a test in a Bayesian setting, a researcher might have a “null” model where the quantity of interest (say $\Delta \equiv \mu1 - \mu2$) has a prior distribution concentrated at or near zero, and one or more “alternative” models with more diffuse priors perhaps centered at $\Delta1 \neq 0$. The researcher could then set a subjective prior probability for each model and use the data to find the posterior probability. Alternatively, the researcher could specify a single model with priors for the different parameters, and then calculate the posterior probability that the quantity of interest is in a particular range considered to be effectively null. In any case, this posterior probability is clearly not the same thing as the PHP. Because the PHP is a function of the $p$-value, it does not have a clear Bayesian interpretation and cannot substitute for a Bayesian analysis.

In summary, when trying to interpret a failure to reject an important $H0$, we have argued that using a Bayesian analysis or a confidence interval is better in many situations than using PPHP, although PPHP is often better than doing nothing about the issue of sample size. However, using PHP itself is considerably worse than doing nothing. Bayesian analysis is promising but still unfamiliar to many investigators, reviewers and readers. Thus, at least in the short term, confidence intervals may often be the best alternative.

## Are Power Calculations for Secondary Data Analysis of an Existing Dataset Post-Hoc?

Researchers sometimes wish to evaluate power for doing a proposed new analysis on an existing dataset. This is an unusual application of power because the dataset is already observed and recorded and it is too late to choose a sample size. The metaphorical marble is in hand and out of the jar even though it has not been examined yet. From this perspective, it would seem as though power analyses for existing datasets are by definition always post hoc. The question, “Does this dataset contain enough evidence to reject $H0$?” is the very question a significance test itself was designed to answer. There does not seem to be further scientific benefit to asking something like “Would the average hypothetical dataset of this size contain enough evidence to reject $H0$?” because in fact the researcher already has the only real dataset. Why not just forego power analysis and simply analyze the data, judging the precision of the results using confidence intervals? While this argument makes some sense, a case has also been made in favor of power analysis for secondary data, as described below.

Someone might wish to do a power analysis simply to decide whether it is worth the bother of analyzing a particular dataset. This rationale is a bit dubious, since it may take more work to get the information needed for a power analysis than to actually do the test with modern software. However, the question of whether an analysis is worthwhile may be more important if the dataset is not publicly available or requires extensive cleaning and merging. Another reason for potentially considering power in secondary data analysis is to try to avoid Type I or Type II errors. Bierman (2007) recommended that power analysis is needed for secondary data in order to avoid Type II errors, just as with new investigations. In a way, this advice is unhelpful, because there is nothing a secondary data analyst can do to get a larger dataset if it is determined that the current dataset is inadequate. However, the intended meaning seems to be that, if the probability of providing a meaningful answer is not high enough, then one should not waste time and risk misunderstandings by doing the analysis at all. This perspective may often be wise, although there may be situations in which it is not — particularly if one is interested in an exploratory analysis rather than a conclusive test, or if one’s only options are using this dataset or abandoning an interesting question. Such a situation is similar to that faced by someone doing a prospective power analysis for a study on ways to treat a serious but rare disease: one’s sample size is unavoidably small, but the study may still be worthwhile. The methodological literature is unclear on how to address these concerns. In some situations, it may be good to avoid doing underpowered analyses, in order to reduce the risk of irreproducible results. However, in other situations, it may be better to proceed with exploratory analyses, even though they might not have enough power to detect modest-sized effects. In such cases it is important to label underpowered, post-hoc, and/or purely exploratory analyses as such (Simmons et al. 2011; Munafò et al. 2017), but exploratory analyses should still be encouraged because exploration is vital for finding new knowledge.

A different pragmatic motivation is that investigators may be required to do a power analysis as a matter of policy by a granting body or a journal reviewer, regardless of whether or not the study involves the collection of new data. In that case, investigators must either make some kind of statement about power, or some explanation for their choice not to do so.

This question of whether secondary data power analysis are post hoc might seem pedantic or abstract. However, it becomes very practical when one has to choose parameters (e.g., effect size, variance, base rate, etc.) to use in a power formula. Ordinarily, power calculations are useful when seeking to collect a new dataset, and researchers use existing literature to find plausible values of the needed parameters. However, if the dataset is available, then it would seem more reasonable to use estimates from the real dataset to fill in for the parameters, rather than choosing values from past studies which may not be directly comparable. Yet once one opens the dataset and gets the parameter estimates, one already has the information needed to calculate the test statistic and *p*-value, so why should a researcher calculate power at all?

For example, consider power for testing the significance of a logistic regression coefficient for a test of association of two dichotomous variables. Let $p_{ab} = p(X = a,\ Y = b)$. Then from Demidenko (2007),

$$\text{Power} = \Phi\left(z_{1-\alpha/2} - \frac{\beta\sqrt{n}}{\sqrt{V}}\right) \tag{4}$$

Where $V = \frac{1}{p_{00}} + \frac{1}{p_{01}} + \frac{1}{p_{10}} + \frac{1}{p_{11}}$ and $\beta = \log\frac{p_{00}p_{11}}{p_{01}p_{10}}$ In (4), $\beta$ acts as a measure of effect size like Δ, and V acts as a measure of noise similar to $\sigma^2$. Notice that it is necessary to specify two quantities, namely $V = \frac{1}{p_{00}} + \frac{1}{p_{01}} + \frac{1}{p_{10}} + \frac{1}{p_{11}}$ and $\beta = \log\frac{p_{00}p_{11}}{p_{01}p_{10}}$, and that all one needs in order to specify both quantities are estimates of the four proportions $p_{ab}$. It seems tempting to use the data to get these estimates. However, given this information, one could also reconstruct the contingency table and do the logistic regression and significance test directly. Since the results of the test have now been implicitly determined, it then becomes difficult to explain why one is doing a power analysis at all, other than "the reviewer told me that I have to." In other words, in this case the seemingly prospective power analysis is arguably just PHP or at best PPHP, because the $p_{ab}$ determine not only the nuisance variance parameter but also the effect size $\beta$ of interest.

The situation is more complicated if the investigator's main goal is not to test a particular hypothesis but to construct a good predictive model. This may occur in multiple regression analysis, or in more complex models such as classification and regression trees. It can also occur in approaches such as cluster analysis, latent class analysis, and exploratory factor analysis, which are used to summarize large datasets interpretably. In each case, there is a meaningful goal of using the data to describe the relationships among certain variables in an insightful and somewhat generalizable way. Having an adequate sample size is important to achieving this goal. However, it is more difficult to say in advance what sample size is needed, because the models can be complex and depend on many parameters which are difficult to specify in advance. It is hard to tell exactly when the goal has been successfully achieved, because it is not simply a matter of rejecting a particular prespecified $H_0$ about a single parameter of interest. It is indeed possible to estimate the power for choosing a hypothesized "correct" number of number of covariates, classes, or factors (e.g., Dziak et al. 2014), but this is different from the probability that the final result of the analysis would be insightful or useful. Investigators want to know whether their sample is big enough, but the answer to follow-up question "big enough for what?" can be surprisingly tricky to operationalize. In some cases, more methodological research may be needed.

## Conclusions

Prospective power analysis, when planning a new empirical study, is widely recognized as important and is to be strongly recommended. However, PHP, in the sense of calculating the power of a test after having performed the test, is highly invalid and misleading, and researchers should not provide it and should not be asked to provide it in papers or reports. Although PHP seems superficially to address concerns of replicability, it does not in fact do so adequately. Heuristically, PHP involves circular reasoning and can lead to misleading recommendations. Algebraically, PHP is a deterministic function of the *p*-value, so that whenever the obtained *p*-value is greater than the $\alpha$ level (e.g., .05), PHP is always less than 50%. Thus, PHP cannot provide a valid way of deciding whether a larger sample really is practically warranted, or whether the effect size is too small to be worth detecting. It would be better to use confidence intervals, or alternatively Bayesian posterior probabilities, for assessing the practical meaning of a finding, instead of using PHP which simply restates the *p*-value. In some situations such as secondary data analysis, it may become somewhat unclear whether a particular power estimates is prospective or post hoc; more research and discussion on how to plan analyses in these situations could be helpful for researchers.

## Acknowledgments

Funding:

This research was supported in part by grant awards P50 DA010075 and P50 DA039838 from the National Institute on Drug Abuse (National Institutes of Health, United States). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions as mentioned above.

The corresponding author thanks Dr. Joseph L. Schafer for a helpful conversation on this topic, and also thanks Dr. J. Timothy Cannon for review and feedback. The authors thank Amanda Applegate for very helpful editorial review, and also thank the anonymous reviewers for substantially improving the paper.

## References

Amrhein V, Korner-Nievergelt F, & Roth T (2017). The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ, 5:e3544. [PubMed: 28698825]

Anderson SF, Kelley K, & Maxwell SE (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. Psychological Science, 28: 1547–1562. [PubMed: 28902575]

Baril GL, & Cannon JT (1995). What is the probability that null hypothesis testing is meaningless? American Psychologist, 50, 1098–1099.

Bierman AS (2007). Secondary analysis of large survey databases. In Max MB & Lynn J (Eds.), NIH interactive textbook on clinical symptom research: Methods and opportunities National Institute of Health (Accessed January 2007.)

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, & Munafò MR (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14: 365–376. [PubMed: 23571845]

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

Cohen J (1990). Things I have learned (so far). American Psychologist, 45, 1304–1312.

Cohen J (1992). A power primer. Psychological Bulletin, 112, 155–159. [PubMed: 19565683]

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

Demidenko E (2007). Sample size determination for logistic regression revisited. Statistics in Medicine, 26, 3385–3397. [PubMed: 17149799]

Demidenko E (2016). The p-value you can't buy. The American Statistician, 70: 33–38. [PubMed: 27226647]

Detsky AS, & Sackett DL (1985). When was a 'negative' clinical trial big enough? How many patients you needed depends on what you found. Archives of Internal Medicine, 145(4): 709–712. [PubMed: 3985731]

Durrleman S, & Simon R (1990). Planning and monitoring of equivalence studies. Biometrics, 46, 329–36. [PubMed: 2194579]

Dziak JJ, Lanza ST, & Tan X (2014). Effect size, statistical power and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. Structural Equation Modeling, 21: 534–552. [PubMed: 25328371]

Eng J (2004). Sample size estimation: A glimpse beyond simple formulas. Radiology, 230, 606–612. [PubMed: 14990827]

Esarey J (2017, 8 7). Lowering the threshold of statistical significance to $p < 0.005$ to encourage enriched theories of politics [Blog post] Retrieved from https://thepoliticalmethodologist.com/.

Fagley NS (1985). Applied statistical power analysis and the interpretation of nonsignificant results. Journal of Counseling Psychology, 32, 391–396.

Goldstein A (1964). Biostatistics: An introductory text New York: MacMillan.

Guo Y, Logan HL, Glueck DH, & Muller KE (2013). Selecting a sample size for studies with repeated measures. BMC Medical Research Methodology, 13:100. [PubMed: 23902644]

Hoenig JM, & Heisey DM (2001). The abuse of power: The pervasive fallacy of power calculations in data analysis. The American Statistician, 55, 19–24.

Jones LV, & Tukey JW (2000). A sensible formulation of the significance test. Psychological Methods, 5, 411–414. [PubMed: 11194204]

Kauermann G, Carroll RJ (2001). A note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association, 96: 1387–1398.

Kirk RE (2007). Effect magnitude: A different focus. Journal of Statistical Planning and Inference, 137, 1634–1646.

Korn EL (1990). Projecting power from a previous study: Maximum likelihood estimation. The American Statistician, 22, 290–2.

Kraemer HC, & Thiemann S (1987). How many subjects?: Statistical power analysis in research SAGE.

Kruschke JK, (2015). Doing Bayesian Data Analysis, A Tutorial with R, JAGS, and Stan ($2^{nd}$ ed.). Waltham, MA: Academic Press / Elsevier.$^{nd}$

Lakens D (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8, 355–362. [PubMed: 28736600]

Lenth RV (2001). Some practical guidelines for effective sample-size determination (Tech. Rep.) University of Iowa.

Lenth RV (2007). Post-hoc power: Tables and commentary (Tech. Rep.) University of Iowa: Department of Statistics and Actuarial Science.

Lindley DV (1998). Decision analysis and bioequivalence trials. Statistical Science, 13, 136–141.

Lipsey MW, Crosse S, Punkle J, Pollard J, & Stohart G(1985). Evaluation: The state of the art and the sorry state of the science. New Directions for Program Evaluation, 27, 7–28.

Longford NT (2016). Comparing two treatments by decision theory. Pharmaceutical Statistics, 15: 387–395. [PubMed: 27247139]

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, Simonsohn U, Wagenmakers E-J, Ware JJ & Ioannidis JPA (2017) A manifesto for reproducible science. Nature Human Behaviour, 1, 0021.

Norcross JC, Hogan TP, Koocher GP, & Maggio LA (2017). Clinician's guide to evidence-based practices: Behavioral health and addictions (2nd ed.). New York: Oxford.

Peterman RM (1990). The importance of reporting statistical power: the forest decline and acidic deposition example. Ecology, 71, 2024–2027.

Schulz KF, & Grimes DA (2005). Sample size calculations in randomized trials: Mandatory and mystical. Lancet, 365, 1348–53. [PubMed: 15823387]

Simmons JP, Nelson LD, & Simonsohn U (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22: 1359–1366. [PubMed: 22006061]

Szucs D, & Ioannidis JPA (2017a). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. Frontiers in Human Neuroscience, 11:390. [PubMed: 28824397]

Szucs D, & Ioannidis JPA (2017b). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biology, 15(3): e2000797. [PubMed: 28253258]

Thomas L (1997). Retrospective power analysis. Conservation Biology, 11, 276–280.

Vandenbroucke JP, von Elm E, Altman DG, Mulrow PCGD, Pocock SJ, Poole C, et al.(2007). Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. PLoS Medicine, 4(10).

Vickers AJ, & Altman DG (2001) Analysing controlled trials with baseline and follow up measurements. BMJ, 323, 1123–1124. [PubMed: 11701584]

Wasserstein RL, & Lazar NA (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70:2, 129–133.

Yuan K-H, & Maxwell S (2005). On the Post Hoc Power in Testing Mean Differences. Journal of Educational and Behavioral Statistics, 30, 141–167.