Behav Res (2015) 47:913–917
DOI 10.3758/s13428-014-0517-4

A power fallacy

Eric-Jan Wagenmakers · Josine Verhagen · Alexander Ly · Marjan Bakker · Michael D. Lee · Dora Matzke · Jeffrey N. Rouder · Richard D. Morey

Published online: 1 October 2014

Abstract The power fallacy refers to the misconception that what holds on average –across an ensemble of hypothetical experiments– also holds for each case individually. According to the fallacy, high-power experiments always yield more informative data than do low-power experiments. Here we expose the fallacy with concrete examples, demonstrating that a particular outcome from a high-power experiment can be completely uninformative, whereas a particular outcome from a low-power experiment can be highly informative. Although power is useful in planning an experiment, it is less useful—and sometimes even misleading—for making inferences from observed data. To make inferences from data, we recommend the use of likelihood ratios or Bayes factors, which are the extension of likelihood ratios beyond point hypotheses. These methods of inference do not average over hypothetical replications of an experiment, but instead condition on the data that have actually been observed. In this way, likelihood ratios and Bayes factors rationally quantify the evidence that a particular data set provides for or against the null or any other hypothesis.

E.-J. Wagenmakers (✉) · J. Verhagen · A. Ly · M. Bakker · D. Matzke
Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, The Netherlands
e-mail: ej.wagenmakers@gmail.com

M. D. Lee
University of California Irvine, Irvine, CA USA

J. N. Rouder
University of Missouri, Columbia, MO USA

R. D. Morey
University of Groningen, Groningen, Netherlands

Keywords Hypothesis test · Likelihood ratio · Statistical evidence · Bayes factor

It is well known that psychology has a power problem. Based on an extensive analysis of the literature, estimates of power (i.e., the probability of rejecting the null hypothesis when it is false) range at best from 0.30 to 0.50 (Bakker, van Dijk, & Wicherts, 2012; Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafò, 2013; Cohen, 1990). Low power is problematic for several reasons. First, underpowered experiments are more susceptible to the biasing impact of questionable research practices (e.g., Bakker, van Dijk, & Wicherts, 2012; Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafò, 2013; Ioannidis, 2005). Second, underpowered experiments are, by definition, less likely to discover systematic deviations from chance when these exist. Third, as explained below, underpowered experiments reduce the probability that a significant result truly reflects a systematic deviation from chance (Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafò, 2013; Ioannidis, 2005; Sham & Purcell, 2014).

These legitimate concerns suggest that power is not only useful in planning an experiment, but should also play a role in drawing inferences or making decisions based on the collected data (Faul, Erdfelder, Lang, & Buchner, 2007). For instance, authors, reviewers, and journal editors may argue that: (1) for a particular set of results of a low-power experiment, an observed significant effect is not very diagnostic; or (2) for a particular set of results from a high-power experiment, an observed non-significant effect is very diagnostic (i.e., it supports the null hypothesis). However, power is a pre-experimental measure that averages over all possible outcomes of an experiment, only one of which will actually be observed. What holds on average, across a collection of hypothetically infinite unobserved data sets, need not hold for the particular data set we have collected. This difference is important, because it counters the temptation to dismiss all findings from low-power experiments or accept all findings from high-power experiments. Below we outline the conditions under which a low-power experiment is uninformative: when the only summary information we have is the significance of the effect (i.e., $p < \alpha$). When we have more information about the experiment, such as test statistics or, better, the data set itself, a low-power experiment can be very informative.

A basic principle of science should be that inference is conditional on the available relevant data (Jaynes, 2003). If an experiment has yet to be conducted, and many data sets are possible, all possible data sets should be considered. If an experiment has been conducted, but only summary information is available, all possible data sets consistent with the summary information should be considered. But if an experiment has been conducted, and the data are available, then only the data that were actually observed should be incorporated into inference.

When power is useful for inference

As Ioannidis and colleagues have emphasized, low power is of concern not only when a non-significant result is obtained, but also when a significant result is obtained: "Low statistical power (because of low sample size of studies, small effects, or both) negatively affects the likelihood that a nominally statistically significant finding actually reflects a true effect." (Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafò, 2013, p. 1).

To explain this counter-intuitive fact, suppose that all that is observed from the outcome of an experiment is the summary $p < \alpha$ (i.e., the effect is significant). Given that it is known that $p < \alpha$, and given that it is known that the power equals $1 - \beta$, what inference can be made about the odds of $\mathcal{H}_1$ vs $\mathcal{H}_0$ being true? Power is the probability of finding a significant effect when the alternative hypothesis is true, $\Pr(p < \alpha \mid \mathcal{H}_1)$ and, similarly, the probability of finding a significant effect when the null hypothesis is true equals $\Pr(p < \alpha \mid \mathcal{H}_0)$. For prior probabilities of true null hypotheses $\Pr(\mathcal{H}_0)$ and true alternative hypotheses $\Pr(\mathcal{H}_1)$, it then follows from probability theory that

$$\frac{\Pr(\mathcal{H}_1 \mid p < \alpha)}{\Pr(\mathcal{H}_0 \mid p < \alpha)} = \frac{\Pr(p < \alpha \mid \mathcal{H}_1)}{\Pr(p < \alpha \mid \mathcal{H}_0)} \times \frac{\Pr(\mathcal{H}_1)}{\Pr(\mathcal{H}_0)}$$
$$= \frac{1-\beta}{\alpha} \times \frac{\Pr(\mathcal{H}_1)}{\Pr(\mathcal{H}_0)}. \quad (1)$$

Thus $(1 - \beta)/\alpha$ is the extent to which the observation that $p < \alpha$ changes the prior odds that $\mathcal{H}_1$ rather than $\mathcal{H}_0$ is true.

Bakker, van Dijk, and Wicherts (2012) estimated that the typical power in psychology experiments to be 0.35. This means that if one conducts a typical experiment and observes $p < \alpha$ for $\alpha = 0.05$, this information changes the prior odds by a factor of $.35/.05 = 7$. This is a moderate level of evidence (Lee & Wagenmakers, 2013, Chapter 7).[1] When $p < \alpha$ for $\alpha = 0.01$, the change in prior odds is $0.35/0.01 = 35$. This is stronger evidence. Finally, when $p < \alpha$ for $\alpha = 0.005$ (i.e., the threshold for significance recently recommended by Johnson, 2013), the change in prior odds is $0.35/0.005 = 70$. This level of evidence is an order of magnitude larger than that based on the earlier experiment using $\alpha = 0.05$. Note that doubling the power to 0.70 also doubles the evidence.

Equation 1 highlights how power affects the diagnosticity of a significant result. It is therefore tempting to believe that high-power experiments are always more diagnostic than low-power experiments. This belief is false; with access to the data, rather than the summary information that $p < \alpha$, we can go beyond measures of diagnosticity such as power. As Sellke, Bayarri, and Berger (2001, p. 64) put it: "(...) if a study yields $p = .049$, this is the actual information, not the summary statement $0 < p < .05$. The two statements are very different in terms of the information they convey, and replacing the former by the latter is simply an egregious mistake."

Example 1: a tale of two urns

Consider two urns that each contain ten balls. Urn $\mathcal{H}_0$ contains nine green balls and one blue ball. Urn $\mathcal{H}_1$ contains nine green balls and one orange ball. You are presented with one of the urns and your task is to determine the identity of the urn by drawing balls from the urn with replacement. Unbeknownst to you, the urn selected is urn $\mathcal{H}_1$.

In the first situation, you plan an experiment that consists of drawing a single ball. The power of this experiment is defined as $\Pr(\text{reject } \mathcal{H}_0 \mid \mathcal{H}_1) = \Pr(\text{"draw orange ball"} \mid \mathcal{H}_1) = 0.10$. This power is very low. Nevertheless, when the experiment is carried out you happen to draw the orange ball. Despite the experiment having very low power, you were lucky and obtained a decisive result. The experimental data permit the completely confident inference that you were presented with urn $\mathcal{H}_1$.

In the second situation, you may have the time and the resources to draw more balls. If you draw N balls with

[1] While odds lie on a naturally meaningful scale calibrated by betting, characterizing evidence through verbal labels such as "moderate" and "strong" is necessarily subjective (Kass & Raftery, 1995). We believe the labels are useful because they facilitate scientific communication, but they should only be considered an approximate descriptive articulation of different standards of evidence.

replacement, the power of your experiment is $1 - (0.9)^N$. You desire 90 % power and therefore decide in advance to draw 22 balls. When the experiment is carried out you happen to draw 22 green balls. Despite the experiment having high power, you were unlucky and obtained a result that is completely uninformative.

Example 2: the t-test

Consider the anomalous phenomenon of retroactive facilitation of recall (Bem, 2011; Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012). The central finding is that people recall more words when these words are to be presented again later, following the test phase. Put simply, people's recall performance is influenced by events that have yet to happen. In the design used by Bem (2011), each participant performed a standard recall task followed by post-test practice on a subset of the words. For each participant, Bem computed a "precognition score", which is the difference in the proportion of correctly recalled items between words that did and did not receive post-test practice.

Suppose that a researcher, Dr. Z, tasks two of her students to reproduce Bem's Experiment 9 (Bem, 2011). Dr. Z gives the students the materials, but does not specify the number of participants that each should obtain. She intends to use the data to test the null hypothesis against a specific effect size consistent with Bem's results. As usual, the null hypothesis states that effect size is zero (i.e., $\mathcal{H}_0 : \delta = 0$), whereas the alternative hypothesis states that there is a specific effect (i.e., $\mathcal{H}_1 : \delta = \delta_a$). Dr. Z chooses $\delta_a = 0.3$, a result consistent with the data reported by Bem (2011), and plans to use a one-sided t test for her analysis.

A likelihood ratio analysis

The two students, A and B, set out to replicate Bem's findings in two experiments. Student A is not particularly interested in the project and does not invest much time, obtaining only 10 participants. The power of A's experiment is 0.22, which is quite low. Student A sends the data to Dr. Z, who is quite disappointed in the student; nevertheless, when Dr. Z carries out her analysis the t value she finds is 5.0 ($p = .0004$).[2]

Figure 1a shows the distribution of p under the null and alternative hypotheses. Under the null hypothesis, the distribution of p is evenly spread across the range of possible p values (horizontal blue line). Under the alternative that the effect size is $\delta_a = 0.3$, however, p values tend to gather near 0 (upper red curve). The low power of the experiment can be seen by comparing the area under the null hypothesis' blue line to the left of $p = 0.05$ with that under the alternative's red curve; the area is only 4.4 times larger.

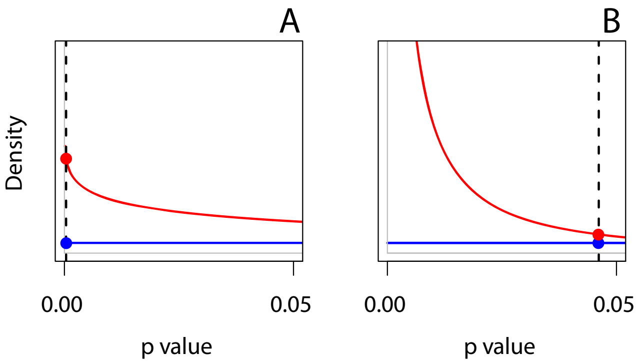

Fig. 1 A dissociation between power and evidence. The panels show distributions of p values under the null hypothesis (*horizontal blue line*) and alternative hypothesis (*red curve*) in Student **a**'s and **b**'s experiments, restricted to $p < 0.05$ for clarity. The *vertical dashed line* represents the p value found in each of the two experiments. The experiment from Student **a** has low power (i.e., the p-curve under $\mathcal{H}_1$ is similar to that under $\mathcal{H}_0$) but nonetheless yields an informative result; the experiment from Student **b** has high power (i.e., the p-curve under $\mathcal{H}_1$ is dissimilar to that under $\mathcal{H}_0$) but nonetheless yields an uninformative result

However, for the observed p value, the information is greater. We can compute the likelihood ratio of the observed data by noting the relative heights of the densities at the observed p value (vertical dashed line). The probability of the observed p value under the alternative is 9.3 times higher under the alternative hypothesis than under the null hypothesis. Although not at all decisive, this level of evidence is generally considered substantial (Jeffreys, 1961, Appendix B). The key point is that it is possible to obtain informative results from a low-power experiment, even though this was unlikely *a priori*.

Student B is more excited about the topic than Student A, working hard to obtain 100 participants. The power of Student B's experiment is 0.91, which is quite high. Dr. Z is considerably happier with Student B's work, correctly reasoning that Student B's experiment is likely to be more informative than Student A's. When analyzing the data, however, Dr. Z observes a t value of 1.7 ($p = 0.046$).[3]

Figure 1b shows the distribution of p under the null and alternative hypotheses. As before, the distribution of p is uniform under the null hypothesis and right-skewed under the the alternative that the effect size is $\delta_a = 0.3$. The high power of the experiment can be seen by comparing the area under the null hypothesis' blue line to the left of $p = 0.05$

[2] In order to obtain a t value of 5 with a sample size of only 10 participants, the precognition score needs to have a large mean or a small variance.

[3] In order to obtain a t value of 1.7 with a sample size of 100, the the precognition score needs to have a small mean or a high variance.

with that under the alternative's red curve; the area is 18.2 times larger.

However, for the observed p value, the information is almost completely uninformative, as the probability of the observed p value under the alternative is only 1.83 times higher under the alternative hypothesis than under the null hypothesis. The key point is that it is possible to obtain uninformative results from a high-power experiment, even though this was unlikely *a priori*.

A Bayes factor analysis

For the above likelihood ratio analysis we compared the probability of the observed p value under the null hypothesis $\mathcal{H}_0 : \delta = 0$ versus the alternative hypothesis $\mathcal{H}_1 : \delta = \delta_a = 0.3$. The outcome of this analysis critically depends on knowledge about the true effect size δ_a. Typically, however, researchers are uncertain about the possible values of δ_a. This uncertainty is explicitly accounted for in Bayesian hypothesis testing (Jeffreys, 1961) by quantifying the uncertainty in δ_a through the assignment of a prior distribution. The addition of the prior distribution means that instead of a single likelihood ratio, we use a weighted average likelihood ratio, known as the Bayes factor. In other words, the Bayes factor is simply a likelihood ratio that accounts for the uncertainty in δ_a.

More concretely, in the case of the t test example above, the likelihood ratio for observed data Y is given by

$$\mathrm{LR}_{10} = \frac{p\,(Y \mid \mathcal{H}_1)}{p\,(Y \mid \mathcal{H}_0)} = \frac{t_{df,\,\delta_a\sqrt{N}}\,(t_{obs})}{t_{df}\,(t_{obs})}, \tag{2}$$

where $t_{df,\,\delta\sqrt{N}}\,(t_{obs})$ is the ordinate of the non-central t-distribution and $t_{df}(t_{obs})$ is the ordinate of the central t-distribution, both with df degrees of freedom and evaluated at the observed value t_{obs}. To obtain a Bayes factor, we then assign δ_a a prior distribution. One reasonable choice of prior distribution for δ_a is the default standard Normal distribution, $\mathcal{H}_1 : \delta_a \sim \mathcal{N}(0, 1)$, which can be motivated as a unit-information prior (Gönen, Johnson, Lu, & Westfall, 2005; Kass & Raftery, 1995). This prior can be adapted for one-sided use by only using the positive half of the normal prior (Morey & Wagenmakers, 2014). For this prior, the Bayes factor is given by

$$\begin{aligned} B_{10} &= \frac{p\,(Y \mid \mathcal{H}_1)}{p\,(Y \mid \mathcal{H}_0)} \\ &= \frac{\int p\,(t_{obs} \mid \delta_a)\,p\,(\delta_a \mid \mathcal{H}_1)\,\mathrm{d}\delta_a}{p\,(t \mid \delta = 0)} \\ &= \frac{\int t_{df,\,\delta_a\sqrt{N}}\,(t_{obs})\mathcal{N}_+(0, 1)\,\mathrm{d}\delta_a}{t_{df}\,(t_{obs})} \\ &= \int \mathrm{LR}_{10}(\delta_a)\mathcal{N}_+(0, 1)\,\mathrm{d}\delta_a. \end{aligned} \tag{3}$$

where $\mathcal{N}_+$ denotes a normal distribution truncated below at 0.

Still intrigued by the phenomenon of precognition, and armed now with the ability to calculate Bayes factors, Dr. Z conduct two more experiments. In the first experiment, she again uses 10 participants. Assuming the same effect size $\delta_a = 0.3$, the power of this experiment is still 0.22. The experiment nevertheless yields $t = 3.0$ ($p = .007$). For these data, the Bayes factor equals $B_{10} = 12.4$, which means that the observed data are 12.4 times more likely to have occurred under $\mathcal{H}_1$ than under $\mathcal{H}_0$. This is a non-negligible level of evidence in support of $\mathcal{H}_1$, even though the experiment had only low power.

In the second experiment, Dr. Z again uses 100 participants. The power of this experiment is 0.91, which is exceptionally high, as before. When the experiment is carried out she finds that $t = 2.0$ ($p = 0.024$). For these observed data, the Bayes factor equals $\mathrm{B}_{10} = 1.38$, which means that the observed data are only 1.38 times more likely to have occurred under $\mathcal{H}_1$ than under $\mathcal{H}_0$. This level of evidence is nearly uninformative, even though the experiment had exceptionally high power.

In both examples described above, the p value happened to be significant at the .05 level. It may transpire, however, that a nonsignificant result is obtained and a power analysis is then used to argue that the nonsignificant outcome provides support in favor of the null hypothesis. The argument goes as follows: "We conducted a high-power experiment and nevertheless we did not obtain a significant result. Surely this means that our results provide strong evidence in favor of the null hypothesis." As a counterexample, consider the following scenario. Dr. Z conducts a third experiment, once again using 100 participants for an exceptionally high power of 0.91. The data yield $t = 1.2$ ($p = .12$). The result is not significant, but neither do the data provide compelling support in favor of $\mathcal{H}_0$: the Bayes factor equals $\mathrm{B}_{01} = 2.8$, which is a relatively uninformative level of evidence. Thus, even in the case that an experiment has high power, a non-significant result is not, by itself, compelling evidence for the null hypothesis. As always, the strength of the evidence will depend on both the sample size and the closeness of the observed test statistic to the null value.

The likelihood ratio analyses and the Bayes factor analyses both require that researchers specify an alternative distribution. For a likelihood ratio analysis, for instance, the same data will carry different evidential impact when the comparison involves $\mathcal{H}_1 : \delta_a = 0.3$ than when it involves $\mathcal{H}_1 : \delta_a = 0.03$. We believe the specification of an acceptable $\mathcal{H}_1$ is a modeling task like any other, in the sense that it involves an informed choice that can be critiqued, probed, and defended through scientific argument. More importantly, the power fallacy is universal in the sense that it does not depend on a specific choice of $\mathcal{H}_1$; the scenarios sketched here are only meant as concrete examples to illustrate the general rule.

Conclusion

On average, high-power experiments yield more informative results than low-power experiments (Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munaf'o, 2013; Cohen, 1990; Ioannidis, 2005). Thus, when planning an experiment, it is best to think big and collect as many observations as possible. Inference is always based on the information that is available, and better experimental designs and more data usually provide more information.

However, what holds on average does not hold for each individual case. With the data in hand, statistical inference should use all of the available information, by conditioning on the specific observations that have been obtained (Berger & Wolpert 1988; Pratt 1965). Power is best understood as a pre-experimental concept that averages over all possible outcomes of an experiment, and its value is largely determined by outcomes that will not be observed when the experiment is run. Thus, when the actual data are available, a power calculation is no longer conditioned on what is known, no longer corresponds to a valid inference, and may now be misleading.

Consequently, as we have demonstrated here, it is entirely possible for low-power experiments to yield strong evidence, and for high-power experiments to yield weak evidence. To assess the extent to which a particular data set provides evidence for or against the null hypothesis, we recommend that researchers use likelihood ratios or Bayes factors. After a likelihood ratio or Bayes factor has been presented, the demand of an editor, the suggestion of a reviewer, or the desire of an author to interpret the relevant statistical evidence with the help of power is superfluous at best and misleading at worst.

Author Note This work was supported by an ERC grant from the European Research Council. Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: EJ.Wagenmakers@gmail.com.

References

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

Bem, D.J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Berger, J.O., & Wolpert, R.L. (1988). In *The likelihood principle*, 2nd edn. Hayward (CA): Institute of Mathematical Statistics.

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

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

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

Galak, J., LeBoeuf, R.A., Nelson, L.D., Simmons, J.P. (2012). Correcting the past: failures to replicate Psi. *Journal of Personality and Social Psychology*, *103*, 933–948.

Gönen, M., Johnson, W.O., Lu, Y., Westfall, P.H. (2005). The Bayesian two–sample *t* test. *The American Statistician*, *59*, 252–257.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.

Jaynes, E.T. (2003). In *Probability theory: the logic of science*. Cambridge: Cambridge University Press.

Jeffreys, H. (1961). In *Theory of probability (3 ed)*. Oxford: Oxford University Press.

Johnson, V.E. (2013). Revised standards for statistical evidence. In *Proceedings of the national academy of sciences of the United States of America* (Vol. 11, pp. 19313–19317).

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Lee, M.D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: a practical course*. Cambridge University Press.

Morey, R.D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121–124.

Pratt, J.W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society B*, *27*, 169–203.

Ritchie, S.J., Wiseman, R., French, C.C. (2012). Failing the future: three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE*, *7*, e33423.

Sellke, T., Bayarri, M.J., Berger, J.O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Sham, P.C., & Purcell, S.M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, *15*, 335–346.