Meta-Psychology, 2022, vol 6, MP.2020.2460
https://doi.org/10.15626/MP.2020.2460
Article type: Original Article

Open data: Not Applicable
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Moritz Heene
Reviewed by: Gelman, A., Martin, S.R., Olvera Astivia, O.
Analysis reproduced by: Jens Fust
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/P6GWF

Power or Alpha? The Better Way of Decreasing the False Discovery Rate

František Bartoš
University of Amsterdam; Faculty of Arts, Charles University

Maximilian Maier
University of Amsterdam

Both authors contributed equally

Abstract

The replication crisis in psychology has led to an increased concern regarding the false discovery rate (FDR) – the proportion of false positive findings among all significant findings. In this article, we compare two previously proposed solutions for decreasing the FDR: increasing statistical power and decreasing significance level α. First, we provide an intuitive explanation for α, power, and FDR to improve the understanding of these concepts. Second, we investigate the relationship between α and power. We show that for decreasing FDR, reducing α is more efficient than increasing power. We suggest that researchers interested in reducing the FDR should decrease α rather than increase power. By investigating the relative importance of both α level and power, we connect the literature on these topics and our results have implications for increasing the reproducibility of psychological science.

Keywords: Power, Significance level, False Discovery Rate, Alpha

The reproducibility of studies in psychology has been questioned in the last few years. Massive replication initiatives found that replicability can be as low as 36% (Open Science Collaboration, 2015; but see Camerer et al., 2018; Ebersole et al., 2016; Klein et al., 2014; Klein et al., 2018 for more optimistic estimates), and many researchers have tried to identify the factors affecting the replicability of studies. While a comprehensive overview of this is beyond the scope of a single article (a whole issue of Perspectives on Psychological Science was dedicated to the problem; Pashler and Wagenmakers, 2012), we focus on statistical power, significance level α and the false discovery rate (FDR, the proportion of false positive findings among all statistically significant findings).[1] While some papers emphasize the importance of increasing statistical power to decrease the FDR (Button et al., 2013; Christley, 2010), others call for decreasing α (Benjamin et al., 2018). However, these two views seem disconnected and it is unclear whether (or under which conditions) researchers should decide to decrease α and when to increase power in order to reduce the FDR. To further explore this disconnect, we reviewed all articles mentioning FDR (or related terms) in the context of power and α in five methods and evidence synthesis journals within psychology (for more details see: https://osf.io/9cfg8/). Out of 106 reviewed articles, nine explicitly stated the

[1] The FDR is sometimes also called False Positive Rate (FPR Benjamin et al., 2018) or False Positive Risk (FPR Colquhoun, 2017).

importance of increasing power to reduce the FDR, while five articles discussed the importance of decreasing α.[2] Notably, only Miller and Ulrich (2019) discussed that both decreasing α and increasing power would reduce FDR. However, the efficiency of those two options was not compared so far.

The current article aims to bridge the discussion over α and power regarding the FDR and investigate the more efficient way of reducing the FDR. To achieve this, we first reiterate the concepts of power, false positives, and false discovery rate. We explain them using intuitive examples to deepen the understanding of these concepts. Next, we examine two possible views and their impact on reducing the FDR. The first view concerns planning a study and deciding on α and power independently. The second view concerns balancing between α and power for a fixed design, where setting α determines power and vice versa.

False Positives and α

In his pivotal book "Statistical Methods for the Research Worker" Fisher (1925) was the first to widely popularize the concept of hypothesis testing and statistical significance to differentiate signal from noise. Neyman and Pearson (1928) introduced the conceptualization of the significance level α as a tool to control the long-term error rates. In other words, a decision from a statistical test with a significance level (i.e., 5%) would not result in more than a rate α of incorrectly rejected true null hypotheses. Thus, α determines the long-term rate of false positives. If researchers set their α to 5%, they will accept the alternative hypothesis when the probability of the data or more extreme data assuming the null hypothesis to be true (the p-value) is below α.

Let us illustrate this concept with an example from Fisher (1935) famous experiment "The Lady tasting tea". Lady Muriel Bristol claims that she can detect whether tea or milk was added first to a cup. To test whether the Lady has these tea tasting abilities, Fisher gives lady Bristol eight cups of tea, in which four of them has milk added first, while the other four have tea added first. Fisher wants to keep his long-term error rate of false positives below 5%. Since the Lady knows that half of the cups are tea first, Fisher focuses only on the number of correctly classified tea-first-cups (because the correctly classified milk-first-cups are dependent on the correctly classified tea-first-cups). How many of the four tea-first-cup cases would the Lady need to classify correctly to convince Fisher of her abilities? The probability of correctly guessing x tea-first-cups in four trials can be obtained using the hypergeometric distribution (Figure 1, left). All four tea-first-cups would be guessed correctly with a probability of 1.43%. So, this event would indicate that it is improbable to see the Lady give all eight correct answers if she has no tea tasting abilities and guessed entirely at random. But what if she makes one mistake? The probability of classifying at least three out of four tea-first-cups correctly by pure guessing is 24.3%. In other words, this would not provide sufficient evidence against her lack of abilities. So, in this case, Fisher would be unable to know whether she can differentiate between the cups. Even if she were guessing entirely at random, she could have achieved at least three out of four correctly guessed tea-first-cups 24.3% of the time.

Power

Neyman and Pearson (1928) introduced the concept of statistical power because of the fundamental asymmetry of controlling Type I error rates without explicitly formalizing Type II error control (the probability of concluding the absence of an effect, when it exists; Lehmann, 1992). Statistical power describes the probability that a statistical test rejects the null hypothesis when it is false. In other words, power refers to the probability of rejecting the null hypothesis, assuming that the hypothesized effect is present. The statistical power of a test depends on α, the sample size, and the magnitude of the true effect. A higher α, a larger sample size, and a larger true effect all contribute to increased statistical power (Cohen, 1992). Power is thus related to false negatives, with higher statistical power decreasing the probability of finding a false negative result.

Let's continue with the previous example but look at it from the other side. Assume that the Lady can distinguish whether the milk or tea was added first. It is a difficult task, and she makes a mistake from time to time. Her probability of classifying the cup correctly is 0.7. The resulting probabilities this time follow a noncentral hypergeometric distribution (Liao and Rosen, 2001; Figure 1, right). Thus, the probability of her classifying all eight cups correctly is now 19%. In other words, if the Lady has the ability to classify correctly in 70% of cases, Fisher would only detect this 19% of the time.

False Discovery Rate

It follows from the previously outlined definition that power does not influence the probability of observing a false-positive result for any single study. However, since negative results are rarely published (Masicampo and Lalande, 2012; Mathur and VanderWeele, 2020; Nelson et al., 1986; Rosenthal, 1979; Rosenthal and Gaito, 1963, 1964; Wicherts, 2017 but see van Aert et al., 2019

[2] Most of the remaining articles focused on correction for multiple testing.

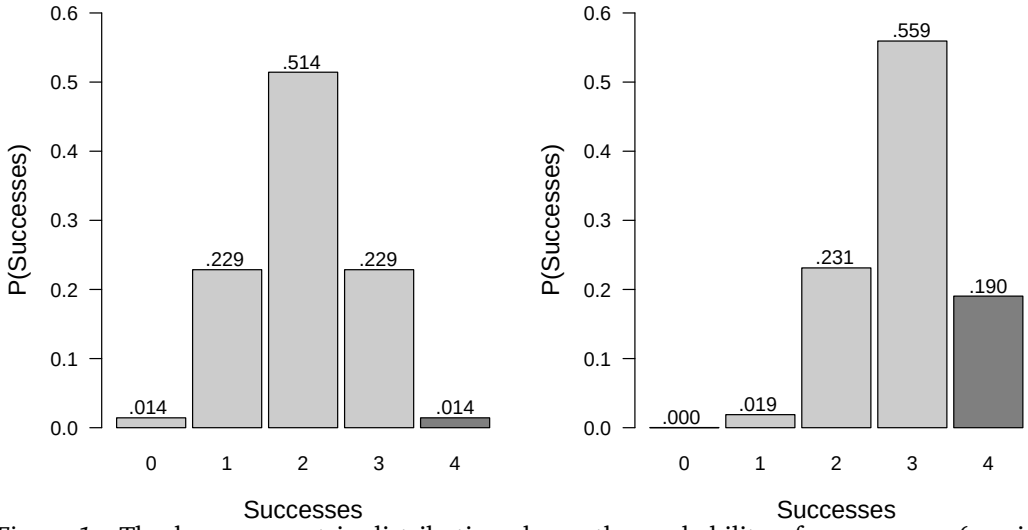

Figure 1. The hypergeometric distribution shows the probability of x successes (x-axis) with the probability of success 0.50 (left) and 0.70 (right). Note that we only display up to four successes. We can think of those bars as the number of tea-first-cups classified correctly. The Lady knows how many (but not which) cups have tea added to them first. Therefore, if she classifies all tea-first-cups correctly, she necessarily also classifies the milk-first-cups correctly. The dark-filled bars correspond to the probability of 4 correct answers.

for contrary evidence), it is more interesting to investigate the proportion of false positives among significant findings, i.e., the false discovery rate (FDR). This proportion depends on the number of true positives (believing that someone possesses the tea tasting abilities when they truly do) and the number of false positives (believing that someone possesses the tea tasting abilities when they do not). While the number of true positives depends on power and the number of true alternative hypotheses, the number of false positives depends on α and the proportion of false hypotheses. So, the FDR connects both previously mentioned concepts, and we illustrate it with our running example.

Her Majesty The Queen decides to start a Royal Tea Tasting Society (RTTS) and requests Fisher to recruit new members based on their tea tasting abilities. Assume that one-fifth of the population possesses such abilities and can identify the order of milk and tea in 70% of cases. The remaining four-fifths do not possess this skill and their answers are equal to random guessing. Fisher decides to use α of 5%; therefore, 0.05×0.80 = 4% of the tests he administers result in false positives. Because he conveniently uses the same set-up as in the previous example, we know that the power of the test is 19%. Therefore, 0.19×0.20 = 3.8% of the tests he administers yield true positives. Subsequently, he introduces all citizens who passed the test to the Queen, who promotes them to members of the RTTS. However, what the Queen does not realize is the fact that 0.04/(0.04+0.038) = 51% of her RTTS members do not possess any tea tasting abilities (the FDR).

As can be deduced from the example, there are two ways to decrease FDR - either increase power and thus the number of true positives, or reduce α and the number of false positives. This relationship is depicted in Equation (1), which illustrates how power and α influence the FDR, with $P(\mathcal{H}_0)$ standing for the proportion of true null hypotheses, α for significance level, and ρ for statistical power,

$$\text{FDR} = \frac{P(\mathcal{H}_0) \times \alpha}{P(\mathcal{H}_0) \times \alpha + (1 - P(\mathcal{H}_0)) \times \rho}. \tag{1}$$

This is the reason why many argue that researchers need to increase the statistical power to reduce the FDR. However, we show in the following paragraphs that reducing α is usually the preferable option by investigating two ways of considering the trade-off between power and α. In the first way, researchers plan a study and independently determine what levels of α and power should be used. In the second, researchers balance between α and power for a fixed design, where setting α determines the power and vice versa.

Determining α and Power Independently

The first view assumes that α and power are set independently.[3] For example, researchers plan a study with

[3]The first case and the following derivations were suggested by Stephen R. Martin in his review (https://osf.io/7kdjn/).

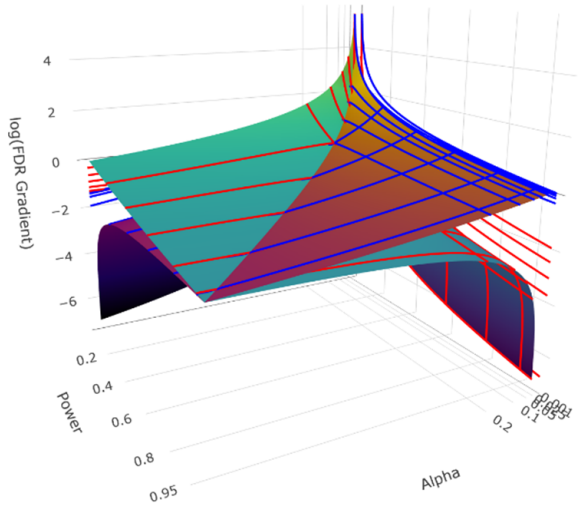

Figure 2. The logarithm of the FDR gradient (z-axis) is dependent on α (Alpha, x-axis) and power (y-axis) for the probability of the null hypothesis being true equal to 0.5. The red surface (with blue lines) depicts the gradient of FDR with respect to α and the green surface (with red lines) depicts the gradient of FDR in respect to power. Note that they intersect when α is equal to power. When α is lower than power (right side), the gradient of FDR with respect to α dominates the gradient with respect to power. An animated version is accessible at https://osf.io/gbtku/.

desired α and power and compute the required sample size for achieving them. Subsequently, we can study how either changing α or power in the planning phase influences the FDR. To do that, we present derivations of Equation (1) with respect to α,

$$\frac{\delta \text{FDR}}{\delta \alpha} = \frac{\rho \times (1 - P(\mathcal{H}_0)) \times P(\mathcal{H}_0)}{(\alpha \times P(\mathcal{H}_0) + \rho \times (1 - P(\mathcal{H}_0)))^2}, \tag{2}$$

and power,

$$\frac{\delta \text{FDR}}{\delta \rho} = \frac{-\alpha \times (1 - P(\mathcal{H}_0)) \times P(\mathcal{H}_0)}{(\alpha \times P(\mathcal{H}_0) + \rho \times (1 - P(\mathcal{H}_0)))^2}, \tag{3}$$

Equations (2) and (3) connect the change in α or power to change in FDR. Since the denominators are the same and $P(\mathcal{H}_0)$ is bound to be between 0 and 1, the comparison of Equation (2) and (3) shows that the gradient of FDR with respect to power will dominate the gradient of FDR with respect to α as long as power is larger than α (Figure 2).

This is generally true because α is the lower bound on power, unless a one-sided test is used and the effect is in the opposite direction. Then, power is lower than α and the gradient of FDR with respect to α dominates the gradient of FDR with respect to power. In addition, when a two-sided test is used but the power is low, many significant results will be in the opposite direction (Type S error; Gelman and Carlin, 2014). Including those into the FDR would further change the results. Compelling visualizations that support this claim are also available in the online materials (https://osf.io/gbtku/) and a more detailed discussion of this approach can be found in the open review (https://osf.io/sp95d/). Overall, this indicates that for all conditions that are typically encountered in hypothesis testing, the gradient with respect to alpha will dominate the gradient with respect to power.

In other words, when designing a study, planning a lower α has a larger effect than planning higher power, as long as power is kept higher than α. So, if Fisher wanted to mitigate the proportion of members of RTTS with no tea tasting abilities before the experiment was conducted, the best solution would be to decrease the α as much as possible.

Trading α and Power

The second view goes one step further. If we assume that researchers are operating with limited resources (i.e., a limited number of participants or time), then α determines power or vice versa. In other words, for a fixed design, researchers can either set α and power can be expressed as a function of α, or researchers can set a desired power, and α can be expressed as a function of power. Equation (4) shows the relationship of power (ρ), on the left side, to α, in the case of a two-tailed independent samples z-test. In addition, sample size n and effect size d are needed to determine the parameter μ of the normal distribution of expected z-statistics under the alternative hypothesis. The significance level α determines the upper and lower cut-off value used for significance testing through a quantile function of the standard normal distribution Φ^{-1}. The cut-off is subsequently used in the cumulative probability density function of the normal distribution Φ_μ with mean μ and standard deviation equal to 1, determine the probability of obtaining more extreme z-values than those equal to α,

$$\rho = 1 - \Phi_\mu(\Phi^{-1}(1 - \alpha/2)) + \Phi_\mu(\Phi^{-1}(\alpha/2)). \tag{4}$$

The μ parameter of the cumulative density function of the normal distribution for a two-sample independent z-test is dependent only on the effect size d and the number of participants n split equally into the groups (Equation (5)). More participants or larger effect size means that the distribution of z-statistics has a higher mean μ,

$$\mu = d\frac{\sqrt{n}}{2}. \tag{5}$$

Equations (4) and (5) are also depicted for a concrete example with $n = 100$, $d = 0.5$ and $\alpha = .05$ (Figure 3).

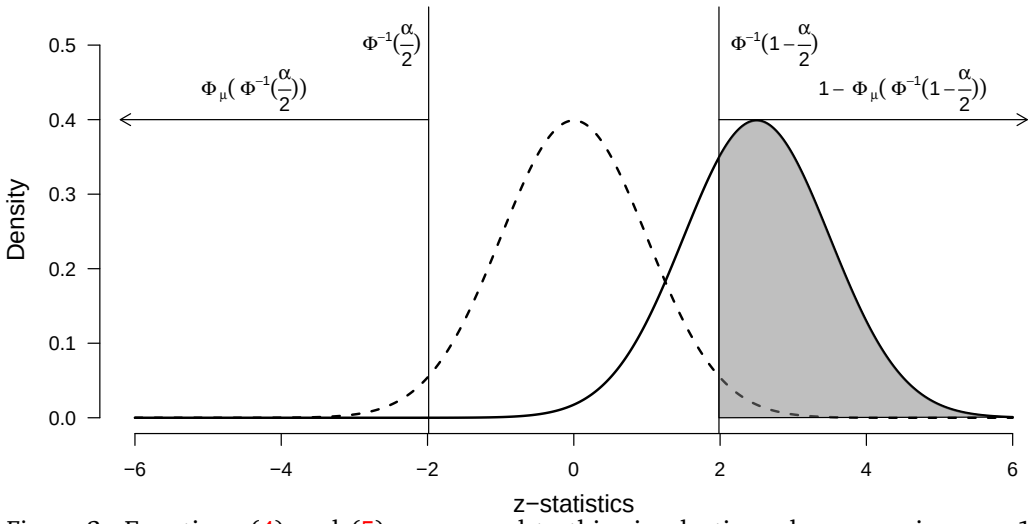

Figure 3. Equations (4) and (5) correspond to this visualzation when assuming $n = 100$, $d = 0.5$ and $\alpha = .05$. The vertical lines correspond to the cut-off z-statistic computed using a quantile function of the normal distribution under the null hypothesis (dashed line). The full line corresponds to the expected distribution of z-statistics under the alternative hypothesis with the grey-filled area corresponding to the power computed using a cumulative density function.

If α is decreased, the vertical lines placed at the cut-off z-statistic determined by the quantile function of the normal distribution move further apart from the center and thus reduce the grey-filled area corresponding to the power. On the other hand, one could also increase α and thus increase the area corresponding to power.

So, given constant sample size and effect size, researchers are faced with two possibilities: they can either (a) increase α, reducing the cut-off and thus achieving higher power; or (b) decrease α and subsequently lower the power. We know that there is a convention to set α in statistical tests to 5%. However, there is no reason why α should remain constant at this fixed value. Fisher (1956) explained that the 5% should be disregarded whenever there are other substantial reasons to determine α. More recently, scientists again called for a more flexible adaption of α (Lakens et al., 2018).

In other words, in psychological science that operates with limited resources, there is always a trade-off that needs to be made between avoiding false positives and detecting true positives. If Fisher wants to mitigate the proportion of members of RTTS with no tea tasting abilities (assuming he has a constrained budget), he is faced with two options. On the one hand, he can decrease α and lower the number of false positives with the cost of decreased power and fewer true positives. On the other hand, he can also increase the power and increase the number of true positives at the cost of increasing α leading to more false positives. The important question is, which is more efficient in lowering the FDR: lowering α or increasing power? We show that for a two-sided z-test and for one-sided z-test with true effect in the predicted direction given a constant sample size, decreasing α leads to lower FDR than increasing statistical power. Figure 4 shows this relationship on an example with an independent samples z-tests for the proportion of true null hypothesis $P(\mathcal{H}_0) = 0.5$, effect size $d = 0.5$ and sample size $n = 100$ (50 per group).

Similar results can be obtained for different sample sizes, effect sizes, proportions of null hypotheses being true and statistical tests (code to generate 3D plots across different μs can be found at https://osf.io/uszxk/). There is always a decrease in FDR with decreasing α but for two exceptions. First, if the null hypotheses are either all false or all true (which would include effect size equal to 0), then the proportion is 1 or 0 respectively, independent of power and α. Second, for one-sided tests where the true effect is opposite to the expected direction, the FDR will increase with reducing α. However, these two situations should be relatively rare in practice; therefore, reducing α is usually the most efficient way to decrease the FDR.

For a more formal analysis we also calculated the gradient of the FDR with regards to α (see Supplementary Materials at https://osf.io/svu7r). This elaborates the conclusion that α is more efficient in reducing the FDR, since the derivative is positive for all values of α apart from one-sided tests with an effect in the opposite direction. 3D plots showing the derivative for different noncentrality parameters (ncps) can be found at

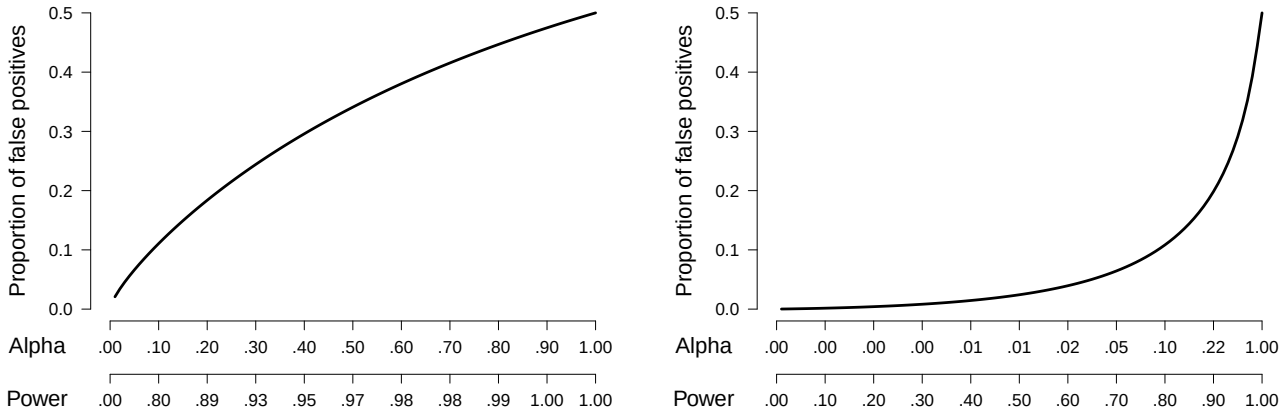

Figure 4. Trading off between power and α with $P(\mathcal{H}_0) = 0.5$, $d = 0.5$ and $n = 100$ (50 per group) results in the displayed FDR. The double x-axis shows α with its corresponding power, scaled according to α in the left chart and according to power in the right chart. To plot the relationship for other statistical tests, see https://osf.io/uwkqz/.

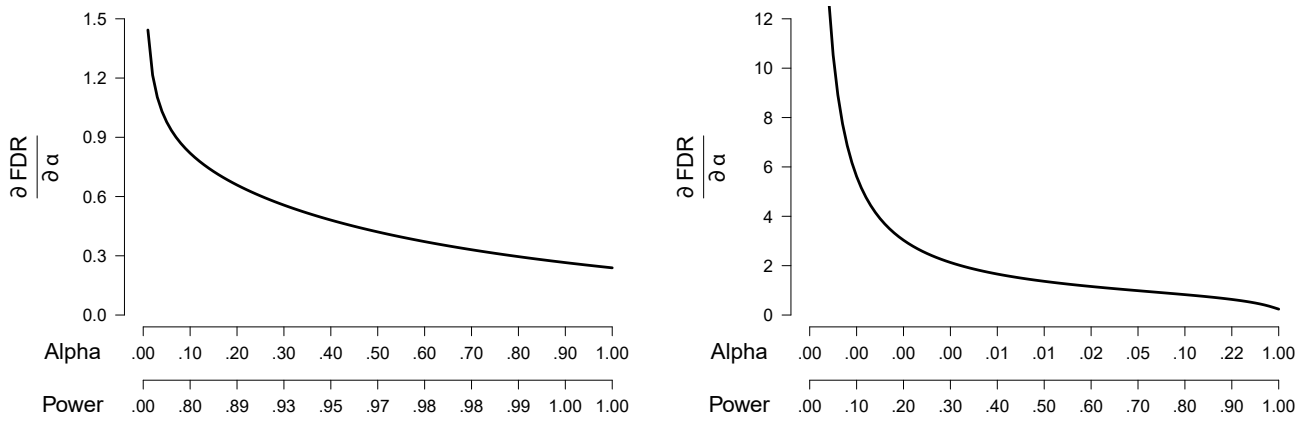

Figure 5. The figure displays the gradient of FDR with respect to α (and corresponding power) from a trade-off between power and α with $P(\mathcal{H}_0) = 0.5$, $d = 0.5$ and $n = 100$ (50 per group). The double x-axis shows α with its corresponding power, scaled according to α in the left chart and according to power in the right chart.

https://osf.io/uszxk/. Figure 5 shows the gradient an independent samples z-tests for the proportion of true null hypothesis $P(\mathcal{H}_0) = 0.5$, effect size $d = 0.5$ and sample size $n = 100$ (50 per group).

An expected objection is that instead of the trade-off by increasing α, one can achieve an increase in power by increasing sample size. As explained before, there is no apparent reason for keeping α constant with increasing sample size. Instead, one can keep the power fixed and use the higher sample size to decrease α. Figure 6 shows that keeping the power constant and decrease α by increasing the sample size is more efficient in lowering the FDR.

Again, a similar pattern can be observed irrespective of the starting sample size, α, power, effect size, and proportion of true null hypotheses. The decrease in FDR is stronger when using the increase in sample size to reduce α rather than increase the power.

Discussion

Our analysis shows that reducing α is usually more effective in reducing the false discovery rate than increasing power. Researchers striving to reduce the false discovery rate should reduce their α instead of increasing power. This is not only true when planning a study and deciding on the levels of α and power, but also when balancing power and α at a constant sample size or when increasing sample size and considering whether to "spend" the additional participants on increasing power or reducing α.

Our conclusion is similar to the long-standing literature on α adjustments for controlling the false discovery rate in multiple testing (e.g., Benjamini & Hochberg,

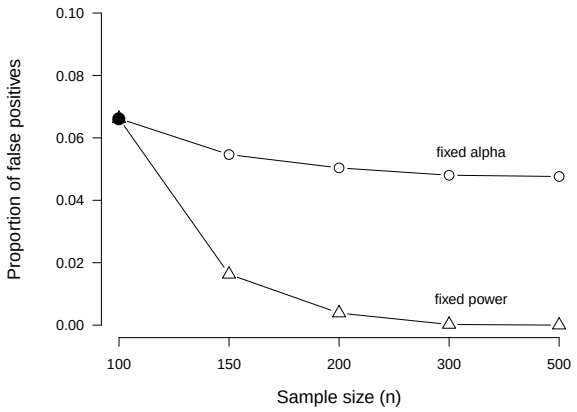

Figure 6. The displayed FDR results when either keeping the power (triangles) or α (circles) fixed while increasing sample size. The filled circle marks the starting point at $n = 100$ (50 per group) with $P(\mathcal{H}_0) = 0.5$, $d = 0.5$, resulting in power = 0.70 and $\alpha = 0.05$.

1995). However, its main goal is to keep the false discovery rate for a set of tests below a certain threshold rather than trading α and power in respect to the FDR.

We also need to consider several limitations of our analyses. In case of one-sided tests, reducing α is only more beneficial if the true effect is in the expected direction. In case of two-sided tests, incorporating type S errors into the definition of FDR increases the effectiveness of power if it is close to α. However, both of these scenarios are not plausible under common conditions. In addition, for balancing α and power, we only present results for the two-sample z-test and assuming that the assumptions of the statistical test (e.g., homoskedasticy and normal distribution) are fulfilled. While the relation between power and α and FDR for a variety of other tests can be found at https://osf.io/uwkqz/ and is in line with our analysis, a formal proof that the proposed relationship is holding for all tests under different conditions is not presented in this paper. More research is needed to generalize our results to more kinds of tests and settings. We also only analyze the effect of α and power, while an additional issue causing non-replicability can be a low prior probability of the tested hypotheses (Benjamin et al., 2018; Hoogeveen et al., 2020; Ioannidis, 2005), which plays a direct role in the FDR formula.

In addition, we want to emphasize that we are still advocates of high power for several reasons.[4] First, high power is crucial for avoiding Type II errors. Controlling Type I errors is often perceived as more important than controlling Type II errors (e.g., Cohen, 1956); however, in some contexts, Type II errors might be more problematic (Fiedler et al., 2012). For example, consider researchers first investigating a new, potentially groundbreaking treatment for depression. Here, the Type II error of not detecting the effectiveness of the treatment might be more costly than concluding that the treatment is effective when it is not. This error (and consequently abandoning this line of research) would mean missing an opportunity to improve the lives of people with depression. Another example might be replication studies, where the primary focus is to test whether a previously reported effect is there, with a lesser concern of inflating FDR. Here, high power is crucial to avoid such Type II errors. In addition, low power and conditioning on significance leads to an overestimation of effect sizes (Type M error) and to effect size estimates in the wrong direction (Type S error; Gelman and Carlin, 2014). For these reasons, high powered studies are crucial for cumulative science. Therefore, we recommend that in practice, researchers think about their inferential goals, weighing the costs of both Type I and Type II errors, to determine an optimal α and power (Lakens et al., 2018; Maier & Lakens, 2022; Miller & Ulrich, 2019; Mudge et al., 2012). If an important goal is to reduce the FDR, our analyses show that reducing α is more effective than increasing power.

Last but not least, we want to point out that the actual α level is often higher than the nominal α level due to questionable research practices, such as optional stopping or failure to report all dependent variables (John et al., 2012; Simmons et al., 2011; Wicherts, 2017). Therefore, finding ways to prevent these practices using tools such as preregistration (Nosek et al., 2018) and registered reports (Chambers et al., 2015) is probably one of the most critical tasks psychological science is facing. Some researchers also argue that we should abandon the framework of statistical testing and instead focus solely on summarizing the full information about effect size estimates (McShane et al., 2019).

Conclusion

We strove for two objectives in this paper. Firstly, we reiterated over α, power, and false discovery rate, hopefully improving the understanding of these concepts. Secondly, we compared two previously proposed solutions to decreasing the false discovery rate. Our results show that with respect to the false discovery rate, it is usually more effective to decrease α than to increase statistical power. We suggest that researchers interested in reducing the false discovery rate focus on reducing α.

[4] And we do not fear that our article will lead to a decrease in power since six decades of articles calling for an increase in statistical power had no visible impact (Smaldino & McElreath, 2016).

Author Contact

František Bartoš; f.bartos96@gmail.com; Department of Psychological Methods, University of Amsterdam; Department of Arts, Faculty of Arts, Charles University; ORCID: 0000-0002-0018-5573
Maximilian Maier; maximilianmaier0401@gmail.com; Department of Psychological Methods, University of Amsterdam; ORCID: 0000-0002-9873-6096

Acknowledgments

We would like to thank Marie Delacre, Jiří Štipl, and Franziska Nippold for helpful comments and suggestions on previous versions of this manuscript.

Conflict of Interest and Funding

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Author Contributions

Both authors contributed equally to all stages of the research process and writing the manuscripts.

Open Science Practices

This article earned the Open Materials badge for making the materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2.

Christley, R. M. (2010). Power and error: Increased risk of false positive results in underpowered studies. *The Open Epidemiology Journal*, *3*(1). http://dx.doi.org/10.2174/1874297101003010016

Cohen, J. (1956). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, *4*(12), 171085. https://doi.org/10.1098/rsos.171085

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*(6), 661–669. https://doi.org/10.1177/1745691612462587

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Hafner Publishing.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M

(magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E.-J. (2020). Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*(3), 267–285. https://doi.org/10.1177/2515245920919667

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahnık, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., et al. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lehmann, E. (1992). Introduction to Neyman and Pearson (1933) On the problem of the most efficient tests of statistical hypotheses. *Breakthroughs in statistics* (pp. 67–72). Springer.

Liao, J. G., & Rosen, O. (2001). Fast and stable algorithms for computing and sampling from the noncentral hypergeometric distribution. *The American Statistician*, *55*(4), 366–369. https://doi.org/10.1080/17470218.2012.711335

Maier, M., & Lakens, D. (2022). *Justify your alpha: A primer on two practical approaches* (No. *2*).

Masicampo, E., & Lalande, D. R. (2012). A peculiar prevalence of *p*-values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1371/journal.pone.0208631

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLoS One*, *14*(1), e0208631. https://doi.org/10.1371/journal.pone.0208631

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PloS One*, *7*(2), e32734. https://doi.org/10.1371/journal.pone.0032734

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, *41*(11), 1299. https://doi.org/10.1037/0003-066X.41.11.1299

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 175–240. https://doi.org/10.1093/biomet/20A.3-4.263

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, *55*(1), 33–38. https://doi.org/10.1080/00223980.1963.9916596

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports*, *15*(2), 570. https://doi.org/10.2466/pr0.1964.15.2.570

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

van Aert, R. C., Wicherts, J. M., & Van Assen, M. A. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PloS One*, *14*(4), e0215052. https://doi.org/10.1371/journal.pone.0215052

Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals*, *7*(12), 90–119. https://doi.org/10.3390/ani7120090