PLOS ONE

RESEARCH ARTICLE

The quest for an optimal alpha

Jeff Miller[1]*, **Rolf Ulrich**[2]

1 Department of Psychology, University of Otago, Dunedin, New Zealand, **2** Research Group for Cognition and Perception, Department of Psychology, University of Tübingen, Tübingen, Germany

* miller@psy.otago.ac.nz

Abstract

Researchers who analyze data within the framework of null hypothesis significance testing must choose a critical "alpha" level, α, to use as a cutoff for deciding whether a given set of data demonstrates the presence of a particular effect. In most fields, $\alpha = 0.05$ has traditionally been used as the standard cutoff. Many researchers have recently argued for a change to a more stringent evidence cutoff such as $\alpha = 0.01$, 0.005, or 0.001, noting that this change would tend to reduce the rate of false positives, which are of growing concern in many research areas. Other researchers oppose this proposed change, however, because it would correspondingly tend to increase the rate of false negatives. We show how a simple statistical model can be used to explore the quantitative tradeoff between reducing false positives and increasing false negatives. In particular, the model shows how the optimal α level depends on numerous characteristics of the research area, and it reveals that although $\alpha = 0.05$ would indeed be approximately the optimal value in some realistic situations, the optimal α could actually be substantially larger or smaller in other situations. The importance of the model lies in making it clear what characteristics of the research area have to be specified to make a principled argument for using one α level rather than another, and the model thereby provides a blueprint for researchers seeking to justify a particular α level.

Check for updates

OPEN ACCESS

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

Editor: Yun Li, University of North Carolina at Chapel Hill, UNITED STATES

Received: July 27, 2018

Accepted: November 20, 2018

Published: January 2, 2019

Copyright: © 2019 Miller, Ulrich. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (Statistical Modeling in Psychology, GRK 2277). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The statistical methods traditionally used in psychology, medicine, economics, and many other empirical disciplines have recently come under intense scrutiny, primarily because a large number of published results appear to reflect chance findings—so-called *false positives* (FPs)—rather than replicable scientific phenomena [1–5]. There have long been concerns that FP rates might be unacceptably high due to a combination of publication bias [6], the rareness of true effects within certain research areas [7, 8], and inappropriate data analysis methods [2, 4, 9], as well as outright fraud [10]. Such concerns have recently been intensified by empirical evidence, both from surveys indicating that researchers do engage in practices known to increase FP rates (e.g., [11–13]; but see [14]), and from the detection of statistical signs that published results have been contaminated by such practices [15–18]. Most tellingly, various systematic attempts to replicate published results have ended with disappointingly low replication rates (e.g., [19–22]; but see [23]). In light of this evidence, numerous strategies for reducing the worryingly high rate of published FPs have been proposed [24–26], and there is good

agreement that common scientific practices and processes can be improved in a number of ways.

One particular strategy for reducing the rate of FPs is at present hotly debated; namely, the strategy of reducing the critical α level for concluding that an effect is real. In contrast to the $\alpha = 0.05$ level that was suggested by Fisher [27] and has been standard for many years [28], various authors have recently argued that much smaller α levels should be used [29–32]. For example, in an article with 72 authors, Benjamin et al. argued that researchers should change to using $\alpha = 0.005$ rather than $\alpha = 0.05$, because this change in α would be expected to reduce the rate of FPs [33]. Benjamin et al. also argued that "a change to P = 0.005. . .would immediately improve the reproducibility of scientific research in many fields" (p. 6). Contrary to this claim, however, changing from $\alpha = 0.05$ to $\alpha = 0.005$ can actually decrease the probability of a successful replication if the same α level is used for all studies. As a numerical example, consider the case of a two-sample t-test with $n = 60$ participants per sample and a true effect size of $d = 0.5$ that is present with a base rate of $\pi = 0.3$. The probability of successful replication would be 0.76 for $\alpha = 0.05$ but only 0.54 for $\alpha = 0.005$, illustrating that decreasing α can decrease the probability of successful replication.

Others, however, have argued against the move to reduce α. In a reply to Benjamin et al. signed by 88 authors, Lakens et al. noted that a reduction in α would also have various negative consequences [34]. Perhaps most importantly, decreasing α would decrease statistical power and thereby increase the rate of *false negatives* (FNs)—that is, the proportion of studies that fail to find conclusive evidence for an effect that actually is present [14, 35].

Statistical significance at the conventionally agreed α level is a major factor in determining what findings are regarded as having been firmly-enough established to warrant publication [36–38], so it is clearly very important to determine the optimal level. The current debate about α, however, illustrates the complexity of determining its optimal value [39, 40]. Indeed, there are good reasons to believe that no single α level is optimal for all research contexts [34], and in some contexts there are strong arguments for increasing the α level to a value larger than 0.05 [41]. At this point, the only agreement concerning the choice of α level is that researchers within a given area should make it carefully—but how are they to do that?

The purpose of the present article is to show exactly what is necessary to provide a principled justification for a particular α level. Using well-known principles of statistical decision theory [42] within the context of a simple mathematical model suggested previously [43], we identify the parameters of a research scenario that must be considered when choosing the optimal α level for that scenario, and we indicate how the effects of those parameters can be combined quantitatively. To illustrate this model, we then show how it can be used to determine whether $\alpha = 0.005$ or $\alpha = 0.05$ would work better within a particular research scenario, given the required information about that scenario's parameters. We conclude that no definitive case for any particular α level has yet been made, because advocates of particular α levels have never specified—even approximately—the key research parameters whose values are needed to identify the optimal α. In addition, although it is universally acknowledged that many factors must be taken into account when choosing α, no quantitative models have been used to compare the overall costs and benefits of different α levels, with proponents of different viewpoints relying instead on rather subjective justifications such as "We believe that efficiency gains [of a change to $\alpha = 0.005$] would far outweigh losses" (p. 8, [33]). To provide an objective basis for the debate, in the following sections we show how a simple model based on the principles of statistical decision theory can be used to quantify the costs and benefits of various α levels, as is required for researchers to choose the optimal one.

1 Statistical fundamentals

The tradeoff between FPs and FNs can be formalized within a simple model in which the overall research scenario is regarded as a collection of studies testing different null hypotheses [21]. Some null hypotheses are false, and we refer to the proportion of these as the *base rate* of true effects, denoted π. The remaining null hypotheses, with proportion $1 - \pi$, are true, at least to a good approximation. In each study, the null hypothesis is rejected or not rejected, depending on whether a statistical analysis produces significant results at the chosen α level. Thus, studies testing true null hypotheses may produce either FPs or true negative (TN) outcomes, whereas studies testing false null hypotheses may produce either FNs or true positive (TP) outcomes. The probabilities of these four outcomes are

$$\Pr(FP) = (1-\pi)\cdot\alpha \quad (1)$$

$$\Pr(TN) = (1-\pi)\cdot(1-\alpha) \quad (2)$$

$$\Pr(FN) = \pi\cdot\beta \quad (3)$$

$$\Pr(TP) = \pi\cdot(1-\beta), \quad (4)$$

where $1 - \beta$ is the statistical power of the test of each false null hypothesis. This power depends on the α level, the size of the true effect, d, and on the sample size, n_s. The rate of false positives (R_{fp}) and rate of false negatives (R_{fn}) are then

$$R_{fp} = \frac{\Pr(FP)}{\Pr(FP) + \Pr(TP)} \quad (5)$$

$$R_{fn} = \frac{\Pr(FN)}{\Pr(FN) + \Pr(TN)}. \quad (6)$$

Fig 1 illustrates how these rates change with the researcher's α level, showing results for two different sample sizes (n_s), three different effect sizes (d), and a wide range of base rates (π). Critically, for every combination of sample size, effect size, and base rate, the rate of FPs is higher with $\alpha = 0.05$ than with $\alpha = 0.005$. In contrast, the rate of FNs is always higher for $\alpha = 0.005$ than for $\alpha = 0.05$. Thus, these two types of decision errors trade off against one another as α changes, and a quantitative model incorporating the frequencies and costs of these errors must be used to choose α.

2 Choosing between $\alpha = 0.05$ and $\alpha = 0.005$

The costs and benefits of using alternative α levels can be assessed quantitatively using standard decision-theory methods [42, 43]. Any given empirical study will produce one of four possible outcomes (i.e., TP, FP, TN, FN) with the probabilities just described [i.e., Pr(*TP*), Pr(*FP*), Pr(*TN*), Pr(*FN*)]. Each of the four outcomes has its own individual *informational payoff value*, and these values may be denoted as $\mathcal{P}_{tp}$, $\mathcal{P}_{fp}$, $\mathcal{P}_{tn}$, and $\mathcal{P}_{fn}$, respectively. The units of these informational payoffs are entirely arbitrary, so it is convenient to fix $\mathcal{P}_{tp} = 1$ and scale the other payoffs relative to that. On this scale, for example, $\mathcal{P}_{fp} = -2$ means that the informational harm to a research area of one FP exactly offsets the informational benefit of two TPs. These individual outcome payoffs would vary across research areas, and it would usually only be possible to estimate them rather subjectively. For example, within a certain research area, researchers might regard the information gains associated with TPs and TNs as representing

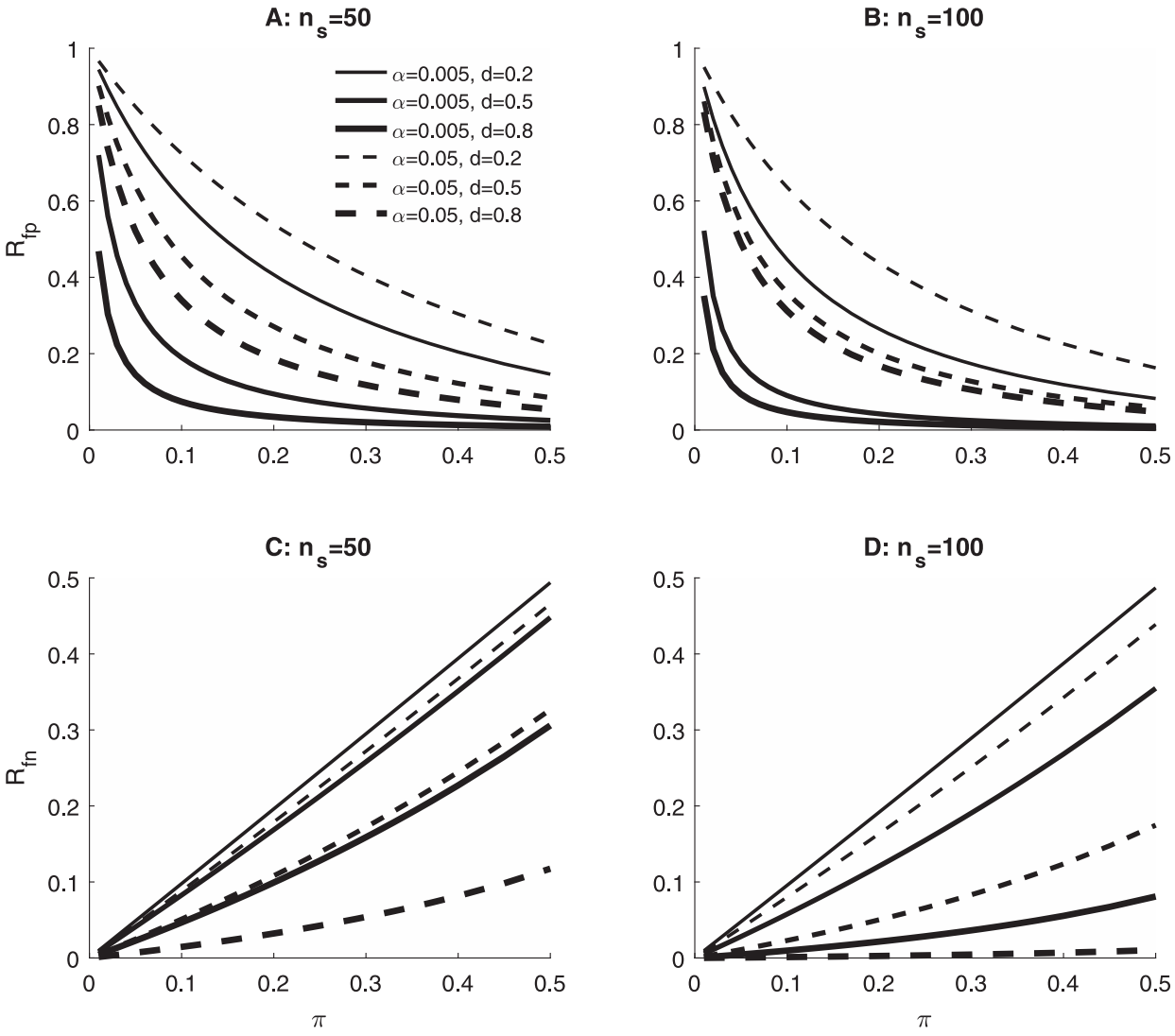

Fig 1. Rates of false positives and false negatives in different research scenarios. R_{fp} (A, B) and R_{fn} (C, D) as functions of the α level used in testing the null hypothesis, the base rate of true effects across studies (π), the size of the true effects when they are present (d), and the total sample size of the study (n_s). Computations were carried out for studies analyzed with one-tailed two-sample t-tests (i.e., samples of $n_s/2$ in each group). Effect size d is the difference between the group means divided by the common within-group standard deviation, $d = (\mu_2 - \mu_1)/\sigma$.

https://doi.org/10.1371/journal.pone.0208631.g001

payoffs of +1 and +0.2, whereas the losses associated with FPs and FNs could be estimated to be -2 and -0.5. The average informational payoff for a single study is simply the weighted average of the four individual outcome payoffs:

$$\mathcal{P}_1 = \Pr(TP) \cdot \mathcal{P}_{tp} + \Pr(FP) \cdot \mathcal{P}_{fp} + \Pr(TN) \cdot \mathcal{P}_{tn} + \Pr(FN) \cdot \mathcal{P}_{fn}. \quad (7)$$

Finally, the total payoff associated with all of the studies conducted within the research scenario is

$$\mathcal{P}_T = k \cdot \mathcal{P}_1, \quad (8)$$

where k is the number of studies conducted within that scenario.

Researchers in a given research scenario must make two separate choices that could influence their expected payoffs, and their goal is to make these choices in a manner that will

maximize that payoff. One choice is the sample size (n_s) to be used in each study. Larger studies have greater statistical power ($1 - \beta$), but they are more time-consuming and expensive to conduct, so increasing n_s decreases the total number of studies (k) that can be conducted. The other choice is the α level, which is currently being debated. For simplicity, we and others discuss the choice of α level as if it were entirely up to the researchers. In practice, though, the researchers' choice of α level may be heavily constrained by the editorial policies of the journals in which they hope to publish their results [36]. In principle, though, the researchers' problem is to choose the particular values of n_s and α that produce the maximum total payoff, $\mathcal{P}_T$.

Fig 2 illustrates the consequences of the researchers' choices by showing the expected total payoff as a function of sample size (n_s) and α level for several example research scenarios

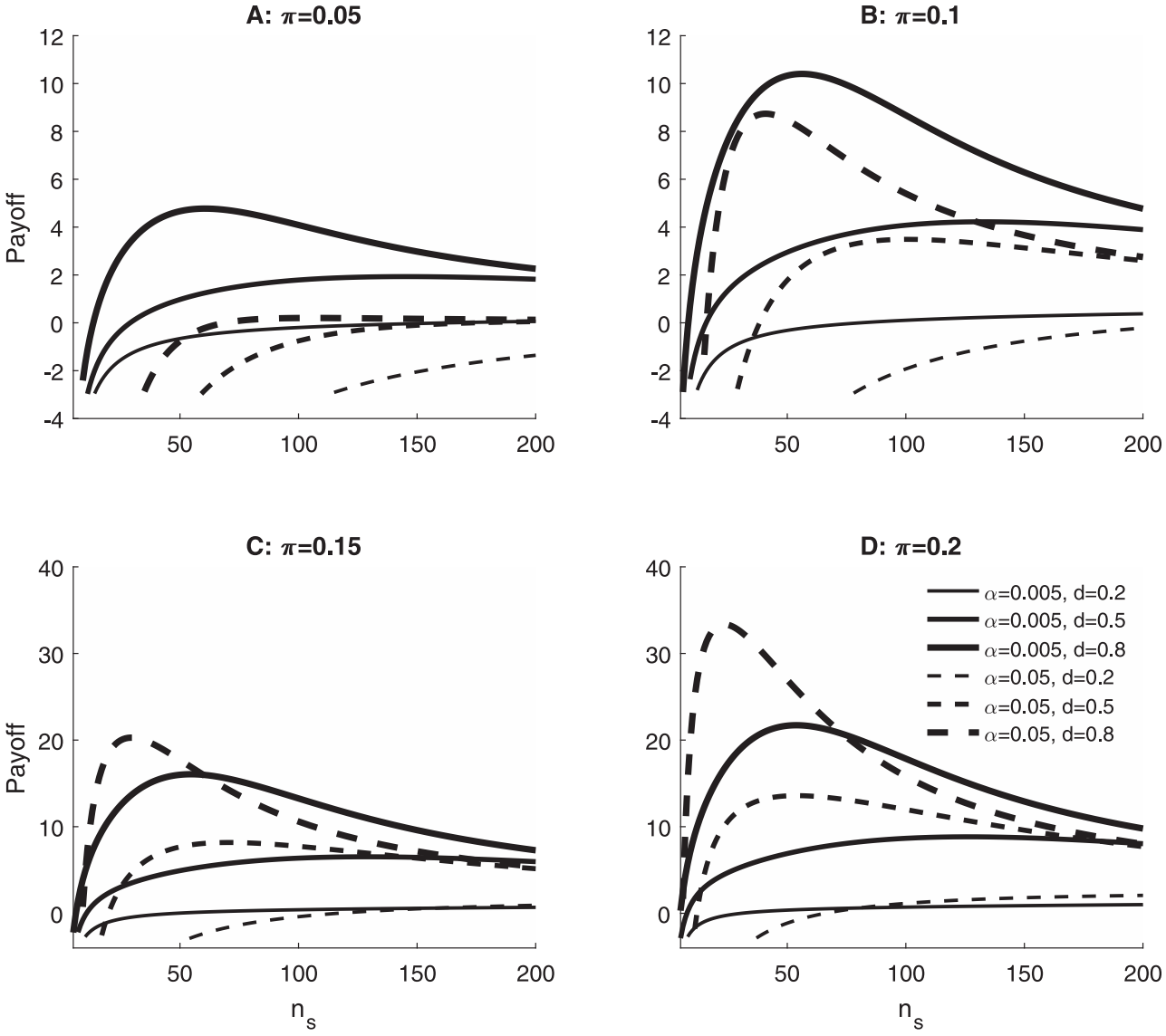

Fig 2. Expected payoffs in different research scenarios. Expected total payoff, ($E[\mathcal{P}_T]$, ordinate) as a function of α level and sample size (n_s) for research scenarios differing in the size of a true effect when it is present ($d = 0.2$, 0.5, or 0.8) and in the base rate probability that the true effect is present (π). The range of base rates, 0.05–0.20, spans approximately the range 0.024–0.167 used by Benjamin et al. in their computational examples [33]. Payoffs were computed from Eq 8 using individual outcome payoffs of $\mathcal{P}_{tp} = 1$, $\mathcal{P}_{fp} = -1$, $\mathcal{P}_{tn} = 0$, and $\mathcal{P}_{fn} = 0$ and assuming a total sample size of 10, 000 across all studies (i.e., $k = 10, 000/n_s$). Computations were carried out for studies analyzed with one-tailed two-sample t-tests (i.e., samples of $n_s/2$ in each group). Effect size d is the difference between the group means divided by the common within-group standard deviation, $d = (\mu_2 - \mu_1)/\sigma$.

https://doi.org/10.1371/journal.pone.0208631.g002

differing in the base rate (π) and size (d) of true effects. Researchers seek to maximize their payoff, of course, so they would be advised to use $\alpha = 0.005$ with any combination of parameters (i.e., π, d, and n_s) for which the solid line is above the dashed line, but to use $\alpha = 0.05$ with combinations for which the solid line is below the dashed line. In addition, though, researchers can choose the sample size, so they should also choose the sample size that leads to the highest payoff. In Fig 2B with $\pi = 0.1$ and $d = 0.5$, for example, the highest possible expected payoff across all sample sizes is approximately 4.22, which is obtained with $\alpha = 0.005$ and $n_s = 135$. Thus, $\alpha = 0.005$ is preferable to $\alpha = 0.05$ in this situation. In contrast, in Fig 2C with $\pi = 0.15$ and $d = 0.5$, the highest payoff is approximately 8.18, obtained with $\alpha = 0.05$ and $n_s = 70$, so $\alpha = 0.05$ would be preferable in this case.

The contrast between the two numerical examples just discussed has profound implications for the current controversy over the best choice of α level. If $\alpha = 0.005$ is better when the base rate is less than $\pi = 0.10$ but $\alpha = 0.05$ is better when the base rate is greater than $\pi = 0.15$, then it follows that *researchers must know the base rate of true effects*—at least approximately—before they can choose the right α level. This shows that advocates of a particular α level should specify the range of base rates being assumed and acknowledge that other α levels would be appropriate for other base rates. To our knowledge this has never been done, although Benjamin et al. did support their argument for $\alpha = 0.005$ partly by presenting evidence for a base rate of approximately 10% [33].

For the present purposes, another important point illustrated by Fig 2 is that the choice between $\alpha = 0.05$ and $\alpha = 0.005$ can depend on the sample size. For example, with $d = 0.8$ in Fig 2C and 2D, the payoff is higher for $\alpha = 0.05$ at some sample sizes but higher for $\alpha = 0.005$ at other sample sizes (i.e., the thick solid and dashed lines cross within both panels). This dependence of α preference on sample size is also quite relevant to the debate about α levels. It implies that there is no single best α across all sample sizes—even within a given research scenario. Again, this implies that advocates of a particular α level must specify the sample sizes to which their recommendations apply as well as the base rates of true effects.

Finally, the choice between $\alpha = 0.05$ and $\alpha = 0.005$ also depends strongly on the exact quantitative payoffs associated with TPs, FPs, TNs, and FNs. To illustrate that, Fig 3 shows how the total payoffs available with $\alpha = 0.05$ and $\alpha = 0.005$ depend on the individual payoffs associated with false positives ($\mathcal{P}_{fp}$) and false negatives ($\mathcal{P}_{fn}$), as well as the base rate of true effects (π), and the size of the effect when it is present (d). Each plotted total payoff value is the maximum (i.e., across all possible values of sample size, n_s) for the indicated combination of parameters, so the figure is computed assuming that researchers have chosen the optimal sample size for each scenario. Again, $\alpha = 0.005$ should be preferred with any combination of parameters for which the solid line is above the dashed line, but it is better to use $\alpha = 0.05$ with combinations for which the solid line is below the dashed line. Comparing the different panels, it is easy to see that the cross-over points for $\alpha = 0.05$ versus $\alpha = 0.005$ depend heavily on the individual payoffs associated with the various outcomes. As was true with base rates and sample sizes, this implies that advocates of a particular α level must provide the values of individual outcomes to which their recommendations apply and acknowledge that other α levels could be appropriate for other values.

Several general lessons about the relative merits of $\alpha = 0.05$ and $\alpha = 0.005$ can be seen in Fig 3, and these help to sharpen intuitions about exactly when each of the two α levels—together with its optimal sample size—would be preferable. First, for the values of $\mathcal{P}_{fp}$ and $\mathcal{P}_{fn}$ examined here, $\alpha = 0.005$ yields larger payoffs when the base rate of true effects is smaller than approximately 0.1, whereas $\alpha = 0.05$ yields larger payoffs when the base rate is larger than approximately 0.4, which provides some boundaries for the use of each α level. Second, for

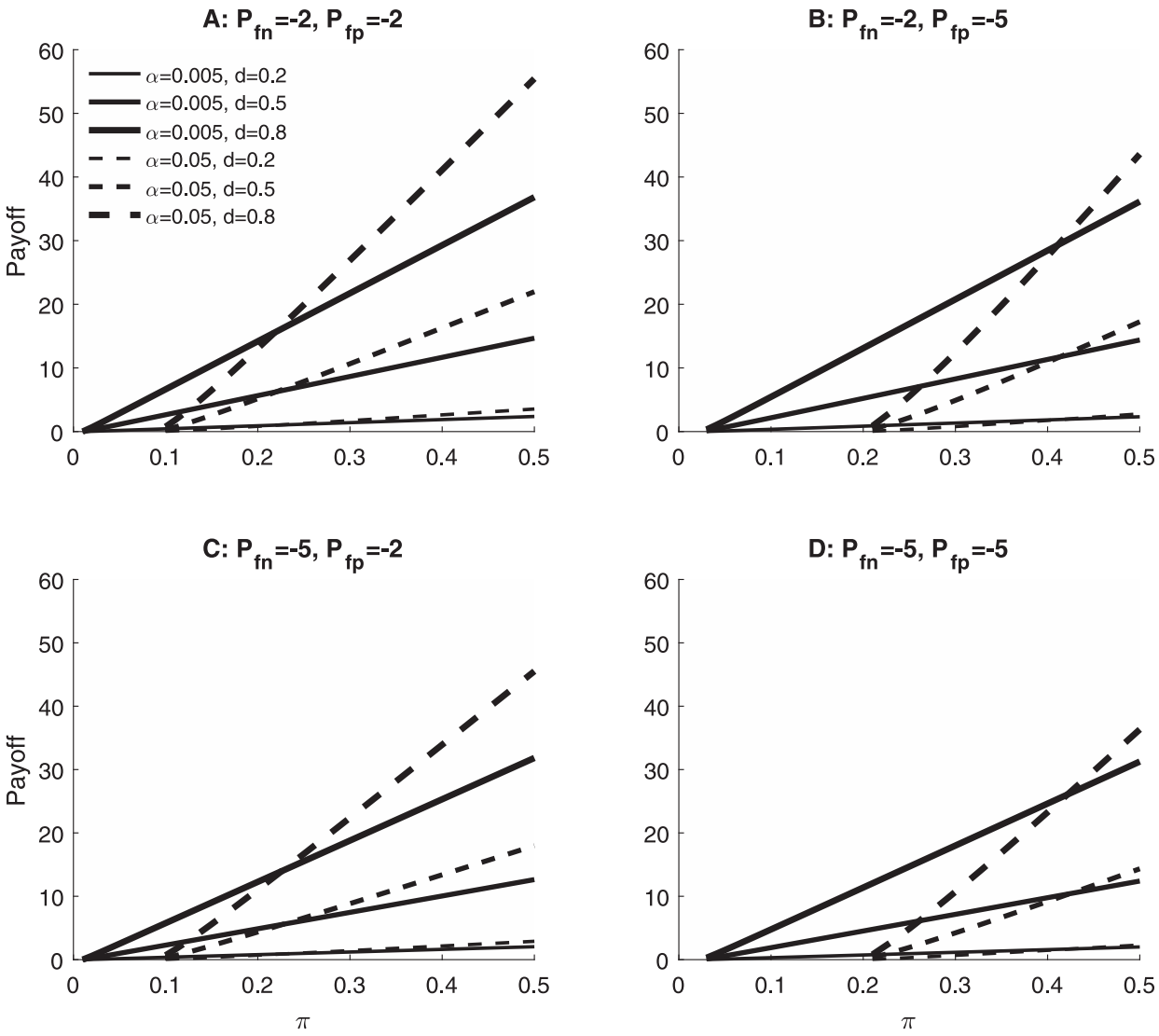

Fig 3. Maximum expected payoffs at optimal sample sizes. The maximum expected total payoff (ordinate), taken across all possible values of n_s, that could be achieved for each combination of α level, base rate probability that a true effect is present (π), the size of the effect when it is present (d = 0.2, 0.5, or 0.8), the payoff associated with false positives ($\mathcal{P}_{fp}$), and the payoff associated with false negatives ($\mathcal{P}_{fn}$). Payoffs were computed as in Fig 2 using individual outcome payoffs of $\mathcal{P}_{tp} = 1$ and $\mathcal{P}_{tn} = 0$. Computations used the same statistical test and definition of d as in Fig 2.

https://doi.org/10.1371/journal.pone.0208631.g003

intermediate base rates (i.e., $0.1 < \pi < 0.4$), the cross-over points at which researchers should switch between the two α levels are quite sensitive to the cost associated with FPs, as can be seen by comparing Fig 3A with 3B. Qualitatively, this is exactly as expected: To the extent that FPs are relatively costly (e.g., $\mathcal{P}_{fp} = -5$ as opposed to $\mathcal{P}_{fp} = -2$), $\alpha = 0.005$ tends to be preferred over $\alpha = 0.05$ because it produces fewer of them. Third, and perhaps somewhat surprisingly, the 0.05/0.005 cross-over points are not very sensitive to the cost associated with FNs, as can be seen by comparing Fig 3A with 3C or Fig 3B with 3D. This is presumably because the base rate of true effects, π, is low, which means that FNs are rare so their cost is not too important. The situation would be reversed if the base rate were high, because in that case FPs would be rare and their cost could be relatively unimportant compared to that of FNs. Fourth, and also somewhat surprisingly, the 0.05/0.005 cross-over points do not seem to be affected much

by the true effect size d. In Fig 3A, for example, the solid and dashed lines cross at a base rate of about $\pi = 0.23$ for all three d values, and the cross-over base rates are also fairly constant across d values in Fig 3B–3D. Illustrative computations in S1 Appendix "Supplementary analysis of other possible α levels" show even more clearly that the optimal α level depends very little on the true effect size d. In summary, then, the optimal α value depends most heavily on the base rate of true effects and secondarily on the relative payoffs of the individual outcomes, especially on the cost of an FP when the base rate is low and on the cost of an FN when the base rate is high.

3 General discussion

By viewing empirical hypothesis testing in a decision-theoretic framework with fixed total resources (i.e., fixed total participants tested, $k \cdot n_s$), it is possible to calculate precisely how researchers' expected total scientific payoffs depend on their choices of α levels and sample sizes within any given research scenario. It is important to examine these total payoffs to understand which research scenario parameters must be considered and to see how the size of the payoff is jointly determined by the various parameter values. There is wide agreement that scientists in any field should consider their α levels carefully [33, 34, 44, 45], and it seems essential to use an objective formalism to compare the expected scientific payoffs of different α levels.

The present approach differs from previous statistical decision models, because these have usually considered the problem of choosing either α level or sample size while keeping the other value fixed. For example, the optimum choice of α has been investigated for a fixed sample size in terms of minimizing the expected number and/or cost of errors [39, 46]. Similarly, within the context of hypothesis testing, the sample size is usually chosen to achieve sufficient statistical power for a fixed α level [47]. The choice of sample size has also been analyzed outside of the hypothesis testing context, with an emphasis on maximizing overall economic, medical, or environmental benefits [48–51], but these analyses have had no clear implications for the choice of α.

The present approach highlights the fact that the optimal choices of α level and sample size depend in a complicated fashion on numerous parameters. Because α and sample size are the only parameters that are usually under the researchers' control, researchers should strive to make optimal choices for them to improve the use of scientific resources [1, 52, 53]. The other parameters (i.e., π, d, $\mathcal{P}_{tp}$, $\mathcal{P}_{fp}$, $\mathcal{P}_{tn}$, and $\mathcal{P}_{fn}$) are essentially inherent in the research area and are thus outside of the researchers' control, but their values must be considered nonetheless. The computations shown in Figs 2 and 3 reflect research scenarios in which there was either an effect of a given fixed size (i.e., $d = 0.2$, 0.5, or 0.8) or there was no effect at all. Analogous computations were carried out for scenarios in which the true effect size varied randomly, and the results were fairly similar. These computations and the differences from the present results are reported in S2 Appendix "Supplementary analysis of varying effect sizes".

Critically, the optimal choices of α level and sample size depend strongly on the values of these other, "out of control" parameters. This presents a challenge for researchers who would like to determine the optimal α level and sample size using Eq 8, because it is essential to obtain good estimates of their values. The standard expected value model underlying Eq 8 is valuable partly because it clarifies exactly which parameters must be estimated to argue for a particular α level or sample size. In addition, considerable insight can be gained from total payoff computations by making rough estimates and performing "what if" calculations, as is done in many scientific areas where parameter estimates are difficult to obtain. For example, economists use models to project future economic growth and activity, despite the fact that future economic

conditions (i.e., parameter values) are unknown because conditions can change abruptly. Similarly, models of global climate change and of endangered species population sizes are used to make ball-park calculations and to inform decision-makers despite major uncertainties about key parameter values.

From the present results, it appears that the base rate of true effects, π, is the parameter with the strongest influence on the optimal choices of α level and sample size (e.g., Fig 3; also see Figs A–C in S1 Appendix). In the case of α level, for example, switching from the current standard $\alpha = 0.05$ to the proposed new $\alpha = 0.005$ could very well increase payoffs in scenarios where the base rate is lower than approximately 0.1, but it could equally well decrease payoffs if the base rate is higher than approximately 0.4. For scenarios with base rates within the range of 0.1-0.4, the choice between $\alpha = 0.05$ and $\alpha = 0.005$ would be heavily influenced by the relative costs of FPs and FNs.

We have focussed on comparing the payoffs for $\alpha = 0.05$ and $\alpha = 0.005$ as two specific values currently being advocated, but these are only two of the infinitely many possible α levels that could be used. In principle, it is possible to determine exactly which α level leads to the highest expected payoff, whether it is one of these two α levels or not. S1 Appendix "Supplementary analysis of other possible α levels" illustrates how this can be done and presents illustrative computations in which the optimal α level varies gradually from approximately 0.001 to 0.12. Among other things, the analysis in this supplement strongly reinforces the earlier suggestions—based on Fig 3—that the base rate has a large effect on the optimal α level whereas the true effect size d has little or no effect.

It is crucial to consider the realistic base rate carefully for each research area. Mudge et al. argued on logical grounds that researchers should assume a base rate of $\pi = 0.5$ in the absence of any relevant information [39], but many have argued that available information hints at base rates much lower than this. In particular, when there have been attempts to replicate previous findings in an area, the base rate can be estimated from the probability of successful replication. From replication rates reported recently [22], for example, it appears that the overall base rate of true effects is approximately 10% within a broad area of experimental psychology represented by the sample of replicated studies [43, 44]. Lower replicability in certain domains of biomedical research suggest that their base rates might be even lower [2, 24]. On the other hand, base rates might be much higher in research areas where there is better prior information about the mechanisms under investigation.

Several authors have suggested that researchers can get reasonable estimates of base rate from prior area-specific knowledge [2, 8]. These estimates obviously depend a great deal on the strength of the theoretical and empirical results suggesting that the tested effect would be present (i.e., the information that led the researchers to test for the effect in the first place). When researchers only have weak grounds for suspecting the presence of a certain effect, they could use an appropriately low estimate of the base rate—perhaps 0.1 or less. In contrast, when an effect is predicted by a detailed theory that has fared well in many previous tests, it would seem appropriate to use a much larger estimate—perhaps 0.9. A high estimate would also be appropriate when, for example, the effect was a minor extension or variation of a phenomenon that had previously been clearly demonstrated. Indeed, the dependence of the base rate estimate on prior knowledge has been tacitly acknowledged by advocates of stringent α levels like Benjamin et al., who proposed using $\alpha = 0.005$ only for the "discovery of new findings . . . [but not] for confirmatory or contradictory replications of existing claims" (p. 6, [33]).

Given the importance of estimating the base rate and given the uncertainties about how that can be done, we propose that experienced researchers can estimate the base rate in their own research areas by looking at the long-run relative frequency of getting significant results across many of their own experiments. The probability of a significant result in a study, p_{sig}, is

a function of the base rate of true effects π within the area, the α level, and statistical power $1 - \beta$:

$$p_{sig} = \pi \cdot (1 - \beta) + (1 - \pi) \cdot \alpha. \quad (9)$$

This equation can be solved for the base rate, yielding

$$\pi = \frac{p_{sig} - \alpha}{1 - \beta - \alpha}, \quad (10)$$

which allows individual researchers to estimate their true base rates from their own estimated values of α, power $1 - \beta$, and p_{sig}. As an example, suppose a researcher uses $\alpha = 0.05$, conducts studies with power approximately $1 - \beta = 0.55$, and finds significant results in approximately half of all studies (i.e., $p_{sig} = 0.5$). Using Eq 10, the researcher can estimate that the base rate of true effects across his or her past studies has been approximately 90%. Eq 10 can also be used to estimate a lower bound for the base rate when power is unknown. The left size of Eq 10 is minimal when $\beta = 0$, which implies

$$\pi \geq \frac{p_{sig} - \alpha}{1 - \alpha}. \quad (11)$$

Thus, for the same researcher with $\alpha = 0.05$ and $p_{sig} = 0.5$, the base rate must be at least 47%, regardless of the power level.

In addition to base rates, the optimal α levels for different scenarios also depend on the individual payoffs associated with the four possible hypothesis testing outcomes, $\mathcal{P}_{tp}$, $\mathcal{P}_{fp}$, $\mathcal{P}_{tn}$, and $\mathcal{P}_{fn}$. If the researchers working in a given field share a common sense of the approximate relative benefits and costs of these outcomes, then the agreed values would be helpful in working out the optimal α level. From informal discussions with colleagues, however, we believe that there are sometimes great disagreements about these values, with estimates of the FP cost varying by as much as two orders of magnitude (e.g., -2 to -200). When individual outcome payoffs are perceived so differently, it is only natural that researchers would prefer different α levels. Thus, the present analysis shows that a convincing case for a given α level must include quantitative assessments—together with supporting evidence—of the costs and benefits of the specific individual outcomes (i.e., TPs, FPs, TNs, FNs). Arguably, these assessments should come from observers who are not directly at the research coal face (e.g., journal editors, granting agencies), since the researchers themselves may have vested interests in reaching positive versus negative decisions.

In the end, the question of which α level researchers should use simply cannot be answered without a detailed quantitative model incorporating not only the researcher's choices of α level and sample size, but also the underlying characteristics of the research scenario and the costs and benefits of reaching the different possible correct and incorrect conclusions. To that end, traditional statistical decision models can be adapted to models of the research process [43], and we suggest that advocates of any particular α level should use such models—in conjunction with estimates of base rates and payoffs—to give their arguments a firm objective foundation.

Supporting information

S1 Appendix. Supplementary analysis of other possible α levels.
(PDF)

S2 Appendix. Supplementary analysis of varying effect sizes.
(PDF)

Acknowledgments

We are grateful to Denis Szucs for providing the full distribution of effect-size estimates summarized by Szucs and Ioannidis (2017), to Patricia Haden for advice on the design of the object-oriented software used for our computations as well as comments on earlier versions of the article, and to Lisa Harlow, Victor Mittelstädt, and Daniel Lakens for helpful comments on earlier versions of the article. This work was supported by the Deutsche Forschungsgemeinschaft (Statistical Modeling in Psychology, GRK 2277).

Author Contributions

Conceptualization: Jeff Miller, Rolf Ulrich.

Formal analysis: Jeff Miller, Rolf Ulrich.

Investigation: Jeff Miller, Rolf Ulrich.

Methodology: Jeff Miller, Rolf Ulrich.

Software: Jeff Miller, Rolf Ulrich.

Writing – original draft: Jeff Miller, Rolf Ulrich.

Writing – review & editing: Jeff Miller, Rolf Ulrich.

References

1. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biology. 2015; 13(6):e1002165. https://doi.org/10.1371/journal.pbio.1002165 PMID: 26057340
2. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005; 294(2):218–228. https://doi.org/10.1001/jama.294.2.218 PMID: 16014596
3. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014; 15(1):1–12. https://doi.org/10.1093/biostatistics/kxt007 PMID: 24068246
4. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011; 22(11):1359–1366. https://doi.org/10.1177/0956797611417632 PMID: 22006061
5. Vul E, Harris C, Winkielman P, Pashler HE. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on Psychological Science. 2009; 4(3):274–290. https://doi.org/10.1111/j.1745-6924.2009.01125.x PMID: 26158964
6. Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. Journal of the American Statistical Association. 1959; 54(285):30–34. https://doi.org/10.2307/2282137
7. Tannock IF. False-positive results in clinical trials: Multiple significance tests and the problem of unreported comparisons. Journal of the National Cancer Institute. 1996; 88(3-4):206–207. https://doi.org/10.1093/jnci/88.3-4.206 PMID: 8632495
8. Wacholder S, Chanock S, Garcia-Closas M, El ghormli L, Rothman N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. Journal of the National Cancer Institute. 2004; 96(6):434–442. https://doi.org/10.1093/jnci/djh075 PMID: 15026468
9. Vasey MW, Thayer JF. The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. Psychophysiology. 1987; 24:479–486. https://doi.org/10.1111/j.1469-8986.1987.tb00324.x PMID: 3615759
10. Stroebe W, Postmes T, Spears R. Scientific misconduct and the myth of self-correction in science. Perspectives on Psychological Science. 2012; 7(6):670–688. https://doi.org/10.1177/1745691612460687 PMID: 26168129
11. Agnoli F, Wicherts JM, Veldkamp CLS, Albiero P, Cubelli R. Questionable research practices among Italian research psychologists. PLoS ONE. 2017; 12(3):1–17. https://doi.org/10.1371/journal.pone.0172792
12. Héroux ME, Loo CK, Taylor JL, Gandevia SC. Questionable science and reproducibility in electrical brain stimulation research. PLoS ONE. 2017; 12(4):1–11.

13. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth-telling. Psychological Science. 2012; 23:524–532. https://doi.org/10.1177/0956797611430953 PMID: 22508865
14. Fiedler K, Schwarz N. Questionable research practices revisited. Social Psychological and Personality Science. 2015; 7(1):45–52. https://doi.org/10.1177/1948550615612150
15. Francis G. Publication bias and the failure of replication in experimental psychology. Psychonomic Bulletin & Review. 2012; 19(6):975–991. https://doi.org/10.3758/s13423-012-0322-y
16. Francis G. Too good to be true: Publication bias in two prominent studies from experimental psychology. Psychonomic Bulletin & Review. 2012; 19(2):151–156. https://doi.org/10.3758/s13423-012-0227-9
17. Francis G. The frequency of excess success for articles in *Psychological Science*. Psychonomic Bulletin & Review. 2014; 21(5):1180–1187. https://doi.org/10.3758/s13423-014-0601-x
18. Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-drawer. Journal of Experimental Psychology: General. 2014; 143(2):534–547. https://doi.org/10.1037/a0033242
19. Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016; 351(6280):1433–1436. https://doi.org/10.1126/science.aaf0918 PMID: 26940865
20. Gorroochurn P, Hodge SE, Heiman GA, Durner M, Greenberg DA. Non-replication of association studies: "Pseudo-failures" to replicate? Genetics in Medicine. 2007; 9:325–331. https://doi.org/10.1097GIM.0b013e3180676d79 PMID: 17575498
21. Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005; 2(8):e124 (696–701). https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
22. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349(6251):aac4716–1–aac4716–8.
23. Gilbert DT, King G, Pettigrew S, Wilson TD. Comment on "Estimating the reproducibility of psychological science". Science. 2016; 351(6277):1037–1037. https://doi.org/10.1126/science.aad7243 PMID: 26941311
24. Begley CG, Ioannidis JPA. Reproducibility in science: Improving the standard for basic and preclinical research. Circulation Research. 2015; 116(1):116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819 PMID: 25552691
25. Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017; 1(0021):1–9.
26. Nosek BA, Spies JR, Motyl M. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science. 2012; 7(6):615–631. https://doi.org/10.1177/1745691612459058 PMID: 26168121
27. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver & Boyd; 1925.
28. Cowles M, Davis C. On the origins of the .05 level of statistical significance. American Psychologist. 1982; 37(5):553–558. https://doi.org/10.1037/0003-066X.37.5.553
29. Aczel B, Palfi B, Szaszi B. Estimating the evidential value of significant results in psychological science. PLoS ONE. 2017; 12(8):1–8. https://doi.org/10.1371/journal.pone.0182651
30. Colhoun HM, McKeigue PM, Smith GD. Problems of reporting genetic associations with complex outcomes. Lancet. 2003; 361(9360):865–872. https://doi.org/10.1016/S0140-6736(03)12715-8 PMID: 12642066
31. Johnson VE. Revised standards for statistical evidence. Proceedings of the National Academy of Sciences. 2013; 110(48):19313–19317. https://doi.org/10.1073/pnas.1313476110
32. Schimmack U. The ironic effect of significant results on the credibility of multiple-study articles. Psychological Methods. 2012; 17(4):551–566. https://doi.org/10.1037/a0029487 PMID: 22924598
33. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2018; 2:6–10. https://doi.org/10.1038/s41562-017-0189-z
34. Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify Your Alpha: A Response to "Redefine Statistical Significance".; 2017.
35. Fiedler K, Kutzner F, Krueger JI. The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. Perspectives on Psychological Science. 2012; 7(6):661–669. https://doi.org/10.1177/1745691612462587 PMID: 26168128
36. Bakker M, Van Dijk A, Wicherts JM. The rules of the game called psychological science. Perspectives on Psychological Science. 2012; 7(6):543–554. https://doi.org/10.1177/1745691612459060 PMID: 26168111

37. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA. 2016; 315(11):1141–1148. https://doi.org/10.1001/jama.2016.1952 PMID: 26978209

38. Masicampo EJ, Lalande DR. A peculiar prevalence of *p* values just below .05. Quarterly Journal of Experimental Psychology. 2012; 65(11):2271–2279. https://doi.org/10.1080/17470218.2012.711335

39. Mudge JF, Baker LF, Edge CB, Houlahan JE. Setting an optimal *α* that minimizes errors in null hypothesis significance tests. PLoS ONE. 2012; 7(2):e32734. https://doi.org/10.1371/journal.pone.0032734 PMID: 22389720

40. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. JAMA. 2018; 319(14):1429–1430. https://doi.org/10.1001/jama.2018.1536 PMID: 29566133

41. Michaels R. Confidence in courts: A delicate balance. Science. 2017; 357(6353):764–764. https://doi.org/10.1126/science.aao3967 PMID: 28839066

42. Raiffa H, Schlaifer R. Applied statistical decision theory. Boston, MA, US: Division of Research, Graduate School of Business Administration, Harvard University; 1961.

43. Miller JO, Ulrich R. Optimizing research payoff. Perspectives on Psychological Science. 2016; 11(5):664–691. https://doi.org/10.1177/1745691616649170 PMID: 27694463

44. Johnson VE, Payne RD, Wang T, Asher A, Mandal S. On the reproducibility of psychological science. Journal of the American Statistical Association. 2017; 112(517):1–10. https://doi.org/10.1080/01621459.2016.1240079 PMID: 29861517

45. Mudge JF, Martyniuk CJ, Houlahan JE. Optimal alpha reduces error rates in gene expression studies: a meta-analysis approach. BMC Bioinformatics. 2017; 18(1):312. https://doi.org/10.1186/s12859-017-1728-3 PMID: 28637422

46. DeGroot MH. Probability and statistics. Menlo Park, NJ, US: Addison-Wesley; 1975.

47. Cohen J. Statistical power analysis for the behavioral sciences. New York, NY: Academic Press; 1977.

48. Canessa S, Guillera-Arroita G, Lahoz-Monfort JJ, Southwell DM, Armstrong DP, Chadès I, et al. When do we need more data? A primer on calculating the value of information for applied ecologists. Methods in Ecology and Evolution. 2015; 6(10):1219–1228. https://doi.org/10.1111/2041-210X.12423

49. Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. Health Economics. 1996; 5(6):513–524. https://doi.org/10.1002/(SICI)1099-1050(199611)5:6<513::AID-HEC237>3.0.CO;2-9 PMID: 9003938

50. Ferrier PM, Buzby JC. The economic efficiency of sampling size: The case of beef trim. Risk Analysis. 2013; 33(3):368–384. https://doi.org/10.1111/j.1539-6924.2012.01874.x PMID: 22830311

51. Willan AR. Optimal sample size determinations from an industry perspective based on the expected value of information. Clinical Trials. 2008; 5(6):587–594. https://doi.org/10.1177/1740774508098413 PMID: 19029207

52. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009; 374(9683):86–89. https://doi.org/10.1016/S0140-6736(09)60329-9 PMID: 19525005

53. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014; 383(9912):166–175. https://doi.org/10.1016/S0140-6736(13)62227-8 PMID: 24411645