Justify Your Alpha: A Practical Guide

Daniel Lakens[1]

[1] Eindhoven University of Technology, The Netherlands

The default use of an alpha level of 0.05 is sub-optimal for several reasons. Decisions based on data can be made more efficiently by choosing an alpha level that minimizes the combined Type 1 and Type 2 error rate or that optimizes the informational value of a study. In addition, in studies with very high statistical power *p*-values lower than the alpha level can actually be support for the null hypothesis, instead of for the alternative hypothesis. Because it is difficult to abandon a bad practice without providing an alternative, this manuscript explains three approaches that can be used to justify your alpha. The first approach is based on the idea to either minimize or balance Type 1 and Type 2 error rates. The second approach is based on optimizing the informational value of studies or aiming for high posterior probability of hypotheses after conducting a statistical test. The third approach lowers the alpha level as a function of the sample size. Software is provided to perform the required calculations. All approaches have their limitations (e.g., the challenge of specifying relative costs and priors, or the slightly arbitrary nature of how the alpha level should decrease as the sample size increases) but they nevertheless provide a clear improvement compared to current practices. The use of alpha levels that have a better justification should improve statistical inferences and increase the efficiency of scientific research.

Keywords: hypothesis testing, Type 1 error, Type 2 error, statistical power
Word count: 4890 words

Researchers often rely on data to decide how to act. In a Neyman-Pearson approach to hypothesis testing (Neyman & Pearson, 1933) studies are designed such that erroneous decisions that determine how we act are controlled in the long run at some desired maximum level. If resources were infinite we could collect so much data that the chance of making a wrong decision is incredibly small. But resources are often limited, which means that researchers need to decide how to choose the rate at which they are willing to make errors. After data is collected researchers can incorrectly act as if there is an effect when there is no true effect (a Type 1 error) or incorrectly act as if there is no effect when there is a true effect (a Type 2 error). For any fixed sample size and true effect size a reduction in the Type 1 error rate will increase the Type 2 error rate and vice versa.

The question how error rates should be set in any study requires careful consideration of the relative costs of a Type 1 error or a Type 2 error. Regrettably researchers rarely provide such a justification, and predominantly use a Type 1 error rate of 5%. One reason researchers rely on norms instead of rationally chosen error rates is the absence of explanations how to justify a different alpha level than the normative use of 0.05. This article explains why error rates need to be justified, and provides three practical guidelines that can be used to justify the alpha level by (1) balacing or minizing error rates, (2) optimizing informational value or aiming for specific posterior probabilities, or (3) lowering the alpha level as a function of the sample size.

Why do we use a 5% alpha level and 80% power?

As a young scholar we might naively assume that when all researchers do something, there must be a good reason for such an established practice. An important step towards maturity as a scholar is the realization that this is not the case. Neither Fisher nor Neyman, two statistical giants largely responsible for the widespread reliance on hypothesis tests in the social sciences, recommended the universal use of any specific threshold. Fisher (1935) writes: “It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result.” Similarly, Neyman and Pearson (1933) write:

All code used to create this manuscript is provided in an OSF repository at https://osf.io/h5e2q/.

Correspondence concerning this article should be addressed to Daniel Lakens, ATLAS 9.402, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

"From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator."

Even though in theory alpha levels should be justified, in practice researchers tend to imitate others. Fisher writes in 1926: "Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level". This sentence is preceded by the statement "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point)." Indeed, in his examples Fisher often uses an alpha of 0.01. Nevertheless, researchers seem to have copied the value Fisher preferred, instead of his more important take-home message that the significance level should be set by the experimenter. The default use of an alpha level of 0.05 seems to originate from the early work of Gosset on the *t*-distribution (Cowles & Davis, 1982), who believed that a difference of two standard deviations (a z-score of 2) was sufficiently rare.

The default use of 80% power (or a 20% Type 2, or beta (b) error) is similarly based on personal preferences by Cohen (1988), who writes: "It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value, the value .80 be used. This means that β is set at .20. This arbitrary but reasonable value is offered for several reasons (Cohen, 1965, pp. 98-99). The chief among them takes into consideration the implicit convention for α of .05. The β of .20 is chosen with the idea that the general relative seriousness of these two kinds of errors is of the order of .20/.05, i.e., that Type I errors are of the order of four times as serious as Type II errors. This .80 desired power convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns in his specific research investigation to choose a value ad hoc."

We see that conventions are built on conventions: the norm to aim for 80% power is built on the norm to set the alpha level at 5%. However, the real lesson Cohen was teaching us is to determine the relative seriousness of Type 1 and Type 2 errors, and to balance both types of errors when a study is designed. If a Type 1 error is considered to be four times as serious as a Type 2 error, the *weighted* error rates in the study are balanced. Instead of imitating the values chosen by Cohen, researchers should aim to imitate the approach he used to justify error rates.

Why error rates should be justified

In 1957 Neyman wrote: "it appears desirable to determine the level of significance in accordance with quite a few circumstances that vary from one particular problem to the next." The mindless application of null hypothesis significance tests, including setting the alpha level at 5% for all tests, has been criticized for more than half a century (Bakan, 1966; Gigerenzer, 2018). But it is difficult to abandon a mediocre research practice without an alternative.

There are three main reasons to abandon the universal use of a 5% alpha level. The first reason to carefully choose an alpha level is that decision making becomes more efficient. If researchers use hypothesis tests to choose how to act while balancing error rates it is typically possible to make decisions more efficiently by setting the alpha level at a different value than 0.05. If we aim to either minimize or balance Type 1 and Type 2 error rates for a given sample size and effect size, the alpha level should be set not based on convention, but by weighting the relative costs of both types of errors.

Second, an alpha of 5% often results in studies of low informational value. For example, when finding a non-significant result in a low powered study, the researcher cannot learn a lot from his experiment. Adjusting the alpha level in a way that both significant and non-significant findings are informative is important to mitigate this problem.

The third reasons is that as the statistical power increases, some *p*-values below 0.05 (e.g., $p = 0.04$) can be more likely when there is *no* effect than when there *is* an effect. This is known as Lindley's paradox (Cousins, 2017; Lindley, 1957). The distribution of *p*-values is a function of the statistical power (Cumming, 2008), and the higher the power, the more right-skewed the distribution becomes (i.e. the more low *p*-values are observed). When there is no true effect *p*-values are uniformly distributed, and 1% of observed *p*-values fall between 0.04 and 0.05. When the statistical power is extremely high, not only will most *p*-values fall below 0.05, most will also fall below 0.01. In Figure 1 we see that with high power very low *p*-values are more likely to be observed when there *is* an effect than when there is *no* effect (e.g., the black curve representing *p*-values when the alternative is true falls above the dashed horizontal line for a *p*-value of 0.01). But observing a *p*-value of 0.04 is more likely when the null hypothesis is true than when the alternative hypothesis is true and we have very high power (the horizontal dashed line falls above the black curve for *p*-values larger than 0.025).

Researchers often want to interpret a significant test result as 'support' for the alternative hypothesis. If so, it makes sense to choose the alpha level such that when a significant *p*-value is observed, the *p*-value is actually more likely when the alternative hypothesis is true than when the null hypothesis is true. This means that when statistical power is very high (e.g., the sample size is very large), the alpha level should be reduced. For example, if the alpha level in Figure 1 is lowered to 0.02 then the alternative hypothesis is more likely

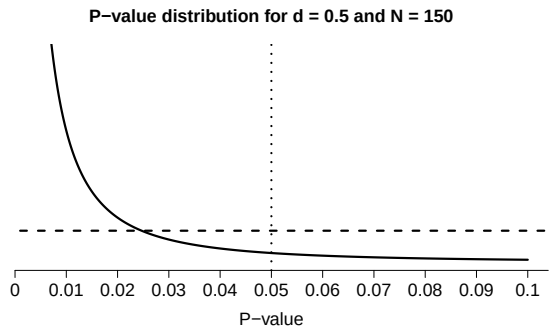

Figure 1. *P*-value distributions for a two-sided independent *t*-test with N = 150 and d = 0.5 (black curve) or d = 0 (horizontal dashed line) which illustrates how *p*-values just below 0.05 can be more likely when there is no effect than when there is an effect.

than the null hypothesis for all significant *p*-values that would be observed.

Minimizing or Balancing Type 1 and Type 2 Error Rates

If both Type 1 as Type 2 errors are costly, then it makes sense to optimally reduce both errors as you design studies. This would make decision making overall most efficient. Researchers can choose to design a study with a statistical power and alpha level that minimize the *combined error rate*. For example, assuming H0 and H1 are a-priori equally probable, with a 5% alpha level and a statistical power of 80% the combined error rate is (5 + 20)/2 = 12.5%. If the statistical power is 99% and the alpha is 5% the combined error rate is (1 + 5)/2 = 3%. As shown below, this approach can be extended to incorporate different prior probabilities of H0 and H1. Mudge, Baker, Edge, and Houlahan (2012) show that by choosing an alpha level based on the relative weight of Type 1 errors and Type 2 errors, and assuming beliefs about the prior probability that H0 and H1 are correct, decisions can be made more efficiently compared to the default use of an alpha level of 0.05.

Winer (1962) writes: "The frequent use of the .05 and .01 levels of significance is a matter of convention having little scientific or logical basis. When the power of tests is likely to be low under these levels of significance, and when Type 1 and Type 2 errors are of approximately equal importance, the .30 and .20 levels of significance may be more appropriate than the .05 and .01 levels." The reasoning here is that a design that has 70% power for the smallest effect size of interest would not balance the Type 1 and Type 2 error rates in a sensible manner. Similarly, in huge datasets it might be possible to achieve very high levels of power for all effect sizes that are still considered meaningful. If such a study has 99% power for effect sizes of interest, and thus a 1% Type 2 error rate, but uses a 5% alpha level, it also suffers from a lack of balance. A common example of this latter scenario is the default use of a 5% alpha level in meta-analyses, which often have extremely high power for any effect size that would be considered meaningful, and where it seems sensible to lower the alpha level considerably.

Researchers can decide to either balance Type 1 and Type 2 error rates (e.g., setting both at 5%), or minimize error rates. For any given sample size and effect size of interest there is an alpha level that minimizes the combined error rates. Because the chosen alpha level also influences the statistical power, and the Type 2 error rate is therefore dependent upon the Type 1 error rate, minimizing or balancing error rates requires an iterative procedure. Imagine researchers plan to perform a study which will be analyzed with an independent two-sided *t*-test. They initially plan to collect 50 participants per condition, and set their smallest effect size of interest to d = 0.5. They think a Type 1 error is just as severe as a Type 2 error, and believe H0 is just as likely to be true as H1. The combined error rate is minimized when they set α to 0.13 (see Figure 2, dotted line), which will give the study a Type 2 error rate of β = 0.166 to detect effects of d = 0.5. The combined error rate is 0.148, while it would have been 0.177 if the alpha level was set at 5%[1]. Instead of choosing an example that is more extreme, I think it is important to point out that a better justified alpha level not necessarily means a hugely different alpha level. The absolute benefits are more noticeable for study designs where Type 2 error rates are high when a 5% alpha level would be used, or where the Type 2 error is very small and the alpha level can be reduced from 0.05. Similarly, Figure 2 shows that if one were to use an alpha level of 0.005 following recommendations of Benjamin et al. (2018) when collecting a sample size of 100 per group to detect an effect of d = 0.5, the study would be much less efficient than it could be if a higher alpha level is chosen. The more important take-home message is that, perhaps counter-intuitively, decision making is sometimes slightly more efficient after *increasing* the alpha level from the default of 0.05.

Weighing the Relative Cost of Errors. Cohen (1988) considered a study design with a 5% Type 1 error rate and a 20% Type 2 error rate balanced. The reason for this was that instead of weighing both types of errors equally, he felt "Type I errors are of the order of four times as serious as Type II errors". To determine the relative costs of Type 1 and Type 2 errors researchers should perform a cost-benefit analysis. As an example from another discipline, (**???**) quantify the relative costs of Type 1 errors when testing whether native species in Australia are declining. They come to the conclusion that when it comes to the Koala population, given

[1] For the same scenario, balanced error rates are α = 0.149 and β = 0.149.

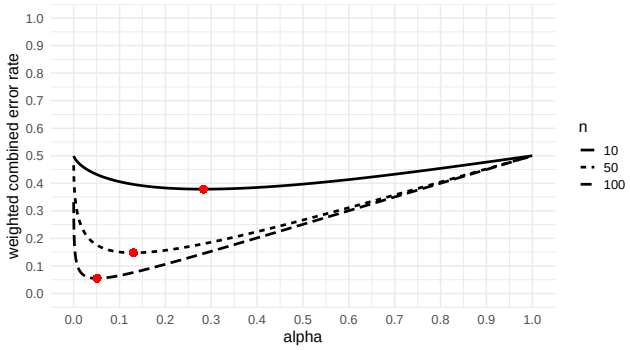

Figure 2. Weighted combined error rate (y-axis) for an independent *t*-test with n = 10, n = 50, and n = 100 per group and a smallest effect of interest of d = 0.5, for all possible alpha levels (x-axis).

its great economic value, a cost-benefit analysis indicates the alpha level should be set to 1. In other words, one should always act as if the population is declining, because the relative cost of a Type 2 error compared to a Type 1 error is extremely large.

Although it can be difficult to formally quantify all factors that determine the costs of Type 1 and Type 2 errors, there is no reason to let the perfect be the enemy of the good. In research questions where there are no direct applications of the research findings, relative costs might be largely subjective. If you have a research strategy where you always follow up on an initial study testing a theoretical prediction with a replication and extension study, and want to innovate, you might want to weigh a Type 2 error more strongly than a Type 1 error. If you are relatively sure policies will be changed based on the outcome of your single study, or it is unlikely many researchers will have the resources to replicate your study, a Type 1 error rate might be weighed more strongly than a Type 2 error. There are no right or wrong answers, but you need to think through this question when you design a study.

If we adapt our calculations for the example above where researchers who plan to collect 50 participants per condition to detect a d = 0.5 effect, but now weigh the cost of Type 1 errors 4 times as much as Type 2 errors, balanced error rates are $\alpha = 0.0655$ and $\beta = 0.262$. You will notice this is exactly the scenario Cohen describes, where a 5% Type 1 error rate and 20% Type 2 error rate is deemed a balanced design because Type 1 errors are weighed 4 times as much as Type 2 errors.

One could also aim to minimize error rates in the latter scenario by setting α to 0.0388 and β to 0.342. Remember the cost of a Type 1 error is twice as large as the cost of a Type 1 error. Imagine a Type 1 error has a cost of 400 and a Type 2 error has a cost of 100. We perform 100 studies, 50 where H0 is true and 50 where H1 is true. We will make 1.94 Type 1 errors and 17.12 Type 2 errors, with a respective cost of 775.1 and 1712, respectively. Increasing or decreasing the alpha level for this design will only increase the total cost of errors, assuming H0 and H1 are equally probable. Figure 3 visualizes the weighted combined error rate for this study design across the all possible alpha levels.

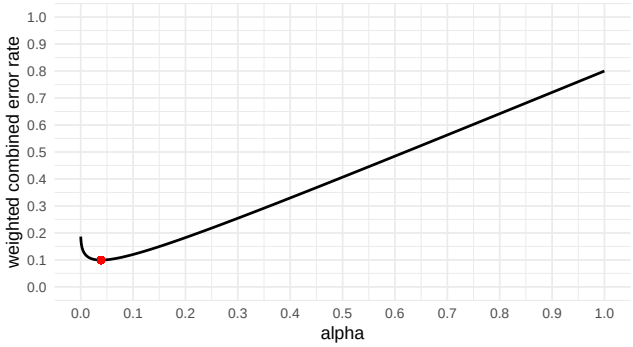

Figure 3. Weighted combined error rate (y-axis) for an independent *t*-test with n = 50 per group and a smallest effect of interest of d = 0.5, where Type 1 errors are weighed 4 times as much as Type 2 errors, for all possible alpha levels (x-axis).

Incorporating prior probabilities. Miller and Ulrich (2019) explain how the choice for an optimal alpha level depends not just on the relative costs of Type 1 and Type 2 errors, but also on the base rate of true effects. In the extreme case, all studies a researcher designs test true hypotheses. In this case, there is no reason to worry about Type 1 errors, because a Type 1 error can only happen when the null hypothesis is true. Therefore, you can set the alpha level to 1 without any negative consequences. If the base rate of true hypotheses is very low, in the long run you will make many more Type 1 errors than Type 2 errors. For example, if you perform 1000 studies, and the base rate of true effects is 10%, with 5% Type 1 error rate and a 5% Type 2 error rate you will observe $1000 \times 0.9 \times 0.05 = 45$ Type 1 errors, and $1000 \times 0.1 \times 0.05 = 5$ Type 2 errors. If possible, researchers can take their expectations about the long run frequency of Type 1 and Type 2 errors into account when choosing an alpha level that either balances or minimizes the error rates.

If you want to balance or minimize error rates in the long run, you should lower the alpha level as the prior probability of a null effect increases, or increase the alpha level as the prior probability of a true effect increases. Because the base rate of true hypotheses is unknown, this step requires subjective judgment. This can not be avoided, because one always makes assumptions about base rates, even if the assumption is that a hypothesis is equally likely to be true as false (with both H1 and H0 having a 50% probability). Assuming equal prior probabilities for H1 and H0, we saw above that balanced error rates assuming d = 0.5 and collection n = 50 per

condition in a *t*-test would imply $\alpha = 0.149$ and $\beta = 0.149$. If you assume the null hypothesis is 10 times more likely than the alternative hypothesis (or the alternative hypothesis is 0.1 times as likely as the null hypothesis) then balanced error rates would require setting α to 0.0356 and β to 0.356. If you believe the alternative hypothesis is twice as likely to be true as the null hypothesis, balancing error rates in the long run would mean increasing the alpha level and increasing the power by choosing $\alpha = 0.213$ and $\beta = 0.107$.

The two approaches (balancing error rates or minimizing error rates) typically yield quite similar resuls. Where minimizing error rates might be slightly more efficient, balancing error rates might be slightly more intuitive (especially when the prior probability of H0 and H1 are equal). Note that although there is always an optimal choice of the alpha level, there is always a range of values that yield quite similar weighted error rates. It is useful to plot weighted error rates as in Figure 3. To balance or minimize error rates, researchers need to carefully consider the relative cost of Type 1 and Type 2 errors, the prior probability the null hypothesis is true, and the effect size they want to detect (Mudge et al., 2012), because these factors are used to calculate the weighted combined error rates *w*:

$$\frac{(cost_{T1T2} \times \alpha + prior_{H1H0} \times \beta)}{prior_{H1H0} + cost_{T1T2}} \tag{1}$$

For example, imagine Type 1 errors are weighted 4 times as much as Type 2 errors ($cost_{T1T2} = 4$) and the alternative hypothesis is believed to be 2 times as likely as the null hypothesis ($prior_{H1H0} = 10$). With $\alpha = 0.213$ and $\beta = 0.107$ the weigthed combined error rate is 0.142.

Probability-Based Alpha Level Justifications

A second approach to justifying alpha levels is based on the posterior probability after conducting a statistical test, or the change in belief between prior and posterior. This combines Frequentist and Bayesian approaches for statistical inference (for a discussion of Bayesian - Frequentist compromises in statistical inference see e.g., Good, 1992; Colquhoun, 2017). Given α, β, and the prior probability of a hypothesis, we can calculate the posterior probability after a significant *p*-value using Bayes rule as in equation 2.

$$p(H_1/+) = \frac{(1-\beta) \times p(H_1)}{(1-\beta) \times p(H_1) + \alpha \times (1-p(H_1))} \tag{2}$$

In the same way we can calcualte the probability of the hypothesis being true after a non-significant *p*-value (equation 3).

$$p(H_1/-) = \frac{\beta \times p(H_1)}{\beta \times p(H_1) + (1-\alpha) \times (1-p(H_1))} \tag{3}$$

These two equations give rise to different ways of justifying alpha levels. In the next sections, we show how to justify alpha based on (1) optimizing the overall informational value of a study, (2) the posterior probability after either a significant or non-significant result and (3) a power analysis indicating the sample size and alpha needed to arrive at posterior probabilities specified for both a significant and non-significant result.

Optimizing Informational Value. The goal of this approach is setting alpha in a way that optimizes the amount of learning by conducting a statistical test. Learning is defined as the expected correct change between the probability of a hypothesis before and after conducting a statistical test (δ L). From equations 2 and 3 it is possible to calculate the expected correct change in belief or learning with a given α. First, we calculate the difference between prior and posterior for a significant and non-significant result as in equations 2 and 3. Then we determine the probability of the four possible outcomes, true positive ($p(H-1) * (1-\beta)$, true negative ($(1-p(H_1) * (1-\alpha))$, false positive ($(1-p(H_1)) * \alpha$), and false negative ($p(H_1) * \beta$). The change in belief after a false negative or a false positive is a form of "unlearning" since it makes the belief in a hypothesis less correct. Therefore, the expected correct change in belief (δ L) can be calculated by subtracting the change after a false negative and false positive from the change after a true negative or true positive (equation 4).

$$\begin{aligned} \delta L = p(H_1) \times (1-\beta) \times (p(H_1/+) - p(H_1)) + \\ (1-p(H_1)) \times (1-\alpha) \times (p(H_1) - p(H_1/-)) - \\ (1-p(H_1)) \times \alpha \times (p(H_1/+) - p(H_1)) - \\ p(H_1) \times \beta \times (p(H_1) - p(H_1/-)) \end{aligned} \tag{4}$$

Let us clarify this idea using an example. A group of researchers studies a new animal Stroop paradigm to test cognitive control in children. They display sad and happy animal faces under the words "sad" and "happy". Participants need to classify the emotion of the animal as fast as possible. However, they are uncertain whether the children can identify the animal emotions correctly and their pilot study had mixed results. Therefore, the prior probability of their hypothesis is 0.2 and the expected effect size is $d = 0.2$. They test 200 participants and analyze their data using a within-sample *t*-test. In this context, the significance level with the highest δL is $\alpha = 0.04$ with a power of 0.77 and an expected learning of 0.18 Thus, the research group uses this alpha level for their first study. To their positive surprise the result is significant. They calculate the posterior probability of 0.83

according to equation 2. See Figure 4 for expected learning as a function of α level.

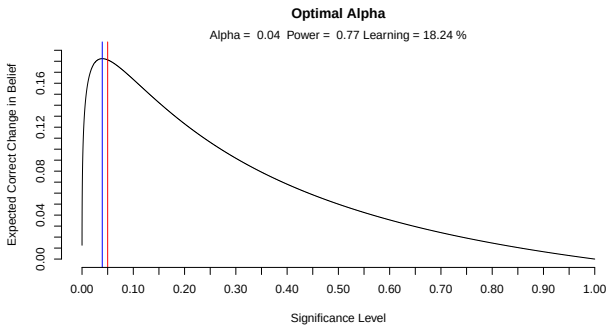

Figure 4. Expected learning for different alpha levels. The blue line indicates the optimal alpha, the red line alpha 0.05

To corroborate the finding more strongly, the researchers decide to conduct a replication with the same sample size. This time their ideal alpha level is 0.21 with a power of 0.94 and an expected learning of (Figure 5).

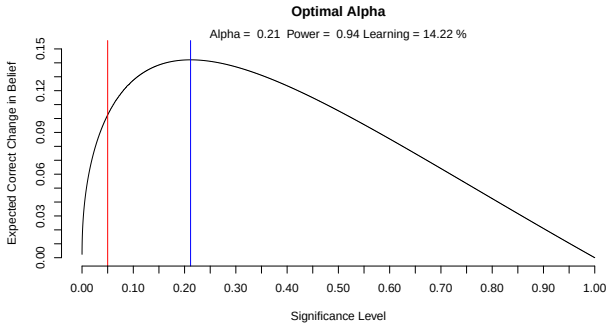

Figure 5. Expected learning for different alpha levels in the replication study. The blue line indicates the optimal alpha of stroop1$alpha, the red line alpha 0.05. Note that the higher prior probability results in higher optimal alpha levels

The example shows, how optimizing informational value can be the method of choice, especially in basic research where the costs of different errors are often difficult to asses. In addition, when the prior probability of a hypothesis is high, it is often beneficial to increase the signficance level. By increasing the significance level the researchers can expect to learn 1.39 as much by conducting the replication compared to the conventional significance level. Note the similarity to minimizing error rates, where a higher prior probability also justifies using a higher alpha level.

Aiming for Posterior Probabilities. Researchers usually want to avoid false positives. Specifying the α level in a way that the posterior probability after conducting a statistical test is above a certain threshold (e.g., 95%) based on equation 2 is a way to formally control the number of positive findings. Note that the idea of interpreting *p*-values and α levels based on the posterior probability of a significant result has been proposed in the past (e.g., Colquhoun, 2019, 2017; Oberauer & Lewandowsky, 2019). However, since it is an important aspect of the proposed framework of probability-based approaches to α level justification we present it again here.

As an example, Bem (2011) tries to show that people are capable of precognition, in other words, anticipating future events that they could not have predicted through any known inferential process. Since precognition seems to be a phenomenon with low prior probability and the existence of the effect would have several unlikely implications (e.g., it is unlikely that casinos would be as lucrative if players could feel the future), we assume the prior probability of precognition to be only 1%. In the first study reported in Bem (2011) on precognitive detection of erotic stimuli, data from 100 students is analyzed using a within-sample *t*-test. Results indicate that participants correctly identified the future positions of erotic pictures more frequently than expected by chance ($t(99) = 2.51$, $p = .01$, $d = 0.25$). Let us consider which α level would be needed to make the existence of precognition 50% likely after finding a positive result. Based on the formula in equation 1, the required alpha level would be at least $\alpha = 0.003$ with a power of 0.29. In other words, Bem (2011) would have needed to test his predictions against an extremely low alpha level, to achieve sufficient probability after a significant result and his experiment does not provide strong evidence for the existence of precognition.

While researchers tend to focus strongly on controlling false positives, it has been argued that false negatives, pose an evenly large problem to science (Fiedler, Kutzner, & Krueger, 2012). Especially in contexts where the costs of not detecting an important effect are large, researchers want to be certain that an effect of the smallest effect size of interest is absent if we do not find a significant result. In this case, it is useful to set the significance level in a way that the probability of the hypothesis as in equation 3 is below a certain value after a non-significant test.

Let us illustrate this with an adapted version of the example from Field, Tyre, Jonzén, Rhodes, and Possingham (2004).[2] A group of ecologists observes two small panda populations in eastern Australia. Near the habitat of one of the populations a new chemical company is built. Although the company ensures that the pandas are in no danger, the researchers are concerned that the toxic waste poses a threat to the panda population. Their prior belief that the waste is harmful is 70%. Therefore, they use a contingency test to investigate whether the survival of the pandas near the company is reduced compared to the other panda group. They estimate that

[2]Field et. al(2004) conduct a simulation-based power analysis based on the idea of a virtual ecologist, however, the code for this anylsis is not available

toxic waste might reduce the panda population by about 20%. Since pandas are a species of extreme ecological and economical value (e.g., Hamilton, Lunney, & Matthews, 2000), the researchers are more concerned about false negatives than false positives. Therefore, they want to set their significance level in a way, that the probability of the waste threatening the health of the pandas after a negative result is only 10%. Based on equation 2, they would need to use an α of at least 0.45 with a power of 0.97 to achieve this. The example shows that when we want to be reasonably certain that the null hypothesis is true after testing we often need to use high power to reduce type 2 errors. This often implies high α levels, especially given small sample size. If the effect is not significant using this α level, this indicates a certain probability of the null hypothesis to be true, an idea similar to equivalence testing (e.g., Lakens, Scheel, & Isager, 2018).

Probability-Based Power Analysis. However, usually researchers want to control the probability of both type one and type two errors. In this context, they can specify the prior, the intended posterior probability after a significant result, and the intended probability after a non-significant result. We developed a probability-based power analysis (PAPA) that then estimates the minum required sample size as well as the needed significance threshold to achive the desired posterior probabilities.

To explain PAPA let us imagine a group of researchers that want to conduct a new study on the framing of persuasion messages to increase compliance with social distancing to control the spread of COVID-19. Since small effect sizes can already safe lives they put their smallest effect size of interest to d = 0.15. Their prior belief in the effect of the tested framing is 50%. Because of the high stakes situation, they want to be at least 95% sure that the effect is present after finding a positive result and 80% sure that it is absent after a negative result. PAPA indicates that they would need a sample size of 678 per group with an α of 0.04and power of 0.76. The example shows that for researchers that want to control both errors and have the capacities to increase their sample accordingly, PAPA should be the method of choice to enable strong inference.

Discussion Probability-Based Approaches. We have now presented three ways to justify alpha levels based on the posterior probability or the change in probability. The first method, optimizing the informational value of a study, is especially useful in theoretical research, when the researcher has no strong opinion regarding which type of error is more costly. If the researcher wants to specifically avoid false positives or false negatives the approach of aiming for specific posterior probabilites is the most suitable. However, usually researchers care about both, false positives and false negatives, here a probability-based power analysis will help to justify the sample size and α level needed to control both types of errors. PAPA is based on calculating the alpha level needed to achieve sufficient posterior probability after a significant result and then repeating this process over different sample sizes until the probability after a significant and a non-significant result are in line with the specifications of the researcher.

An interesting question is also, how this approach relates to Bayesian statistical inference. While equation 2 and 3 are based on Bayes theorem, our appraoch is still a form of significance testing. There are certain advantages to Bayesian procedures such as the ability to stop conditional on the evidence and to quantify evidence for the null hypothesis and the alternative continously (Wagenmakers, Morey, & Lee, 2016). However, our approach might be more easy to learn for researchers already familiar with classical statistics and does not require to specify a parameter prior distribution under the alternative hypothesis. In practice researchers might want to report 'a B for every *p*', or a Bayes factor for every *p*-value (Dienes & Mclatchie, 2017), and take any disagreement between these two approaches (which should be quite rare in practice according to Jeffreys, 1939) as a reason to be cautious when drawing conclusions from the data.

Lowering the Alpha Level as a Function of the Sample Size

Sometimes a researcher might feel it is too challenging to specify relative costs, prior probabilities, and the effect size of interest, and be tempted to fall back to the use of an alpha level of 0.05. For such cases a less philosophically coherent approach exists where the alpha level is lowered as a function off the sample size. Although still built on some simple conventions, it is offered here as an alternative approach to justify the alpha level that, although not perfect, is still an improvement over current practices.

This approach is discussed most extensively by Leamer (1978). He writes "The rule of thumb quite popular now, that is, setting the significance level arbitrarily to .05, is shown to be deficient in the sense that from every reasonable viewpoint the significance level should be a decreasing function of sample size." This was already recognized by Jeffreys (1939), who discusses ways to set the alpha level in the Neyman-Pearson approach to statistics: "We should therefore get the best result, with any distribution of alpha, by some form that makes the ratio of the critical value to the standard error increase with n. It appears then that whatever the distribution may be, the use of a fixed *P* limit cannot be the one that will make the smallest number of mistakes."

To understand this recommendation it is important to distinguish between statistical inferences based on error control and inferences based on likelihoods. An alpha level of 5% will limit incorrect decisions to a desired maximum (in the

long run, and when all test assumptions are met). However, from a likelihood perspective it is possible that the observed data is much more likely when the null hypothesis is true, than when the alternative hypothesis is true, even when the observed p-value is smaller than 0.05. This situation, known as Lindley's paradox, is visualized in Figure 1. It emerges because the critical value of a frequentist test approaches a limit as the sample size increases (e.g., t = 1.96 for a two-sided t-test). In Bayesian statistics the same Bayes factor requires a larger test statistic when the sample size is larger (see Rouder, Speckman, Sun, Morey, and Iverson (2009) and Zellner (1971)).

To prevent Lindley's paradox when using frequentist statistics one would need to lower the alpha level as a function of the statistical power. Good (1992) notes: 'we have empirical evidence that sensible P values are related to weights of evidence and, therefore, that P values are not entirely without merit. The real objection to P values is not that they usually are utter nonsense, but rather that they can be highly misleading, especially if the value of N is not also taken into account and is large.' Based on the observation by Jeffrey's (1939) that, under specific circumstances, the Bayes factor against the null hypothesis is approximately inversely proportional to $(\sqrt{N})$, Good (1982) suggests a standardized p-value to bring p-values in closer relationship with weights of evidence:

$$p_{stan} = p\sqrt{\frac{N}{100}} \tag{5}$$

where p is the observed p-value and N is the total sample size. This formula standardizes the p-value to the evidence against the null hypothesis that would be observed if p_{stan} was the tail area probability observed in a sample of 100 participants. It is perhaps easier to think about a standardized alpha level:

$$\alpha_{stan} = \frac{\alpha}{\sqrt{\frac{N}{100}}} \tag{6}$$

With 100 participants α and α_{stan} are the same, but as the sample size increases beyond 100 observations the alpha level decreases. A higher critical test-statistic needs to be observed to still consider a finding 'statistically significant' as the sample size increases, which prevents Lindley's paradox. For example, when researchers would otherwise by default have used α = 0.05, they lower their alpha level for a sample size of 500 to an α_{stan} of 0.022, and use this alpha level to test against when the 500 observations are collected. Although in principle the function to calculate the standardized p-value can also be used when sample sizes are *smaller* than 100, which would return a standardized alpha level *larger* than 0.05, this is not the intended use.

As mentioned by Rouder et al. (2009): "NP testing can be made consistent by allowing Type I error rates to decrease toward zero as the sample size increases. How this rate should decrease with sample size, however, is neither obvious nor stipulated by the statistical theory underlying NP testing." Indeed, whereas lowering the alpha level as the sample size increases is a valid approach from a likelihood perspective, the specific way to do this as proposed by Good (1982) is itself largely based on the heuristic that starting to lower the alpha level when the number of total observations increases above 100 is largely arbitrary. There is no special reason to standardize the $p$value or alpha level to 100 observations, or for that matter to choose a 5% alpha level as a starting point. The reason to nevertheless recommend this approach to justify alpha levels is that it is straightforward to implement and quite likely to succesfully prevent Lindley's paradox for most studies in psychology. At the same time, in future research statisticians could develop more principled approaches to lowering the alpha level as a function of the sample size.

Discussion

Editors and reviewers should *always* ask authors to justify their choice of error rates, whenever researchers use data to make decisions about the presence or absence of an effect. As Skipper, Guenther, and Nass (1967) remarks: "If, in contrast with present policy, it were conventional that editorial readers for professional journals routinely asked:"What justification is there for this level of significance?" authors might be less likely to indiscriminately select an alpha level from the field of popular eligibles." In a Neyman-Pearson approach to statistics, the alpha level should be set before the data is collected. When reviewers of for example a Registered Report would ask authors to justify their alpha level, it would be convenient if they can recommend some approaches to do so. The current manuscript hopefully helps to fill this gap.

If a power analysis can be performed (i.e., whenever a desired or expected effect size of interest can be specified) and relative costs and priors can be specified, researchers can design efficient by minimizing or balancing error rates, optimizing informational value, or aiming for specific posterior probabilities. If it is difficult to specify these aspects of the study, researchers can fall back to the approach where the alpha level is reduced as a function of the sample size.

A Shiny app is avaiable that allows users to perform the calculations recommended in this article. It can be used to minimize or balance α and β, which works by specifying the effect size of interest and the sample size, as well as an analytic power calculation. The effect size should be determined as in a normal a-priori power analysis (preferably based on the smallest effect size of interest, for recommendations, see Albers and Lakens (2018)), and the sample size can be increased until the error rates are deemed acceptable. Alter-

natively, researchers lower the alpha level as a function of the sample size by spacifying their sample size. Whichever approach is used, it is strongly recommended to preregister the alpha level that researchers plan to use before the data is collected.

Because of the strong overreliance on a 5% error rate when designing studies, we have seen relatively few people attempt to justify their alpha level. As researchers start to justify their alpha, we will hopefully see the development of good examples and best practices for psychological science. It might be a challenge to get started, but the two approaches presented in the current article are one way to move away from the mindless use of a 5% alpha level. There is a lot to gain, and justifying alpha levels should improve our statistical inferences and increase the efficiency of the research we perform.

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/S41562-017-0189-Z

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

Colquhoun. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, *4*(12), 171085.

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, *73*(sup1), 192–201.

Cousins, R. D. (2017). The JeffreysLindley paradox and discovery criteria in high energy physics. *Synthese*, *194*(2), 395–432. https://doi.org/10.1007/s11229-014-0525-z

Cowles, M., & Davis, C. (1982). On the origins of the. 05 level of statistical significance. *American Psychologist*, *37*(5), 553.

Cumming, G. (2008). Replication and *p* Intervals: *P* Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Dienes, Z., & Mclatchie, N. (2017). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, 1–12. https://doi.org/10.3758/s13423-017-1266-z

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.

Field, S., Tyre, A. J., Jonzén, N., Rhodes, J. R., & Possingham, H. P. (2004). Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters*, *7*(8), 669–675.

Fisher, R. A. (1935). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, 2515245918771329. https://doi.org/10.1177/2515245918771329

Good, I. J. (1982). C140. Standardized tail-area probabilities. *Journal of Statistical Computation and Simulation*, *16*(1), 65–66. https://doi.org/10.1080/00949658208810607

Good, I. J. (1992). The Bayes/Non-Bayes Compromise: A Brief Review. *Journal of the American Statistical Association*, *87*(419), 597. https://doi.org/10.2307/2290192

Hamilton, C., Lunney, D., & Matthews, A. (2000). An economic evaluation of local government approaches to koala conservation. *Australian Journal of Environmental Management*, *7*(3), 158–169.

Jeffreys, H. (1939). *Theory of probability* (1st ed). Oxford [Oxfordshire] : New York: Clarendon Press ; Oxford University Press.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (1 edition). New York usw.: Wiley.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.

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

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *231*(694-706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Skipper, J. K., Guenther, A. L., & Nass, G. (1967). The Sacredness of .05: A Note concerning the Uses of Statistical Levels of Significance in Social Science. *The American Sociologist*, *2*(1), 16–18.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Winer, B. J. (1962). *Statistical principles in experimental design*. New York : McGraw-Hill.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: Wiley.