Justify Your Alpha: A Primer on Two Practical Approaches

Maximilian Maier[1] & Daniël Lakens[2]

[1] University of Amsterdam, The Netherlands

[2] Eindhoven University of Technology, The Netherlands

This manuscript is a preprint and has not yet been peer reviewed.

Author Note

This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research. All code used to create this manuscript is provided at https://github.com/Lakens/justify_alpha_in_practice. Information about the JustifyAlpha R package and Shiny app is available at https://lakens.github.io/JustifyAlpha/index.html

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

Abstract

The default use of an alpha level of 0.05 is suboptimal for two reasons. First, 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. Second, it is possible that in studies with very high statistical power *p*-values lower than the alpha level can be more likely when the null hypothesis is true, than when the alternative hypothesis is true (i.e., Lindley's paradox). This manuscript explains two approaches that can be used to justify a better choice of an alpha level than relying on the default threshold of 0.05. The first approach is based on the idea to either minimize or balance Type 1 and Type 2 error rates. The second approach lowers the alpha level as a function of the sample size to prevent Lindley's paradox. An R package and Shiny app is provided to perform the required calculations. Both approaches have their limitations (e.g., the challenge of specifying relative costs and priors), but can offer an improvement to current practices, especially when sample sizes are large. The use of alpha levels that have a better justification should improve statistical inferences and can increase the efficiency and informativeness of scientific research.

Keywords: hypothesis testing, Type 1 error, Type 2 error, statistical power

Word count: 7418 words

Justify Your Alpha: A Primer on Two Practical Approaches

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 enough data to make the chance of a wrong decision incredibly small. But resources are 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). Given the same number of observations, 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%. In the past the strong convention to use a 5% alpha level might have functioned as a de facto prespecification of the alpha level, which needs to be before the data is analyzed (Uygun-Tunç, Tunç, & Lakens, 2021). Nowadays researchers can transparently preregister a statistical analysis plan in an online repository, which makes it possible to specify more appropriate but less conventional alpha levels. Even though it is possible to preregister non-conventional alpha levels, there is relatively little practical guidance on how to choose an alpha level for a study. This article explains why error rates need to be justified, and provides two practical approaches that can be used to justify the alpha level. In the first approach, the Type I and Type II error rates are balanced or minimized, and in the second approach the alpha level is lowered as a function of the sample size.

Why Do We Use a 5% Alpha Level and 80% Power?

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. Ronald A. Fisher (1971) 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: “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. R. A. Fisher (1926) notes: “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; Kennedy-Shaffer, 2019), 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 beta is set at .20. This value is offered for several reasons (Cohen, 1965, pp. 98-99). The chief among them takes into consideration the implicit convention for alpha of .05. The beta 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%. The real lesson we should take away from Cohen 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 with a 5% Type 1 error rate and a 20% Type 2 error rate.

Justifying the Alpha Level

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." Despite this advice, the mindless application of null hypothesis significance tests, including setting the alpha level at 5% for all tests, is so universal that is has been criticized for more than half a century (Bakan, 1966; Gigerenzer, 2018). The default use of a 5% alpha level might have been difficult to abandon, even if it was a mediocre research practice, without an alternative approach in which alpha levels are better justified.

There are two 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 (Mudge, Baker, Edge, & Houlahan, 2012). If researchers use hypothesis tests to make dichotomous decisions from a methodological falsificationist approach to statistical inferences (Uygun-Tunç, Tunç, & Lakens, 2021), and have a certain maximum sample size they are willing or able to collect, it is typically possible to make decisions more efficiently by choosing error rates such that the combined Type 1 and Type 2 error rate is minimized. 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 cost of both types of errors.

The second reason 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 likely it becomes that small *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 small *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 (H0) is true than when the alternative hypothesis (H1) is true and we have very high power (the horizontal dashed line falls above the black curve for *p*-values larger than ~0.025).

Although it is not necessary from a Neyman-Pearson error-statistical perspective, researchers often want to interpret a significant test result as evidence for the alternative hypothesis. In other words, in addition to controlling the *error rate*, researchers might be interested in interpreting the *relative evidence* in the data for null hypothesis over 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

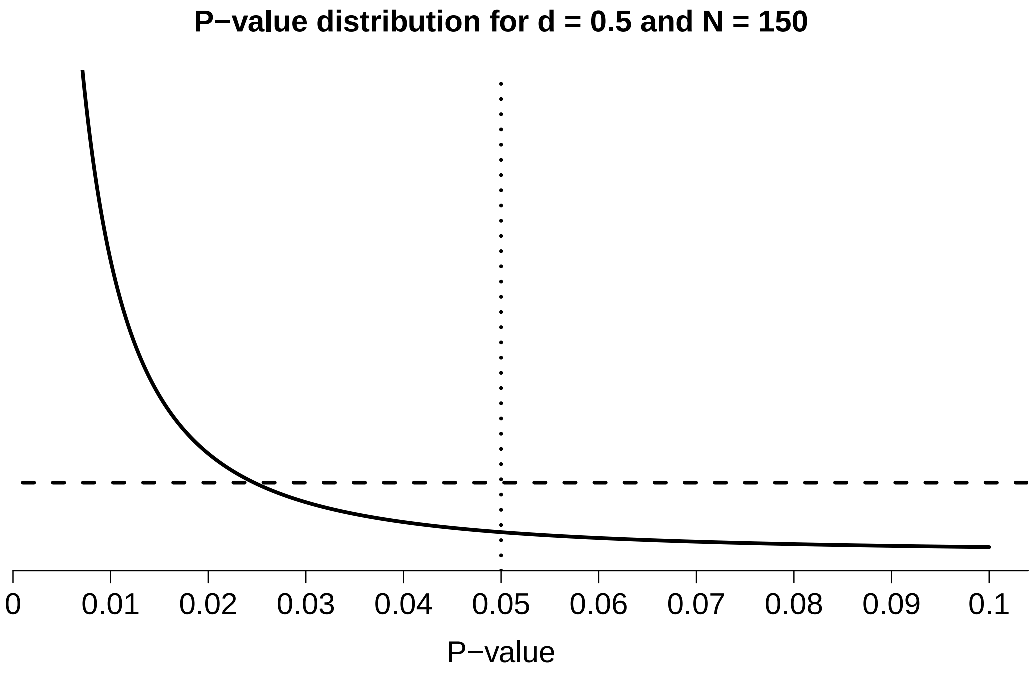

Figure 1. *P*-value distributions for a two-sided independent *t*-test with N = 150 and d = 0.5 (black curve) or d = 0 (horizontal dashed line) which illustrates how *p*-values just below 0.05 can be more likely when there is no effect than when there is an effect.

the alternative hypothesis is more likely than the null hypothesis for all significant *p*-values that would be observed. This approach to justifying the alpha level can be seen as a frequentist/Bayesian compromise (Good, 1992). The error rate is controlled, but the alpha level is also set at a value that guarantees that whenever we reject the null hypothesis, the data is more likely under the alternative hypothesis, than under the null.

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 leads to studies where you make decisions most efficiently. Researchers can choose to design a study with a statistical power and alpha level that minimizes the *weighted combined error rate*. For example, a researcher designs an experiment where they assume H0 and H1 are a-priori equally probable (the prior probability for both is 0.5). They set the Type 1 error rate to 0.05 and collect sufficient data such that the statistical power is 0.80. The weighted combined error rate is 0.5 (the probability H0 is true) $\times$ 0.05 (the probability of a Type 1 error) + 0.5 (the probability that H1 is true) $\times$ 0.20 (the probability of a Type 2 error) = 0.125. This weighted combined error rate might be lower if a different choice for Type 1 and Type 2 errors was made.

Assume that in the previous example data will be analyzed in an independent t-test, and the researcher was willing to collect 64 participants in each condition to achieve the 0.05 Type 1 error rate and 0.8 power. The researcher could have chosen to set the alpha level in this study to 0.1 instead of 0.05. If the Type 1 error rate is 0.1, the statistical power (given the same sample size of 64 per group) would be 0.88. The weighted combined error rate is now $(0.5 \times 0.1 + 0.5 \times 0.12) = 0.11$. In other words, increasing the Type 1 error rate from 0.05 to 0.1 reduced the Type 2 error rate from 0.2 to 0.12, and the combined error rate from 0.125 to 0.11. In the latter scenario, our total probability of making an erroneous decision has become 0.015 smaller. As shown below, this approach can be extended to incorporate scenarios where the prior probability of H0 and H1 differ. Mudge, Baker, Edge, and Houlahan (2012) and Kim and Choi (2021) 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 than when the default alpha level of 0.05 is used.

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, and perhaps more importantly, one should carefully reflect on the choice of the alpha level when an experiment achieves very high statistical power for all effect sizes that are considered meaningful. If a study has 99% power for effect sizes of interest, and thus a 1% Type 2 error rate, but uses the default 5% alpha level, it also suffers from a lack of balance. This latter scenario is quite common in meta-analyses, where researchers by default use a 0.05 alpha level, while the meta-analysis often has very high power for all effect sizes of interest. It is also increasingly common when analyzing large existing data sets, or when collecting thousands of observations online. In such cases where power for all effects of interest is very high, it is sensible to lower the alpha level for statistical tests to reduce the weighted combined error rate, and increase the severity of the test.

Researchers can decide to either balance Type 1 and Type 2 error rates (e.g., designing a study such that the Type 1 and Type 2 error rate are equal), or minimize the weighted combined error rate. For any given sample size and effect size of interest there is an alpha level that minimizes the weighted combined Type 1 and Type 2 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 optimization procedure.

As an example, imagine a researcher who plans to perform a study which will be analyzed with an independent two-sided *t*-test. They will collect 50 participants per condition, and set their smallest effect size of interest to Cohen's d = 0.5. They think a Type 1 error is just as costly as a Type 2 error, and believe H0 is just as likely to be true as H1. The weighted combined error rate is minimized when they set alpha to 0.13 (see Figure 2, dotted line), which will give the study a Type 2 error rate of beta = 0.166 to detect effects of d = 0.5. The weighted combined error rate is 0.148, while it would have been 0.177 if the alpha level was set at 5%[1].

We see that increasing the alpha level from the normative 5% level to 0.13 reduced the weighted combined error rate - any larger or smaller alpha level would increase the weighted combined error rate. The reduction in the weighted combined error rate is not huge, but we have reduced the overall probability of making an error. More

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

importantly, we have chosen an alpha level based on a justifiable principle, and clearly articulated the relative costs of a Type 1 and Type 2 error. Perhaps counter-intuitively, decision making is sometimes slightly more efficient after *increasing* the alpha level from the default of 0.05, because a small increase in the Type 1 error rate can lead to a larger decrease in the Type 2 error rate. Had the sample size been much smaller, such as n = 10, the solid line in Figure 2 shows that the weighted combined error rate will always be high, but it is minimized if we increase the alpha level to alpha to 0.283. If the sample size had been n = 100, the optimal alpha level to minimize the weighted combined error rate (still assuming H0 and H1 have equal probabilities, and Type 1 and Type 2 errors are equally costly) is 0.0509 (the long-dashed line in Figure 2).

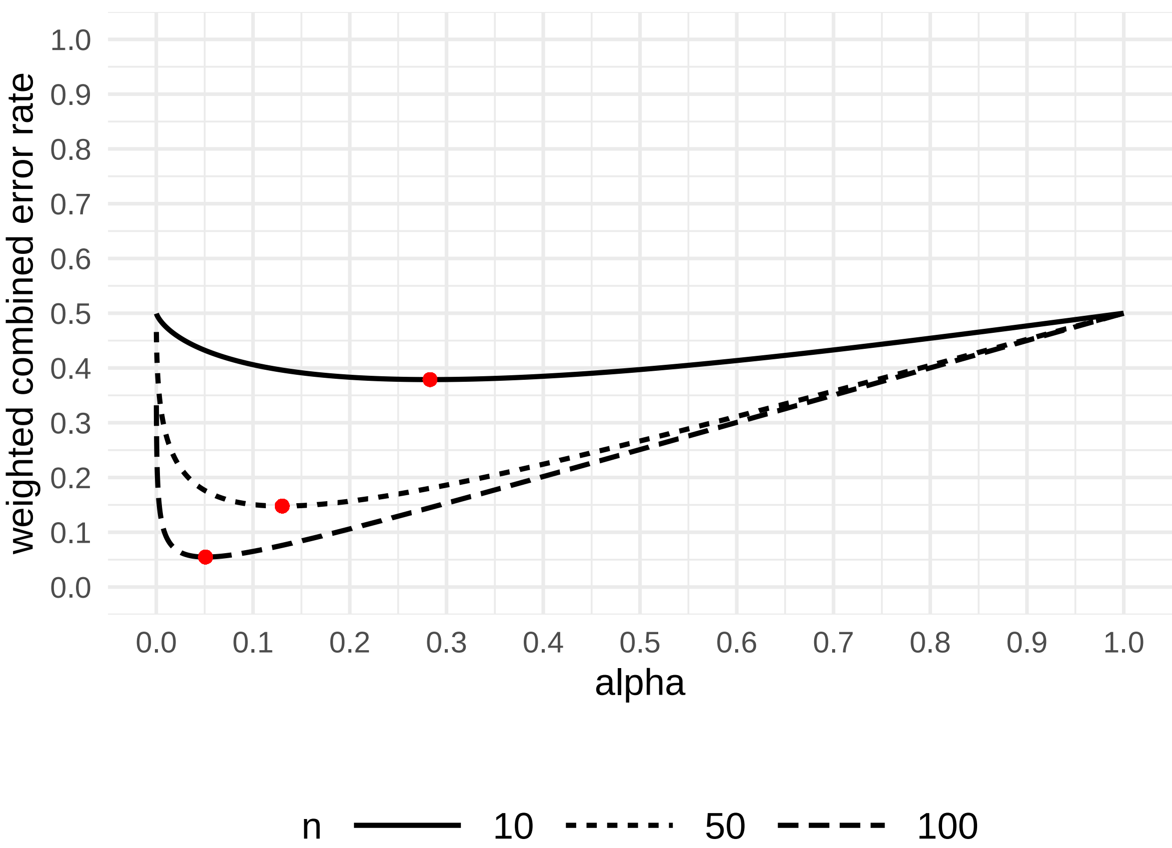

Figure 2. Weighted combined error rate (y-axis) for an independent *t*-test with n = 10, n = 50, and n = 100 per group and a smallest effect of interest of d = 0.5, for all possible alpha levels (x-axis).

Weighing the Relative Cost of Errors

Cohen (1988) recommended a study design with a 5% Type 1 error rate and a 20% Type 2 error rate. He personally felt "Type I errors are of the order of four times as serious as Type II errors." However, some researchers have pointed out, following Neyman (1933), that false negatives might be more severe than false positives (Fiedler, Kutzner, & Krueger, 2012). The best way to determine the relative costs of Type 1 and Type 2 errors is by performing a cost-benefit analysis. For example, Field, Tyre, Jonzén, Rhodes, and Possingham (2004) quantify the relative costs of Type 1 errors when testing whether native species in Australia are declining. They find that when it comes to the Koala population, given 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 practically infinite.

Although it can be difficult to formally quantify all relevant factors that influence the costs of Type 1 and Type 2 errors, there is no reason to let the perfect be the enemy of the good. In practice, even if researchers don't explicitly discuss their choice for the relative weight of Type 1 versus Type 2 errors, they make a choice in every hypothesis test they perform, even if they simply follow conventions (e.g., a 5% Type 1 error rate and a 20% Type 2 error rate). It might be especially difficult to decide upon the relative costs of Type 1 and Type 2 errors when there are no practical applications of the research findings, but even in these circumstances, it is up to the researcher to make a decision (Douglas, 2000). It is therefore worth reflecting on how researchers can start to think about the relative weight of Type 1 and Type 2 errors.

First, if a researcher only cares about not making a decision error, but the researcher does not care about whether this decision error is a false positive or a false negative, Type 1 and Type 2 errors are weighed equally. Therefore, weighing Type 1 and Type 2 errors equally is a defensible default, unless there are good arguments to weigh false positives more strongly than false negatives (or vice versa). When deciding upon whether there is a reason to weigh Type 1 and Type 2 errors differently, researchers are in essence performing a multiple criterion decision analysis (Edwards, Miles Jr., & Winterfeldt, 2007), and it is likely that treating the justification of the relative weight of Type 1 and Type 2 errors as a formal decision analysis would be a massive improvement over current research practices. A first step is to determine the objectives of the decision that is made in the hypothesis test, assign attributes to measure the degree to which these objectives are achieved, within a specific time-frame (Clemen, 1997), and finally to specify a value function.

In a hypothesis test we do not simply want to make accurate decisions, but we want to make accurate decisions given the resources we have available (e.g., time and money). Incorrect decisions have consequences, both for the researcher themselves, as for scientific peers, and sometimes for the general public. We know relatively little about the actual costs of publishing a Type 1 error for a researcher, but in many disciplines the costs of publishing a false claim are low, while the benefits of an additional publication on a resume are large. However, by publishing too many claims that do not replicate, a researcher risks gaining a reputation for publishing unreliable work. There are additional criteria to consider. A researcher might plan to build on work in the future, as might peers. The costs of experiments that follow up on a false lead might be much larger than the cost to reduce the possibility of a Type 1 error in an initial study, unless replication studies are cheap, will be performed anyway, and will be shared with peers. However, it might also be true that the hypothesis has great potential for impact if true, and the cost of a false negative might be substantial whenever it closes off a fruitful avenue for future research. A Type 2 error might be more costly than a Type 1 error, especially in a research field where all findings are published and people regularly perform replication studies to identify Type 1 errors in the literature (Fiedler, Kutzner, & Krueger, 2012).

Another objective might be to influence policy, in which case the consequences of a Type 1 and Type 2 error should be weighed by examining the relative costs of implementing a policy that does not work against not implementing a policy that works. The second author once attended a presentation by policy advisor who decided whether new therapies would be covered by the national healthcare system. She discussed Eye Movement Desensitization and Reprocessing (EMDR) therapy. She said that, although the evidence for EMDR was weak at best, the costs of the therapy (which can be done behind a computer) are very low, it was applied in settings where no good alternative therapies were available (e.g., inside prisons), and risk of negative side-effects was basically zero. They were aware of the fact that there was a very high probability that the claim that EMDR was beneficial might be a Type 1 error, but the cost of a Type 1 error was deemed much lower than the cost of a Type 2 error.

Imagine a researcher plans to collect 64 participants per condition to detect a d = 0.5 effect, and weighs the cost of Type 1 errors 4 times as much as Type 2 errors. To minimize error rates, the Type 1 error rate should be set to 0.0327, which will make the Type 2 error rate 0.252. If we would perform 20000 studies designed with these error rates, and assume H0 and H1 are equally likely to be true, we would observe 0.5 (the prior probability that H0 is true) × 0.0327 (the alpha level) × 20000 = 327 Type 1 errors, and 0.5 (the prior probability that H1 is true) × 0.252 (the Type 2 error rate) × 20000 = 2524 Type 2 errors. Since we weigh Type 1 errors 4 times as much as Type 2 errors, we multiple the cost of the 327 Type 1 errors by 4, which makes 4 × 327 = 1308, and to keep the weighted error rate between 0 and 1, we also multiply the 10000 studies where we expect H0 to be true by 4, such that the weighted combined error rate is (1308 + 2524)/(40000 + 10000) = 0.0766. Figure 3 visualizes the weighted combined error rate for this study design across the all possible alpha levels, and illustrated the weighted error rate is smallest when the alpha level is 0.0327.

If the researcher had decided to *balance* error rates instead of *minimizing* error rates, we recognize that with 64 participants per condition we are exactly in the scenario Cohen (1988) described. When Type 1 errors are considered 4 times as costly as Type 2 errors, 64 participants per condition yield a 5% Type 1 error rate and a 20% Type 2 error rate. If we would increase the sample size, The Type 1 and Type 2 error rates would remain in a balanced 1:4 ratio, but both error rates would be smaller. With

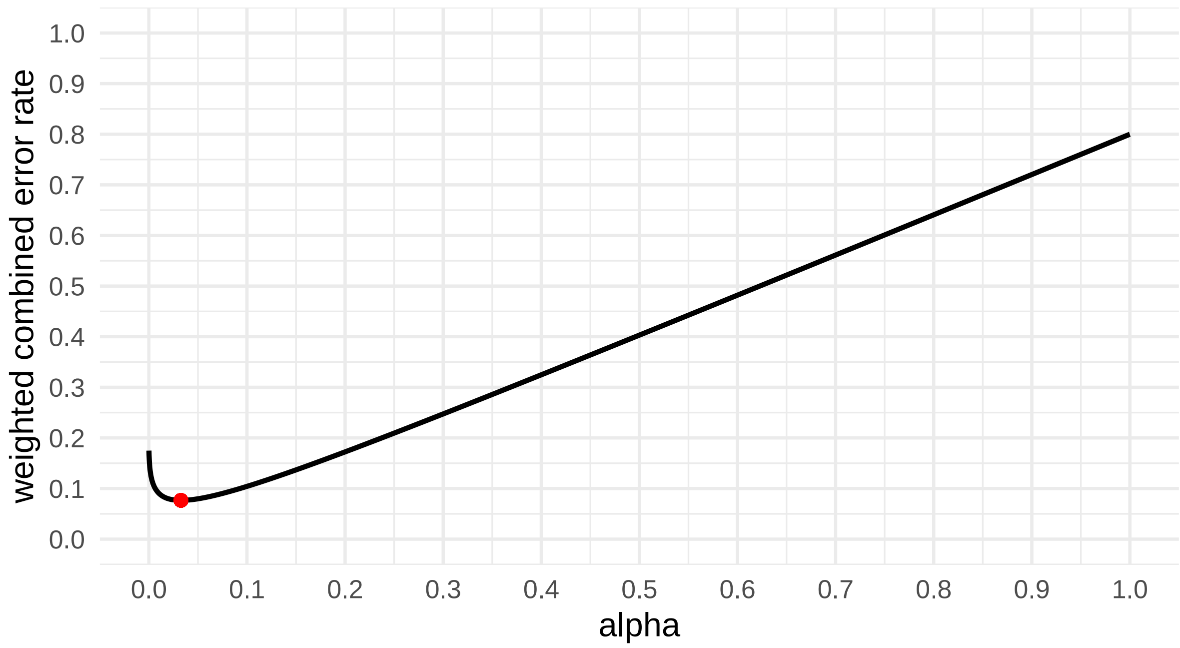

Figure 3. Weighted combined error rate (y-axis) for an independent *t*-test with n = 64 per group and a smallest effect of interest of d = 0.5, where Type 1 errors are weighed 4 times as much as Type 2 errors, for all possible alpha levels (x-axis).

a smaller sample size, both error rates would be larger.

Incorporating Prior Probabilities

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 (Miller & Ulrich, 2019). 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. On the other hand, if the base rate of true hypotheses is very low, you are more likely to test a hypothesis where H0 is true, and the probability of observing a false positive becomes a more important consideration. Whatever the prior probabilities are believed to be, researchers always need to specify the prior probabilities of H0 and H1. Researchers should take their expectations about the probability that H0 and H1 are true into account when evaluating costs and benefits.

For example, let's assume a researcher performs 1000 studies. The researcher expects 100 studies to test a hypothesis where H1 is true, while the remaining 900 studies test a hypothesis where H0 is true. This means H0 is believed to be 9 times more likely than H1, or equivalently, that the relative probability of H1 versus H0 is 0.1111:1. However, the researcher decides to ignore these prior probabilities and designs a study that has the normative 5% Type 1 error rate and a 20% Type 2 error rate. The researcher should expect to observe 0.9 (the prior probability that H0 is true) $\times$ 0.05 (the alpha level) $\times$ 1000 = 45.00 Type 1 errors, and 0.1 (the prior probability that H1 is true) $\times$ 0.2 (the Type 2 error rate) $\times$ 1000 = 20.00 Type 2 errors, for a total of 65.00 errors.

However, the total number of errors do not tell the whole story, as Type 1 errors are weighed four times more than Type 2 errors. We therefore need to compute the weighted combined error rates w taking the relative cost of Type 1 and Type 2 errors into account, and the prior probabilities of H0 and H1, which can be done with the following formula from Mudge, Baker, Edge, and Houlahan (2012):

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

For the previous example, the weighted combined error rate is (4 $\times$ 0.05 + 0.1111 $\times$ 0.2) / (0.1111 + 4) = 0.054. If the researcher had taken the prior probabilities into account when deciding upon the error rates, a lower combined error rate can be achieved. With the same sample size (64 per condition) the combined weighted error rate was not as small as possible, optimally balanced error rates (maintaining the 4:1 ratio of the weight of Type 1 versus Type 2 errors) would require setting alpha to 0.011 and the Type 2 error rate to 0.402. The researcher should now expect to observe 0.9 (the prior probability that H0 is true) $\times$ 0.011 (the alpha level) $\times$ 1000 = 9.89 Type 1 errors, and 0.1 (the prior probability that H1 is true) $\times$ 0.402 (the Type 2 error rate) $\times$ 1000 = 40.16 Type 2 errors. The weighted error rate is 0.0216.

Because the prior probability of H0 and H1 influence the expected number of

Type 1 and Type 2 errors one will observe in the long run, the alpha level should be lowered as the prior probability of H0 increases, or equivalently, the alpha level should be increased as the prior probability of H1 increases. Because the base rate of true hypotheses is unknown, this step requires a 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), which is often unlikely in practice. In the previous example, it would also have been possible minimize (instead of balance) the error rates, which is achieved with an alpha of 0.00344 and a beta of 0.558, for a total of 58.86 errors, where the weighted error rate is 0.0184.

The two approaches (balancing error rates or minimizing error rates) typically yield quite similar results. 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 is equal). Note that although there is always an optimal choice of the alpha level, there is always a range of values for the alpha level that yield quite similar weighted error rates, as can be seen in Figure 3.

Sample Size Justification when Minimizing or Balancing Error Rates

So far we have illustrated how to perform what is known as a *compromise power analysis* where the weighted combined error rate is computed as a function of the sample size, the effect size, and the desired ratio of Type 1 and Type 2 errors (Erdfelder, Faul, & Buchner, 1996). However, in practice researchers will often want to justify their sample size based on an *a-priori power analysis* where the required sample size is computed to achieve desired error rates, given an effect size of interest (Lakens, 2021). It is possible to determine the sample size at which we achieve a certain desired weighted combined error rate. This requires researchers to specify the effect size of interest, and relative cost of Type 1 and Type 2 errors, the prior probabilities of H0 and H1, whether error rates should be balanced or minimized, and the desired weighted combined error rate.

Imagine a researcher is interested in detecting an effect of Cohen's d = 0.5 with a two sample *t*-test. The researcher believes Type 1 errors are equally costly as Type 2 errors, and believes a H0 is equally likely to be true as H1. The researcher desires a minimized weighted combined error rate of 5%. Figure 4 shows the optimal alpha level, beta, and weighed combined error rate as a function of sample size for this situation. We can optimize the weighted combined error rate as a function of the alpha level and sample size through an iterative procedure, which reveals that a sample size of 105 participants in each independent condition is required to achieve the desired weighted combined error rate. In the specific cases where the prior probability of H0 and H1 are equal, this sample size can also be computed directly with common power analysis software by entering the desired alpha level and statistical power. In this example, where Type 1 and Type 2 error rates are weighted equally, and the prior probability of H0 and H1 is assumed to be 0.5, the sample size is identical to that required to achieve an alpha of 0.05 and a desired statistical power for d = 0.5 of 0.95. Note that it might be difficult to specify the desired *weighted* combined error rate for a power analysis when Type 1 and Type 2 errors are not weighed equally, and/or H1 and H0 are not equally probable.

Lowering the Alpha Level as a Function of the Sample Size

Formally controlling the costs of errors can be a challenge, as it requires researchers to specify relative cost of Type 1 and Type 2 errors, prior probabilities, and the effect size of interest. Due to this complexity, researchers might be tempted to fall back to the heuristic use of an alpha level of 0.05. Fisher (1971) referred to the default alpha level of 0.05 as a "convenient convention," and believe it suffices as a low enough threshold to make scientific claims in a scientific system where we have limited resources and value independent replications (Uygun-Tunç, Tunç, & Lakens, 2021).

However, there is a well known limitation of using a fixed alpha level that has lead statisticians to recommend choosing an alpha level as a function of the sample size. To understand the argument behind this recommendation, it is important to distinguish

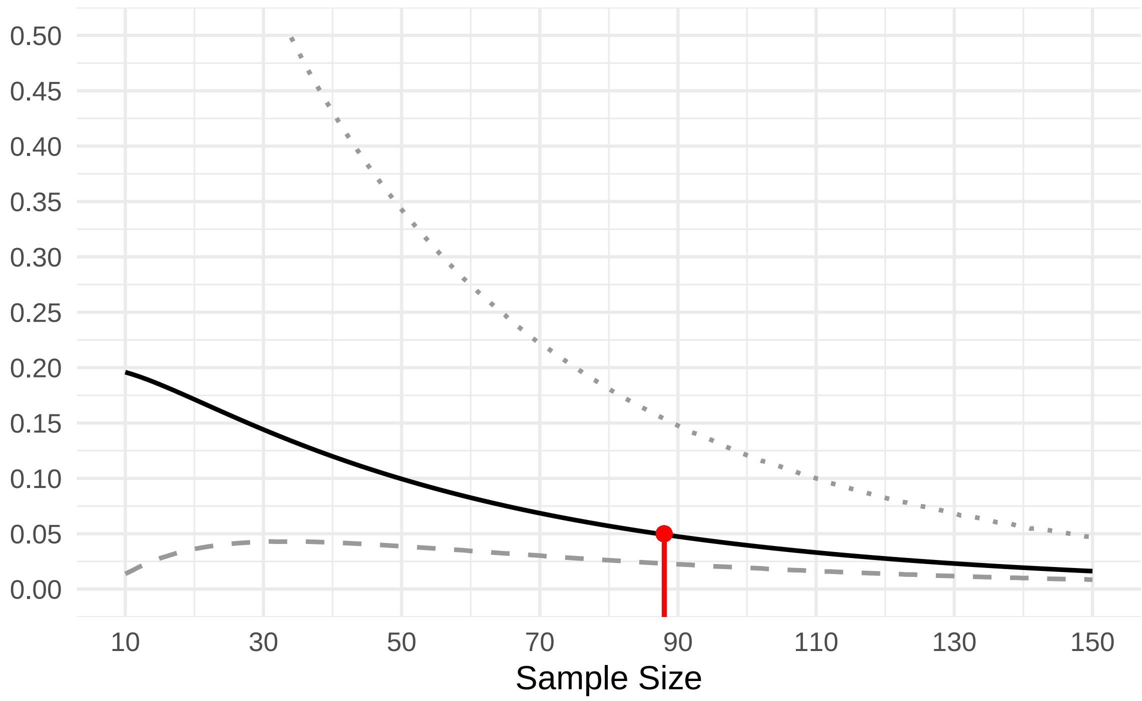

Figure 4. Weighted combined error rate (solid black line), alpha (lower grey dashed line), and beta (upper grey dotted line) for an independent t-test as a function of sample size when the alpha level is justified based on the goal to minimize the error rate at each sample size. The sample size corresponding to the red dot is the minimum required sample size to achieve a 5% weighted combined error rate.

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.

To prevent situations where a frequentist rejects the null hypothesis based on $p <$ 0.05, when the evidence in the test favors the null hypothesis over the alternative hypothesis, it is recommended to lower the alpha level as a function of the sample size. The need to do so is discussed 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." Similarly, 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."

Lindley's paradox emerges because in frequentist statistics the critical value of a test approaches a limit as the sample size increases (e.g., $t = 1.96$ for a two-sided t-test with an alpha level of 0.05). It does not emerge in Bayesian hypothesis tests because the critical value (e.g., a BF > 10) requires a larger test statistic as the sample size increases (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Zellner, 1971). A Bayes Factor of 1 implies equal evidence for H0 and H1. To prevent Lindley's paradox when using frequentist statistics one would need to adjust the alpha level in a way that the likelihood ratio (also called the Bayes factor) at the critical test statistic is not larger than 1. With such an alpha level, a significant p-value will always be at least as likely if H1 is true than if H0 is true, which avoids Lindley's paradox. Faulkenberry (2019) and Rouder, Speckman, Sun, Morey, and Iverson (2009) developed Bayes factors for t-tests and Analysis of Variance (ANOVA) which can calculate the Bayes factor from the test statistic and degrees of freedom. We developed a Shiny app that lowers the alpha level for a t-test or ANOVA, such that the critical value that leads researchers to reject H0 is also high enough to guarantee (under the assumption of the priors) that the data provide relative evidence in favor of H1.

There are two decisions that should be made when desiring to prevent Lindley's paradox, the first about the prior, and the second about the threshold for the desired evidence in favor of H1. Both Leamer (1978) and Good (1992) offer their own suggstions. We rely on a unit information prior for the ANOVA and a Cauchy prior with scale 0.707 for *t*-tests (although the package allows users to adjust the r scale). Both of these priors are relatively wide, which makes them a conservative choice when attempting to prevent Lindley's paradox. The choice for this prior is itself a 'convenient convention,' but the approach extends to other priors researchers prefer, and researchers can write custom code if they want to specify a different prior. A benefit of the chosen defaults for the priors is that, in contrast to previous approaches that aimed to calculate a Bayes factor for every *p*-value (Colquhoun, 2017, 2019), researchers do not need to specify the effect size under the alternative hypothesis. This lowers the barrier of adopting this approach in situations where it is difficult to specify a smallest effect size of interest or an expected effect size.

A second decision is the threshold of the Bayes factor used to lower the alpha level. Using a Bayes factor of 1 formally prevents Lindley's paradox. It does mean that one might reject the null hypothesis when the data provide just as much evidence for H1 as for H0. Although it is important to note that researchers will often observe *p*-values well below the critical value, and thus, in practice the evidence in the data will be in favor of H1 when H0 is rejected, researchers might want to increase the threshold of the Bayes factor that is used to lower the alpha level to prevent weak evidence (Jeffreys, 1939). This can be achieved by setting the threshold to a larger value than 1 (e.g., BF $>$ 3). The Shiny app allows researchers to adjust the alpha level in a way that a significant *p*-value will always provide moderate (BF $>$ 3) or strong (BF $>$ 10) evidence against the null hypothesis.

To illustrate this approach to justifying the alpha level as a function of the sample size, imagine a researcher collected 150 observations in a within subjects design where they aim to test a directional prediction in a dependent *t*-test. For any sample size and choice of prior, a *p*-value is directly related to a Bayes factor. Figure 5 shows the relationship of two-sided *p*-values and Bayes factors using a Cauchy prior with a r-scale

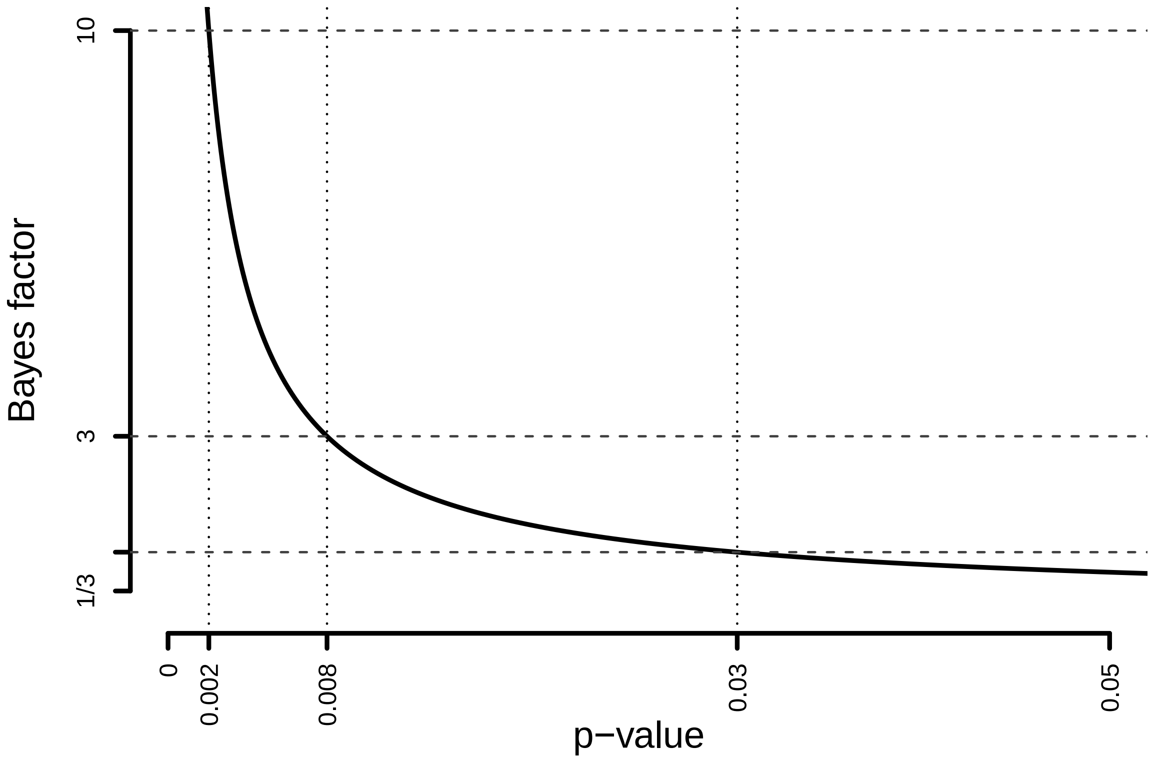

Figure 5. Relationship between *p*-value and Bayes factor for a one-sample *t*-test with 150 participants using a Cauchy prior.

of 0.707 given a sample size of 150 for a within subjects *t*-test. To avoid Lindley's paradox, the researcher would need to use an alpha level of 0.0302 for the one-sided *t*-test, given the chosen prior, as this choice for an alpha level guarantees that a significant *p*-value will correspond to evidence in favor of H1.

For small sample sizes it is possible to guarantee that a significant result is evidence for the alternative hypothesis using an alpha level that is higher than 0.05. It is not recommended to use the procedure outlined in this section to *increase* the sample size above the conventional choice of an alpha level (e.g., 0.05). This approach to the justification of an alpha level assumes researchers first want to control the error rate, and as a secondary aim want to prevent Lindley's paradox by reducing the alpha level as a function of the sample size where needed. Figure 6 shows the alpha levels for different values of N for between and within subjects *t*-test. We can see that particularly for within subjects t-tests the alpha level rapidly falls below 5% as the sample size increases.

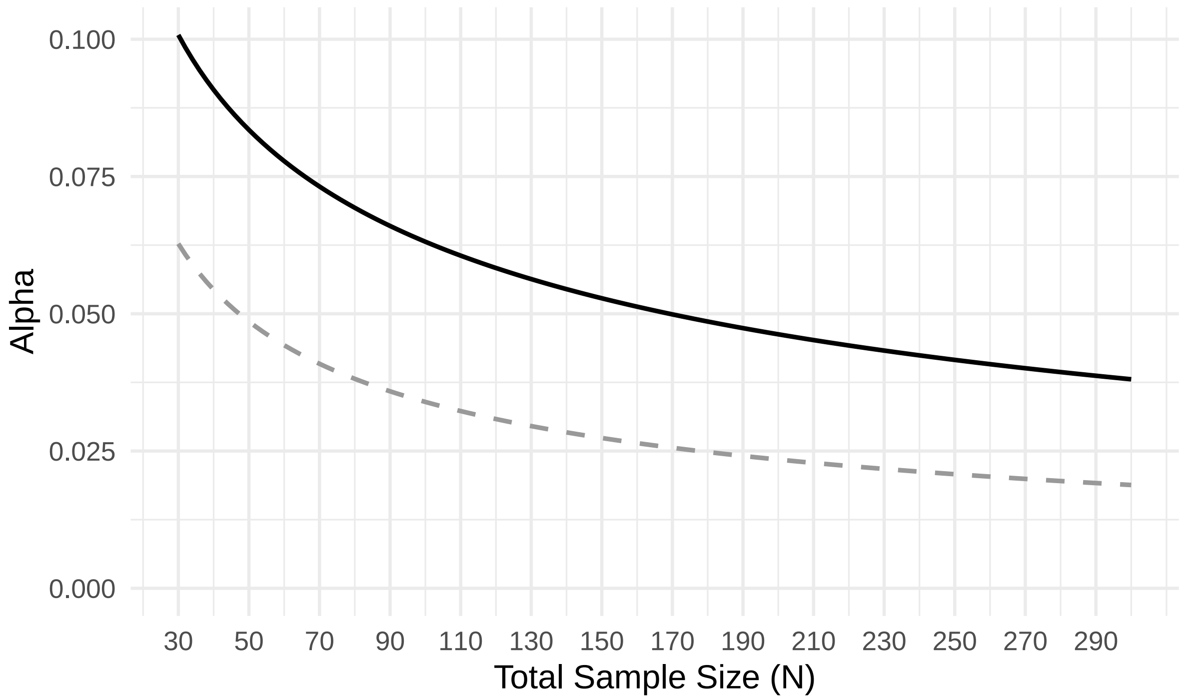

Figure 6. Optimal alpha level for within (grey dashed line) and between-sample (solid black line) two-sided t-tests.

Discussion

As the choice of error rates is an important decision in any hypothesis test, authors should always be expected to justify their choice of error rates whenever they use data to make decisions about the presence or absence of an effect. As Skipper, Guenther, and Nass (1967) remark: “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." It should especially become more common to lower the alpha level when analyzing large data sets, or when performing meta-analyses, whenever each test has very high power to detect any effect of interest. Researchers should also consider increasing the alpha level when the combination of the effect size of interest, the sample size, the relative cost of

Type 1 and Type 2 errors, and the prior probability of H1 and H0 mean this will improve the efficiency of decisions that are made.

When should we minimize or balance error rates and when should we avoid Lindley's paradox? In practice, it might be most convenient to minimize or balance error rates whenever there is enough information to conduct a power analysis, and if researchers feel comfortable specifying the relative cost of Type 1 and Type 2 errors, and the prior probabilities of the null and alternative hypothesis. If researchers do not feel they can specify these parameters, they can fall back on the approach to lower the alpha level as a function of the sample size to prevent Lindley's paradox. The first approach is most attractive to researchers who follow a strict Neyman-Pearson approach, while researchers interested in a compromise between frequentist and Bayesian inference might be drawn more strongly towards the second approach (Good, 1992).

A Shiny app is available that allows users to perform the calculations recommended in this article. It can be used to minimize or balance alpha and beta by specifying the effect size of interest and the sample size, as well as an analytic power function. 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 Lakens (2021)). Alternatively, researchers can lower the alpha level as a function of the sample size by specifying only their sample size. In a Neyman-Pearson approach to statistics the alpha level should be set before the data is collected. Whichever approach is used, it is strongly recommended to preregister the alpha level that researchers plan to use before the data is collected. In this preregistration, researchers should document and explain all assumptions underlying their decision for an alpha level, such as beliefs about prior probabilities, or choices for the relative weight of Type 1 and Type 2 errors.

Throughout this manuscript we have reported error rates rounded to three decimal places. Although we can compute error rates to many decimals, it is useful to remember that the error rate is a long run frequency, and in any finite number of tests (e.g., all the tests you will perform in your lifetime) the observed error rate varies somewhere around the long run error rate. The weighted combined error rate might be quite similar across a range of alpha levels, or when using different justifications (e.g., or balancing versus minimizing alpha levels in a cost-benefit approach) and small differences between alpha levels might not be noticeable in a limited number of studies in practice. We recommend to preregister alpha levels up to three decimals, while keeping in mind there is some false precision in error rates with too many decimals.

Because of the strong norms to use a 5% error rate when designing studies there are relatively few examples of researchers who attempt to justify the use of a different alpha level. Within specific research lines researchers will need to start to develop best practices to decide how to weigh the relative cost of Type 1 and Type 2 errors, or quantify beliefs about prior probabilities. It might be a challenge to get started, but the two approaches illustrated here provide one way to move beyond the mindless use of a 5% alpha level, and make more informative decisions when we test hypotheses.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Clemen, R. T. (1997). *Making Hard Decisions: An Introduction to Decision Analysis* (2 edition). Belmont, Calif: Duxbury.

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

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

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, *73*(sup1), 192–201. https://doi.org/10.1080/00031305.2018.1529622

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--558. https://doi.org/10.1037/0003-066X.37.5.553

Cowles, M., & Davis, C. (1982). On the origins of the. 05 level of statistical significance. *American Psychologist*, *37*(5), 553--558. https://doi.org/10.1037/0003-066X.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

Douglas, H. E. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579. https://doi.org/10.1086/392855

Edwards, W., Miles Jr., R. F., & Winterfeldt, D. von (Eds.). (2007). *Advances in decision analysis: From foundations to applications*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511611308

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Faulkenberry, T. J. (2019). Estimating evidential value from analysis of variance summaries: A comment on ly et al.(2018). *Advances in Methods and Practices in Psychological Science*, *2*(4), 406–409. https://doi.org/10.1177/2515245919872960

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*(661–669), 661–669. https://doi.org/10.1177/1745691612462587

Field, S. A., 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. Retrieved from 10.1111/j.1461-0248.2004.00625.x

Fisher, R. A. (1926). Introduction to "The arrangement of field experiments." *Journal of the Ministry of Agriculture*, *33*, 503–513.

Fisher, R. A. (1926). Introduction to "The arrangement of field experiments." *Journal of the Ministry of Agriculture*, *33*, 503–513.

Fisher, Ronald A. (1971). *The Design of Experiments* (9 edition). New York: Macmillan Pub Co.

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

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

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

Kennedy-Shaffer, L. (2019). Before p < 0.05 to Beyond p < 0.05: Using History to Contextualize p-Values and Significance Testing. *The American Statistician*, *73*(sup1), 82–90. https://doi.org/10.1080/00031305.2018.1537891

Kim, J. H., & Choi, I. (2021). Choosing the level of significance: A decision-theoretic approach. *Abacus*, *57*(1), 27–71. https://doi.org/10.1111/abac.12172

Lakens, D. (2021). *Sample Size Justification*. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

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.

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

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.

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.

Uygun-Tunç, D., Tunç, M. N., & Lakens, D. (2021). *The epistemic and pragmatic function of dichotomous Claims based on statistical hypothesis tests*. PsyArXiv. https://doi.org/10.31234/osf.io/af9by

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.