Scientists regularly need to make dichotomous decisions when they perform lines of research. Should a pilot study be performed or not? When multiple possible manipulations or measures are available, which should be used for the next study? Should the design of a study include a possible moderator, or can it be ignored? Should a research line be continued or abandoned? These decisions come with costs and benefits for the scientist and for society when bad decisions lead to research waste. 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 by using an extremely low alpha level while still achieving very high statistical power. However, because resources are limited, researchers need to decide how to choose the rate at which they are willing to make errors (Wald, 1949). After data are 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). With 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 of 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 an alpha level 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 was useful given that the alpha level needs to be decided on before the data are analyzed (Uygun-Tunç et al., 2021). Nowadays, researchers can transparently preregister a 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 an alpha level 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 was useful given that the alpha level needs to be decided on before the data are analyzed (Uygun-Tunç et al., 2021). Nowadays, researchers can transparently preregister a 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 an alpha level 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 was useful given that the alpha level needs to be decided on before the data are analyzed (Uygun-Tunç et al., 2021). Nowadays, researchers can transparently preregister a 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 an alpha level 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 was useful given that the alpha level needs to be decided on before the data are analyzed (Uygun-Tunç et al., 2021). Nowadays, researchers can transparently preregister a 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 an alpha level 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 was useful given that the alpha level needs to be decided on before the data are analyzed (Uygun-Tunç et al., 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. Although it is possible to pre-register nonconventional 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 cost of Type 1 and Type 2 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 toward 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) wrote: “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” (p. 13). Likewise, Neyman and Pearson (1933) wrote: “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” (p. 296)

Although in theory, alpha levels should be justified, in practice, researchers tend to imitate others. R. A. Fisher (1926) noted: “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” (p. 504). 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)” (same page, p. 504). Indeed, in his examples, Fisher often used an alpha of .01. Nevertheless, researchers 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 .05 can already be found in work of Gosset on the $t$ distribution (Cowles & Davis, 1982; Kennedy-Shaffer, 2019), who believed that a difference of 2 $SD$ (a $z$ score of 2) was sufficiently rare.

The default use of 80% power (or a 20% Type 2, or beta, error) is similarly based on personal preferences by Cohen (1988), who wrote:

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 1 errors are of the order of four times as serious as Type 2 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. (p. 56)

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%. This normative use of statistics was criticized in a statement by the American Statistical Association (Wasserstein & Lazar, 2016), who wrote: “We teach it because it’s what we do; we do it because it’s what we teach” (p. 129). 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 4 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” (p. 12). 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 it 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 et al., 2012). If researchers use hypothesis tests to make dichotomous decisions from a methodological falsificationist approach to statistical inferences (Uygun-Tunç et al., 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 cost of
Type 1 and Type 2 errors 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 on the basis of convention but by weighting the relative cost of both types of errors.

The second reason is most relevant for large data sets (Harford, 2014). As the statistical power increases, some p values below .05 (e.g., p = .04) can be more likely when there is no effect than when there is an effect. This is known as Lindley’s paradox (Bartlett et al., 1957; Cousins, 2017; Jeffreys, 1935, 1936a, 1936b; Lin et al., 2013; Lindley, 1957) or sometimes the Jeffreys-Lindley paradox (Spanos, 2013) because Harold Jeffreys discussed the paradox long before Lindley (Wagenmakers & Ly, 2021). 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 .04 and .05. When the statistical power is extremely high, not only will most p values fall below .05, but also most will fall below .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 .01). But observing a p value of .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, as illustrated by the fact that the density of the p-value distribution is higher under H0 than under H1 at .04 in Figure 1.

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 the alternative hypothesis over the null 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 .02, then 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 at the same time, the alpha level is set to a value that guarantees that whenever we reject the null hypothesis, the data are more likely under the alternative hypothesis than under the null.

**Minimizing or balancing Type 1 and Type 2 error rates**

If both Type 1 and Type 2 errors are costly, then it makes sense to optimally reduce both errors as you design studies. This idea is well established in applied statistics (Cornfield, 1969; DeGroot, 1975; Kim & Choi, 2021; Lindley, 1953; Mudge et al., 2012; Pericchi & Pereira, 2016) and leads to studies in which 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 in which they assume H0 and H1 are a priori equally probable (the prior probability for both is .5). They set the Type 1 error rate to .05 and collect sufficient data such that the statistical power is .80. The weighted combined error rate is .5 (the probability H0 is true) × .05 (the probability of a Type 1 error) + .5 (the probability that H1 is true) × .20 (the probability of a Type 2 error) = .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 would be analyzed in an independent t test and the researcher was willing to collect 64 participants in each condition to achieve the .05 Type 1 error rate and .80 power. The researcher could have chosen to set the alpha level in this study to .1 instead of .05. If the Type 1 error rate is .1, the statistical power (given the same sample size of 64 per group) would be .88. The weighted combined error rate is now (2 × .1 + 5 × .12) = .11. In other words, increasing the Type 1 error rate from .05 to .1 reduced the Type 2 error rate from .20 to .12 and the combined error rate from .125 to .11. In the latter scenario, the total probability of making an erroneous decision has become
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. Likewise, 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 of interest. 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, in which researchers by default use a .05 alpha level even though 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 in which 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., by 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 on the Type 1 error rate, minimizing or balancing error rates requires an iterative optimization procedure.

As an example, imagine researchers who plan to perform a study that 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 $H_0$ is just as likely to be true as $H_1$. The weighted combined error rate is minimized when they set alpha to .13 (see Fig. 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 .148, whereas it would have been .177 if the alpha level was set at 5%.

We see that increasing the alpha level from the normative 5% level to .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 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 counterintuitively, decision-making is sometimes slightly more efficient after increasing the alpha level from the default of .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 .283. If the sample size had been 100, the optimal alpha level to minimize the weighted combined error rate (still assuming $H_0$ and $H_1$ have equal probabilities and Type 1 and Type 2 errors are equally costly) is .0509 (the long-dashed line in Fig. 2).
**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” (p. 56). However, some researchers have pointed out, following Neyman and Pearson (1933), that false negatives might be more severe than false positives (Field et al., 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 et al. (2004) quantified the relative costs of Type 1 errors when he tested whether native species in Australia were declining. In this example, the H1 is that the koala population is declining, and the H0 is that the koala population is not declining. The Type 1 error would be to decide that the koala population is declining when in fact it is not; a Type 2 error would be to decide that the koala population is not declining when in fact it is. Field et al. concluded 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 with a Type 1 error is too high. Note that in this example, the decision not to collect data is deterministically dominant (Clemen, 1997). The alpha of 1 shows that the results of the data collection will not influence future decisions in any way—it is always beneficial to intervene. This is arguably rare but not incredibly rare. If you are bitten by an animal, it is possible to observe the animal for 10 days to see whether it has rabies before you decide to go to the doctor for a rabies shot, but given the costs and benefits, it is more cost-efficient to assume the animal has rabies and get a rabies shot. In psychology, it is possible that accurate pilot studies to determine which of two possible manipulations has a larger effect size will require a larger sample than if one designs a study conservatively powered for the manipulation that given a personal prior is believed to have the smallest effect size. There are similar situations in which researchers might decide to skip a pilot study and immediately perform the main experiment because this is the most efficient choice.

An applied example in which the decision is not deterministically dominant can be found in Viamonte et al. (2006), who evaluated the benefits of a computerized intervention aimed at improving speed of processing to reduce car collisions in people age 75 or older. They estimated that the risk of getting into an accident for these older drivers is 7.1%. The cost of a collision was estimated to be $22,000, or $22,000 × 0.071 = $1,562.84 per driver in the United States. Furthermore, they estimated that the intervention can prevent accidents for 86% of drivers. Therefore, the probability of a collision after intervention is now (1 – 0.86) × 0.071 = 0.00994. The total cost of completing the intervention was estimated to be $274.50. When the intervention is implemented, some drivers will still get into a collision, so the total cost of the intervention and collisions is $493.30 per driver ($274.50 + 0.00994 × $22,000).

We can implement the intervention when it does not actually work, making a Type 1 error. The waste is $274.50 per driver because this is what the intervention costs even if it offers no benefits. If the intervention works but is not implemented, we make a Type 2 error, and the amount of money that is not saved is $1,562.84 (the cost of doing nothing) – $493.30 (the cost if the intervention was implemented), for a waste of $1,069.54 per driver. This means that the relative cost of a Type 1 error compared with a Type 2 error is 274.50 / 1,069.54 = 0.257, or the waste in money after a Type 1 error is 3.896 times (1,069.54 / 274.50) worse than a Type 2 error. This ratio reflects that the intervention is relatively cheap and therefore a Type 1 error is not that costly, whereas the potential savings if collisions are prevented is relatively large. Of course, quantifying costs and benefits comes with uncertainties. The intervention might prevent more or less accidents, the risks of an accident for drivers of 75 years or older might be greater or smaller, and so on. Sensitivity analyses can be used to compute a range of the ratio of the costs of Type 1 and Type 2 errors (see Viamonte et al., 2006).

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 do not 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 on 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 cares only 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 on 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 et al., 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, 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, whereas 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. In addition, 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 in which all findings are published and people regularly perform replication studies to identify Type 1 errors in the literature (Fiedler et al., 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. D. Lakens once attended a presentation by a policy advisor who decided whether new therapies would be covered by the national health-care 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 in which 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 .0327, which will make the Type 2 error rate .252. If we would perform 20,000 studies designed with these error rates and assume H0 and H1 are equally likely to be true, we would observe .5 (the prior probability that H0 is true) + .0327 (the alpha level) × 20,000 = 327 Type 1 errors and .5 (the prior probability that H1 is true) × .252 (the Type 2 error rate) × 20,000 = 2,524 Type 2 errors. Because 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 $4 \times 327 = 1,308$, and to keep the weighted error rate between 0 and 1, we also multiply the 10,000 studies in which we expect H0 to be true by 4 such that the weighted combined error rate is $(1,308 + 2,524) / (40,000 + 10,000) = .0766$. Figure 3 visualizes the weighted combined error rate for this study design across all the possible alpha levels and illustrates the weighted error rate is smallest when the alpha level is .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 yields a 5% Type 1 error rate and a 20% Type 2 error rate. If we 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 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, 1997).
In the extreme case, in all studies, a researcher designs H1 as true. In this case, there is no reason to worry about Type 1 errors because a Type 1 error can happen only 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 H1s is very low, you are more likely to test a hypothesis in which H0 is true. Therefore, 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, assume a researcher performs 1,000 studies. The researcher expects 100 studies to test a hypothesis in which H1 is true, whereas the remaining 900 studies test a hypothesis in which 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 .9 (the prior probability that H0 is true) \times .05 (the alpha level) \times 1,000 = 45.00 Type 1 errors and .1 (the prior probability that H1 is true) \times .2 (the Type 2 error rate) \times 1,000 = 20.00 Type 2 errors, for a total of 65.00 errors.

However, the total number of errors does not tell the whole story because Type 1 errors are weighed 4 times more than Type 2 errors. We therefore need to compute the weighted combined error rates, \( w \), by 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 et al. (2012):

\[
\frac{(\text{cost}_{TT2} \times \alpha + \text{prior}_{H10} \times \beta)}{\text{prior}_{H10} + \text{cost}_{TT2}}.
\]

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 on 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, and optimally balanced error rates (maintaining the 4:1 ratio of the weight of Type 1 vs. Type 2 errors) would require setting alpha to .011 and the Type 2 error rate to .402. The researcher should now expect to observe .9 (the prior probability that H0 is true) \times .011 (the alpha level) \times 1,000 = 9.89 Type 1 errors and .1 (the prior probability that H1 is true) \times .402 (the Type 2 error rate) \times 1,000 = 40.16 Type 2 errors. The weighted error rate is .0216.

Because the prior probability of H0 and H1 influences 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 cannot 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). In the previous example, it would also have been possible to minimize (instead of balance) the error rates, which is achieved with an alpha of .00344 and a beta of 0.558, for a total of 58.86 errors, in which the weighted error rate is .0184.

The two approaches (balancing error rates or minimizing error rates) typically yield quite similar results. Whereas 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.

**Increasing the alpha level above .05**

Many empirical sciences have recently been troubled by a replication crisis (Camerer et al., 2016; Open Science Collaboration, 2015), which has in part been caused by inflated alpha levels because of p-hacking (Simmons et al., 2011), publication bias, and low statistical power (Lindsay, 2015). In light of this low replicability, a potential concern about allowing researchers to justify their alpha level is that researchers can decide to increase the alpha level above the .05 threshold. This could increase the rate of false positives published in the literature compared with when an alpha level of .05 remains the norm. An increase of the alpha level should be deemed acceptable only when authors can justify that the costs of the increase in the Type 1 error rate is sufficiently compensated by the benefit of decreased Type 2 error rate. Furthermore, researchers should explicitly accompany claims by their error rates throughout an article, especially when the alpha level is increased, and readers of claims made with higher alpha levels should understand such claims are made with greater uncertainty and could very well be false.

There are circumstances under which optimal error rates will require an increase of the alpha level, which will also increase the number of false positives in the literature. Assuming the goal of scientists is to efficiently generate reliable knowledge, the proposal to increase the alpha level (and thus to increase the Type 1 error
rate in the literature) should be adopted only if the cost of an increase in Type 1 errors is compensated in some way. So far, we have focused only on how the increase in the Type 1 error rate will lead to a greater reduction in the Type 2 error rate, which all else being equal, should improve decision-making in hypothesis tests. In practice, it might be a challenge to reach agreement on the weight of Type 1 and Type 2 errors among different stakeholders. For example, whereas a team of researchers might believe a Type 1 error and Type 2 error is equally costly, an editor of a journal might weigh Type 1 errors more than Type 2 errors. We should also consider the possibility that researchers try to opportunistically specify the relative cost of Type 1 and Type 2 error rates to increase their alpha level and increase the probability of finding a “significant” effect.

Nevertheless, in some cases, it can be justified to increase the alpha level above the .05 threshold. These will usually be cases in which (a) the study will have implications for relevant direct decision-making (as in the above EMDR example), (b) a cost-benefit analysis is provided that gives a clear rationale for relatively high costs of a Type 2 error, (c) the probability of H1 being false is relatively low, and (d) it is not feasible to reduce overall error rates by collecting more data. In these cases, it will often be desirable to justify the alpha level during the first phase of a Registered Report so that the higher alpha level that will be used in a study can be discussed transparently during peer review. At the same time, given the complexity of weighing the costs and benefits of research, it is understandable if some journals consider such discussions too great a burden for reviewers. If so, these journals could indicate that they limit deviations from an alpha level of .05 only in cases in which researchers increase the severity of their test by lowering the alpha level.

Journals might also prefer to use a default alpha level of .05 to reduce the burden on readers to examine at which alpha level claims in their journal are made. Especially if an increase in alpha levels was not evaluated by peers during the first phase of a Registered Report, the evaluation of whether this alpha level was appropriate is left to readers. In practice, the use of a higher alpha level will require readers to keep track of the fact that the claim of the presence of an effect was less severely tested than it would have been with a default alpha instead of keeping track of the fact that claims of the absence of an effect were less severely tested than they would have been when the statistical power had been higher (i.e., by increasing the alpha level). In a science in which people focus only on significant effects and treat all significant effects as equally well supported, increasing alpha levels could lead to a sense of false certainty about a body of work. If the practice to increase alpha levels becomes popular, it will be important to examine whether varying alpha levels are taken into account when interpreting and discussing research findings and how negative side effects can be mitigated.

Finally, the use of a high alpha level might be missed if readers skim an article. We believe this can be avoided by having each scientific claim accompanied by the alpha level under which it was made. Scientists should be required to report their alpha levels prominently, usually in the abstract of an article alongside a summary of the main claim. The correct interpretation of a hypothesis test was never to label an effect as “significant” or “nonsignificant” but to reject effects implied by the null model with a specific error rate. Replacing “the effect was significant” with “we reject an effect size of 0 with a 10% error rate” might end up improving the interpretation of hypothesis tests. Note that by explicitly reporting the alpha level alongside a claim, it will also become more visible when researchers lower their alpha level, and this practice will therefore clearly communicate whenever readers should be impressed by the fact that a claim passed an even more severe test than if a traditional alpha level of .05 would have been used.

**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 in which 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 et al., 1996). However, in practice, researchers will often want to justify their sample size according to an a priori power analysis in which 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, the 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 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.
the specific cases in which 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, in which 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 .05 and a desired statistical power for \( d = 0.5 \) of .95.

**Lowering the alpha level to avoid Lindley’s paradox**

Formally controlling the costs of errors can be a challenge because it requires researchers to specify the relative cost of Type 1 and Type 2 errors, prior probabilities, and the effect size of interest. Given this complexity, researchers might be tempted to fall back on the heuristic use of an alpha level of .05. Fisher (1971) referred to the default alpha level of .05 as a “convenient convention,” and one can argue it suffices as a low enough threshold to make scientific claims in a scientific system in which we have limited resources and value independent replications (Uygun-Tunç et al., 2021).

However, there is a well-known limitation of using a fixed alpha level that has led statisticians to recommend choosing an alpha level as a function of the sample size. This suggestion of a flexible decision criterion was already mentioned by the statistician Harold Jeffreys in a letter he wrote to Fisher in 1934 (Wagenmakers & Ly, 2021). Jeffreys later stated more explicitly that the critical value should increase with the sample size:

The results show that the probability that such a term is needed is increased or decreased according as the coefficient is more or less than a certain multiple of its standard error; the multiple needed, however, increases with the number of observations. (Jeffreys, 1936b, p. 345)

To understand the argument behind 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 are 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 .05. This situation, known as Lindley’s paradox, is visualized in Figure 1.

To prevent situations in which a frequentist rejects the null hypothesis on the basis of \( p < .05 \) 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 was discussed extensively by Leamer (1978). He wrote,

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. (p. 92)

The same point was already recognized by Jeffreys (1939), who discussed 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. (p. 397)

Likewise, Good (1992) noted:

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. (p. 600)

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
A BF contrasts the probability of the data under the competing hypotheses considered. When comparing H1 with H0, it is given by Equation 2:

\[
\frac{p(\text{data}|H_1)}{p(\text{data}|H_0)} \quad (2)
\]

Note that Equation 2 shows a crucial difference between \( p \) values and BFs: A \( p \) value depends only on the probability of the data or more extreme data under H0, whereas the BF takes both H0 and H1 into account.

A BF of 1 implies equal evidence for H0 and H1. Although any discretization inevitably results in loss of information, as a rule of thumb, BFs between 3 and 10 imply moderate evidence for H1, and BFs larger than 10 imply strong evidence (Jeffreys, 1939; Lee & Wagenmakers, 2013). To prevent Lindley’s paradox when using frequentist statistics, one would need to adjust the alpha level such that 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 as if H0 is true, which avoids Lindley’s paradox. Rouder et al. (2009) and Faulkenberry (2019) developed BFs for \( t \) tests and analysis of variance (ANOVA), which can calculate the BF 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) offered their own suggestions. 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 BF 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 in which it is difficult to specify a smallest effect size of interest or an expected effect size.

A second decision is the threshold of the BF used to lower the alpha level. Using a BF 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 we 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 BF that is used to lower the alpha level to prevent weak evidence (Jeffreys, 1939). This can be achieved by setting the threshold to a value larger 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 in which the aim was to test a directional prediction in a dependent \( p \) test. For any sample size and choice of prior, a \( p \) value is directly related to a BF. Figure 5 shows the relationship of two-sided \( p \) values and BFs using a Cauchy prior with an \( r \)-scale 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 .0502 for the one-sided \( t \)-test, given the chosen prior, because this choice for an alpha level guarantees that a significant \( p \) value will correspond to evidence in favor of H1.

To give a practical example of how the alpha level can be justified to prevent Lindley’s paradox, we can reexamine a study by Pennycook and Rand (2019), who investigated sharing of misinformation on social media. They reported that Clinton supporters were better able than Trump supporters, \( F(1, 798) = 28.95, p < .001 \). However, given the large number of observations, which likely provide very high power for all effect sizes that would be considered large enough to be meaningful, one could have decided to reduce the alpha level so that any observed significant \( p \) value can also be interpreted as evidence for the alternative hypothesis. If the authors had justified their alpha level as a function of their sample size as described above, they would have set the alpha level to .010. Calculating the precise \( p \) value of \( 9.77 \times 10^{-8} \) shows their result is still significant using this more stringent alpha level. Pennycook and Rand could have designed a study in which the choice of the alpha level would have prevented significant results from being evidence for the
null hypothesis. Note that by choosing an alpha level that prevents Lindley’s paradox, the study would also have more balanced error rates (Wagenmakers & Ly, 2021), thereby improving optimal decision-making. By lowering the alpha level at the expense of a relatively modest drop in statistical power, the authors would have more severely tested their hypothesis. Given the observed $p$ value, the study would have provided even more impressive support for their prediction because of the smaller Type 1 error rate.

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 .05. It is not recommended to use the procedure outlined in this section to increase the alpha level above the conventional choice of an alpha level (e.g., .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 as needed. Figure 6 shows the alpha levels for different values of $N$ for between- and within-subjects $t$ tests. We can see that particularly for within-subjects $t$ tests, the alpha level rapidly falls below 5% as the sample size increases.

**When to Minimize Alpha Levels and When to Avoid Lindley’s Paradox**

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 have a decent empirically justified estimate of prior probabilities of the null and alternative hypotheses. This is more likely for applied research, as in the case of the test of an intervention for older drivers discussed previously. When a study has direct policy implications, the costs of Type 1 error (the policy being implemented although it does not work) compared with a Type 2 error (the policy is not implemented even though it does work) can often be assessed through cost-benefit analysis. Note that the approach that tries to minimize or balance error rates will in practice also reduce the alpha level as a function of sample size and should therefore avoid Lindley’s paradox in most applied cases (although it does not guarantee to do so). 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. This might often be the more feasible approach in basic research.
In addition, the two approaches differ regarding their underlying philosophy of science. The first is based on decision theoretical developments that build on a Neyman-Pearson approach and might, therefore, be more attractive to researchers whose inferential philosophy is based on statistical decision theory. The second approach, on the other hand, offers a Bayes–non-Bayes hybrid combining frequentist and Bayesian statistics, which might be more attractive to researchers who care about both statistical schools (Good, 1992).

Discussion

Because 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 et al. (1967) remarked:

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. (p. 18)

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.

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, the sample size, and an analytic power function. The effect size should be determined as in a normal a priori power analysis (preferably according to 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 are collected. Whichever approach is used, it is strongly recommended to preregister the alpha level that researchers plan to use before the data are 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.

In this article, we presented two ways of justifying alpha levels, the first based on minimizing or balancing the relative costs of errors and the second based on avoiding Lindley’s paradox. Additional approaches to justifying the alpha level have been presented, such as Bayarri et al. (2016), who proposed to justify the alpha level according to the strength of evidence (\(1 - \beta\) / alpha). We look forward to the development of additional approaches and hope that in the future, researchers will have multiple tools in their statistical toolbox to justify alpha levels.

Throughout this article, 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., balancing vs. 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 preregistering 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 informed decisions when we test hypotheses.

Transparency

Action Editor: Alexa Tullett
Editor: Daniel J. Simons
Author Contributions
M. Maier and D. Lakens jointly wrote and edited the manuscript. D. Lakens focused more strongly on minimizing or balancing error rates, and M. Maier focused more strongly on justifying the alpha level as a function of sample size. D. Lakens and M. Maier both contributed to the R package and the Shiny app.

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

Funding
This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

Open Practices
Open Data: not applicable
Open Materials: https://github.com/Lakens/justify_alpha_in_practice
Preregistration: not applicable
All materials have been made publicly available via Github and can be accessed at https://github.com/Lakens/justify_alpha_in_practice. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

**ORCID iDs**
Maximilian Maier https://orcid.org/0000-0002-9873-6096
Daniël Lakens https://orcid.org/0000-0002-0247-239X

**Acknowledgments**
We thank Eric-Jan Wagenmakers for feedback on a previous version of the manuscript that helped improve the section on the Jeffreys-Lindley paradox. A preprint of this article is available at https://doi.org/10.31234/osf.io/ts4r6.

**Supplemental Material**
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.

**Note**
1. For the same scenario, balanced error rates are $\alpha = 0.149$ and $\beta = 0.149$.

**References**
Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 425–437. https://doi.org/10.1037/h0020412
Barlett, P. L., Jordan, M. I., & Mcauliffe, J. D. (1957). Comment on d. V. Lindley's statistical paradox. *Biometrika, 44*, 533–534.
Bayarri, M., Benjamini, Y., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology, 72*, 90–103. https://doi.org/10.1016/j.jmp.2015.12.007
Camerer, C. F., Dreber, A., Forsell, E., Ho, T-H., Huber, J., Johannesson, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 357*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918
Clemen, R. T. (1997). *Making hard decisions: An introduction to decision analysis* (2nd ed.). Duxbury.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science, 4*(12), Article 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*(1), 192–201. https://doi.org/10.1080/00031305.2018.1529622
Cornfield, J. (1969). The Bayesian outlook and its application. *Biometrics, 25*(4), 617–657.
Cousins, R. D. (2017). The Jeffreys-Lindley 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 0.05 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
DeGroot, M. (1975). *Probability and statistics*. Addison-Wesley.
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, R. F., Jr., & von Winterfeldt, D. (Eds.). (2007). *Advances in decision analysis: From foundations to applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611308
Erdfelder, E., Faul, F., & Buchner, A. (1996). GPPOWER: 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/10.1177/251524919872960
Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from $\alpha$-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 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. https://doi.org/10.1111/j.1461-0248.2004.00625.x
Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, Fisheries and Horticulture, 7*, 192–201. https://doi.org/10.1017/CBO9780511611308
Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*(2), 198–218. https://doi.org/10.1177/2515249198771329
Good, I. J. (1992). The Bayes-non-Bayes compromise: A brief review. *Journal of the American Statistical Association, 87*(419), 597–606. https://doi.org/10.2307/2290192
Harford, T. (2014). Big data: A big mistake? *Significance, 11*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2014.00778.x
Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In B. J. Green (Ed.), *Theory of probability*. In B. J. Green (Ed.), *Mathematical foundations of the Cambridge philosophical society* (Vol. 31, pp. 203–222). Cambridge University Press.
Jeffreys, H. (1936a). Further significance tests. In *Mathematical foundations of the Cambridge philosophical society* (Vol. 32, pp. 416–445). Cambridge University Press.
Jeffreys, H. (1936b). On some criticisms of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 22*(146), 337–359.

Jeffreys, H. (1939). *Theory of probability*. Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

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*(1), 82–90. https://doi.org/10.1080/00031305.2018.1537891

Kim, J. H. (2020). Decision-theoretic hypothesis testing: A primer with R package OptSig. *The American Statistician, 74*(4), 370–379. https://doi.org/10.1080/00031305.2020.1750484

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). *Sample size justification*. *PsyArXiv*. https://doi.org/10.31234/osf.io/9d3yf

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

Lin, M., Lucas, H. C., Jr., & Shmueli, G. (2013). Research commentary—too big to fail: Large samples and the p-value problem. *Information Systems Research, 24*(4), 906–917.

Lindley, D. V. (1953). Statistical inference. *Journal of the Royal Statistical Society: Series B (Methodological), 15*(1), 906–917.

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

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

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

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlanah, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLOS ONE, 7*(2), Article e23734. https://doi.org/10.1371/journal.pone.0023734

Neuman, J. (1957). “Inductive Behavior” as a basic concept of philosophy of science. *Revue de L’Institut International de Statistique / Review of the International Statistical Institute, 25*(3–4), 7–22. https://doi.org/10.2507/1401671

Neuman, J., & Pearson, E. S. (1953). 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.1953.0009

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

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pericchi, L., & Pereira, C. (2016). Adaptive significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics, 30*(1), 70–90.

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417539

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.

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science, 80*(1), 73–93. https://doi.org/10.1086/668875

Uygur-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/a99by

Viamonte, S. M., Ball, K. K., & Kilgore, M. (2006). A cost-benefit analysis of risk-reduction strategies targeted at older drivers. *Traffic Injury Prevention, 7*(4), 352–359. https://doi.org/10.1080/15389580600791362

Wagenmakers, E. J., & Ly, A. (2021). *History and nature of the Jeffreys-Lindley paradox* [Manuscript submitted for publication]. PsyArXiv. https://doi.org/10.31234/osf.io/af9by

Wagenmakers, E. J., & Ly, A. (2021). *History and nature of the Jeffreys-Lindley paradox* [Manuscript submitted for publication]. PsyArXiv. https://doi.org/10.31234/osf.io/af9by

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics, 20*(2), 165–205. https://doi.org/10.1214/aoms/1177730300

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

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

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. John Wiley & Sons.