There is a contentious debate brewing in the social sciences. A growing number of professionals have begun advocating for a more stringent p-value with which to define statistical significance in null hypothesis testing. The affirmative side of this debate proposes that the standard p-value of < 0.05 be changed to < 0.005, whereas the opposing view is entirely against this proposition.

Admittedly, the < 0.005 significance value is somewhat arbitrary, but so too is the conventional < 0.05 significance value. Nevertheless, Benjamin and colleagues (2017) propose that a trade-off must be made between so-called type I and type II errors if the significance threshold is to be changed. A type I error is considered to be a false positive, whereby the null hypothesis is rejected (when it should have been retained) and the alternate hypothesis is retained. In contrast, a type II error is considered to be a false negative, whereby the null hypothesis is retained when it should have been rejected. Benjamin et al., (2017) suggest that by changing the p-value to < 0.005, the rate of false positives would be reduced. A reduction in false positive is particularly crucial in psychological disciplines because the rate of replication in these kinds of studies has been found to be nearly double for studies using a p-value of < 0.005, as compared to studies using a p-value of only $< .05$.

The affirmative view proposes that results that meet a significance threshold of $< .05$ ought not to be outright rejected. Instead, results that are between 0.05 and 0.005 should be treated as suggestive. That is, further evidence should be collected via additional studies to assess the credibility of the findings. In contrast, the opposing side proposes that changing the p-value to < 0.005 would cause false negatives to increase to an unacceptable rate. That is, concluding an effect does not exist, when in fact it does.

Unfortunately, in order to maintain adequate power to uncover true effects, transitioning to a p-value of < 0.005 would require an increase in sample size of nearly 70%. The implication, of course, is that fewer studies could be conducted for various reasons (e.g., budgetary constraints, type of experimental design, etc.). Moreover, it is possible that, even after conducting such expensive research, the false negative rate is still present at an alarmingly high rate. A response to this criticism, of course, is that by changing the p-value to < 0.005, the scientific community as a whole will save resources by not conducting studies based on false pretenses (i.e., false positives), and an increase in sample size will increase power resulting in fewer false negatives.

In addition, the opposing view advises that even if the scientific community as a whole were to adopt a more conservative p-value, this change would not address other problems associated with hypothesis testing. These problems include: multiple hypothesis testing, p-hacking, publication bias, low power, among others, which, according to the opposing perspective, are debatably larger issues that must be addressed first. Another concern is that, over the course of various statistical tests, it is highly likely that at least one test will end up statistically significant just due to chance. Given these concerns, the opposing view concludes that changing the significance threshold is not a real solution to the multitude of problems researchers are being confronted with in various scientific fields. Instead, the scientific community ought to replace null hypothesis testing by focusing on effect sizes and confidence intervals, or by using other statistical methods (e.g., Bayesian statistics) that are arguably more resistant to problems like p-hacking.

Even so, the affirmative side concurs that reducing the significance threshold will not directly solve any of these issues. On the contrary, reducing the significance threshold

is not meant to solve these problems; it is meant to be part of a larger overarching approach to improving the credibility of the sciences. Not to mention, there is a lack of consensus in the scientific community regarding a suitable replacement for null hypothesis testing, resulting in even more contention.

Even with these differences, both sides agree that the significance threshold should be dependent on the type of research being conducted. For example, genetics research uses a significance threshold of $5x10^{-8}$, and some areas of physics use a p-value threshold that approximates $3x10^{-7}$. The main contention in this debate is that the affirmative view proposes that the default standard ought to be shifted from $p < 0.05$ to $p < 0.005$, whereas the opposing view disagrees with this shift.

According to Benjamin et al., (2017) an immediate benefit of using a lower threshold is that reproducibility of findings will be improved. These researchers contend that studies that do not reach this new threshold are not unimportant. Rather, these studies ought to be published if they address scientifically important questions and are scientifically rigorous. As mentioned prior, according to this perspective, studies that fall into this particular category would be deemed suggestive, such that additional evidence must be gathered to determine whether an effect truly exists. Unfortunately, given the prominent belief that studies with negative results ought not to be published (called the 'file-drawer problem'), it would be quite challenging to publish studies that fall in the suggestive category. In order to circumvent this problem, journals would have to be encouraged to publish studies that did not uncover statistically significant results, but still addressed important scientific questions. This, of course, is easier said than done.

Ultimately, I believe this debate can be reduced down to one simple question: Which is worse? Asserting there is an effect when there is none, or asserting there is no effect when there is? My belief is that it is far worse to assume there is an effect when there is none. We see this most clearly in clinical studies where medications – and the effects of using them – must be studied over and over again before the medication can be approved for public usage. Why should we require any less of a standard for other scientific research? Requiring a more conservative significance threshold would help limit these type I errors. While it is true that such studies would require more resources – at least at the very beginning – I believe the benefit that the scientific community would gain, in terms of increased reproducibility and scientific integrity, far outweigh the costs.

Sir Francis Bacon was one of the first to formalize the concepts that underpin the scientific method. More specifically, Bacon was grounded in the belief that we have a bad way of thinking and in order to overcome this thinking, all proper scientific thought ought to proceed through our sensory knowledge of the world. This is, in essence, the foundation of the principles behind the modern-day scientific method. Scientists ought to formulate a hypothesis based on sensory knowledge (e.g., what they see, hear, smell, etc.) and then test that hypothesis, after which this cycle continues based on the new information gained. In order to be true scientists, we ought to continue in the footsteps of Sir Francis Bacon and re-evaluate the evidence before us, before coming to valid conclusions. Benjamin et al., (2017) conclude their critique by reminding us all that theory and empirical evidence has shown that a lower significance threshold is needed. This is precisely the purpose behind the scientific method – to show scientists a way to move forward when evidence and convention are at an impasse.