Psychonomic Bulletin & Review (2019) 26:430–433
https://doi.org/10.3758/s13423-018-1523-9

THEORETICAL REVIEW

Redefine or justify? Comments on the alpha debate

Jan de Ruiter[1]

Published online: 27 September 2018

Abstract
Benjamin et al. (*Nature Human Behaviour 2,* 6-10, 2017) proposed improving the reproducibility of findings in psychological research by lowering the alpha level of our conventional null hypothesis significance tests from .05 to .005, because findings with p-values close to .05 represent insufficient empirical evidence. They argued that findings with a p-value between 0.005 and 0.05 should still be published, but not called "significant" anymore. This proposal was criticized and rejected in a response by Lakens et al. (*Nature Human Behavior 2,* 168-171, 2018), who argued that instead of lowering the traditional alpha threshold to .005, we should stop using the term "statistically significant," and require researchers to determine and justify their alpha levels before they collect data. In this contribution, I argue that the arguments presented by Lakens et al. against the proposal by Benjamin et al. are not convincing. Thus, given that it is highly unlikely that our field will abandon the NHST paradigm any time soon, lowering our alpha level to .005 is at this moment the best way to combat the replication crisis in psychology.

Keywords Significance · Reproducibility · Alpha · Evidence

A recent proposal by Benjamin et al. (2017) to modify the conventional threshold for claiming "significance" in empirical studies from .05 to .005 has led to a response arguing against this proposal by Lakens et al. (2018). For convenience, I will call the first paper RSS (Redefine Statistical Significance), and the second paper that was a response to it JYA (Justify Your Alpha).

The main point of RSS is that if (and only if) we – researchers, journal editors, and reviewers – want to keep using a threshold p-value (alpha) for calling findings "significant," then lowering that threshold from .05 to .005 will improve the replicability of our findings. It is important to note that the RSS authors do not defend our widespread use of Null Hypothesis Significance Testing (NHST) as such, but that they argue that *if* we keep using NHST, *then* lowering alpha to .005 will cause the general replicability[1] of our findings to improve.

The authors of the JYA paper criticize and reject this proposal. They argue that replacing the arbitrary .05 threshold by an equally arbitrary .005 threshold will not necessarily improve replicability, and instead suggest we abandon the notion of "statistical significance" altogether. According to them, researchers should set and justify their design decisions (which crucially includes the alpha level) before collecting data.

In this contribution I argue that many of the arguments in the JYA paper against the RSS proposal are flawed. In short, there are multiple ways in which the JYA authors argue against particularities of the RSS paper without effectively arguing against the general thrust of its recommendations.

To criticize the proposal in RSS, the JYA paper presents three main theses, discussed below.

Thesis 1: an alpha level of p ≤ .005 does not improve replicability

While the JYA authors agree that a lower alpha threshold should indeed, theoretically, reduce the number of false positives in the literature, they argue that the observed differences in the replicability of the studies in the *Reproducibility Project: Psychology* or RPP (Open Science Collaboration, 2015) are likely due to other factors.

[1] The terms reproducibility and replicability are often used interchangeably and can refer to either the same or different aspects of scientific studies. From here on, I will use the term "replicability" for the probability that a given empirical finding will, when the same experiment is repeated, yield the same finding.

✉ Jan de Ruiter
storkchen@gmail.com

[1] Departments of Computer Science and Psychology, Tufts University, 490 Boston Avenue, Medford, MA 02155, USA

They present three arguments against the claim that a new alpha threshold of .005 will improve replicability: (a) the empirical evidence we have from the RPP project is itself not statistically significant, (b) non-replicability can also be caused by other factors, such as p-hacking, and (c) in the RPP a post-facto alpha level of .005 only led to the replicability of about half of the studies. I will address these arguments in order.

First, the replication rate of the studies in the RPP project that had p-values between .005 and .05 was 24%, whereas the replication rate of the studies with $p < .005$ was 49%. The statistical evidence related to this difference indeed does not cross traditional thresholds, such as $p < .05$ in NHST, or a Bayes Factor > 10 (Etz & Vandekerckhove, 2016; Held & Ott, In Press), neither using NHST ($p = .015$) nor using Bayes Factors ($BF_{10} = 6.84$).

It is remarkable that in making this argument, the JYA authors use criteria that they themselves argue we should abandon. Furthermore, such statistical evidence would be extremely hard to obtain, as every datapoint for such an analysis would be *an entire study*. If we were to insist on this degree of empirical evidence for the claim that higher standards of evidence lead to more replicability, we would in all likelihood never abandon our current practice of using NHST with an alpha of .05, which is a situation both the RSS and the JYA authors are arguing against. But most importantly, we don't *need* empirical data to make the point that a lower alpha level leads to a higher replicability. For a given design and sample plan, a smaller p-value is associated with a larger effect size, and larger effect sizes are more replicable (see also Morey, 2017).

A second argument of the JYA authors is that "lower replication rates for p-values just below .05 are likely confounded by p-hacking (the practice of flexibly analyzing data until the p-value passes the 'significance' threshold) in the original study." It is indeed quite plausible that p-hacking of studies with p-values just below .05 was partly responsible for the low replication rate of these studies. However, p-hacking to obtain a value just below .05 is considerably easier and less conspicuous than p-hacking to get a p-value below .005 (see for instance the lucid presentation by F. Schönbrodt on this topic, https://osf.io/rkya5/). The latter not only involves more work (simply because the target p-value is further away) but we can also expect that the methodological flexibility needed to bring the p-value below such an ambitious target value is more likely to draw the attention of readers and reviewers. Furthermore, p-hacking is often something done *implicitly*, by researchers who may not be aware that their creative and iterative analyses lead to more false positives. Trying to p-hack a result that initially isn't "significant enough" to get a p-value below .005 is likely to involve more conscious effort than it is to p-hack to a value just below .05. This will make it more likely that the researcher becomes aware of the suspect nature of their analyses. Hence, if studies with $p < .05$ are more likely to be p-hacked than studies with $p < .005$, resulting in a higher reproducibility rate for the latter, this is in fact a compelling argument *in favor* of the RSS proposal of lowering our alpha levels for claiming significance.

The final argument by JYA against the claim that an alpha level of .005 lowers replicability is, citing JYA verbatim here, that "the difference in replication rates between studies with $.005 < p \leq .05$ compared with $p \leq .005$ may not be entirely due to the level of evidence." This argument is similar to the one discussed above, but it is different, and it is based on a logical fallacy that can be illustrated with an analogy. Imagine a proposal to reduce the number of traffic accidents by lowering the speed limit, because there is evidence that excessive speed causes a lot of accidents. Opponents of the proposal then argue against it on the grounds that a major factor causing traffic accidents is drinking while driving. While the latter might well be true, it isn't a valid argument. *Both* drinking while driving *and* high speeds increase the number of traffic accidents, but that is not a valid argument against lowering the speed limit. Analogously, the fact that p-hacking *also* increases the number of false positives (as do, for instance, fraud, low power, and measurement error) is not a valid argument against lowering the alpha threshold.

Thesis 2: the arguments for a new alpha of .005 are weak

The first argument the JYA authors provide to support this thesis is to criticize the new alpha level by questioning the claim by RSS that under a Bayesian analysis, a p-value of .05 at best represents a Bayes Factor of 2.5 to 3.4, whereas a p-value of .005 is equivalent to a Bayes Factor between 14 and 26 (note that these are *upper bounds*, meaning that under the assumptions made, with these p-values, evidence levels can never be higher than that). JYA counters this claim by pointing out that the RSS analysis holds for the two-sided testing of point-null hypotheses. If one-sided tests or non-point-null hypotheses were used, JYA argues, this would imply different new alpha levels (a point made by Morey, 2017).

Note first that the assumptions made here by the RSS authors are sensible and plausible because the vast majority of NHST tests in our literature in fact use point-null hypotheses and two-sided tests. One-sided tests are frowned upon by many statisticians (see Royall, 1997, p.116), and are used much less frequently than two-sided tests. Non-point-null hypotheses in NHST are even more rare in the literature.

Second, while this argument by JYA legitimately questions the *value* to which alpha should ideally be lowered, it is important to realize that this is not an argument against *lowering alpha* as such. While the exact level of a proposed lower alpha level can be the subject of interesting discussions, the fact that an alpha level of .05 represents very modest levels of evidence

(a point repeatedly made by RSS, and one that is acknowledged by the JYA authors) is a sufficient reason to lower alpha, independent of the *degree* to which it is lowered. If the JYA authors had argued "Yes, we should lower alpha, but not that much; lowering it to .01 is sufficient," or alternatively, "No, we should lower alpha even further, to .001, otherwise it doesn't help enough," then this discussion would have been very relevant. So the JYA authors appear to argue against lowering alpha to *exactly .005*, not against lowering it as such.

Another JYA-alleged weakness in the argument for an alpha of .005 is that in calculating the false-positive report probability (FPRP), we need to specify, among other things, the ratio of true versus false effects. But first of all, our empirical estimate of this value (e.g., on the basis of replication projects like the RPP) has a high uncertainty, and second, it involves a "reference class" problem: does it refer to all hypotheses in science, those of a specific field, a specific researcher, etc.? Again, for details, see Morey (2017). In this particular context, I don't find either of these arguments convincing, because RSS used their estimate of this ratio as a (necessary) *assumption* to calculate the proportion of expected false-positive results given different alpha levels, and they have reasonable arguments to assume a ball-park value for this parameter on the basis of previous analyses or the RPP project (see also Etz and Vandekerckhove, 2016; Johnson et al., 2017; Wilson and Wixted, 2018). As with the previous point, the discussion about the precise value of the ratio of true to false hypotheses is indeed relevant when discussing what level we want to lower alpha to, but not as a counter-argument against a substantial lowering of alpha.

Thesis 3: an alpha level of .005 might harm scientific practice

The JYA authors mention three concerns. The first is that a lower alpha threshold requires a higher N to achieve the same statistical power and, given limited resources, this might reduce the number of replication studies that are performed. Intriguingly, if we are talking about *exact* replications here, and not about conceptual ones (see Zwaan, Etz, Lucas, & Donnellan, 2017), it would suggest that it is better to "chop up" one's data into many small studies than it is to publish them as one large study. But this should not make a difference. Data are data, and as long as the data come from the same source, evidence is not magically growing by chopping up our data into smaller chunks. The only interpretation under which this *appears* to be true is undesirable: assuming the common situation that we have a normally distributed dependent variable and perform parametric NHST tests, if we have N experimental units (e.g., participants) at our disposal, the probability of getting at least one significant result is highest when we run N/2 studies with each study having two participants (assuming N is even), and lowest when we run only one study with N participants. This is because the Type I error rate of 5% is (by definition) unaffected by the number of experimental units in an experiment, so under the null hypothesis, the probability of getting at least one significant effect in K experiments is $1-(1-\alpha)^K$, which for K = 1 is, obviously, 0.05, and for K = 14 is already higher than 0.5 (see also Bakker, van Dijk, & Wicherts, 2012). Note that his holds independent of the value of α that the researcher is using. But of course while this strategy could increase the number of significant results published, it would also dramatically inflate the proportion of false positives, which is exactly what we are trying to avoid.

Another possible negative consequence, according to the JYA authors, is that the higher N needed to get acceptable levels of statistical power would disproportionally affect researchers who study unique populations or studies where data collection requires substantially more resources. This is of course a genuine concern. But it is hard to see how abandoning the concept of significance and justifying our alpha is going to alleviate this problem. Assuming that in those resource-intensive or unique-population cases, we would choose and justify a higher alpha (choosing a lower alpha would obviously not be helpful), it would mean that our high-resource studies would risk being less replicable. This, in turn, increases the risk that the invested effort by the researcher (and in the case of unique populations, by the experimental participants) was all in vain. Also, the RSS authors did not propose that findings with alpha = .05 should not be published, only that they should not be called "significant." So, under their (RSS's) proposal, studies that do not "reach" conventional levels of significance because of high-resource or otherwise difficult data collection would still be published, but without the label "statistically significant." What can and should in my view be justified in such cases is not the alpha level, but the low statistical power. Nature is cruel: no matter how hard it is to get the data, it will not make our evidence stronger.

A third negative consequence according to the JYA authors is the "risk of exaggerating the focus on single p-values." Whether focusing on single p-values is a necessary evil or avoidable is an interesting and important discussion, but to repeat, the RSS proposal was that *if* we have a criterion for a single p-value to call something statistically significant, *then* we need to at least lower alpha substantially, in order to increase the replicabiltiy amongst studies that we call significant. The JYA authors want to abandon the use of the label "statistically significant" altogether. But that is an entirely different discussion. Either we abandon statistical significance as a criterion, or we don't. If we do so, calculating utility functions expressing the cost / benefit of type I or type II errors in order to determine our alpha level is one option. There are others, for instance the Information-Theoretic, Likelihood, or Bayesian frameworks. But the question of whether the alpha level should be adjusted if we keep using NHST remains a

valid one. And the possibility of using an entirely different paradigm for evaluating the statistical evidence in our studies is not an argument against the RSS proposal. To use the earlier speed-limit analogy: if we were to propose a lower speed limit, pointing out that it would be much better if we all used public transport instead is a beautiful idea, but is in no way a valid argument against lowering the speed limit.

Some other issues

In their conclusion, the JYA authors mention many ways in which the replicability of scientific findings in the social sciences could be improved. These are, among other things, preregistration, intellectual honesty, redundancy (e.g., by replication and validation), avoidance of logical traps, transparency, and accounting for potential sources of error. I wholeheartedly support these arguments. I'm confident that the systematic application of these principles would substantially improve the replicability of our findings, and I support the authors in their striving towards these urgent improvements in our practices.

However, the claim that we should not have any form of threshold criterion for required levels of evidence for making scientific claims, but rather only control our long-term error rates, and leave it "up to scientists to justify the alpha level they decide to use" (JYA, p. 170), is one that I find difficult to support. I believe that it is practical and desirable to have a method for evaluating the amount of evidence represented by a single study. It allows us to evaluate what we know and what we don't know after we have performed a new experiment. If we do research, we want to be able to publish what we have learned from it. Our published studies would not be very useful if they ended only with the bland reassurance that if we were to repeat this particular experiment thousands of times, the proportion of type I errors we would make is equal to, at most, alpha. (Whether such an alpha level was individually justified or conventional in nature would almost be beside the point.)

Summary and conclusions

The RSS proposal (Benjamin et al., 2017) is that *if* we keep using a conventional threshold for statistical significance in the NHST context, *then* we should at least lower our alpha level from .05 to .005. This would, for theoretical reasons, decrease the number of our false-positive findings. The JYA authors (Lakens et al., 2018) present arguments against a lowering of the conventional alpha. They also argue for abandoning the concept of statistical significance altogether, and say that we should instead rely on controlling error rates with alpha levels that are justified by the researcher for every individual study. I analyze a number of JYA's central counterarguments and conclude that they are flawed. Given that there are no signs that the practice of using the Null Hypothesis Significance Testing framework is going to be abandoned any time soon, lowering our conventional alpha levels is at the moment the most realistic proposal for addressing our replication crisis using statistical means.

Acknowledgements The author wishes to thank Alexander Etz, Jason Noble, and Eric-Jan Wagenmakers for their helpful comments on earlier versions of this paper.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour, 1*.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PloS one, 11*(2), e0149794.

Held, L., & Ott, M. (In Press). On p-values and Bayes factors. *Annual Review of Statistics and Its Application*.

Johnson, V. E., Payne, R.D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association 112(517)*, 1-10.

Lakens, D., Adolfi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., ... Bradford, D. (2018). Justify your alpha. *Nature Human Behavior, 2,* 168-171.

Morey, R. (2017). Redefining statistical significance: the statistical arguments [blog post]. Retrieved from https://medium.com/@richarddmorey/redefining-statistical-significance-the-statistical-arguments-ae9007bc1f91

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

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. New York: Routledge.

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science,* 2515245918767122.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences*, 1-50.