Some comments on the “Justify Your Alpha” response to
“Redefine Statistical Significance”

J.P. de Ruiter
Tufts University

Author note:

J.P. de Ruiter, Depts. of Computer Science and Psychology, Tufts University

Correspondence concerning this article should be sent to J.P. de Ruiter, Dept.
of Psychology, Tufts University, 490 Boston Avenue, Medford, MA 02155

Acknowledgement: The author wishes to thank Alexander Etz and Eric-Jan
Wagenmakers for their helpful comments on earlier versions of this paper.

Contact: jp.deruiter@tufts.edu, Tel. 1-617-627-2531

Abstract

Benjamin et al. (2017) proposed to improve 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 evidence. This proposal was criticized and rejected by a commentary by Lakens et al. (2017), who argued that a) the empirical evidence for the effectiveness of such a policy is weak, b) the theoretical arguments for the effectiveness of such a policy are weak, and c) the proposal also has negative consequences for reproducibility. In this contribution, I argue that the arguments by Lakens et al. are either unconvincing, or in fact arguments *in favor* of the proposal by Benjamin et al.

Keywords: Significance, reproducibility, alpha, evidence.

SOME COMMENTS ON THE “JUSTIFY YOUR ALPHA” RESPONSE TO
“REDEFINE STATISTICAL SIGNIFICANCE”

A recent proposal by 77 authors (Benjamin et al., 2017) to increase the conventional threshold for claiming “significance” in empirical studies from .05 or .005 has led to a response by 88 authors arguing against this proposal (Lakens et al., 2017). For convenience, I will call the first paper the RSS (Redefine Statistical Significance) paper, and the second paper that was a response to it the JYA (Justify Your Alpha) paper. In this contribution I aim to demonstrate that many of the arguments in the JYA paper against the RSS proposal are either not convincing, or, in fact, argument *in favor* of the lowering alpha as proposed by RSS.

To support their point, the JYA paper presents three theses, each supported by several arguments.

Thesis A: The evidence that lowering alpha would increase reproducibility is weak.

Their first thesis is that there is “*insufficient evidence that the current standard is in fact a leading cause of non-reproducibility*”, as the RSS paper argues. While the JYA authors agree that a lower alpha level improves reproducibility *in theory* they don’t believe that it will necessarily lead to more reproducibility *in practice*.

To support this thesis the authors of the JYA paper present three arguments. The first is that the RSS paper reported that the reproducibility rate among studies with a .005 threshold was twice as high as those with a .05 threshold, but they (perhaps teasingly) observed that this difference is itself not significant (neither at the .005 level that RSS proposed, nor at the conventional .05 level). The suggestion here is that if the RSS authors want to argue for an alpha level of .005, they better have .005 level evidence that lowering alpha to .005 results in improvement of replicability.

However, the JYA paper’s main point is that alpha levels should not be fixed, but justified, and one of the possible justifications of a higher-than-conventional alpha level is when data collection is very expensive and difficult. Now I can hardly imagine a more expensive data point than an *entire experiment*. So by their (the JYA authors’) own logic, the $p = .015$ from this comparison should be met with some leniency here, and not necessarily be dismissed as insufficient evidence.

More importantly, the argument that there is insufficient evidence for the effectiveness of higher standards of evidence can easily lead to infinite regress. For instance, even if we find a significant difference (according to either alpha level under discussion) between the replicability of studies with an alpha of .05 and .005, we can start asking whether this meta-study is itself reproducible, etc., etc. In other words, it would require unrealistic amounts of work to get conclusive empirical evidence that higher standards of evidence are effective in increasing reproducibility. Therefore, in discussing methodological improvements, we should

prioritize theoretical arguments over empirical ones, just as the authors of the RSS paper did. Otherwise any proposed change of our scientific practices, no matter how urgently needed, could always be resisted on the basis that there is insufficient evidence that it would lead to improvement. This, in turn, would lead to the situation that we never change our practices, which would only be beneficial if we were absolutely certain that our current practices are perfect, which seems rather implausible given, among other things, the replication crisis we are having in the field.

A second argument is that “the lower replication rate for p-values just below .05 is likely confounded by p-hacking (the practice of flexibly analyzing data until the p-value passes the ‘significance’ threshold) in the original study” [JYA, p. 8]. This is indeed quite plausible, but p-hacking to obtain a value below .05 is considerably easier and less conspicuous than p-hacking to get below .005 (see for instance the lucid presentation by F. Schönbrodt on this topic, <https://osf.io/rkya5/>). The latter involves not only more work (simply because the target value is further away) but it is also to be expected that the statistical meanderings necessary to get the p-value below that ambitious a target value are more likely to draw the attention of reviewers. Also, much p-hacking is arguably done *implicitly*, by researchers who may not be aware of the fact that their creative and iterative analyses lead to more false positives. Trying to p-hack a result that isn’t “significant enough” to become $p < .005$ is likely to involve more conscious effort than to $p < .05$, making it more likely that the researcher becomes aware of the suspicious 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 requiring $p < .005$ for claiming significance.

Their third argument is also, in fact, an argument in favor of lowering alpha. According to the JYA authors, the reason that studies with p-values below .005 have been reproduced more often than those with $0.05 < p < .005$ is due to the number of studies with an *even lower* p-value of $p < .0001$. So, in the words of the authors, this suggests that “a relatively small number of studies with p-values much lower than .005 were largely responsible for the 49% replication rate for studies with $p \leq .005$ ”. The pattern here, not entirely surprising, is that the lower the p-value, the higher the percentage of reproduced studies. How this is an argument *against* lowering the threshold for significance is not satisfactorily explained by the authors of the JYA paper.

Thesis B: The arguments for the proposed new value of .005 are weak

The next thesis by the JYA authors is that “even though p-values close to .05 never provide strong ‘evidence’ against the null hypothesis on their own, the argument that p-values provide weak evidence based on Bayes factors has been called into question.” [single quotes in original]. The JYA authors give the following argument for this: “Redefining the alpha level as a function of the strength of relative evidence

measured by the Bayes factor is undesirable, given that the marginal likelihood is very sensitive to different (somewhat arbitrary) choices for the models that are compared.”

First of all, the authors of the RSS paper provided an *upper bound* to the level of evidence for the p-value, under reasonable assumptions, so changing these ‘somewhat arbitrary choices’ are only going to make these evidence levels even lower. In fact, the JYA authors agree that “p-values close to .05 *never* provide strong ‘evidence’ against the null hypothesis on their own, ...” [my emphasis]. If this is not controversial, that is already a sufficient reason to lower the .05 threshold. I have no position worth defending on the precise optimal value of the new lower alpha value, and as noted above, .005 may well turn out to be too lenient. But it is important to note that the existence of disagreement about the optimal lower value of the improved threshold is not by itself a valid reason not to lower the threshold, given the apparent mutual agreement that a p-value just below .05 *never* provides strong evidence for the null. (If the JYA authors had argued for an alpha of 0.001 or 0.01 instead of 0.05, these considerations would of course have been highly relevant). So, the JYA authors do not address the key claim of the RSS paper, namely that it is *impossible* to conduct a plausible analysis for a p-just-below-.05 result and obtain a compelling amount of evidence.

Thesis C: The RSS proposal has negative consequences for reproducibility

The JYA authors argue that lowering the significance threshold to .005 might “harm scientific practice”. Their first concern is that a lower alpha threshold requires more data and that in turn might reduce the number of replication studies. This suggests that it would be better to perform two studies of size N, than to perform one large study with size 2N. From either a Bayesian or a Frequentist meta-analytic perspective this should not make a difference. Data is data, and evidence is not magically growing by chopping up our experimental data into smaller chunks. The only interpretation under which this appears to be true is undesirable: in Null Hypothesis Significance Testing (NHST), if I have N experimental units (e.g., participants) at my disposal, the probability of getting at least one significant result is highest when I run N/2 studies with each 2 participants (assuming N is even), and lowest when I run only one study with N participants. This is because the Type I error rate of 5% is unaffected by the number of participants in an experiment, so under the null hypothesis, the probability of getting at least one significant effect is $1-(1-\alpha)^N$ which for 1 experiment is, obviously, 0.05, and for 14 experiments is already higher than $\frac{1}{2}$ (see also Bakker, van Dijk, & Wicherts, 2012). But of course such a strategy would 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 need for more statistical power, and therefore higher N, would increase our reliance on “cheap” data, for instance by relying (even) more on undergraduate populations

or Mechanical Turk workers. This could disproportionately affect researchers who study unique populations or studies where data collection requires substantially more resources. This is all true, but lowering alpha or keeping it as high as it is now will not alleviate this, because if the researcher faced with resource-intensive data collection did not require these extra resources (in time or money), this would inevitably lead to a lower reproducibility of their study. This, in turn, increases the risk that the invested effort by the researcher (and in some cases by the experimental participants) was all in vain. Nature is cruel: no matter how hard it is to get the data, it will not make the evidence stronger. Note that the RSS authors did not propose that findings with $.005 < \alpha < .05$ should not be published, but only that they do not deserve the label “significant”. To label findings with p values just below .05 with data from Mechanical Turk “suggestive” and studies with the same amount of evidence, but obtained at a much higher cost, “significant” is not desirable, because as mentioned above, the amount of evidence does not grow with our effort to obtain the data, and the same amount of evidence should be described with the same words, independent of how easy or difficult it was to obtain the data.

A third negative consequence according to the JYA authors is “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 if I understand the essence of the RSS proposal correctly, it was that *if* we have a criterion for a single p-value, *then* we need to at least lower alpha substantially. The proposal to abandon essentially arbitrary thresholds is worth discussing (see e.g. Morey, 2015) but it is not, as such, an argument against lowering the existing alpha level used in our current, deeply entrenched, practice of focusing on single p-values to describe and evaluate evidence.

Summary and Conclusion

To summarize, while the authors of the JYA paper have provided ample material for interesting discussions on the nature of statistical evidence and scientific practices, their arguments against the RSS proposal to lower alpha substantially are not convincing. In some cases they have even provided good arguments *in favor* of the RSS proposal (or an even stronger proposal). So the JYA paper does not produce convincing arguments against lowering our default alpha level to .005 as an urgently needed policy to combat our reproducibility crisis.

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.
- Lakens, D., Adolphi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., . . . Zwaan, R. (2017). Justify your alpha: A response to “Redefine statistical significance”. Retrieved from *psyarxiv.com/9s3y6*.
- Morey, R. (2015). On verbal categories for the interpretation of bayes factors. *BayesFactor*. Retrieved from <https://bayesfactor.blogspot.nl/2015/01/on-verbal-categories-for-interpretation.html>