Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-019-00440-1

The Alpha War

Edouard Machery[1]

Published online: 01 June 2019

Abstract
Benjamin et al. *Nature Human Behavior* 2 (1), 6–10 (2018) proposed decreasing the significance level by an order of magnitude to improve the replicability of psychology. This modest, practical proposal has been widely criticized, and its prospects remain unclear. This article defends this proposal against these criticisms and highlights its virtues.

Psychology, epidemiology, and more than a few other sciences are in a crisis: Many published results, including classic, textbook ones, appear not to replicate, and whole empirical literatures are now under suspicion. The reputation of psychology, and of a few other sciences, has dimmed considerably, and much soul searching is going on in various quarters of science.

We should rejoice! The on-going upheaval is a unique occasion to reform dubious scientific practices and improve science. Indeed, psychology, the poster child of the current crisis in science, already shows signs of fragile, timid, but encouraging reforms. Despite undeniable conservatism, proposals to improve psychology are put forward, discussed in journals and on blogs, and sometimes implemented by audacious editors.

The present article discusses one of these proposals: Cut the conventional significance level by an order of magnitude (Benjamin et al. 2018; see also Benjamin et al. 2017). This modest, practical proposal has been widely criticized, and its prospects remain unclear. This article defends this proposal against these criticisms and highlights its virtues.

In Section 1, I review some aspects of the current crisis in science, focusing mostly on psychology. In Section 2, I present the $p < .005$ proposal and the rationale behind it. Section 3 briefly dismisses a few misguided objections. In Sections 4–6, I discuss the criticisms raised against the efficacy of this proposal. Section 4 focuses on the objections against the claim that a significance level set at .05 is too lenient. Section 5 examines the objections against the claim that decreasing the significance

✉ Edouard Machery
Machery@pitt.edu

[1] University of Pittsburgh, Pittsburgh, PA, USA

level would improve the replicability of psychology. Section 6 examines further challenges against this claim. Section 7 turns to the actionability of our proposal.

1 The Replication Crisis

We should have seen it coming! It was right in front of our eyes. "It"? The replication crisis—the failed replications of many published results. Why? The power of psychological experiments—the probability of rejecting the null hypothesis if it is false—has remained unacceptably low for decades (e.g., Cohen 1962; Sedlmeier and Gigerenzer 1989; Fraley and Vazire 2014). For an effect size equal to $r = .2$ (a typical effect size in psychology), the average power of psychological experiments in many journals in social and personality psychology is lower than .5 (Fraley and Vazire 2014). That is, if the null hypothesis happens to be false, a scientist eager to find the truth would be better off throwing a coin rather than running an experiment. Speak of a waste of resources! But if the power of experiments is low, the proportion of false positives among significant results can be substantial (Ioannidis 2005; Button et al. 2013; Colquhoun 2014; Fraley and Vazire 2014). This proportion, which we will call the "false discovery rate" (Colquhoun 2014), is a function of the prior probability that the null hypothesis is true (the number of true null hypotheses under test), represented here by φ, the significance level (α), and the power (1-β):

$$\frac{\alpha\varphi}{\alpha\varphi + (1-\beta)(1-\varphi)} \quad (1)$$

A prior probability of 0 means that psychologists always test true alternative hypotheses (equivalently that all the null hypotheses under test are false), a prior probability of 1 that they only test true null hypotheses. For many prior probabilities of the null hypothesis, a large proportion of significant results are false positives. Now, add to this conclusion the fact that negative results, which would contradict those false positives, are not published.[1] The outcome is a literature with a very high number of false positives that are not undermined by true negatives. When replications happen and are published, or at least made public, then many of them should fail. A replication crisis ensues.

Indeed, numerous attempts at replicating results published in psychology, epidemiology, and other sciences have failed. Most famously, Nosek and colleagues (Open Science Collaboration 2015) attempted to replicate 100 psychological experiments drawn from a leading journal in cognitive psychology—*Journal of Experimental Psychology: Language, Memory, and Cognition*—and two top journals in social psychology—*Journal of Personality and Social Psychology* and *Psychological Science*.[2] There is no perfect way of assessing whether a replication attempt succeeds, but various measures converged to suggest a low rate of replication: Only 36.1% successfully replicated, as measured by the number of significant results in the replications reported in Open Science Collaboration (2015),[3] and 41% as measured by the

[1] Supposing that these true negatives are ever observed, an unlikely outcome if people heavily engage in questionable research practices that ensure reaching the significance level (Simmons et al. 2011)

[2] *Psychological Science* publishes articles in various areas of psychology.

[3] That is, the proportion of replicated studies whose *p*-values is below .05.

proportion of 95% confidence intervals of the effect sizes in the replications that included the original effect sizes (for a useful discussion of this result, see Etz and Vandekerckhove 2016).[4] While it is not possible to estimate the rate of false positives in psychology from the Open Science Collaboration's findings, at the very least we can confidently say that a surprisingly large number of results are unlikely to replicate and that some of them are likely to be false positives. Sadly, this conclusion is not limited to psychology. The same pessimism applies to epidemiology, including oncology (Begley and Ellis 2012; Baker and Dolgin 2017), ecology (Lemoine et al. 2016), and economics (Chang and Li 2015).

While low power and publication bias are two important sources of the replication crisis, they are unfortunately not the only ones. P-hacking—the reliance on questionable research practices to increase the probability of obtaining a *p*-value below the significance level (Simmons et al. 2011)—is probably common in psychology and other sciences and results in a real type-I error probability that is substantially higher than the nominal significance level.

2 The $P < .005$ Proposal

Psychologists have been at the forefront of the reform of scientific practices in response to the replication crisis. Many reform proposals have been put forward (from banning null hypothesis significance testing, to replacing classical statistics with Bayesian statistics, to preregistration, to pre-data collection peer review, to changing the incentives at work in science, to increase scientific transparency), and, while some of them are of dubious value, others are extremely promising. Two main features make them promising. First, they are *efficacious*: That is, they are likely to reduce the proportion of false positives in the scientific literature; they are a useful means to address the replication crisis. Second, reform proposals must be *actionable*: They are the kind of reforms that can be put in practice. Of course, efficacy and actionability are graded dimensions; thus, the more efficacious and actionable, the better a proposal. Preregistration meets these two criteria. It considerably limits p-hacking, and while taking preregistration seriously requires some real changes in the planning of experiments and data analyses and in the assessment of science (e.g., reviewers must consult preregistrations and check manuscripts against them), it does not require a complete reorganization of the scientific process.

Benjamin et al. (2018) have put forward one of the most discussed reform proposals in recent years.[5] It is striking in its simplicity: reduce the significance level by one order of magnitude. That is, it proposes to designate empirical results as statistically significant only if their *p*-value is below .005 instead of .05 as is customary. As we put it (2018, 6):

> [A] leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with P < 0.05 results in a high rate of false positives even in the absence of other

[4] See also Trafimow (2018) for discussion of how to measure replicability.
[5] We do not claim originality for many of the points put forward in our paper.

> experimental, procedural and reporting problems. For fields where the threshold for defining statistical significance for new discoveries is P < 0.05, we propose a change to P < 0.005. This simple step would immediately improve the reproducibility of scientific research in many fields.

Benjamin et al. (2018) proceed in two steps: First, we provide an argument against using .05 as the conventional significance level. Second, we provide two arguments for using .005 as the new conventional significance level. Let's start with the argument against using .05 as the conventional significance level. The gist of the argument is straightforward: a *p*-value = .05 provides too little evidence against the null hypothesis. To develop this argument, we go through the following two steps: We identify how much evidence a *p*-value = .05 provides; we then show that this amount is insufficient evidence. Let's look at the first step first. Assuming that Bayes factors provide a measure of evidence, Benjamin and colleagues' task is to map a .05 *p*-value onto a Bayes factor. As we acknowledge, however, there is not a one-to-one mapping between *p*-values and Bayes factors: Because a *p*-value only depends on the null hypothesis, while the Bayes factor depends on the null hypothesis and on the alternative hypothesis, a given *p*-value can be mapped onto distinct Bayes factors depending on the choice of the alternative hypothesis. We examine four mappings, all of which pits a hypothesis centered at 0 with a hypothesis that is symmetric around 0 (Fig. 1 of Benjamin et al. 2018). For present purposes, one of these mappings is of particular interest (the black line in Fig. 1 of Benjamin et al. 2018 reproduced in Fig. 1).

The null hypothesis is a point hypothesis: $H_0 : X_1 \sim N(0, 1)$. H_0 states that the observed data point is drawn from a normal distribution centered at 0 with a variance equal to 1. In a two-sided z-test, a *p*-value of .05 would correspond to a value of X_1 equal to 1.96 or − 1.96. The alternative hypothesis is the following: $H_1 : X_1 \sim N(1.96,1)$ with probability .5 and $X_1 \sim N(-1.96,1)$ with probability .5. H_1 is so chosen because the alternative hypothesis with the highest likelihood is the one that predicts the observed

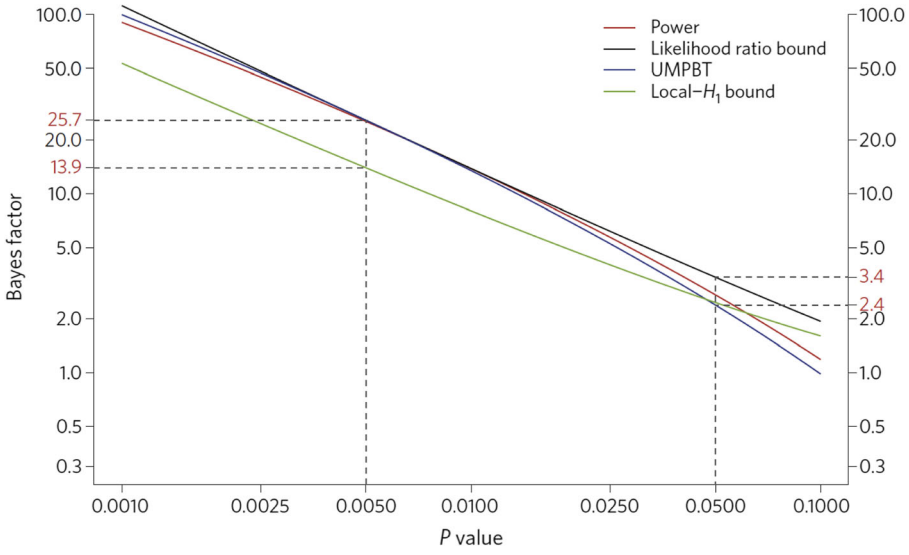

Fig. 1 The 4 mappings between *p*-values and Bayes factors (Fig. 1 of Benjamin et al. 2018)

data point, namely $X_1 = 1.96$. This hypothesis is symmetric around zero in order to set up a Bayesian counterpart of a classical two-sided test.[6] Since the likelihood function for a normal distribution $N(\theta, \sigma^2)$ is $L(\theta, \sigma^2; X_1 = y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)}$, the Bayes factor BF_{10} is equal to $\left(.5 \times e^{\left(-\frac{1}{2}(1.96-1.96)^2\right)} + .5 \times e^{\left(-\frac{1}{2}(1.96+1.96)^2\right)}\right) / e^{\left(-\frac{1}{2}(1.96-0)^2\right)}$, which is about 3.4. Thus, given our Bayesian representation of a two-sided test, a .05 p-value corresponds at most to a Bayes factor equal to 3.4: Other alternative hypotheses that are symmetric around 0 would result in lower Bayes factors.[7]

The second step of the argument is to show that a Bayes factor equal to 3.4 provides too little evidence to reject the null. We do this in three different ways: We appeal to the conventional classifications of the strength of evidence as measured by the Bayes factor, and we note that according to these a Bayes factor around 3.00 amounts to negligible evidence; further, we note that consumers of null hypothesis testing probably take reaching the significance level to provide more evidence against the null; third, and crucially, we note that assuming prior odds of 1 to 10, a Bayes factor equal to 3.4 results in posterior odds of 3:1 in favor of the null hypothesis. Let's unpack the third argument a bit, which is the most compelling. Let's suppose a scientist does bold and innovative research, as is likely given current scientific incentives. Her hypothesis is then improbable: Let's say that if the scientist were to wager a dollar, she would win ten dollars if her hypothesis were true. She rejects the null hypothesis and accepts her own hypothesis on the basis of an associated probability equal to .05. Despite having rejected the null, she should nonetheless bet against her own hypothesis except for bets when she would at least win three dollars for every wagered dollar if her hypothesis were true.

The conclusion follows: The maximum amount of evidence a .05 p-value can correspond to still isn't sufficient evidence against the null hypothesis.

The first argument for setting the conventional significance level at .005 follows a similar strategy. One may think we should identify the lower bound of the mapping between a p-value and a Bayes factor—the lowest amount of evidence a .005 p-value could provide—but this is not a successful strategy: The Bayes factor converges to 0 as the alternative hypothesis states a mean value increasingly away from the observed sample means. Rather, we examine four different alternative hypotheses to characterize several possible Bayes factors (see legend of Fig. 1 and Supplementary materials of Benjamin et al. 2018). Given these alternative hypotheses, a .005 p-value corresponds to a Bayes factor for H_1 between 14 and 26.

The second argument follows a different line. It focuses on the proportion of false positives among significant results (which is what we've called the "false discovery rate" above). As we have seen earlier, the false discovery rate depends on the prior probability that the null hypothesis is true, the significance level, and the power of the test (Fig. 2 of Benjamin et al. 2018). For many priors and power levels, a significance level set at .05 results in a large proportion of false positives among significant results (Ioannidis 2005), and a significance level set at .005 reduces this proportion considerably (Benjamin et al. 2018, 2): "[T]he false positive rate [i.e., the false discovery rate] is

[6] The Bayes Factor argument does not rely on the ascription of probabilities to hypotheses about parameter values: No probability is assigned to the null model or to the alternative model.

[7] This is only the case when the standard deviation is the same for the null and alternative models. Thanks to Justin fisher for noting this point.

greater than 33% with prior odds of 1:10 and a *p*-value threshold of 0.05, regardless of the level of statistical power. Reducing the threshold to 0.005 would reduce this minimum false positive rate to 5%." To support the plausibility of this estimate, we note that the replication rate reported in Open Science Collaboration (2015) was twice as large (50% vs. 24%) for *p*-values below .005 compared to *p*-values between .005 and .05. Retrospectively, this comparison is not compelling since p-hacking likely explained much of this difference: Psychologists p-hack up to the significance level, but not lower (as Lakens and colleagues (2018) correctly note).[8]

Importantly, we do not propose to use a .005 significance level as a criterion for publication: "[T]his proposal is about standards of evidence, not standards for policy action nor standards for publication. Results that do not reach the threshold for statistical significance (whatever it is) can still be important and merit publication in leading journals if they address important research questions with rigorous methods. This proposal should not be used to reject publications of novel findings with $0.005 < P < 0.05$ properly labelled as suggestive evidence. We should reward quality and transparency." (Benjamin et al. 2018, 8). Rather, we propose that $p < .05$ be labeled as "suggestive," while those lower than .005 be labeled as significant. We agree that *any* good scientific paper, irrespective of its *p*-value, should be published, and we do not endorse using the significance level as a publication filter. Finally, we limit our proposal to "claims of discovery of new effects": Replications could do with a .05 significance level.[9] In my view, our proposal really amounts to a call for caution for *p*-values between .005 and .05: Great caution should be exercised when a claim is only supported by a *p*-value in this interval. Finally, it should go without saying (but in light of some of the responses to our work it is worth saying) that we do not take empirical discoveries to be established once and for all on the basis of a single *p*-value below .005. Rather, the significance level functions as "a norm of assertion": It identifies when one can legitimately claim to have discovered something. Such discovery claims are naturally defeasible and must be supported by further studies.

Our proposal is an efficacious and actionable scientific reform. Consider efficacy first. We do not claim that reducing the significance level will solve all the problems that plague psychology and other experimental sciences: Our proposal "complements — but does not substitute for — solutions to these other problems [p-hacking, publication bias, etc.], which include good study design, ex ante power calculations, preregistration of planned analyses, replications, and transparent reporting of procedures and all statistical analyses conducted" (Benjamin et al. 2018, 8). But we do claim that the replication crisis would be mitigated by reducing the significance level by an order of magnitude: "This simple step would immediately improve the reproducibility of scientific research in many fields" (Benjamin et al. 2018, 8). The analysis of the false discovery rate and of the Open Science Collaboration's (2015) data is meant to support this claim. (On the second point, see however the remark above.)

Furthermore, the proposal is actionable: "changing the *p*-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve

[8] Morey (2018) also criticizes this empirical argument, but he misunderstands its point. The argument does not aim to show that there are fewer failed replications for original p's ≤ .005 than for .005 < original p's ≤ .05, but rather to give a sense of how much replicability could increase following a reduction of the significance level.
[9] This restriction applies only to direct replications and not to conceptual replications (for a criticism of this distinction, see however Machery n.d.).

broad acceptance" (Benjamin et al. 2018, 6). It does not require scientists to learn new statistics, and it boils down to a different classification scheme of empirical results.

Decreasing the significance level by an order of magnitude would have a substantial impact on psychology. Of the 100 studies examined in OSC 2015, 7 reported a *p*-value near .05 and more than a third a *p*-value larger than .01. None of these studies would have been in a position to report a discovery if the significance level had been set at .005.

3 Red Herrings

The remainder of this article examines the objections raised by the critics of Benjamin et al. (2018). Many of them miss the mark: Some seem to misunderstand the point of our article; others criticize us for asserting things that we painstakingly deny. Let's review briefly these flawed criticisms.

A few responses to Benjamin et al. (2018) repeat well-known criticisms of null hypothesis significance testing (e.g., Trafimow et al. 2018; McShane et al. 2018). Trafimow et al. (2018, 2) state that "Benjamin et al. (2018) did argue that using a 0.005 cutoff would fix much of what is wrong with significance testing," but our goal was not to "fix much of what is wrong with significance testing"; in fact, we explicitly denied it (2018, 8): "Many of us agree that there are better approaches to statistical analyses than null hypothesis significance testing." Our goal was to put forward an actionable proposal that would improve the replicability of psychology given its current statistical practices.

Trafimow and colleagues also seem to believe that nothing short of a complete reorganization of experimental, statistical, and publication practices can address the replication crisis. They write that "If significance testing—at any *p*-value threshold—is as badly flawed as we will maintain it is (…), these reasons [i.e., the actionability of our proposal] are clearly insufficient to justify merely changing the cutoff" (2018, 2). I will not address here their, in my mind, exaggerated criticisms of null hypothesis significance testing (for a useful discussion, see García-Pérez 2017), and I will merely assert that critics of classical statistics often impute to null hypothesis significance testing vices that would arise for any way of doing inferential statistics.[10] But it is important to note that radicalism may in practice be the other face of conservatism: If drastic changes in experimental, statistical, and publication practices are to happen at all, they will rather happen slowly, very slowly. Meanwhile, actionable steps have to be taken to improve day-to-day scientific practices.

Some critics also impute to us views we went out of our way to deny. Mayo (2017a) writes that our paper "ignores or downplays what almost everyone knows is the real cause of non-reproducibility: cherry-picking, p-hacking, hunting for significance, selective reporting, multiple testing and other biasing selection effects." In the second line of our paper, we acknowledge these issues and the need to address them, and we come back to such need several times in the remainder of the article. Perhaps, Mayo disagrees with our claim that setting the significance level at .05 is "a *leading* cause of non-

[10] It is for example surprising that critics of null hypothesis significance testing fail to see that even in the absence of a cutoff scientists would engage in practices that exaggerate how much evidence they have for their pet hypotheses.

reproducibility" (my emphasis), but, first, we didn't say that it is the only cause, not even *the* leading cause; second the false discovery rate argument *does* suggest that a .05 significance level is an important cause of non-reproducibility. True, Mayo criticizes this argument (as do Lakens et al. (2018)), but her criticisms can be resisted (see below).

McShane et al. (2018, 14) write that "the *p*-value is relevant to the question of how easily a result could be explained by a particular null model, but there is no reason this should be the crucial factor in publication" (see also Lakens et al. 2018, 2; Trafimow et al. 2018). As noted above, we explicitly rejected using the significance level as a publication filter. McShane and colleagues' remark is also a bit odd: Admittedly, obtaining a significant result is almost a necessary condition for publication (Fanelli 2010), but for all that it just isn't the case that psychologists take the *p*-value to be "the crucial factor in publication." Most articles submitted to psychology journals report values below the current conventional significance level, but only a few are selected for publication, suggesting that factors other than statistical significance—the interest of the theoretical hypotheses, the nature of the experimental design, the control of potential confounds, etc.—are taken into consideration during the editorial process (a point also made by Gelman 2017b, which provides a friendlier assessment of Benjamin et al. 2018).

In a similar spirit to McShane and colleagues, Amrhein and Greenland (2018, 4) write that "reliable scientific conclusions require information to be combined from multiple studies and lines of evidence. To allow valid inference from literature syntheses, results must be published regardless of statistical significance, with the *p*-value presented as a continuous summary" (see also Argamon 2017; Amrhein et al. 2018; Lakens et al. 2018; Trafimow et al. 2018). It is unclear why Amrhein and Greenland think we (or anyone, really) would disagree with the need to combine information from several studies, and unsurprisingly they fail to quote any passage of our article suggesting we would. Also, Amrhein and Greenland must not browse the *Journal of Personality and Social Psychology*, say, very often: Despite a resurgence of the one-study paper in recent years, for decades the multiple-studies paper format has dominated scientific publishing in psychology (for debate about this format, see, e.g., Greenwald 1976 and Wegner 1992). Systematic literature reviews as well as metaanalyses also are a central part of contemporary psychology as of other fields (e.g., epidemiology; see Ioannidis 2016): Hundreds of metaanalyses are written every year in psychology (Guilera et al. 2013), and they are often among the most influential articles, as measured by their citation rates.

Morey (2017) seems to object to the idea that there should be "statistical standards of evidence for *claiming new discoveries*" (my emphasis): As he puts it, "Discovery of a new effect is a matter for a research programme, not a single experiment. There is no statistical criterion that can establish a discovery." Like various discussions of this point, Morey's criticism suffers from imprecision about what "establishing" amounts to. In a sense, his point is trivially true, and no one denies it: Establishing *beyond any reasonable doubt* that a phenomenon is real requires many experiments of different kinds. And that is a matter of "a research programme." But this is compatible with having statistical criteria (i.e., *necessary*, but not sufficient, conditions) for asserting, "making a claim," that a phenomenon has been, *defeasibly*, discovered (see further discussion below).

The claim that the significance level should be understood as a norm of assertion for the discovery of an effect raises the question of what an effect is. The issue is particularly challenging given that null hypotheses in psychology are probably typically false (few manipulations have truly no effect—Meehl 1990): Do all these cases count as effects to be discovered? While addressing this question would require a full-length article, the following points may be useful. Because effects are not defined at the level of statistical hypotheses, but at the level of substantial hypotheses, not all false null hypotheses will correspond to the existence of an effect. The relation between the manipulation and the measurement (i.e., the dependent variable) must be of the right kind (right sign, meaningful size), for it to provide evidence for an effect. For this reason (and for others too), finding a significant result cannot be sufficient for asserting that an effect has been found. For all that, one might take it to be necessary: Only then has one enough evidence to assert that an effect has been discovered.

In the remainder of this paper, I turn toward more substantial concerns with the $P < .005$ proposal.

4 Against the Bayes Factor Arguments

The Bayes Factor arguments that were summarized above have been criticized on several grounds. As we have seen the argument against the .05 significance level and the first argument for the .005 level rest on mapping *p*-values onto a Bayesian measure of evidence. Lakens and colleagues object to this mapping in four different ways. First, they write (2018, 168; see also Morey 2018): "Given that the marginal likelihood is sensitive to different choices for the models being compared, redefining alpha levels as a function of the Bayes factor is undesirable." The argument here is elliptic, but the following points should be relevant. The fact that the Bayes Factor depends on the specification of the null and alternative hypotheses does not in itself undermine its relevance for assessing how much evidence a .05 *p*-value corresponds to since the *p*-value too depends on the specification of a model (namely the null model), and since evidence is plausibly a comparative notion (data provide evidence between two competing hypotheses). This fact just shows that one must be careful in deciding how to represent a two-sided test in a Bayesian framework (more on this below). Furthermore, if it is granted that classical two-sided tests must be mapped onto a comparison between a point null hypothesis and a symmetric hypothesis around 0, the argument against the .05 level rests on the identification of an upper bound for the Bayes Factor: No other alternative model can result in a larger Bayes Factor. Admittedly, the first argument for the .005 significance level does depend on the specification of the alternatives, but Lakens and colleagues do not explain why the four alternatives hypotheses we examined are unreasonable nor do they propose any other alternative hypothesis.

What Lakens and colleagues seem to really object to is the very set-up of our analysis: It pits a point null hypothesis against a hypothesis that is symmetric around 0 (2018, 2; see also Mayo 2017b; Morey 2018): "one-sided tests or Bayes factors for non-point null models would imply different alpha thresholds." A one-sided *p*-value equal to .05 would indeed correspond to a different Bayes factor since a one-sided test would be represented differently from a two-sided test. A one-sided test is not a test of existence (it does not examine whether an effect is real), but a test of direction: It

examines whether the effect is positive or negative. In a Bayesian framework, it can be represented as comparing the area under the posterior distribution below 0 to the area above 0. Let's suppose that a single data point is sampled from a normal distribution with a known standard deviation: $X_1 \sim N(\theta, 1)$. Let's suppose also that the prior probability of θ is symmetric around 0, say $N(0, 1)$. Suppose that $X_1 = 1.64$. The one-tailed test rejects the null hypothesis $H_0 : X_1 \sim N(0, 1)$, and thus the directional hypothesis H_- that X_1 is taken from a normal distribution whose means is smaller than 0, at the .05 significance level ($p = .05$). The posterior probability distribution conditional on $X_1 = 1.64$ is $N(.82, .5)$. The Bayes factor is then $BF_{-+} = \frac{\int_{-\infty}^{0} N(.82,.5)}{\int_{0}^{\infty} N(.82,.5)}$, which is equal to .0505/.9495 or .053. The observed data point provides substantial evidence for the hypothesis that the mean of the normal distribution is larger than 0. Generally, if the prior of the mean is symmetric and the variance of the observations is known, a one-sided p-value equal to .05 will approximately correspond to an area under the posterior distribution below 0 equal to 5%, and the Bayes factor will be roughly equal to 1/19 (Marsman and Wagenmakers 2017).

While Lakens et al. are right that a one-sided test would result in a different Bayes factor than the Bayes factors we have reported, there is a good reason for mapping p-values onto Bayes factors comparing a null hypothesis and a symmetric alternative hypothesis: Psychologists almost always use two-sided tests, and the null hypothesis is rejected if the data are sufficiently extreme on either side of the null hypothesis (sufficiently large or small). The goal of Benjamin et al.'s proposal was to assess how much evidence a p-value in this kind of two-sided tests corresponds to, and it is strange to fault them for not considering other, rarely used tests. Lakens and colleagues may think that one-sided tests should replace two-sided tests, but they don't say it and a fortiori don't argue for it. More in the spirit of their critique, they may think that one-sided tests are sometimes appropriate, depending on the details of the research question. While this is certainly true, in the absence of preregistration we should be wary of adding more degrees of freedom to the analysis of data. Of course, one-sided tests can be preregistered. In this case, the very kind of Bayesian analysis that challenges the use of a .05 significance level for two-sided tests validates its use for preregistered one-sided tests.

Morey (2018) acknowledges that psychologists use two-sided tests, but rejects our representation of these tests by means of a symmetric alternative hypothesis around 0 against a hypothesis centered at 0. On his view, two-sided tests only appear to test a null model; rather, they really are a pair of one-sided tests with a penalty for looking at the sign after having seen the data (because the null model would have been rejected if the sign had been reversed). As he puts it,

> It is taken for granted (…) that the purpose of significance testing is to test a point null hypothesis against a two-sided alternative. The authors translate this into a two-sided Bayesian test with a point null. But if one doesn't assent to their interpretation of the purpose of the p-value, then the link between the p-value and their particular assessments of the Bayesian evidence falls apart, and the argument with it. Consider, for instance, if instead of viewing the 'two-sided' significance test as such, we understood it as two one-sided tests with a correction (doubling the p-value). (…) If the purpose of a significance test is testing one sign

> against the other, then (…) the Bayes factor—not the *p*-value—can appear to overstate the evidence. Wagenmakers has argued that *p*-values are intended as point-null tests against two-sided alternatives, and hence we should evaluate them as such, but it seems like building a reform around the expectations of badly-trained scientists might be a bad idea.

So, the crucial issue concerns the proper interpretation of classical two-sided tests (for a related point, see also Mayo 2017b): Do they genuinely test the null model or are they really two one-sided tests in order to determine the direction of the effect?[11]

The issue has a long history in statistics (e.g., Cox 1977) and remains unresolved. Morey gives several arguments for his preferred interpretation, but none is decisive. He claims that "it certainly aligns with many of its proponents' views." This point is however immediately undermined when he concedes that the vast majority of psychologists do not interpret two-sided tests this way, dismissing them as "badly-trained scientists." He also notes that under his preferred interpretation the *p*-value can be given a Bayesian interpretation under some conditions, but this is a strange remark given that Morey's goal is to challenge the relevance of a Bayesian approach to assessing the strength of evidence provided by a *p*-value.

Third, and more to the point, Morey rightly notes that for a classical statistician it is immaterial whether a two-sided test tests the null model or is really a pair of two one-sided tests since in the latter case the directional hypothesis is rejected if and only if the null model is rejected: The *p*-value is the same under either interpretation and the classical statistician has equal reason to reject the null or to reject the relevant directional hypothesis. By contrast, as we have seen, it can make an enormous difference to the Bayesian statistician whether a test contrasts a null model with an alternative hypothesis or two directional hypotheses. The conclusion then is that the particular representation of two-sided tests chosen by Benjamin et al. (2018) can't claim to be a natural representation of classical two-sided tests in a Bayesian framework. It is based on a distinction that makes no difference for the classical statistician, but is meaningful for the Bayesian statistician. As Morey (2018) puts it, "there isn't really about a disagreement between *p*-values and Bayes factors; it is really a disagreement between Bayesian evidence when a point null is included and Bayesian evidence when it is not. (…) It is the case that Bayesian evidence computed with respect to a particular set of hypotheses will differ from Bayesian evidence computed with respect to another set of hypotheses. The wisdom of incorporating a null hypothesis will depend on the situation, and Bayesian analysts can disagree."

Morey's argument is powerful, but it can be resisted. Whether one has Bayesian or classical leanings, there is a genuine distinction between examining whether a hypothesized correlation (or a causal relation) is real (a reality question) and whether the correlation is positive or negative (or whether a variable influences or hinders another—a direction question) *even if* from a classical-statistical point of view it makes no difference which question is asked. The same *p*-value could then be used to answer two genuinely different questions, and, if it is meaningful to assess the amount of evidence corresponding to a *p*-value equal to .05 using the Bayes factor (more on this

[11] Morey (2018) also shows that according to his own representation of two-sided tails as pairs of one-sided tails, a *p*-value equal to .02 provides substantial evidence for a directional hypothesis.

below), which question is asked calls for different Bayesian representations. The fact that psychologists do not see a two-sided test as a pair of one-sided tests justifies representing their tests as Benjamin et al. did: Psychologists are asking reality questions. Morey would probably reply that psychologists' attitude toward two-sided tests should be ignored because the best interpretation of these tests usually is as pairs of one-sided tests: As he puts it, "Under what circumstances would a scientist care about an effect but not its sign?" (Remember that it makes no difference to the *p*-value which question is asked, while making a difference for the Bayes factor.) Morey would surely concede that testing whether ESP is real would justify caring about whether an effect is real, but he would insist that most research situations aren't like that. However, keeping in mind the fact that statistical hypotheses are typically derived from competing theories explains why often psychologists primarily care about the reality of an effect. Typically psychologists compare two theories, one, but not the other, predicting a (causal or not) relation between two or more variables. Showing the reality of this relation is sufficient to undermine one of the two competing theories.

Lakens and colleagues also make the following remark (2018, 169): "When a test yields BF = 25, the data are interpreted as strong relative evidence for a specific alternative (for example, mean = 2.81), while a $P \leq 0.005$ only warrants the more modest rejection of a null effect without allowing one to reject even small positive effects with a reasonable error rate." But Lakens and colleagues seem to be comparing apples and oranges: that is, a Bayes Factor pitting a point null hypothesis centered at 0 against a particular alternative hypothesis with a *p*-value computed by reference to a null model centered at some small, but positive value, *d*. A proper comparison would compare this *p*-value with a Bayes Factor pitting a null hypothesis centered at *d* against a particular alternative hypothesis.[12]

Finally, they "question the idea that the alpha level at which an error rate is controlled should be based on the amount of relative evidence indicated by Bayes factors" (2018, 2). They do not elaborate on this point, but the idea may be that given the differences between the meaning of alpha and beta in significance testing (specifying error rates) and of the Bayes factor (measuring comparative evidence) it makes no sense to calibrate one by means of the other. Indeed, a committed classical statistician could simply reject the relevance of a measure of evidence to decide which significance level to adopt. However, one needs not be a card-carrying Bayesian to see the relevance of Bayes factors. Statistical "syncretism" (Greenland 2010), according to which statistical practices should be justifiable from various statistical perspectives, is sufficient to justify the appeal to Bayes factors in our argument.[13]

[12] Given that many null hypotheses are literally false (there is very often a tiny effect), Lakens and colleagues' remark challenges the common assumption that by rejecting a point null hypothesis one is also entitled to conclude that the effect is at most negligible (Machery 2014).

[13] One may question this appeal to syncretism since the choice of a .005 level is only justified on Bayesian grounds. A true syncretic approach would instead justify it on Bayesian *and* on frequentist grounds. However, first, the appeal to syncretism is meant to undermine the idea that Bayesian considerations are always irrelevant to a frequentist. Even is no frequentist justification is provided, a syncretist can't dismiss the relevance of Bayesian considerations. Second, the argument from the false discovery rate can be given a frequentist interpretation: It examines the frequency of false positives among significant results for various possible base rate of true null hypotheses, exactly as we would do when we assess whether a medical test is sufficiently sensitive.

5 Against the False Discovery Rate Argument

Several critics have objected to the false-discovery rate argument, a crucial part of the claim for the efficacy of our proposal.

5.1 Taking P-Hacking into Account

An important objection is that the false-discovery rate argument ignores various research practices, which, when taken into account, would negate the advantage of decreasing the significance level. Crane (n.d.) has recently claimed that "once P-hacking is accounted for the perceived benefits of the lower threshold all but disappear, prompting two main conclusions: (i) The claimed improvements to false discovery rate and replication rate in Benjamin et al. (2017) are exaggerated and misleading. (ii) There are plausible scenarios under which the lower cutoff will make the replication crisis worse." Crane computes the false-discovery rate as a function of power for a .05 and .005 significance level and for three rates of p-hacking (0, 5%, and 15% of significant results are p-hacked). He then shows that the reduction in the false discovery rate caused by a reduction of the significance level from .05 to .005 depends on the frequency of p-hacking. For a power equal to .8, "if 15% of all *p*-values are hacked, then the false positive rate would decrease from 0.75 to 0.71, just a 5% improvement, as a result of the lower cutoff."

Let's look at Crane's model in more detail. Crane starts with the usual false 500 discovery rate:

$$\frac{\alpha\varphi}{\alpha\varphi + (1-\beta)(1-\varphi)} \tag{1}$$

Where φ is the proportion of true null hypotheses. (1) is supplemented by assuming that a proportion h of all the computed p-values is p-hacked (and P-hacking is taken to be always successful):

$$\frac{\alpha\varphi(1-h) + h}{\alpha\varphi(1-h) + (1-\beta)(1-h)(1-\varphi) + h} \tag{2}$$

Crane then introduces a parameter π, called "persistence," that determines how many of the p-hacked p-values below .05, but above .005, are decreased to a value below .005 by increased p-hacking when the significance level is reduced to .005. Introducing π result in a false discovery rate defined as[14]

$$\frac{(\alpha/c)\varphi(1-h) + h\pi}{(\alpha/c)\varphi(1-h) + (1-\beta)(1-h)(1-\varphi) + h\pi} \tag{3}$$

Crane then goes on to estimate the rate of p-hacking comparing the reproducibility rate predicted by (1) with the results reported in Open Science Collaboration (2015), and

[14] This presentation simplifies slightly Crane's presentation, but nothing of importance is lost (see eq. 9 in Crane, n.d.).

proposes that h ranges between .05 and .15. We should not take this estimate seriously, however: It is computed assuming a power equal to .80, which is unrealistically high for psychology. Be it as it may, Crane then shows that for many levels of persistence the false discovery rate remains high when the significance level is decreased to .005. For instance, when $h = .05$, $\beta = .2$, and $\pi \approx .25$, the false discovery rate is equal to 20% (Fig. 3 of Crane n.d.). Furthermore, if $\pi = 1$, the reduction of the false discovery rate is much smaller than claimed by Benjamin et al. for even moderate levels of p-hacking (Fig. 4 of Crane n.d.). Finally, unsurprisingly, if power does not remain constant when the significance level is reduced, there are values of π and of β at $\alpha = .005$ such that the false discovery rate increases with the new significance level.

Crane rightly highlights the fact that the reduction in the false discovery rate reported in Benjamin et al. does not take into account p-hacking or a reduction in power, and thus that the benefits to be obtained from a reduction of the significance level depends on a host of further scientific practices (more on this below). We should nonetheless resist Crane's skeptical take on Benjamin et al.'s proposal. Crane asserts that "Without compelling evidence to the contrary, we should expect P-hacking to continue just as it is currently." On the contrary, we should expect p-hacking to be considerably reduced by a stringent decrease in the significance level and thus persistence to be quite low (see also de Ruiter 2019). Simonsohn et al. (2015) have shown that while p-hacking is relatively easy when $\alpha < .05$, it gets much harder for lower significance levels. Committed p-hackers would still be able to p-hack, no doubt, but amateur p-hackers would not. And because the border between committed p-hacking and fraud is really tenuous, few would be motivated to p-hack their way to p's below .005. If persistence is low, then Crane's analysis confirms Benjamin et al.'s conclusion. In fact, we can turn Crane's argument on its head: Reducing the significance level would also reduce the frequency of p-hacking, thereby reducing the rate of false discoveries even more than what Benjamin et al. (2018) advertised. If $h = .15$ when $\alpha = .05$ and $\pi = 0$, assuming $\beta = .2$, the false discovery rate is reduced by a factor of 12.5!

5.2 Doubts about the False Discovery Rate

Mayo (2017a) raises several concerns, but most seem to rest on misreading our proposal. First, she writes: "Their argument either turns on committing the fallacy [i.e., "the fallacy of transposing the conditional"] or holding a Bayesian measure as a gold standard from which to judge a frequentist error probability." The fallacy of transposing the conditional is the confusion of the probability of the data conditional on a hypothesis with the posterior probability. Of course, we do not commit this fallacy: The first argument reports the Bayes factor corresponding to a .05 p-value under a controversial, but defensible, way of representing a two-sided test, while the second argument examines the frequency of false positives among significant results assuming various prior odds. So, if we do anything wrong, then, on Mayo's views, it must be that we are "holding a Bayesian measure as a gold standard from which to judge a frequentist error probability." Here, Mayo seems to think that we take the false discovery rate to measure evidence (as does Morey 2018): She writes that "that the measure they recommend, the positive predictive value (PPV), or posterior probability of H1, actually greatly exaggerates the evidence" (2017a). In fact, only Bayes factors are taken to measure evidence in Benjamin et al. (2018). The false discovery rate is used to justify a particular choice of significance level given various possible prior odds.

Nothing more. Mayo also summarizes our argument by saying that "The criticism is that, at least if we accept these Bayesian assignments of priors, the posterior probability on H0 will be larger than the *p*-value." In effect Mayo seems to identify our argument to Lindley's paradox (Lindley 1957), but while both our argument and Lindley's paradox do require not assigning zero probability to the null model, the similarity stops there.

Lakens and colleagues challenge the false discovery rate argument as follows (2018, 169; see also Morey 2018): "Without stating the reference class for the 'base rate of true nulls' (for example, does this refer to all hypotheses in science, in a discipline or by a single researcher?), the concept of 'prior odds that the alternative hypothesis is true' has little meaning. Furthermore, there is insufficient representative data to accurately estimate the prior odds that researchers examine a true hypothesis, and thus, there is currently no strong argument based on FPRP [i.e., false positive report probabilities] to redefine statistical significance." These two problems call for the same response: It does not really matter how the reference class is specified and what exactly the prior odds are; what matters is that the false discovery rate improves substantially for various natural ways of specifying the reference class and for plausible prior odds. It is natural to understand the reference class in Benjamin et al. as being the class of hypotheses tested in cognitive, social, and personality psychology (the fields the journals examined by OSC (2015) belong to), but other plausible reference classes would work as well. Furthermore, Benjamin et al. examine the false discovery rate for a .05 and a .005 significance level given three possible prior odds, and we show that decreasing the significance level would substantially improve the false discovery rate for all of them. The prior odds at which our proposal would not make much of a difference are not plausible and remarkably Lakens and colleagues do not argue that they are.

Morey's (2018) arguments have a similar flavor, and do not fare better. He notes that the choice of a hypothesis "cannot be analogized to drawing random 'hypotheses' from an urn, some of which are true and some are false." Clearly scientists do not choose their hypotheses at random, but nor do physicians choose their patients at random from a given class of possible patients, some of which have a disease, while others don't. Nonetheless, it is still useful to idealize the choice of patients as being randomly drawn from an urn, and refers to the base rate of a disease in a population. Furthermore, Morey (2018) objects to the use of the notion of power (He asks: "Is there a 'power' associated with a 'field'?"), but he does not explain what could be the issue with the notion of the average power of an area of science. He then objects to the use of Bayes theorem to compute the false discovery rate (again very allusively), but there is surely nothing wrong in computing a rate from a base rate and the miss rate of a test.

6 Against the Efficacy of the $P < .005$ Proposal

The third line of argument against the efficacy of Benjamin et al.'s proposal contends that we didn't take into account the side effects of their proposal. There are two distinct versions of this concern.

6.1 Its Efficacy Can't Be Known

Lakens and colleagues (2018) insist that we can't really know whether decreasing the significance level by an order of magnitude would improve the reproducibility of psychology (see also Zollman 2017). "[I]n practice" as they say, it depends on "unknowns," i.e., on how reducing the significance level would impact other factors that contribute to the replication crisis. McShane and colleagues concur (McShane et al. 2018, 10): "[W]e have no idea whether implementation of the proposed 0.005 threshold would improve or degrade the state of science as we can envision both positive and negative outcomes resulting from it."

It may be tempting, but flawed, to respond that, everything else being equal, our proposal would improve the reproducibility of psychology. Even if that is true, we can be certain that not everything would be equal if Benjamin et al.'s proposal were implemented, and our critics would be right to be unimpressed by this line of response. A better response concedes that the efficacy of this proposal isn't knowable with certainty, while insisting that plausibility arguments can be brought to bear on the question. Indeed, cutting down the significance level would plausibly undermine some of the possible causes of the replication crisis: As we've seen, p-hacking would be abated by our proposal.

6.2 It is Inefficacious because of Likely Vicious Side Effects

A somewhat different objection is to argue that decreasing the significance level by an order of magnitude is likely to produce specific vicious side effects (e.g., Malinsky 2017). Argamon (2017) holds that that "While tightening the standards for statistical significance might slightly reduce the number of irreproducible results, the proposal will reinforce pernicious ideas that prevent the scientific community from adopting better methodologies."[15] In particular, it is supposed to reinforce the importance of a significance level in scientific practice (see also the conclusion of Crane n.d.): "Reducing the *p*-value threshold as proposed will have the effect of doubling down on the supposed importance of statistical significance and will only reinforce the problematic idea that a study is either 'in' or 'out'."[16] Argamon's characterization of the role of the significance level as a cutoff is vague. A cutoff could work as a necessary[17] condition for publication, for taking a phenomenon to be definitely established (that is: an effect is real if and only if an internally valid and methodologically sound study obtains a significant result), or for asserting, defeasibly, that some effect is real. To repeat, Benjamin et al. do not hold that the significance level should be used as a publication filter (in fact, if scientists were to follow our recommendation to the letter, the significance level would stop being used as a publication filter), and we do not take effects to be definitively established by single studies.[18] So, Benjamin et al. only endorse the third type of cutoff, and it is not clear

[15] This type of objection would undermine various other proposals that take for granted the null hypothesis significance testing framework (e.g., preregistration).

[16] Argamon also flirts with the everything-or-nothing attitude that we criticized earlier when we discussed Trafimow et al. (2018).

[17] Noone thinks it is sufficient.

[18] Our proposal is entirely consistent with a metaanalytic approach, and it is unclear why, as Lakens et al. (2018, 169) assert, our proposal would "divert attention from the cumulative evaluation of findings, such as converging results of multiple (replication) studies."

what is wrong with it. Perhaps Argamon rejects the very idea that scientists need to decide in a dichotomous manner that an effect has been found, but this radical idea is hard to take seriously. Surely, Argamon would agree that we need to be able to claim that, e.g., a treatment is a fraud (Giner-Sorolla 2018). Furthermore, nearly all sciences involve mechanisms that decide whether to take a phenomenon to be real. Science is a social procedure that involves mechanisms by which phenomena are accepted: This is true in particle physics (5 sigmas significance level), epidemiology (consensus conferences and reports), climate science (Intergovernmental Panel on Climate Change), psychiatry (development of the DSM), etc. Perhaps Argamon thinks more reasonably that dichotomous decisions are part and parcel of science, but that they shouldn't be based on statistical criteria. This moderate position is not fully satisfying: If decisions are going to be made, they will involve cutoffs that turn continuous differences into dichotomies, explicit cutoffs are surely preferable to implicit ones, and conventional cutoffs to subjective ones. And if there are any cutoffs, why couldn't we use explicit, conventional statistical cutoffs? What is so peculiar about statistical measures that would make it unacceptable to have such cutoffs? Additionally, it is dubious that the existence of a significance level is the source of everything that goes wrong in psychology (see also Crane n.d.): Some sciences such as particle physics with its five-sigma significance level and GWAS-based genetic in recent years use statistical cutoffs, but are doing just fine. Perhaps Argamon and others would acknowledge the necessity of explicit cutoffs, but respond that these need to be determined on a case-by-case basis as a function of the costs of various types of errors rather than having conventional cutoffs. We will discuss this possibility below.

Lakens and colleagues (2018, 169) argue that it would result in fewer replication attempts: "[D]esigning larger original studies would leave fewer resources (that is, time, money, participants) for replication studies, assuming fixed resources overall." Because we didn't propose to set the significance level at .005 for replication attempts, but only for original results, attempts to replicate published results won't be made more onerous by the fact that a larger sample size is required to reach a given power for a lower alpha. Now it is true that scientists would have overall less time and fewer resources for replication if the scientific community followed our proposal because original experimentation would require more resources (due to the larger required samples). But why is this a problem? What we care about is replica*bility*: the fact that published scientific findings *can* be successfully replicated. Replica*tion* is a means to assess replicability and to incentivize researchers to improve their research practices (nobody wants their results to fail to replicate), improving thereby the replicability of their research. Our proposal would improve the replicability of significant results, lessening the need for replication.

Perhaps one may think that decreasing the significance level by an order of magnitude would increase the file-drawer problem: More papers would fail to reach the significance level than is currently the case, and scientists may be motivated to store them away, convinced that they would not be selected for publication (e.g., Argamon 2017). Again, we do not recommend using a significance level set at .005 as a publication filter. In response, one may accuse us of naiveté: Were our proposal widely accepted, a critic may object, it is likely that reviewers and journals would use it for deciding which submissions to select for publication, given the traditional bias for publishing significant results (Fanelli 2010). But what this concern shows is merely that our proposal does not fully stand on its own. I will come back to this point in the last section of this paper.

One may also object that our proposal will prevent people from testing bold, unlikely hypotheses and promote safe science. Schimmack (2017) voices this objection as follows (see also Hamlin 2017): "the proposal to require $p < .005$ as evidence for an original, risky study implies that researchers need to invest a lot of resources in a risky study that may provide inconclusive results if it fails to produce a significant result." However, bold hypotheses can be piloted with small samples. Those that survive such piloting would then be tested in resource-intensive experiments. In fact, our proposal would discourage psychologists from presenting what are really pilot tests of off-the-cuff bold ideas, which may well be significant by chance, as planned tests. If the significance level is decreased, pilot, exploratory studies and confirmatory studies will have very different sample sizes, and it will be harder to pass ones for the others.

Trafimow et al. (2018), Lakens and colleagues (2018), and Amrhein and Greenland (2018) assert that decreasing the significance level by an order of magnitude would increase the inflation of effect size because only effect sizes that are by chance very large would result in a *p*-value lower than .005. Amrhein and Greenland (2018, 4) write that "rejections due to $P > 0.05$ will remain, and rejections due to $P > 0.005$ will now also occur, leading to more intense P hacking and selective reporting, with increased bias in reported effects." We have already explained why psychologists would not p-hack with increased fervor, and I have discussed the issue of selective reporting above; I now focus on the claim about effect size. Inflation of effect size does not follow from the use of a significance level, but from the use of a significance level as a publication filter. Since we recommend that all the competent studies be published, whether or not significant at the .005 level, our proposal should not increase effect size inflation. Furthermore, we are expecting psychologists to increase their sample size so as to reach an acceptable level of power. By increasing the sample size, scientists would get more accurate, thus less inflated estimates, of the true effect sizes (assuming again no publication filter). We see again that our proposal does not fully stand on its own: Its efficacy is conditional on further improvements in scientists' research practices, in the present case, on an increase of power and the elimination of publication filters.[19]

Finally, many, including McShane et al. (2018, 19), Amrhein and Greenland (2018), Amrhein et al. (2017), and Lakens et al. (2018) are concerned about the trade-off between controlling for false positives (the focus of our proposal) and controlling for false negatives. Everything else being equal, if we control for the former, we increase the probability of false negatives. However, we do not take things to be equal: We do expect scientists to increase their sample size and power to remain constant. Perhaps however this is a naïve expectation. I discuss this issue in Section 7.

[19] What if the significance level is used as publication filter? We then need to distinguish the situation where the null hypothesis is true and those where the null hypothesis is false. When the null hypothesis is true, effect size inflation increases with a decreased significance level, even if the sample size increases to maintain power constant. However, when the null is false, such increase need not be the case. *P*-values are right skewed when the null is false and the extent of the skew depends on the sample size for constant population parameters. So, if decreasing the significance level results in an increase in sample size, a larger number of *p*-values may be significant for a smaller significance level. As a result, effect size inflation may decrease rather than increase.

7 Against the Actionability of the $P < .005$ Proposal

As we saw earlier, one of the selling points of Benjamin and colleagues' proposal is its actionability. This section discusses objections to its alleged actionability.

7.1 Alpha Flexibility and the Varying Cost of Research

The most obvious concern with Benjamin and colleagues' proposal is that it will require a substantial increase in sample size to keep power constant (e.g., Esarey 2017; Schimmack 2017; Hamlin 2017; Mayo 2017b; Lakens et al. 2018; Schmalz 2018). Since power is already unacceptably low in psychology, psychologists would have to collect much more data than they are used to in order to reach a reasonable level of power (at least .8) for an alpha set at .005. Benjamin and colleagues report that for most common tests the sample size would have to increase by 70%. Schmalz (2018) notes that for an effect size that is typical in psychology (d = .2) and a two-tailed, independent-sample t-test, the sample size would have to be increased from 788 to 1336 to reach a power of .8.

We expect psychologists to increase their samples sizes in response to a decrease in the significance level, but perhaps this expectation is naïve. Psychologists, and possibly other scientists, seem to rarely set their sample size by considering the desired power of their tests; rather, they follow usual practice (Vankov et al. 2014). For instance, Oakes (2017) shows that developmental psychologists follow the typical practice of having between twenty to thirty babies or children per cell instead of considering the power of their test. Furthermore, it can be very difficult to recruit participants and collect independent measurements in some areas of psychology such as developmental psychology, neuropsychology, and cognitive anthropology. More generally, any psychological research that goes beyond recruiting undergraduate students or on-line participants (e.g., longitudinal research programs) may be challenged by the need to increase sample sizes. In some areas of neuroscience, research requires intervening on animals and sometimes killing them: More participants just mean more deaths. For ethical reasons, then, it may not be permissible to increase sample sizes. Similar concerns are found outside psychology, such as epidemiology: Participants may be hard to recruit and involving many participants in some epidemiological studies may raise ethical concerns. In those areas of science, decreasing the significance level would likely result in an increase in the probability of committing a false negative. Or it might disincentivize research in these areas and incentivize instead research using easily accessible samples such as Amazon Turkers (Hamlin 2017; Lakens et al. 2018).

The proposal to reduce the significance level by an order of magnitude thus faces a dilemma: In some areas of science, either its efficacy (because of its side effect: an increase in the probability of a false negative) or its actionability (it can't be acted upon if it requires high power and low alpha) is challenged. This dilemma is related to a more general criticism of the proposal made in Benjamin et al. (2018): It proposes a rule for scientific research without considering the costs involved in setting alpha at a given level for particular research projects (although we do acknowledge that the significance level depends on "a trade-off between type I and type II errors" (2018, 8)). Instead, we should examine how easily the sample size can be increased and what the respective costs of false positives and false negatives are. On this basis, scientists should carefully set their alpha level at a particular, context-sensitive value, and no conventional alpha

level is needed: As Lakens et al. put it (2018, 168) "researchers [should] justify their choice for an alpha level before collecting the data" (see also Trafimow et al. 2018).

No doubt, recruiting participants is harder in some areas of psychology than in others, but, whether or not the significance level is changed, psychologists must in any case increase their sample size substantially. While collecting data from 1336 participants for a two-tailed, independent-sample t-test may seem daunting, the fact is that the sample size of most psychological experiments is not nearly near the 788 participants that would be needed to reach a .8 power with the current .05 significance level. In the areas of psychology that do not use on-line, easily accessible samples, psychologists must reform their data collection methods to reach such numbers (and this, again, independently of any change in the significance level). And once this reorganization has taken place, the increase in sample size required by a smaller significance level should not be that daunting. Furthermore, the claim that this results in an unavoidable trade-off between false positives and false negatives probabilities is exaggerated. At least in some areas of psychology, the challenges involved in recruiting participants can be alleviated by multi-lab collaborations, the future of psychology (more on this below), and switching to within-subjects designs can increase power.

Still, it should be conceded that these remedies won't be available everywhere, including in comparative psychology and cognitive anthropology as well as in epidemiology outside psychology. But even there we should resist the suggestion to let scientists set up their own significance level by considering the costs of false positives and false negatives as well as the difficulty of recruiting participants. The replication crisis is at least partly driven by the degrees of freedom available to scientists when they analyze data, and adding yet another degree of freedom is unlikely to be helpful. Furthermore, it is dubious that, given current career incentives, scientists will choose to make things harder for themselves when their competitors can choose a more lenient significance level. Thus, letting scientists choose their own alpha level is a recipe for status quo. Additionally, the significance level plays a role in the communication of results to outsiders, including scientists in other disciplines and non-scientists such as policy makers. It tells outsiders that some result has been vetted to a degree judged satisfying by the scientific community; different standards would be extremely confusing for them (Bright 2017).

Perhaps one may propose to set the alpha level by field or area of research in order to address the concern of increased degrees of freedom in data analysis while respecting the varying costs of false positives and false negatives: lower in cognitive psychology, say, than in the areas of neuroscience that involve intervening on, to say nothing of killing, animals. The prudent thing here would be to respond that the proposal made in Benjamin et al. (2018) only applies to areas of science that do not suffer from such constraints. Indeed, Benjamin et al. (2018, 3) concede that "the significance threshold […] should depend on the prior odds that the null hypothesis is true, the number of hypotheses tested, the study design, the relative cost of type I versus type II errors, and other factors that vary by research topic." However, we can say something stronger (see also Benjamin et al. 2017). Remember that we do not recommend treating the .005 significance level as a requirement for publication; rather, when they are the products of good scientific practices, results should be published independently of their p-value, and, when $p < .05$, they should be identified as "suggestive." Rather than claiming that reaching the .05 level, but not the .005 level constitutes a scientific discovery, it is better to publish them, acknowledge that they are merely suggestive, and insist that in some areas of science we can only do with

suggestive results. The alternative leads scientists to take too much comfort in results that have a substantial probability of being non-replicable. The right thing to do when it is hard to gather evidence is not to relax standards, but rather to acknowledge that less evidence is the best we can do in some areas.

7.2 The Needed Convergence of Reforms

It should now be clear that the success of our proposal depends on at least two complementary reforms: the increase in psychological experiments' sample size (to reach a power of .80 for an alpha of .005), and thus an increased in power, and a reform of the current publication practices (the significance level should not be used as a publication filter). (P-hacking is expected to decrease as a side effect of a more stringent significance level.) We did acknowledge the complementarity of our proposal and of other attempts at improving psychology, but we should have been explicit about this entanglement.

A critic could reasonably hold that this entanglement weakens the actionability of our proposal since it now appears that the proposed reform isn't simply a matter of changing a norm of assertion. However, psychologists are now aware of the importance of power, and sample size seems to be increasing (Fraley and Vazire 2014), although not nearly as much as needed. Multi-lab collaborations have appeared in recent years, and pre-data collection peer review is gaining traction. In any case such reforms must happen for psychology to improve: Publication filters result in inflated effect sizes and misleading meta-analyses, and they motivate p-hacking; small sample sizes result in low power, wasted resources, and high false discovery rates.

7.3 Against Small-Scale Science

Peters (2017) has insightfully remarked that "The way such a redefinition [of the alpha level] would shake up the entire landscape, basically provide a 'soft reset' to the way much of science is done and communicated, seems to be ignored." Decreasing the significance level by an order of magnitude might indeed usher a host of (in my mind excellent) reforms of psychological research. In fact, as insightfully noted by Gelman (2017a), the efficacy of Benjamin et al.'s proposal derives in part from its indirect effects. In particular, it would create new incentives, which would ultimately promote complementary reforms: Psychologists would be less inclined to p-hack since p-hacking becomes less efficient, and they would increase power in order to reach a lower significance level.

Evolutionary biologists distinguish two reproductive strategies: R-strategy (parents produce many offspring but invest few resources in each of them) and K-strategy (parents produce few offspring, but invest more resources in each of them). By analogy, we can distinguish two (obviously idealized) research strategies: small-scale strategies and big-scale strategies. Small-scale empirical scientists run many (comparatively) cheap experiments, and try to publish as much as possible. Each scientist has his or her own lab, and the post-docs and graduate students within each lab conduct experiments. Graduate students are expected to conduct several studies to complete their PhD. Funding goes to a lab PI for projects that will be mostly conducted in her lab. Big-scale empirical scientists run few expensive experiments, all of which involve several labs. No scientist is expected to understand the whole scientific process (from experimental design to data analysis), but rather scientists specialize on a particular aspect of

the experimental process. Graduate students are not expected to conduct several experiments, but contribute their expertise to large-scale projects. Funding goes to collaborations that involve dozens or hundreds of scientists. Psychology of course follows a small-scale strategy, particle physics a large-scale one.

Benjamin et al.'s (2018) proposal would push psychology away from its nineteenth century small-scale roots and prompts it to become a more large-scale enterprise. Decreasing the significance level would compel psychologists to increase their sample size, a step beyond the contemporary friendly exhortations, and thus to develop new strategies for collecting data. Multi-lab collaborations and long-term, large projects would be encouraged instead of lab-based, short-term experiments. Graduate students could not expected to conduct several studies given the difficulty of collecting data from the required sample sizes. Funding for research, graduate education, and post-doctoral position would then be likely to go to large-scale projects.

8 Conclusion

No crisis should go to waste. The soul-searching episode psychology is going through is a unique opportunity to rectify unacceptable, but nonetheless prevalent research practices. More ambitiously, it is time for psychology to change its identity and become more of a large-scale science. Benjamin et al.'s (2018) proposal to decrease the significance level by an order of magnitude can be instrumental in such transformation by compelling scientists to develop new strategies to collect data. This is one of the two aspects of our proposal. The other aspect is less ambitious, but equally important and perhaps more urgent: This reform proposes an actionable, but efficacious way of improving the replicability of psychology. It is reasonable to be concerned with its efficacy since it could have detrimental side effects. It is surely not impossible to exclude those, but doom-and-gloom pessimism about the efficacy of our proposal is unjustified. It also reasonable to challenge its actionability since our proposal is really entangled with other reforms of scientific practices; however these reforms must take place. The objections against decreasing the significance level that were reviewed in this article, while often reasonable, are thus not decisive, and should not stop us from implementing it.

Acknowledgements I owe the expression "alpha war" to Simine Vazire. Thanks to John Doris, Felipe Romero, and two reviewers for very helpful feedback.

References

Amrhein, V., and S. Greenland. 2018. Remove, rather than redefine, statistical significance. *Nature Human Behaviour* 2: 4.

Amrhein, V., F. Korner-Nievergelt, and T. Roth. 2017. The earth is flat (p> 0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ* 5: e3544.

Amrhein, V., D. Trafimow, and S. Greenland. 2018. Abandon statistical inference. *PeerJ Preprints* 6: e26857v1. https://doi.org/10.7287/peerj.preprints.26857v1.

Argamon, S. E. (2017). Don't strengthen statistical significance—Abolish it. https://www.americanscientist.org/blog/macroscope/dont-strengthen-statistical-significance-abolish-it.

Baker, M., and E. Dolgin. 2017. Cancer reproducibility project releases first results. *Nature* 541: 269–270.

Begley, C.G., and L.M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483: 531–533.

Benjamin, D., Berger, J., Johannesson, M., Johnson, V., Nosek, B., & Wagenmakers, E. J. (2017). Précis by Dan Benjamin, Jim Berger, Magnus Johannesson, Valen Johnson, Brian Nosek, and EJ Wagenmakers. http://philosophyofbrains.com/2017/10/02/should-we-redefine-statistical-significance-a-brains-blog-roundtable.aspx.

Benjamin, D.J., J.O. Berger, M. Johannesson, B.A. Nosek, E.–J. Wagenmakers, R. Berk, K.A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C.D. Chambers, M. Clyde, T.D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A.P. Field, M. Forster, E.I. George, R. Gonzalez, S. Goodman, E. Green, D.P. Green, A. Greenwald, J.D. Hadfield, L.V. Hedges, L. Held, T.–H. Ho, H. Hoijtink, J.H. Jones, D.J. Hruschka, K. Imai, G. Imbens, J.P.A. Ioannidis, M. Jeon, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S.E. Maxwell, M. McCarthy, D. Moore, S.L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T.H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F.D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D.J. Watts, C. Winship, R.L. Wolpert, Y. Xie, C. Young, J. Zinman, and V.E. Johnson. 2018. Redefine statistical significance. *Nature Human Behavior* 2 (1): 6–10.

Bright, L. K. (2017). Supporting the redefinition of statistical significance. http://sootyempiric.blogspot.com/2017/07/supporting-redefinition-of-statistical.html.

Button, K.S., J.P. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S. Robinson, and E.R. Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Review Neuroscience* 14: 365376. https://doi.org/10.1038/nrn3475.

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say 'usually not'. https://doi.org/10.17016/FEDS.2015.083. Available at SSRN: https://ssrn.com/abstract=2669564 or https://doi.org/10.2139/ssrn.2669564

Cohen, J. 1962. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 65: 145–153.

Colquhoun, D. 2014. An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science* 1 (3): 140216.

Cox, D.R. 1977. The role of significance tests. *Scandinavian Journal of Statistics* 4: 49–63.

Crane, H. (n.d.). Why 'redefining statistical significance' will not improve reproducibility and could make the replication crisis worse.

de Ruiter. 2019. Redefine or justify? Comments on the alpha debate. *Psychonomic Bulletin & Review* 26 (2): 430–433.

Esarey, J. (2017). Lowering the threshold of statistical significance to $p < 0.005$ to encourage enriched theories of politics. https://thepoliticalmethodologist.com/2017/08/07/in-support-of-enriched-theories-of-politics-a-case-for-lowering-the-threshold-of-statistical-significance-to-p-0-00

Etz, A., and J. Vandekerckhove. 2016. A Bayesian perspective on the reproducibility project: Psychology. *PLoS One* 11 (2): e0149794.

Fanelli, D. 2010. "Positive" results increase down the hierarchy of the sciences. *PLoS One* 5 (4): e10068.

Fraley, R.C., and S. Vazire. 2014. The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One* 9 (10): e109019.

García-Pérez, M.A. 2017. Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement* 77: 631–662.

Gelman, A. (2017a). When considering proposals for redefining or abandoning statistical significance, remember that their effects on science will only be indirect! http://andrewgelman.com/2017/10/03/one-discussion-redefining-abandoning-statistical-significance/.

Gelman, A. (2017b). Response to some comments on "abandon statistical significance." http://andrewgelman.com/2017/10/02/response-comments-abandon-statistical-significance/.

Giner-Sorolla, R., (2018). Justify your alpha … for its audience. https://approachingblog.wordpress.com/2018/03/28/justify-your-alpha-to-an-audience/.

Greenland, S. 2010. Comment: The need for syncretism in applied statistics. *Statistical Science* 25 (2): 158–161.

Greenwald, A.G. 1976. An editorial. *Journal of Personality and Social Psychology* 33: 1–7.

Guilera, G., M. Barrios, and J. Gómez-Benito. 2013. Meta-analysis in psychology: A bibliometric study. *Scientometrics* 94 (3): 943–954.

Hamlin, K. (2017). Commentary by Kiley Hamlin. http://philosophyofbrains.com/2017/10/02/should-we-redefine-statistical-significance-a-brains-blog-roundtable.aspx.

Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Medicine* 2 (8): e124.

Ioannidis, J.P.A. 2016. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly* 94 (3): 485–514.

Lakens, D., F.G. Adolfi, C.J. Albers, F. Anvari, M.A.J. Apps, S.E. Argamon, T. Baguley, R.B. Becker, S.D. Benning, D.E. Bradford, E.M. Buchanan, A.R. Caldwell, B. Van Calster, R. Carlsson, S.-C. Chen, B. Chung, L.J. Colling, G.S. Collins, Z. Crook, E.S. Cross, S. Daniels, H. Danielsson, L. DeBruine, D.J. Dunleavy, B.D. Earp, M.I. Feist, J.D. Ferrell, J.G. Field, N.W. Fox, A. Friesen, C. Gomes, M. Gonzalez-Marquez, J.A. Grange, A.P. Grieve, R. Guggenberger, J. Grist, A.-L. van Harmelen, F. Hasselman, K.D. Hochard, M.R. Hoffarth, N.P. Holmes, M. Ingre, P.M. Isager, H.K. Isotalus, C. Johansson, K. Juszczyk, D.A. Kenny, A.A. Khalil, B. Konat, J. Lao, E.G. Larsen, G.M.A. Lodder, J. Lukavský, C.R. Madan, D. Manheim, S.R. Martin, A.E. Martin, D.G. Mayo, R.J. McCarthy, K. McConway, C. McFarland, A.Q.X. Nio, G. Nilsonne, C.L. de Oliveira, J.-J.O. de Xivry, S. Parsons, G. Pfuhl, K.A. Quinn, J.J. Sakon, S.A. Saribay, I.K. Schneider, M. Selvaraju, Z. Sjoerds, S.G. Smith, T. Smits, J.R. Spies, V. Sreekumar, C.N. Steltenpohl, N. Stenhouse, W. Świątkowski, M.A. Vadillo, M.A.L.M. Van Assen, M.N. Williams, S.E. Williams, D.R. Williams, T. Yarkoni, I. Ziano, and R.A. Zwaan. 2018. Justify your alpha. *Nature Human Behaviour* 2 (3): 168–171.

Lemoine, N.P., A. Hoffman, A.J. Felton, L. Baur, F. Chaves, J. Gray, Q. Yu, and M.D. Smith. 2016. Underappreciated problems of low replication in ecological field studies. *Ecology* 97 (10): 2554–2561.

Lindley, D.V. 1957. A statistical paradox. *Biometrika* 44: 187–192.

Machery, E. 2014. Significance testing in neuroimagery. In *New waves in the philosophy of mind*, ed. J. Kallestrup and M. Sprevak, 262–277. Palgrave Macmillan.

Machery, E. (n.d.). What is a replication?.

Malinsky, D. (2017). Significant moral hazard. https://sootyempiric.blogspot.com/2017/08/significant-moral-hazard.html.

Marsman, M., and E.J. Wagenmakers. 2017. Three insights from a Bayesian interpretation of the one-sided *p*-value. *Educational and Psychological Measurement* 77 (3): 529–539.

Mayo, D. (2017a). Commentary by Deborah Mayo. http://philosophyofbrains.com/2017/10/02/should-we-redefinestatistical-significance-a-brains-blog-roundtable.aspx.

Mayo, D. (2017b). Why significance testers should reject the argument to "redefine statistical significance", even if they want to lower the *p*-value. https://errorstatistics.com/2017/12/17/why-significance-testers-should-reject-the-argument-to-redefine-statistical-significance-even-if-they-want-to-lower-the-p-value/.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2018). Abandon statistical significance. April 9, 2018.

Meehl, P.E. 1990. Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports* 66: 195–244.

Morey, E. (2017). When the statistical tail wags the scientific dog. Should we 'redefine' statistical significance? https://medium.com/@richarddmorey/when-the-statistical-tailwags-the-scientific-dog-d09a9f1a7c63.

Morey, E. (2018). Redefining statistical significance: The statistical arguments. https://medium.com/@richarddmorey/redefining-statistical-significance-the-statistical-arguments-ae9007bc1f91.

Oakes, L.M. 2017. Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy* 22 (4): 436–469.

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

Peters, G. J. (2017). Appropriate humility: Choosing sides in the alpha wars based on psychology rather than methodology and statistics. https://sciencer.eu/2017/08/appropriate-humility-choosing-sides-in-the-alpha-wars-based-on-psychology-rather-than-methodology-and-statistics/.

Schimmack, U. (2017). What would Cohen say? A comment on p < .005. https://replicationindex.wordpress.com/2017/08/02/what-would-cohen-say-a-comment-on-p-005/.

Schmalz, X. (2018). By how much would we need to increase our sample sizes to have adequate power with an alpha level of 0.005? http://xeniaschmalz.blogspot.ca/2018/02/by-how-much-would-we-need-to-increase.html?

Sedlmeier, P., and G. Gigerenzer. 1989. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105: 309–316.

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

Simonsohn, U., J.P. Simmons, and L.D. Nelson. 2015. Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General* 144 (6): 1146–1152.

Trafimow, D. 2018. An a priori solution to the replication crisis. *Philosophical Psychology* 31: 1188–1214.

Trafimow, D., V. Amrhein, C.N. Areshenkoff, C. Barrera-Causil, E.J. Beh, Y. Bilgiç, R. Bono, M.T. Bradley, W.M. Briggs, H.A. Cepeda-Freyre, S.E. Chaigneau, D.R. Ciocca, J. Carlos Correa, D. Cousineau, M.R. de

Boer, S.S. Dhar, I. Dolgov, J. Gómez-Benito, M. Grendar, J. Grice, M.E. Guerrero-Gimenez, A. Gutiérrez, T.B. Huedo-Medina, K. Jaffe, A. Janyan, A. Karimnezhad, F. Korner-Nievergelt, K. Kosugi, M. Lachmair, R. Ledesma, R. Limongi, M.T. Liuzza, R. Lombardo, M. Marks, G. Meinlschmidt, L. Nalborczyk, H.T. Nguyen, R. Ospina, J.D. Perezgonzalez, R. Pfister, J.J. Rahona, D.A. Rodríguez-Medina, X. Romão, S. Ruiz-Fernández, I. Suarez, M. Tegethoff, M. Tejo, R. van de Schoot, I. Vankov, S. Velasco-Forero, T. Wang, Y. Yamada, F.C. Zoppino, and F. Marmolejo-Ramos. 2018. Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology* 9, article 699. https://doi.org/10.3389/fpsyg.2018.00699.

Vankov, I., J. Bowers, and M.R. Munafò. 2014. On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology* 67 (5): 1037–1040.

Wegner, D.M. 1992. The premature demise of the solo experiment. *Personality and Social Psychology Bulletin* 18 (4): 504–508.

Zollman, K. (2017). Commentary by Kevin Zollman. http://philosophyofbrains.com/2017/10/02/should-we-redefinestatistical-significance-a-brains-blog-roundtable.aspx.