OPINION ARTICLE

Retract $p < 0.005$ and propose using JASP, instead [version 1; referees: 1 approved]

Jose D. Perezgonzalez $^{1}$, M. Dolores Frías-Navarro $^{2}$

$^{1}$Business School, Massey University, Manawatu Campus, P.O.Box 11-222, Palmerston North, 4442, New Zealand

$^{2}$Department of Methodology of the Behavioural Sciences, Faculty of Psychology, University of Valencia, Avenida Blasco Ibáñez 21, Valencia, 46010, Spain

Abstract

Seeking to address the lack of research reproducibility in science, including psychology and the life sciences, a pragmatic solution has been raised recently: to use a stricter $p < 0.005$ standard for statistical significance when claiming evidence of new discoveries. Notwithstanding its potential impact, the proposal has motivated a large mass of authors to dispute it from different philosophical and methodological angles. This article reflects on the original argument and the consequent counterarguments, and concludes with a simpler and better-suited alternative that the authors of the proposal knew about and, perhaps, should have made from their Jeffreysian perspective: to use a Bayes factors analysis in parallel (e.g., via JASP) in order to learn more about frequentist error statistics and about Bayesian prior and posterior beliefs without having to mix inconsistent research philosophies.

Corresponding author: Jose D. Perezgonzalez (j.d.perezgonzalez@massey.ac.nz)

Author roles: Perezgonzalez JD: Conceptualization, Funding Acquisition, Writing – Original Draft Preparation, Writing – Review & Editing; Frías-Navarro MD: Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Perezgonzalez JD and Frías-Navarro MD. Retract $p < 0.005$ and propose using JASP, instead [version 1; referees: 1 approved] F1000Research 2017, 6:2122 (doi: 10.12688/f1000research.13389.1)

Copyright: © 2017 Perezgonzalez JD and Frías-Navarro MD. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: This publication was financially supported by Massey University Research Fund (MURF) Mini Grant 2017 for Open Access publication, Massey University, New Zealand.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Argument**

Seeking to address the lack of research reproducibility due to the high rate of false positives in the literature, Benjamin *et al.* (2017a); Benjamin *et al.* (2017b) propose a pragmatic solution which “aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance” (also Savehn, 2017): to use a stricter $p < 0.005$ standard for statistical significance when claiming evidence of new discoveries.

The proposal is subject to several constraints in its application: (1) to claims of discovery of new effects (thus, not necessarily to replication studies); (2) when using null hypothesis significance testing (arguably Fisher’s approach, perhaps even Neyman-Pearson’s, but excluding other $p$-value-generating approaches such as resampling); (3) in fields with too flexible standards (namely 5% or above); (4) when the prior odds of alternative-to-null hypothesis is in the range 1-to-5, to 1-to-40 (stricter standards are required with lower odds); (5) for researcher’s consumption (thus, not a standard for journal rejection, although “journals can help transition to the new statistical significance threshold”; also, “journals editors and funding institutions could easily enforce the proposal”, Wagenmakers, 2017; and “its implementation only requires journal editors to agree on the new threshold”, Machery, 2017); (6) while still keeping findings with probability up to 5% as suggestive (and merit publication if so “properly labelled”); (7) despite many of the proponents believing that the proposal is nonsense, anyway (that is, it is a quick fix, not a credible one; also Ioannidis in Easwaran, 2017; Resnick, 2017; Wagenmakers, 2017; Wagenmakers & Gronau, 2017).

The main analyses supporting the proposal were the following: (a) a calibration of $p$-values to Bayes factors under certain plausible alternatives (their Figure 1, Benjamin *et al.*, 2017b, p. 2); (b) an estimated false positive rate under certain plausible prior odds of alternative-to-null hypotheses as a function of power (their Figure 2, Benjamin *et al.*, 2017b, p. 3); and (c) a “critical mass of researchers now endorses[ing] this change” (a.k.a., the 72 co-authors). Meanwhile, the proposal downplays the costs due to increasing sample size in order to ensure enough research power, the potential increase in false negatives and misconduct, and the existence of alternative solutions well known by, at least, some of the authors (despite Machery’s, 2017, claim to the contrary).

Notwithstanding its potential impact, Benjamin *et al.*’s (2017a); Benjamin *et al.*’s (2017b) proposal is not recognizant of the entire engine of science (Figure 1), which has motivated an even larger mass of authors to offer their counter-arguments.

**Counter-arguments**

From Philosophy of science, Mayo (2017a), in particular, provides important counter-arguments against the philosophical standing of Benjamin *et al.*’s proposal. Quoting Greenland *et al.* (2016), Mayo asserts that whether $p$-values exaggerate the evidence against the null hypothesis depends on the philosophical background within which such claim is made: It may seem so from a Bayesian perspective, but both from a frequentist and an error statistics perspective, it is the Bayes factor which exaggerates the evidence (also 2017e). “I find it especially troubling”—she continues—“to space an error statistician…ought

---

**Figure 1.** The engine of science (conceptual illustration).
to use a Bayes Factor as the future gold standard for measuring his error statistical tool…even though Bayes Factors don’t control or measure error probabilities” (2017b). Furthermore, Mayo pinpoints the (old) fallacy of transposing the conditional, whereby the (error) probability of a test is confused with the (posterior) probability of a belief (also Trafimow et al., 2017). And despite “60 years (sic) old…demonstrations [showing] that with reasonable tests and reasonable prior probabilities, the disparity vanishes…they still mean different things” (2017c). In particular, “in an adequate account [of severity testing], the improbability of a claim must be distinguished from its having been poorly tested. (You need to be able to say things like, ‘it’s plausible, but that’s a lousy test of it.’)” (2017d). And “the method for such checking is significance tests!” (2017b).

From Methodology, a large number of critics have faulted Benjamin et al.’s disregard of what is really affecting replication in order to focus on a relatively minor issue. Mayo, for example, lists biasing selection effects, cherry-picking, multiple testing, hunting for significance, optional stopping, violated statistical assumptions, missing or irrelevant statistical-substantive links, and questionable measurements as the main culprits in flawed findings and lack of reproducibility (2017a–e; also Gelman & McShane, 2017; Gelman, 2017b). Lakens et al. (2017) add lack of experimental redundancy, logical traps, research opacity, and poor accounting of sources of error, as well as the risks of reduced generalisability and research breadth were Benjamin et al.’s proposal to succeed. Methodological concerns were also raised by Amrhein & Greenland (2017); Black (2017); Byrd (2017); Chapman (2017); Crane (2017); Ferreira & Henderson (2017); Greenland (2017); Hamlin (2017); Kong (2017); Lew (2017); Llewelyn (2017); Martin (2017); McShane et al. (2017); Passin (2017); Steltenpohl (2017); Trafimow et al. (2017); Young (2017); Zollman (2017); and Morey (2017). Some researchers even propose the use of preregistration as a way of minimizing above problems (Hamlin, 2017; Llewelyn, 2017; van der Zee, 2017).

The pseudoscientific NHST element is the cornerstone of Benjamin et al.’s proposal, as it mixes Fisher’s tests of significance with Neyman-Pearson’s alternative hypotheses, Type II errors and power estimation, and with Bayesian prior probabilities, in order to argue about the false positive rate as the culprit for the lack of replicability. Indeed, the false positive rate argument has been heavily criticized by several authors: Mayo (2017b); also Byrd, 2017) chiefly derides as questionable the prior probabilities given to the hypotheses; Colquhoun (2017) denounces the credibility of Benjamin et al.’s false positive rate calculation for some of the prior probabilities used, which he sees as a rather liberal rate; Chapman (2017); also Byrd, 2017) claims that the prior probability ratios for the hypotheses are not realistic; Llewelyn (2017) raises an argument on prior probabilities and their combination with accumulated evidence in order to estimate the probability of future replication, not finding a more restrictive level of significance an adequate antecedent for increasing such probability; Krueger (2017) adds that, under the new proposal, the proportion of false negatives rises faster than the speed at which false positives drops, leading Phaf (2017) to wonder whether, because of it, “it makes more sense to increase, rather than reduce, the significance level for novel findings”; Kong (2017) states that the false positive rate formula is misleading and finds the expected false positive rate when testing the conventional 5% threshold.

Furthermore, Perezgonzalez (2017) argued that the misinterpretation of p-values as evidence in favour or against a hypothesis has more to do with the pseudoscientific use of NHST than with frequentist testing proper (also Mayo, 2017b; McShane et al., 2017; Amrhein & Greenland, 2017). As Benjamin et al.’s proposal is made within the pseudo-philosophical argument of NHST (e.g., confusing statistical and substantive significance; Mayo, 2017c), a lower threshold of significance does not improve such ‘magical thinking’ (also Argamon, 2017; Diebold, 2017; Greenland, 2017; Krueger, 2017; Phaf, 2017). Lakens et al. (2017); also Morey, 2017; O’Rourke, 2017) equally warn that the proposal exaggerates the focus on single p-values in scientific practice, education, decision making and publication.

Unsupported conclusions are another consequence of pseudoscientific thinking, and McShane et al. (2017) and Amrhein & Greenland (2017) highlight the risks of studies published under the new banner exaggerating effect sizes and overconfidence in significant results while discounting non-significant findings. De Brigard (2017) argues, instead, for the need to be humbler and not generalize to populations, irrespective of statistical significance, effects that may only be present in the sample.

The descriptive data analysis element was briefly touched upon by Mayo (2017a)—who recommends abandoning significance testing in favour of inferring the magnitudes that are well (or poorly) indicated by the data (a.k.a., CIs)—and Lew (2017), Phaf (2017), and McShane et al. (2017)—whom argue for the need to interpret evidence in the context of other findings and theoretical models in lieu of NHST.

Regarding data testing, Mayo (2017a); Mayo (2017b) and Perezgonzalez (2017) have no particular problem with a lower level of significance although both reckon Benjamin et al.’s proposal does not address anything in particular with frequentist testing proper. Mayo equally sees such lowering unnecessary as “you might not want to be too demanding before claiming evidence of [null] model violation” (2017b), while the “lack of replication is effectively uncovered thanks to statistical significance tests” (2017e). Young (2017) reminds us of the appropriate use of significance tests in experimental contexts. McShane et al. (2017) also argues the credibility of uniformly most powerful priors and tests and the loss of its connection to Bayesianism as for justifying Benjamin et al.’s proposal (but see Wagenmakers & Gronau’s counter-argument, 2017).

Other authors recommend continuing using p-values either as part of frequentist tests proper (Bates, 2017) or without significance testing, the latter including Colquhoun (2017); Mayo (2017b, who also proposes estimating false positive rates), Greenland (2017, who proposes transforming them
to bits of information against the null hypothesis), Diebold (2017); Funder (2017); Lakens et al. (2017, who recommend justifying whatever thresholds may be used at the design stage), McShane et al. (2017, who recommend using them as descriptive among other neglected factors—but see Wagenmakers & Gronau’s, 2017, counter-argument, and Gelman’s, 2017a, counter-counter-argument), and Amrhein & Greenland (2017). Argamon (2017) suggests substituting Bayesian statistics, instead.

Addressing replication directly, Chapman (2017) and Trafimow et al. (2017) point out that the problem with replication is not too many false positives but insufficient power. Krueger (2017; also McShane et al., 2017, Trafimow et al., 2017) chides Benjamin et al. for the incoherence of considering replication as order-dependent and inverting the exploratory-confirmatory nature of replication by proposing to make the former more difficult to achieve and the latter more liberal. He further chides them on whether they would disallow their own past findings at the 5% threshold. Lakens et al. (2017; also Crane, 2017) found biases in the analyses done by Benjamin et al., and conclude that there is no [Bayesian] evidence that a lower level of significance improves replicability. They also warn of the risks of fewer availability of replication studies were the proposal to succeed (also Greenland, 2017); Gelman (2017b); Hamlin (2017); Ickes (2017); Kong (2017); Roberts (2017), and Trafimow et al. (2017) propose to stop the proverbial beating around the bushes and perform direct replications as a straightforward measure of replicability.

The hypothesis testing element (namely full Bayesian hypothesis testing) has been seldom touched upon but briefly by Llewelyn (2017), who extended the need for prior probabilities and even more accumulated evidence as pre-requisite for estimating the probability of hypotheses being ‘true’.

In conclusion, despite Benjamin et al.’s best intentions, their proposal only reinforces the NHST pseudoscientific approach with a stricter bright-line threshold for claiming evidence it cannot possibly claim—irrespective of how well it calibrates with Bayes factors under certain conditions—while dismissing potential consequences such as the stifling of research and its publication, an increase in false negatives, an increase in misbehaviour, and a continuation of pseudoscientific practices. Furthermore, theirs is a frequentist solution many of their Bayesian proponents do not even believe in, born out of a false belief of lacking equally simple but better-suited alternatives (Machery, 2017).

In reality, an equally simple and better suited alternative exists, perhaps the only one the authors could be entitled to make from their Jeffreysian perspective, hence our own recommendation: Use JASP, the stand-alone, free-to-download, R-based statistical software with user-friendly interface, developed by Wagenmakers’ team (https://jasp-stats.org). JASP allows for analysing the same dataset using frequentist (Fisher’s tests of significance) and Bayesian tools (Jeffrey’s Bayes factors) without necessarily forcing a mishmash of philosophies of science. JASP allows for the Popperian (also Meehl’s, Mayo’s) approach to severely testing a null hypothesis but also to ascertain the credibility of tests based on the observed data. JASP allows researchers to qualitatively calibrate their significance tests to Bayes factors but also the possibility of calibrating Bayes factors to frequentist tests so as to ascertain their error probabilities. JASP is an existing simple and easy application that shows two interpretations of the same data, which aligns well with the training undertaken by frequentist and Jeffreysian researchers alike while allowing them the opportunity to learn the alternative philosophy—irrespective of which they will choose for publication—and may as well help diminish the misinterpretation of results and the reaching of unsupported conclusions. In so doing, it may even help improve replicability.

Competing interests
No competing interests.

Grant information
This publication was financially supported by Massey University Research Fund (MURF) Mini Grant 2017 for Open Access publication, Massey University, New Zealand.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Argamon SE: New “p < 0.005” standard considered harmful [Web log comment]. 2017. Reference Source
Bates T: Changing the default p-value threshold for statistical significance ought not to be done, and is the least of our problems [Web log post]. 2017. Reference Source
Chapman P: "A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Colquhoun D: "A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Crane H: Why "Redefine Statistical Significance” will not improve reproducibility and could make the replication crisis worse. PayAfxiv Preprints. 2017. Publisher Full Text
De Brigard F: Should we redefine statistical significance. A brains blog roundtable [Web log comment], 2017. Reference Source
Funder D: Thresholds [Web log post], 2017. Reference Source
Funder D: A response to some comments on “Abandon Statistical Significance” [Web log post], 2017a. Reference Source
Funder D: When considering proposals for redefining or abandoning statistical significance, remember that their effects on science will only be indirect! [Web log post], 2017b. Reference Source
Gelman A: A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017a. Reference Source
Greenland S: "A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Ickes W: Thresholds [Web log comment], 2017. Reference Source
Lew M: "A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Liewen H: "A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Martin S: Response to some comments on “Abandon Statistical Significance” [Web log comment], 2017. Reference Source
Mayo D: “A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log post], 2017a. Reference Source
Mayo D: “A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017b. Reference Source
Mayo D: Should we redefine statistical significance? A brains blog roundtable [Web log comment], 2017c. Reference Source
Mayo D: Going round and round again: a roundtable on reproducibility & lowering p-values [Web log post], 2017e. Reference Source
Moore RD: When the statistical tail wags the scientific dog [Web log post], 2017. Reference Source
O'Rourke K: “A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Passin T: "A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Perezgonzalez JD: Better Science - The call for significance of 5% (0.005) [Video file], 2017. Reference Source
Phaf RH: Comment on redefine statistical significance [Web log post], 2017. Reference Source
Resnick B: What a nerdy debate about p-values shows about science — and how to fix it [Web log post], 2017. Reference Source
Roberts B: Thresholds [Web log comment], 2017. Reference Source
Savehnia T: “A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
van der Zee T: Arguing for all the wrong reforms [Web log post], 2017. Reference Source
Wagenmakers EJ, Gronau Q: Redefine statistical significance Part IX: Gelman and Robert join the fray, but are quickly chased by two kangaroos [Web log post], 2017. Reference Source
Young S: “A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment], 2017. Reference Source
Open Peer Review

Current Referee Status: ✔

Version 1

Referee Report 14 December 2017

doi:10.5256/f1000research.14537.r28961

Patrizio Tressoldi
Department of General Psychology, University of Padova, Padova, Italy

This is a timely and well organized synthesis of the debate ignited by Benjamin et al.’s, 2017 influential paper which is still going on. My only suggestions are the followings:
- to add a legend to the acronyms in Figure 1, e.g. ES, CI, etc.;
- to correct the statement that the JASP software allows only a Jeffreys’s Bayes Factor given that in the present .8.4.0 version, the user can also choose informed priors based on Cauchy, Normal or t distributions;
- I think that a final synoptic table related to the information offered by the different statistical indices, i.e., p value, CI, Bayes Factor $H_1/H_0$, Effect Size, etc., could help the readers acknowledge the different informational value among them and consequently to choose those ones which answer better the questions they ask their data.

Is the topic of the opinion article discussed accurately in the context of the current literature? Yes

Are all factual statements correct and adequately supported by citations? Yes

Are arguments sufficiently supported by evidence from the published literature? Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments? Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com