OPINION ARTICLE

REVISED Retract *p* < 0.005 and propose using JASP, instead [version 2; referees: 3 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 Ibañez 21, Valencia, 46010, Spain

v2 **First published:** 12 Dec 2017, **6**:2122 (doi: 10.12688/f1000research.13389.1)
Latest published: 16 Feb 2018, **6**:2122 (doi: 10.12688/f1000research.13389.2)

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 Jeffresian 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.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees		
	1	**2**	**3**
REVISED **version 2** published 16 Feb 2018			
version 1 published 12 Dec 2017	✓ report	✓ report	✓ report

1 **Patrizio Tressoldi**, University of Padova, Italy

2 **Juan Carro Ramos**, University of Salamanca, Spain

3 **Stephen Senn**, Luxembourg Institute Of Health, Luxembourg
University of Sheffield, UK

Discuss this article

Comments (2)

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 2; referees: 3 approved]** *F1000Research* 2018, **6**:2122 (doi: 10.12688/f1000research.13389.2)

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.

First published: 12 Dec 2017, **6**:2122 (doi: 10.12688/f1000research.13389.1)

REVISED **Amendments from Version 1**

Minor changes incorporating reviewers' recommendations:

- [1] The legend in Figure 1 now defines the acronyms in the figure.
- [2] A new reference to Perezgonzalez (2015) now implies that the pseudoscientific label attached to the NHST element (Figure 1) follows from the rhetoric in such reference.
- [3] A second note clarifies that JASP also allows to use Cauchy, Normal and t-distributions as informed priors for the alternative hypothesis.
- [4] There is a new acknowledgement section.
- [5] A worked-out example on the parallel use of Fisher's and Jeffreys's tests has been appended (Supplementary file 1).

See referee reports

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 constrains 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 meriting 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 endors[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.

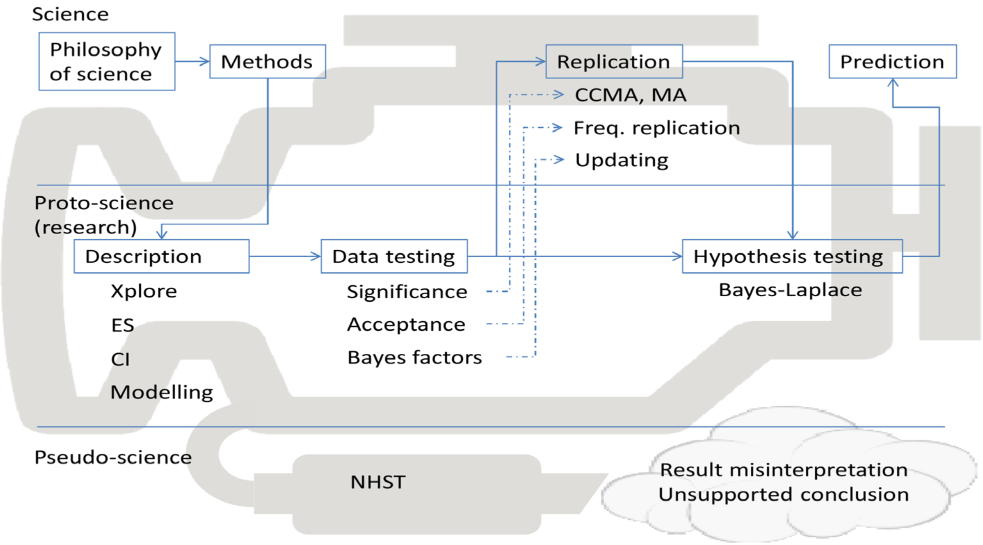

Figure 1. The engine of science (conceptual illustration). Xplore = exploratory data analysis; *ES* = effect size; *CI* = confidence interval, credible interval; *CCMA* = continuous cumulating meta-analysis; *MA* = meta-analysis; *Freq. replication* = frequentist replication; *NHST* = null hypothesis significance testing (as in Perezgonzalez, 2015).

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 spoze an error statistician…ought 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** (as conceptualized in Perezgonzalez, 2015) 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 Amrhein & Greenland, 2017; Mayo, 2017b; McShane *et al.*, 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 a [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 descriptives 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 Jeffresian 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[1]). JASP allows for analysing the same dataset using frequentist (Fisher's tests of significance) and Bayesian tools (Jeffreys's Bayes factors[2]) without necessarily forcing a mishmash of philosophies of science (e.g., see Supplementary File 1). 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 were disclosed.

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.

Acknowledgements

We wish to thank the reviewers, P. Tressoldi, J. C. Ramos, and S. Senn for their reviews and suggestions for improvement.

[1] Worthy alternatives are Morey's R package for Bayes factors (http://bayesfactorpcl.r-forge.r-project.org), Morey's Bayes factor calculator (http://pcl.missouri.edu/bayesfactor), and Dienes's Bayes factor calculator (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm). Even better would be to skip comparing models altogether and go for full Bayesian hypothesis testing (e.g., Kruschke, 2011). Yet none of those alternatives surpass, at present, JASP's interface and its flexibility for parallel analysis.

[2] The latest versions of JASP (e.g. 0.8.4.0; 0.8.5.1) also allows for choosing informed priors based on Cauchy, Normal, and *t*-distributions.

Supplementary material

Supplementary file 1: **Parallel statistical analysis via JASP**. Example of data analysis using both frequentist (Fisher) and Bayesian (Jeffreys) approaches.

Click here to access the data.

References

Amrhein V, Greenland S: **Remove, rather than redefine, statistical significance.** *Nat Hum Behav.* 2017.
Publisher Full Text

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 be done, and is the least of our problems [Web log post].** 2017.
Reference Source

Benjamin DJ, Berger JO, Johannesson M, *et al.*: **Redefine statistical significance.** *PsyArXiv Preprints.* 2017a.
Publisher Full Text

Benjamin DJ, Berger JO, Johannesson M, *et al.*: **Redefine statistical significance.** *Nat Hum Behav.* 2017b; **1**: 0189.
Publisher Full Text

Black J: **Thresholds [Web log comment].** 2017.
Reference Source

Byrd J: **"A megateam of reproducibility-minded scientists" look to lowering the p-value [Web log comment].** 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 "Redefining Statistical Significance" will not improve reproducibility and could make the replication crisis worse.** *PsyArXiv Preprints.* 2017.
Publisher Full Text

De Brigard F: **Should we redefine statistical significance. A brains blog roundatble [Web log comment].** 2017.
Reference Source

Diebold FX: **New p-value thresholds for statistical significance [Web log post].** 2017.
Reference Source

Easwaran K: **Should we redefine statistical significance. A brains blog roundatble [Web log comment].** 2017.
Reference Source

Ferreira F, Henderson JM: **Defending .05: It's not enough to be suggestive [Web log post].** 2017.
Reference Source

Funder D: **Thresholds [Web log post].** 2017.
Reference Source

Gelman A: **Response to some comments on "Abandon Statistical Significance" [Web log post].** 2017a.
Reference Source

Gelman A: **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, McShane B: **Should we redefine statistical significance. A brains blog roundatble [Web log comment].** 2017.
Reference Source

Greenland S: **"A megateam of reproducibility-minded scientists" look to lowering the p-value [Web log comment].** 2017.
Reference Source

Greenland S, Senn SJ, Rothman KJ, *et al.*: **Statistical tests, *p* values, confidence intervals, and power: a guide to misinterpretations.** *Eur J Epidemiol.* 2016; **31**(4): 337–350.
PubMed Abstract | Publisher Full Text | Free Full Text

Hamlin K: **Should we redefine statistical significance? A brains blog roundtable [Web log comment].** 2017.
Reference Source

Ickes W: **Thresholds [Web log comment].** 2017.
Reference Source

Kong X: **Redefine statistical significance? Let's just do science in a scientific way [Web log post].** 2017.
Reference Source

Krueger JI: **Fear of false positives [Web log post].** 2017.
Reference Source

Kruschke JK: **Doing Bayesian data analysis. A tutorial with R and BUGS.** Amsterdam, The Netherlands: Academic Press. 2011.
Reference Source

Lakens D, Adolfi FG, Albers CJ, *et al.*: **Justify your alpha: A response to "Redefine statistical significance".** *PsyArXiv Preprints.* 2017.
Publisher Full Text

Lew M: **"A megateam of reproducibility-minded scientists" look to lowering the p-value [Web log comment].** 2017.
Reference Source

Llewelyn H: **"A megateam of reproducibility-minded scientists" look to lowering the p-value [Web log comment].** 2017.
Reference Source

Machery E: **Should we redefine statistical significance. A brains blog roundatble [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: **New venues for the statistics wars [Web log post].** 2017d.
Reference Source

Mayo D: **Going round and round again: a roundtable on reproducibility & lowering p-values [Web log post].** 2017e.
Reference Source

McShane BB, Gal D, Gelman A, *et al.*: **Abandon statistical significance.** 2017.
Reference Source

Morey 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: **Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing.** *Front Psychol.* 2015; **6:** 223.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Savehn T: **"A megateam of reproducibility-minded scientists" look to lowering the p-value [Web log comment].** 2017.
Reference Source

Steltenpohl CN: **The littlest p: redefining statistical significance [Web log post].** 2017.
Reference Source

Trafimow D, Amrhein V, Areshenkoff CN, *et al.*: **Manipulating the alpha level cannot cure significance testing. Comments on "Redefine statistical significance".** *PeerJ Preprints.* 2017; **5**: e3411v1.
Publisher Full Text

van der Zee T: **Arguing for all the wrong reforms [Web log post].** 2017.
Reference Source

Wagenmakers EJ: **Redefine statistical significance Part I: Sleep trolls & red herrings [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

Zollman K: **Should we redefine statistical significance? A brains blog roundtable [Web log comment].** 2017.
Reference Source

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 1

Referee Report 12 January 2018

doi:10.5256/f1000research.14537.r28963

✓

Stephen Senn [1,2]
[1] Competence Center for Methodology and Statistics, Luxembourg Institute Of Health, Strassen, Luxembourg
[2] School of Health and Related Research, University of Sheffield, Sheffield, UK

I consider this is a valid contribution to the debate on this topic but like many such contributions it reflects a particular viewpoint, my own contributions[1] in this field are no exception, and like many contributors, and, again, I am probably also guilty, there is minimal recognition that their own views are not reasonable from other perspectives. In particular (see below) they are over-fond of the adjective *pseudoscientific* and their use of it without further justification is, in my view, a form of pseudo-argumentation.

Particular objections that I have are the following.

1. They state: "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." This, in my opinion, is an unsubstantiated and misleading jibe. First, although this claim is often made, it is not correct that NHST is some sort of monstrous hybrid of Neyman-Pearson (N-P) and Fisherian views. It very naturally uses common elements that belong in both.(2) For instance, Fisher, although usually associated with P-values, introduced the habit of tabulating percentage points in *Statistical Methods for Research Workers*. Secondly, as Lehmann, a classic proponent of the N-P view explained in *Testing Statistical Hypotheses*, P-values are a practical way of allowing scientist who may not agree on appropriate Type I error rates to share results efficiently. Furthermore, although I have little enthusiasm for what Benjamin et al are doing, I think that fact that NHST seems to represent some sort of a fusion of Fisher and NP (a point that is much exaggerated and of little relevance) is the least important aspect of what they are doing. If they had rejected everything the Fisher proposed and described what they were doing as trying to convince diehard N-P acolytes that the common 5% level was better changed to 0.5%, it would have zero technical effect on what they are proposing and would not make it either better or worse. Whatever problems or virtues there are with the Benjamin et al proposal, the Fisher, N-P fusion is a complete red herring.
2. Will JASP and Bayes Factors (BF) really save statistical inference? I doubt it. One of the problems with the Jeffreys approach is the huge influence that the lump of probability on the 'null' has on the posterior inference. NHST is far less dependent on this and, in a context in which I work, that of drug development, the practical challenge is to avoid recommending a treatment that is worse. it is a strength of NHST that one does not have to worry too much about whether one is talking about precise(or point hypotheses) on the one hand or dividing ones on the other. Of course, Bayesians could regard this as not a strength but a weakness, on the lines that if it *does* make a difference to

Bayesian posterior statements (and the difference can be enormous) it *ought* to in the NHST context. (It is a very common habit of Bayesians to abrogate the right of judging other systems by theirs.) However, I also note that amongst the Bayesian commentators on the recent ASA statement, some dismissed precise hypotheses as being completely irrelevant.

3. Like many commentators, they appear to accept much of the replication crisis at face value. However, there are four aspects here that are important. i) Gaming can effect any system (as they recognise) ii) Much of the discussion has taken it as obvious that failure to replicate is a problem of the original study *only*, rather than a shared problem. However, it can only be the former if the second study is perfect and that includes not being subject to sampling variation itself. What thus becomes important is to judge how well studies of infinite size are predicted, not those that just happen to be the same size as the original (2). iii) It assumes that what is important is the extent to which the P-value predicts the P-value but there is no claim under NHST that this is so. What is relevant is the extent to which it predicts the sign of the true difference Furthermore, to shackle the Bayesian approach with the same standard (what's sauce for the goose is sauce for the gander) would require that a high BF predicted a further high BF (*using data from the subsequent study only*) with high probability. JASP on its own will not deliver this. iv) Finally, if the only replication that matters is failure to replicate 'significance' then the proposal of Benjamin et al is nowhere near radical enough. 'Significance' must never be granted. On the other hand, if false negatives are also a problem, and not just false positives, then the solution has to be rethought from the beginning. (To be fair to the authors, they do cite Krueger's, 2017 warning about this but it is unclear to me to what extent they take it on board in what they are proposing.)

However, putting aside my objections to some of the polemics of the article, the idea of supplementing P-values by other inferential statistics strikes my as being sensible and I certainly approve their suggestion of keeping Bayes factors separate from P-values. In fact, I have no objection to Bayes Factors provided that their very tentative nature is recognised and I certainly sign up to the idea of the utility of having different ways of looking at data[34].

References

1. Senn S: Two cheers for P-values?. *J Epidemiol Biostat*. 2001; **6** (2): 193-204; discussion 205 PubMed Abstract

2. Senn S: A comment on replication,p-values and evidence S.N.Goodman,Statistics in Medicine 1992;11:875-879. *Statistics in Medicine*. 2002; **21** (16): 2437-2444 Publisher Full Text

3. Gigerenzer G, Marewski J: Surrogate Science. *Journal of Management*. 2015; **41** (2): 421-440 Publisher Full Text

4. Senn, SJ: You may believe you are a Bayesian but you are probably wrong.Rationality, Markets and Morals. 2011. 48-66.

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

Partly

Are all factual statements correct and adequately supported by citations?

Partly

Are arguments sufficiently supported by evidence from the published literature?

Partly

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

Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Biostatistics, Drug Development, Statistical Inference

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.

Referee Report 21 December 2017

doi:10.5256/f1000research.14537.r28966

✓ **Juan Carro Ramos**
University of Salamanca, Salamanca, Spain

The authors discuss the article by Benjamin et al., 2017 and their proposal of p <.005 as a possible solution to the lack of research reproducibility due to the high rate of false positives in the literature. The article reviews the current debate topics on NHST and contributes to the reflection on how to advance towards a Science that allows accumulating valid scientific knowledge. Suggestions: Write down at the end of the article a simple example where the result is interpreted with NHST and with the Bayes factor and the utility of JASP is presented.

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.

Author Response 22 Dec 2017

Jose Perezgonzalez, Massey University, New Zealand

Thank you very much for your prompt review. I will append an example to address your suggestion as soon as a current research using a double frequentist/Jeffresian approach is concluded.

Competing Interests: No competing interests.

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 $_{H1/H0}$, 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.

Author Response 22 Dec 2017

Jose Perezgonzalez, Massey University, New Zealand

Thank you very much for your prompt review. I will upload a new version incorporating two of the suggestions made (plus proper acknowledgement). I will append an example to address the third suggestion as soon as a current research using a double frequentist/Jeffresian is finished.

Competing Interests: No competing interests.

Discuss this Article

Version 1

Author Response 22 Dec 2017

Jose Perezgonzalez, Massey University, New Zealand

Thanks for your comment and links. Indeed, the contribution you made to Mayo's blog back in July (Colquhoun, D., 2017, July 26; “A megateam of reproducibility-minded scientists” look to lowering the p-value [Web log comment]. Retrieved from https://errorstatistics.com/2017/07/25/a-megateam-of-reproducibility-minded-scientists-look-to-lowering-the-p- was the first questioning the entire foundation of Benjamin et al.'s analysis. I think it is very illustrative and prompts a reflection to the quick fix-all solutions that appear here and there in the literature.

I believe Benjamin et al. would have made more sense proposing a double analysis frequentist/Jeffresian (it prompts you to know what's the knowledge each approach warrants) than proposing a reduction of significance just because it calibrates with Bayes Factors under certain conditions (which suggests *p* -values and *BF* are different tools measuring the same thing, like the Celsius and Fahrenheit temperature scales).

Of course, this doesn't mean that the double analysis is the quick fix-all solution, though. I think Gelman (or is it Senn?) has also been wondering why *BF* when we could go for a full Bayesian analysis instead. I also understand Mayo's point focused on error statistics. Yet all those perspective are focusing on the problem from different points-of-views, and such points-of-view need to be understood in order to make sense of each claim.

In a nutshell, next time I go to the doctor for a test, I want that test to be pretty good at reducing false positives (Fisher). I also want it to be pretty good at reducing false negatives (Neyman-Pearson; BF). And not only that, I want such test validated under the frequentist approach. And yet, I don't want my diagnosis been determined by the test. Indeed, I want my doctor to make the diagnosis (Bayes). Hopefully, she will recommend further tests or follow ups (some form of frequentist replication, or Bayesian updating). I need (and want) all those approaches!

Competing Interests: No competing interests.

Reader Comment 21 Dec 2017

David Colquhoun, UCL, UK

I fear that the use of Bayes; factors alone is not the answer.

I'll restrict my comments to the case where we are testing a point null hypothesis because in this case Bayes' factors are just likelihood ratios, so much easier to understand.

The problem with likelihood ratios is that their use could result in the execution of an innocent person. See http://www.dcscience.net/2016/03/22/statistics-and-the-law-the-prosecutors-fallacy/ (in particular, the island problem).

I take it that the main aim of a statistical test is to avoid making a fool of yourself by claiming that an effect exists when in fact your observation is just chance. The probability of this happening is the *false positive risk (FPR).* As you have explained, a large proportion of researchers still think that this is what the p-value tells you, quite wrongly,

The problem lies in the fact that, in order to calculate the FPR you need to know something about the prior probability that the effect is real (as well as the likelihood ratio). Such information is very rare, and people will disagree about it.

The best way to circumvent this dilemma seems to me to follow a suggestion of Robert Matthews and to calculate the prior probability that would be needed to reduce your FPR to, say, 5% (that's what so many people think, mistakenly, that p = 0.05 achieves). The calculations are described in
http://rsos.royalsocietypublishing.org/content/4/12/171085
and they can be done without using our R scripts using our web calculator, at
http://fpr-calc.ucl.ac.uk/ (but please check the Notes tab before using it, if you haven't read the paper).

For example, one finds that, for a well-powered experiment in which you observe p = 0.05, you'd have to be 87% sure that there was a real effect before you did the experiment in order to achieve an FPR of 5%. It would be up to you to persuade editors and readers that this was a reasonable assumption. Clearly it isn't.

If you observe p = 0.005 then you'd need a prior probability of 0.4 in order to achieve an FPR of 5%. Again it is up to you to persuade people that it's a reasonable assumption. It might well be deemed reasonable for a plausible hypothesis, but if you were testing a homeopathic pill, it would be absurdly high.

The likelihood ratio can be regarded as the odds on there being a real effect when the prior odds are 1, i.e. a prior probability of 0.5. Since it is hardly ever justifiable to assume a prior larger than this. the likelihood ratio gives a measure of the minimum FPR -that for a plausible hypothesis. For example, if you observe p = 0.05, then the minimum FPR is 27%. If you observe p = 0.005, the minimum FPR is 3.4%.

But if the hypothesis were implausible, with a prior probability of 0.1, then p = 0.05 would correspond to an FPR of 76% -disastrously high. And if you observed p = 0.005, the FPR would still be 24%. Even p = 0.001 gives an FPR greater than 5% in this case: to achieve an FPR of 5% you'd need to observe p = 0.00045.

So I would conclude that, awkward though it is, it is asking for trouble to rely on likelihood ratios alone. You simply can't ignore prior probabilities just because you don't know them. Hence my suggestion to use the reverse Bayes approach, i.e. to calculate the prior needed to achieve an acceptable FPR. Of course people may disagree about how plausible this prior is. Sadly there is no way to do inductive inference in an entirely objective way,
e.g, https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

Even if you are willing to believe that the chance your hypothesis is right, is as high as 50:50, so likelihood ratios would imply the FPR, it still seems better to express the result as an FPR, because probabilities are a much more familiar measure than likelihood ratios.

Competing Interests: Only having written about the topic myself