Manipulating the Alpha Level Cannot Cure Significance Testing
Comments on "Redefine Statistical Significance"

David Trafimow[1¤], Valentin Amrhein[2], Corson N. Areshenkoff[3], Carlos Barrera-Causil[4], Eric J. Beh[5], Yusuf Bilgiç[6], Roser Bono[7], Michael T. Bradley[8], William M. Briggs[9], Héctor A. Cepeda-Freyre[10], Sergio E. Chaigneau[11], Daniel R. Ciocca[12], Juan Carlos Correa[13], Denis Cousineau[14], Michiel R. de Boer[15], Subhra Sankar Dhar[16], Igor Dolgov[1], Juana Gómez-Benito[7], Marian Grendar[17], James Grice[18], Martin E. Guerrero-Gimenez[12], Andrés Gutiérrez[19], Tania B. Huedo-Medina[20], Klaus Jaffe[21], Armina Janyan[22,23], Ali Karimnezhad[24], Fränzi Korner-Nievergelt[25], Koji Kosugi[26], Martin Lachmair[27], Rubén Ledesma[28], Roberto Limongi[29], Marco Tullio Liuzza[30], Rosaria Lombardo[31], Michael Marks[1], Gunther Meinlschmidt[32,33,34], Ladislas Nalborczyk[35,36], Hung T. Nguyen[37], Raydonal Ospina[38], Jose D. Perezgonzalez[39], Roland Pfister[40], Juan José Rahona[27], David A. Rodríguez-Medina[41], Xavier Romão[42], Susana Ruiz-Fernández[27], Isabel Suarez[43], Marion Tegethoff[44], Mauricio Tejo[45], Rens van de Schoot[46,47], Ivan Vankov[22], Santiago Velasco-Forero[48], Tonghui Wang[49], Yuki Yamada[50], Felipe C. M. Zoppino[12] & Fernando Marmolejo-Ramos[51]

Authorship order is alphabetical, except for the first, second, and last author.

1. Department of Psychology, New Mexico State University, U.S.A.
2. Zoological Institute, University of Basel, Basel, Switzerland
3. Centre for Neuroscience Studies, Queens University, Ontario, Canada
4. Faculty of Applied and Exact Sciences, Metropolitan Technological Institute, Medellín, Colombia
5. School of Mathematical & Physical Sciences, University of Newcastle, Australia
6. Department of Mathematics, State University of New York at Geneseo, U.S.A.
7. Quantitative Psychology Unit, Faculty of Psychology, University of Barcelona, Barcelona, Spain
8. Department of Psychology, Faculty of Arts, University of New Brunswick, Canada
9. Independent Researcher, New York, U.S.A.
10. School of Psychology, Benemérita Universidad Autónoma de Puebla, México

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

11. Center for Cognition Research, CINCO, School of Psychology, Universidad Adolfo Ibáñez, Santiago, Chile
12. Oncology Laboratory, IMBECU, CCT CONICET Mendoza, Argentina
13. School of Statistics, Faculty of Sciences, National University of Colombia, Medellín, Colombia
14. School of Psychology, University of Ottawa, Ottawa, Canada
15. Department of Health Sciences, Vrije Universiteit Amsterdam and Amsterdam Public Health research institute, Amsterdam, The Netherlands
16. Department of Mathematics and Statistics, IIT Kanpur, India
17. Biomedical Center Martin, Jessenius Faculty of Medicine, Comenius University, Slovakia, and Institute of Measurement Science, Slovak Academy of Sciences, Slovakia
18. Department of Psychology, Oklahoma State University, U.S.A.
19. Faculty of Statistics, Saint Thomas University, Colombia
20. Department of Allied Health Sciences, College of Health, Agriculture, and Natural Resources, University of Connecticut, U.S.A.
21. Simón Bolívar University, Caracas, Venezuela
22. Department of Cognitive Science and Psychology, New Bulgarian University, Sofia, Bulgaria
23. National Research Tomsk State University, Tomsk, Russia
24. Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, Canada
25. Oikostat GmbH, Ettiswil, Switzerland
26. Department of Education, Yamaguchi University, Japan
27. Leibniz Knowledge Media Research Center, Multimodal Interaction Lab, Tübingen, Germany
28. Consejo Nacional de Investigaciones Científicas y Técnicas, and Universidad Nacional de Mar del Plata, Argentina
29. School of Pedagogy, Pontifical Catholic University of Valparaíso, Chile
30. Department of Medical and Surgical Sciences, “Magna Graecia” University of Catanzaro, Catanzaro, Italy
31. Economics Department, University of Campania “Luigi Vanvitelli”, Capua, Italy

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

32. Department of Psychosomatic Medicine, University Hospital Basel and University of Basel, Basel, Switzerland
33. Division of Clinical Psychology and Cognitive Behavioral Therapy, International Psychoanalytic University, Berlin, Germany
34. Division of Clinical Psychology and Epidemiology, Department of Psychology, University of Basel, Basel, Switzerland
35. Univ. Grenoble Alpes, CNRS, LPNC UMR 5105, F-38000, Grenoble, France
36. Department of Experimental Clinical and Health Psychology, Ghent University, Belgium
37. Faculty of Economics, Chiang Mai University, Thailand
38. Department of Statistics, Computational Statistics Laboratory (CAST), Universidade Federal de Pernambuco, Brazil
39. Business School, Massey University, New Zealand
40. Department of Psychology III, University of Würzburg, Germany
41. School of Psychology, National Autonomous University of Mexico, Mexico
42. CONSTRUCT-LESE, Faculty of Engineering, University of Porto, Portugal
43. Department of Psychology, Universidad del Norte, Barranquilla, Colombia
44. Division of Clinical Psychology and Psychiatry, Department of Psychology, University of Basel, Basel, Switzerland
45. Facultad de Ciencias Naturales y Exactas, Universidad de Playa Ancha, Valparaíso, Chile
46. Utrecht University, Faculty of Social and Behavioural Sciences, Department of Methods and Statistics, Utrecht, The Netherlands
47. North-West University, Optentia Research Focus Area, Vanderbijlpark, South Africa
48. MINES Paristech, PSL Research University, Centre for Mathematical Morphology, France
49. Department of Mathematical Sciences, New Mexico State University, U.S.A.
50. Faculty of Arts and Science, Kyushu University, Japan
51. School of Psychology, The University of Adelaide, Australia

¤ Corresponding author: David Trafimow, Department of Psychology, New Mexico State University, U.S.A. Email: dtrafimo@nmsu.edu

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

Acknowledgments: GM has been acting as consultant for Janssen Research & Development, LLC. MG acknowledges support from VEGA 2/0047/15 grant. RvdS was supported by a grant from the Netherlands organization for scientific research: NWO-VIDI-45-14-006.

One sentence summary: We argue that depending on *p*-values to reject null hypotheses, including a recent call for changing the canonical alpha level for statistical significance from .05 to .005, is deleterious for the finding of new discoveries and the progress of cumulative science.

Many researchers have criticized null hypothesis significance testing, though many have defended it too (see Balluerka, Gómez, & Hidalgo, 2005 for a review). Sometimes, there is a recommendation that the alpha level be reduced to a more conservative value, to reduce the Type I error rate. For example, Melton (1962), the editor of *Journal of Experimental Social Psychology* from 1950–1962, favored an alpha level of .01 over the typical .05 alpha level. More recently, Benjamin and 71 scientists (2017) recommended shifting to .005—consistent with Melton's comment that even the .01 level might not be "sufficiently impressive" to warrant publication (p. 554). In addition, Benjamin et al. (2017) stipulated that the .005 criterion should be for new findings but were vague about what to do with findings that are not new. Though not necessarily endorsing significance testing as the preferred inferential statistical procedure,[1] Benjamin et al. (2017) did argue that using a .005 criterion would fix much of what is wrong with significance testing. Unfortunately, as we will demonstrate, the problems with significance tests cannot be importantly mitigated merely by having a more conservative rejection criterion, and some problems are exacerbated by adopting a more conservative criterion.

We commence with some claims on the part of Benjamin et al. (2017). For example, they wrote "…changing the *p*-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance." If significance testing—at any *p*-value threshold—is as badly flawed as we will maintain it is (see also Amrhein, Korner-Nievergelt, & Roth, 2017; Greenland, 2017), these reasons are clearly insufficient to justify it.

[1] Many of the authors favor Bayesian procedures.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

Consider another claim: "The new significance threshold will help researchers and readers to understand and communicate evidence more accurately." But if researchers have understanding and communication problems with a .05 threshold, it is unclear how using a .005 threshold will eliminate these problems. And consider yet another claim: "Authors and readers can themselves take the initiative by describing and interpreting results more appropriately in light of the new proposed definition of statistical significance." Again, it is not clear how adopting a .005 threshold will allow authors and readers to take the initiative with respect to better data interpretation. Thus, even prior to a discussion of our main arguments, there is reason for the reader to be suspicious of hasty claims with insufficient support.

With the foregoing out of the way, consider that a basic problem with tests of significance is that the goal is to reject the null hypothesis. This goal seems to demand—if one is a Bayesian—that the posterior probability of the null hypothesis should be low given the obtained finding. But the *p*-value one obtains is the probability of the finding (or a more extreme finding) given the null hypothesis (and the assumptions underlying the test), and one would need to make an invalid inverse inference to draw a conclusion about the probability of the null hypothesis given the finding. And if one is a frequentist, there is no way to traverse the logical gap from the probability of the finding given the null hypothesis to a decision about whether one should accept or reject the null hypothesis (Briggs, 2016; Trafimow, 2017). We accept that, by frequentist logic, the probability of a Type I error really is lower if $p = .005$ than if $p = .05$, all else being equal. We also accept the Bayesian argument by Benjamin et al. (2017) that the null hypothesis is less likely if $p = .005$ than if $p = .05$, all else being equal (although determining *p*-values via Bayes Factors is problematic; see Appendix).[2] Finally, we acknowledge that Benjamin et al. (2017) performed a service for science by further stimulating debate about significance testing. But there are important issues Benjamin et al. (2017) seem not to have considered, discussed in the following sections.

[2] Depaoli and van de Schoot (2017) provided a critique showing how it is possible to abuse Bayesian statistics, and provided potential solutions to such abuse. Konijn, van de Schoot, Winter, and Ferguson (2015) suggested a way to use Bayesian statistics to reduce publication bias.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

Regression and Reliability

Trafimow and Earp (2017) argued against the general notion of setting an alpha level to make decisions to reject or not reject null hypotheses, and the arguments retain their force even if the alpha level is reduced to .005. In some ways, the reduction worsens matters. One problem is that *p*-values have a sampling distribution,[3] as do other statistics (Cumming, 2012). Whether the *p*-value obtained in any experiment passes the alpha level is partly a matter of luck (which *p*-value one happens to sample), with the caveat that large effect and sample sizes, and small variation, should decrease *p*-values. Absent the caveat, the researcher is unlikely to re-sample a *p*-value below a significance threshold upon replication, as there may be many more *p*-values above than below the threshold in the *p*-value distribution. Thus, the phenomenon of regression to the mean suggests that the *p*-value obtained in a replication experiment is likely to regress to whatever the mean *p*-value would be if many replications were performed to obtain a distribution of *p*-values for the experiment. How much regression should occur? That depends on the reliability of *p*-values.

Based on data placed online by the Open Science Collaboration (2015; https://osf.io/fgjvw), Trafimow and de Boer (2017) calculated the correlation between *p*-values obtained in the original cohort of studies with *p*-values obtained in the replication cohort, and obtained the dismal value of .004.[4] Clearly, then, the obtained *p*-value in the original study has little to do with the *p*-value obtained in a replication experiment. The best prediction would be a *p*-value for the replication experiment being vastly closer to the mean of the *p*-value distribution than to the *p*-value obtained in the original experiment. Under the null hypothesis, the lower the *p*-value published in the original experiment (e.g., .005 rather than .05), the greater the amount of distance of the *p*-value from the *p*-value mean, implying increased regression to the mean.[5] Thus, even using the .05 value is problematic, with exacerbation using the .005 value (Amrhein & Greenland, 2017). When studies have low power, it is not rare to obtain large sample effects that

[3] For a test of the difference between two normal means, the *p*-value is uniformly distributed on [0,1] under the point null hypothesis. Under a range alternative hypothesis, the distribution may be unknowable.

[4] There are several possible reasons for the low value. These could include the nonlinear relation between *p*-values and effect sizes, mixing cases where the null hypothesis is true (or close to true) with cases where it is not, publication bias, and imperfect replication methodology, as well as random sampling error.

[5] Recall, the *p*-value distribution under the alternative hypothesis often is not knowable.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

are overestimates, and using the .005 threshold instead of .05 would guarantee that statistically significant results are even larger overestimates of population effect sizes (Button et al. 2013).

In addition, from a measurement point of view, where reliability is a prerequisite for validity, the *p*-value correlation (reliability) of .004 obtained by Trafimow and de Boer (2017) indicates that as a basis for binary decisions, *p*-values are incapable of measuring anything validly, including the strength of the evidence (Fisher, 1925; 1973) or the severity of the test (Mayo, 1996).[6] This could be argued to be a good reason not to use *p*-values at all. Alternatively, the dismal *p*-value reliability as evidenced by the Open Science Collaboration could be attributed, in part, to the publication bias caused by having a publishing criterion (Locascio, 2017a). But if one wishes to make such an attribution, although it provides a justification for using *p*-values in a hypothetical scientific universe where *p*-values are more reliable because of a lack of publication bias, the attribution provides yet another important reason to avoid publishing criteria based on *p*-values.

Type I and Type II Errors

Another disadvantage of using any set criterion level for publication is that the relative importance of Type I and Type II errors might differ across studies within or between areas and researchers (Trafimow & Earp, 2017). Setting a blanket level of either .05 or .005, or anything else, forces researchers to pretend that the relative importance of Type I and Type II errors is constant.[7] Benjamin et al. (2017) pointed out that a few areas of science use very low criterion levels to justify their recommendation to reduce to the .005 level, but this justification seems to tacitly admit that a blanket level across many areas is undesirable. It seems obvious that a wide variety of factors can influence the relative importance of Type I and Type II errors, thereby rendering any blanket recommendation undesirable (indeed Miller & Ulrich, 2016, show how these and other factors have a direct bearing on the final research payoff). These factors may include the clarity of the theory or auxiliary assumptions, practical or applied concerns, or experimental rigor. There is an impressive literature attesting to the difficulties in setting a

[6] "Correcting" the correlation for attenuation due to restriction of range, in the original cohort of studies, increases the correlation to .01, which is still low.

[7] Another problem is that for different sample sizes the same *p*-value may imply a different extent of the evidence against the null hypothesis (Royall, 1986).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

blanket recommendation (e.g., Buhl-Mortensen, 1996; Lemons, Shrader-Frechette, & Cranor, 1997; Lemons & Victor, 2008; Lieberman & Cunningham, 2009; Mudge, Baker, Edge, & Houlahan, 2012; Myhr, 2010; Rice & Trafimow 2010). This argument is not a recommendation that every researcher should get to set her own criterion, as that has obvious problems too (as Trafimow & Earp, 2017, showed).[8] Rather, given that blanket and variable criterion levels both are problematic, it is sensible to dispense with significance testing altogether.

Defining Replicability

Yet another disadvantage pertains to what Benjamin et al. (2017) touted as the main advantage of their proposal, that published findings will be more replicable using the .005 than .05 alpha level. This depends on what is meant by "replicate" (see Lykken, 1968, for some definitions). If one insists on the same alpha level for the original study and the replication study, then we see no reason to believe that there will be more successful replications using the .005 level than using the .05 level. In fact, the statistical regression argument made earlier suggests that the regression issue is made even worse using .005 than using .05. Alternatively, as Benjamin et al. (2017) seem to suggest, one could use .005 for the original study and .05 for the replication study. In this case, we agree that the combination of .005 and .05 will create fewer unsuccessful replications than the combination of .05 and .05, for the initial and replication studies, respectively. However, this comes at a high price in arbitrariness. Suppose that two studies come in at $p < .005$ and $p < .05$, respectively. This would count as a successful replication. In contrast, suppose that the two studies come in at $p < .05$ and $p < .005$, respectively. Only the second study would count, and the combination would not qualify as indicating a successful replication. The arbitrariness of declaring the combination of .005 and .05 as being a successful replication, whereas the combination of .05 and .005 is not, adds to the myriad difficulties researchers have interpreting their data. More generally, insisting that setting

[8] In addition to creating new issues of how researchers should decide on the criteria for each experiment, how editors and reviewers should evaluate different criteria proposed by different authors, and losing what many consider to be the point of NHST—which is to have a consistent threshold level across a scientific domain: with variable thresholds, many old problems with NHST remain unsolved, such as the problems of regression to the mean, unreliability of *p*-values, inflation of effect sizes, publication bias, and the general disadvantage of forcing decisions too quickly rather than considering cumulative evidence across experiments.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

a criterion of .005 renders research more replicable demands much more specificity with respect to how to conceptualize replicability. In addition, we do not see a single replication success or failure as definitive. If one wishes to make a strong case for replication success or failure, multiple replication attempts are desirable.[9]

Questioning the Assumptions

The discussion thus far is under the pretense that the assumptions underlying the computation of *p*-values are true. But how likely is this? Berk and Freedman (2003) have made a strong case that the assumptions of random sampling from a population and independence are rarely true. The problems are particularly salient in the clinical sciences, where the falsity of the assumptions, as well as the divergences between statistical and clinical significance, are particularly obvious and dramatic (Bhardwaj, Camacho, Derrow, Fleischer, & Feldman 2004; Ferrill, Brown, & Kyle, 2010; Fethney, 2010; Page, 2014). The problem of likely false assumptions underlying the computation of *p*-values, in combination with the other problems already discussed, render the illusory garnering of truth from *p*-values yet more dramatic.

The Population Effect Size

Let us continue with the significance and replication issues, reverting to the pretense that significance testing assumptions are correct, while keeping in mind that this is unlikely. Consider that as matters now stand using tests of significance with the .05 criterion, the population effect size plays an important role both in obtaining statistical significance (all else being equal, the sample effect size will be larger if the population effect size is larger) and in obtaining statistical significance twice for a successful replication. Switching to the .005 criterion would not lessen the importance of the population effect size, and would increase its importance unless sample sizes increased substantially from those commonly used.[10] And there is good reason to reject that

[9] The present NHST focus should not detract from the importance of the quality of the theory and auxiliary assumptions for replication, as is attested to by recent successful replication studies in cognitive psychology (Zwaan et al., 2017) and social sciences (Mullinix et al., 2015).

[10] In addition, with an alpha level of .005, large effect sizes would be more important for publication, and researchers might lean much more towards "obvious" research than in testing creative ideas where there is more of a risk of weak effects and *p*-values that fail to meet the .005 bar.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

replicability should depend on the population effect size. To see this quickly, consider one of the most important science experiments of all time, by Michelson and Morley (1887). They used their interferometer to test whether the universe is filled with a luminiferous ether that allows light to travel to Earth from the stars. Their sample effect size was very small, and physicists accept that the population effect size is zero because there is no luminiferous ether. Using traditional tests of significance with either a .05 or .005 criterion, replicating Michelson and Morley would be problematic (see Sawilowsky, 2003, for a discussion of this experiment in the context of hypothesis testing). And yet physicists consider the experiment to be highly replicable (see also Meehl, 1967).[11] More generally, an experiment's replicability should not depend on the population effect size. Any proposal that features *p*-value rejection criteria forces the replication probability to be impacted by the population effect size, and should be rejected.

Accuracy of Published Effect Sizes

It is desirable that published facts in scientific literatures accurately reflect reality. Consider again the regression issue. The more stringent the criterion level for publishing, the more distance there is from a finding that passes the criterion to the mean, and so there is an increasing regression effect. Even at the .05 level, researchers have long recognized that published effect sizes likely do not reflect reality, or at least not the reality that would be seen if there were many replications of each experiment and all were published (see Briggs, 2016; Grice, 2017; Hyman, 2017; Kline, 2017; Locascio, 2017a; 2017b; and Marks, 2017 for a recent discussion of this problem). Under reasonable sample sizes and reasonable population effect sizes, it is the abnormally large sample effect sizes that result in *p*-values that meet the .05 (or .005) criterion, as is obvious from the standpoint of statistical regression. Moreover, with typically low sample sizes, statistically significant effects often require overestimates of population effect sizes. Effect size overestimation was empirically verified by the Open Science Collaboration project (2015), where the average effect size in the replication cohort of studies was dramatically reduced from the average effect size in the original cohort (from .403 to .197). Changing to a more stringent .005 criterion merely would result in yet worse effect size

[11] Very likely, a reason null results are so difficult to publish in sciences such as psychology is because the tradition of using *p*-value cutoffs is so ingrained. It would be well to terminate this tradition.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

overestimation (Button et al. 2013). The importance of having published effect sizes accurately reflect population effect sizes contradicts the use of significance tests, at any criterion.

Sample size and Alternatives to Significance Testing

We stress that replication depends largely on sample size, but there are factors that interfere with researchers using the large sample sizes necessary for good sampling precision and replicability. In addition to the obvious costs of obtaining large sample sizes, there may be an underappreciation of how much sample size matters (Vankov, Bowers, & Munafo, 2014), of the importance of incentives to favor novelty over replicability (Nosek, Spies, & Motyl, 2012) and of a prevalent misperception that the complement of *p*-values measures replicability (Cohen, 1994; Thompson, 1996; Greenland et al. 2016). A focus on sample size suggests an alternative to significance testing. Trafimow (2017; Trafimow & MacDonald, 2017) suggested a procedure as follows. The researcher specifies how close she wishes the sample statistics to be to their corresponding population parameters, and the desired probability of being that close. Trafimow's equations can be used to obtain the necessary sample size to meet specifications. The researcher then obtains the necessary sample size, computes the descriptive statistics, and takes them as accurate estimates of population parameters (provisionally on new data, of course).[12] This *a priori* procedure stresses (a) deciding what it takes to believe that the sample statistics are good estimates of the population parameters before data collection rather than afterwards, and (b) obtaining a large enough sample size to be confident that the obtained sample statistics really are within specified distances of corresponding population parameters. The procedure also does not promote publication bias because there is no cutoff for publication decisions.[13]

The larger point is that there are creative alternatives to significance testing that confront the sample size issue much more directly than significance testing does. The "statistical toolbox" (Gigerenzer & Marewski, 2015) further includes, for example, confidence intervals, equivalence

[12] An optimal way to obtain reliable estimation is via robust methods (Erceg-Hurn, Wilcox, & Keselman, 2013; Field & Wilcox, 2017; Huber, 1972; Portnoy & He, 2000; Rousseeuw, 1991; Tukey, 1979).

[13] The foregoing description may make the *a priori* procedure seem to be the same as traditional power analysis, but this is not so. First, the goal of traditional power analysis is to find the sample size needed to have a good chance of obtaining a statistically significant *p*-value. Second, traditional power analysis is strongly influenced by the expected effect size whereas this *a priori* procedure is completely uninfluenced by the expected effect size.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

tests, alternative ways of dealing with *p*-values as continuous indices, Bayesian methods, or information criteria; but none of those tools should replace conventional significance testing as the new magic method giving clear-cut mechanical answers (Cohen, 1994). In fact, inference should not be based on single studies at all (Neyman & Pearson, 1933; Fisher, 1937; Greenland, 2017), nor on replications from the same lab, but on cumulative evidence from multiple independent studies. It is desirable to obtain precise estimates in those studies, but the more important goal may be to publish also our wide confidence intervals and small effects, without which the cumulative evidence will be distorted (Amrhein, Korner-Nievergelt, & Roth, 2017; Amrhein & Greenland, 2017). Along these lines, Briggs (2016) argues for abandoning parameter-based inference and adopting purely predictive, and therefore verifiable, probability models, and Greenland (2017) sees "a dire need to get away from inferential statistics and hew more closely to descriptions of study procedures, data collection [...], and the resulting data."

Conclusion

It seems appropriate to conclude with the basic issue that has been with us from the beginning. Should *p*-values and *p*-value thresholds be used as the main criterion for making publication decisions? The mere fact that researchers are concerned with replication, however it is conceptualized, indicates an appreciation that single studies are rarely definitive and rarely justify a final decision. Thus, *p*-value criteria may not be very sensible. A counterargument might be that researchers often make decisions about what to believe, and using *p*-value criteria formalize what otherwise would be an informal process. But this counterargument is too simplistic. When evaluating the strength of the evidence, sophisticated researchers consider, in an admittedly subjective way, theoretical considerations such as scope, explanatory breadth, and predictive power; the worth of the auxiliary assumptions connecting nonobservational terms in theories to observational terms in empirical hypotheses; the strength of the experimental design; or implications for applications. To boil all this down to a binary decision based on a *p*-value threshold of .05, .01, .005, or anything else, is not acceptable.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

References

Amrhein, V. & Greenland, S. (2017). Remove, rather than redefine, statistical significance. *Nature Hum. Behav.* **1**, 0224.

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

Balluerka, N., Gómez, J., Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology* **1**, 55–77.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T.-H., Hoijtink, H., Jones, J. H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., & Johnson, V. E. (2017). Redefine statistical significance. *Nature Hum. Behav.* **1**, 0189.

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds). *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (2nd Ed, pp. 235–254). Aldine de Gruyter.

Bhardwaj, S., Camacho, F., Derrow, A., Fleischer, A., & Feldman, S. (2004). Statistical significance and clinical relevance. *Archives of Dermatology* **140**, 1520–1523.

Briggs, W. M. (2016). *Uncertainty: The Soul of Modeling, Probability & Statistics*. New York: Springer.

Buhl-Mortensen, L. (1996). Type-II statistical errors in environmental science and the precautionary principle. *Marine Pollution Bulletin* **32**, 528–531.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365376.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist* **49**, 997–1003.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods* **22**, 240–261.

Erceg-Hurn, D. M., Wilcox, R. R., & Keselman, H. J. (2013). Robust statistical estimation. In T. Little (Ed.), *The Oxford Handbook of Quantitative Methods*, Vol. 1, 388–406. New York: Oxford University Press.

Ferrill, M., Brown, D., & Kyle, J. (2010). Clinical versus statistical significance: Interpreting p values and confidence intervals related to measures of association to decision making. *Journal of Pharmacy Practice* **23**, 344–351.

Fethney, J. (2010). Statistical and clinical significance, and how to use confidence intervals to help interpret both. *Australian Critical Care* **23**, 93–97.

Field, A. & Wilcox, R. (2017). Robust statistical methods: a primer for clinical psychology and experimental psychopathology researchers. *Behavior Research and Therapy* **98**, 19–38.

Fisher, R. A. (1925). *Statistical methods for research workers* (1th ed.). Edinburgh: Oliver and Boyd.

Fisher, R. A. (1937). *The design of experiments* (2nd ed.). Edinburgh: Oliver and Boyd.

Fisher,R . A. (1973). *Statistical methods and scientific inference* (3rd ed.). London: Macmillan.

Gigerenzer, G. & Marewski, J. N. (2015). Surrogate science: the idol of a universal method for scientific inference. *Journal of Management* **41**, 421–440.

Greenland, S. (2017). The need for cognitive science in methodology. *American Journal of Epidemiology* **186**, 639–645.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* **31**, 337–350.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

Grice, J. W. (2017). Comment on Locascio's results blind manuscript evaluation proposal. *Basic and Applied Social Psychology*. In press.

Huber, P. J. (1972). Robust statistics: a review. *The Annals of Mathematical Statistics* **43**, 1041–1067.

Hyman, M. (2017). Can 'results blind manuscript evaluation' assuage 'publication bias'? *Basic and Applied Social Psychology*, In press.

Kline, R. (2017). Comment on Locascio, results blind science publishing. *Basic and Applied Social Psychology*, In press.

Konijn, E. A., van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015) Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing, *Communication Methods and Measures* **9**, 280–302.

Lemons, J., & Victor, R. (2008) Uncertainty in river restoration. In Darby, S., Sear, D. (Eds.), *River restoration: managing the uncertainty in restoring physical habitat*. John Wiley & Sons.

Lemons, J., Shrader-Frechette, K., Cranor, C. (1997) The precautionary principle: Scientific uncertainty and type I and type II errors. *Foundations of Science* **2**, 207–236.

Lieberman, M. D., Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience* **4**, 423–428.

Locascio, J. (2017a). Results blind publishing. *Basic and Applied Social Psychology*. In press.

Locascio, J. (2017b). Rejoinder to responses to "results blind publishing." *Basic and Applied Social Psychology*. In press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin* **70**, 151–159.

Marks, M. J. (2017). Commentary on Locascio 2017. *Basic and Applied Social Psychology*. In press.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: The University of Chicago Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics. A methodological paradox. *Philosophy of Science* **34**, 103–115.

Melton, A. (1962). Editorial. *Journal of Experimental Psychology* **64**, 553–557.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

Michelson, A. A., & Morley, E. W. (1887). On the relative motion of earth and luminiferous ether. *American Journal of Science, Third Series, 34,* 203, 233–245. http://history.aip.org/exhibits/gap/PDF/michelson.pdf

Miller, J., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science* **11**, 664–691.

Mudge, J.F., Baker, L.F., Edge, C.B., Houlahan, J.E. (2012). Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests. *PLoS ONE* **7**, e32734.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science* **2**, 109–138.

Myhr, A. I. (2010). A precautionary approach to genetically modified organisms: challenges and implications for policy and science. *Journal of Agricultural and Environmental Ethics* **23**, 501–525.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A* **231**, 289–337.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives in Psychological Science* **7**, 615–631.

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

Page, P. (2014). Beyond statistical significance: Clinical interpretation of rehabilitation research literature. *The International Journal of Sports Physical Therapy* **9**, 72.

Portnoy, S., & He, X. (2000). A robust journey in the new millennium. *Journal of the American Statistical Association* **95**, 1331–1335.

Rice, S., & Trafimow, D. (2010). How many people have to die for a type II error? *Theoretical Issues in Ergonomics Science* **11**, 387–401.

Rousseeuw, P. J. (1991). Tutorial to robust statistics. *Journal of Chemometrics* **5**, 1–20.

Royall, R. M. (1986). The Effect of Sample Size on the Meaning of Significance Tests. *The American Statistician* **40**, 313–315.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

Sawilowski, S. (2003). Deconstructing arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods* **2**, 467–474.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher* **25**, 26–30.

Trafimow, D. (2017). Using the coefficient of confidence to make the philosophical switch from *a posteriori* to *a priori* inferential statistics. *Educational and Psychological Measurement* **77**, 831–854.

Trafimow, D., & de Boer, M. (2017). Measuring the strength of the evidence. Under submission.

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and the use of P values to control the Type I error rate: The domain problem. *New Ideas in Psychology* **45**, 19–27. http://dx.doi.org/10.1016/j.newideapsych.2017.01.002.

Trafimow, D., & MacDonald, J. A. (2017). Performing inferential statistics prior to data collection. *Educational and Psychological Measurement* **77**, 204–219.

Tukey, J. W. (1979). Robust techniques for the user. In R. L. Launer and G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 103–106). Academic Press, New York.

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

Zwaan, R., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2017). Participant Nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin and Review*. doi.org/10.3758/s13423-017-1348-y

Appendix

The Bayes Factor is an approach to model selection that attempts to quantify the posterior probability of one model relative to another, given set of observed data (Kass & Raftery, 1995). Formally, given a model H and observed data D, the posterior probability of H is given by Bayes theorem:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

where $P(D|H)$ is the likelihood (determined by the statistical model), and $P(H)$ is the prior on the model H. Given two competing models H_1 and H_2, the ratio of the posterior probabilities is given by

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}$$

Here, the quantity

$$B = \frac{P(D|H_1)}{P(D|H_2)}$$

which multiplies the ratio of prior probabilities to obtain the posterior, is traditionally called the Bayes factor, and is generally interpreted as reflecting the relative weight of evidence provided by the data for the models H_1 and H_2. Quantities greater than 1 suggest that the data favor H_1, while quantities less than 1 suggest support for H_2.

Ostensibly a form of model comparison, the Bayes factor has been widely employed in the sciences to perform null-hypothesis significance testing by specifying the null model H_1 to be the alternative H_2 with some parameter(s) set to zero (e.g. Wagenmakers, 2007; Wetzels & Wagenmakers, 2012; Ly, Verhagen, & Wagenmakers, 2016). In this way, Bayes factors find use as a sort of "Bayesian hypothesis test". Although increasingly popular, Bayes factors have several problems which complicate their use as a form of hypothesis testing. Most notably, Bayes factors are highly sensitive to the choice of prior in a way that a true Bayesian analysis (one which returns a posterior distribution over the model parameters) is not, as we demonstrate below. Further, this prior sensitivity can often behave unintuitively; for instance, Gelman, et al (2014; pp. 183–184) provide an example in which the goal is to estimate a mean treatment effect within several groups. A Bayes factor is employed to compare a null model in which all groups have the same mean value, which has a normal prior; and a model with no shrinkage, in which

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

the means are independent draws from the same normal prior. In this case, the resulting Bayes factor is highly sensitive to the variance of the prior, and will always select the null model as the prior variance goes to infinity. A researcher, mistakenly believing that they are constructing a non-informative prior, might choose a very large prior variance, unknowingly forcing the Bayes factor to select the null model, regardless of what the data say.

Moreover, in a fully Bayesian model the likelihood begins to dominate the prior as the sample size goes to infinity, as we might expect (since a larger sample provides more information about the model parameters). The prior thus has less effect with increasing sample size. This is not, in general, true of Bayes factors, which retain their prior sensitivity even with large samples.

As an example, consider a set of data which are assumed to be Poisson distributed with rate λ, with competing models $H_1: \lambda \leq 1$ vs $H_2: \lambda > 1$. A simulation study considering different sample sizes (20, 30 and 50) out with 10,000 replicates. The prior distributions for H_1 and H_2, and mean and standard deviations are presented in Table 1.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

n	Prior H_1	Prior H_2	$mean(BF)$	$sd(BF)$
20	Gamma(1,2)	Gamma(3,3)	7.0	2.7
	Gamma(2,2)		202.0	74.7
	Gamma(3,2)		3726.9	1331.9
	Gamma(2,2)	Gamma(1,3)	11570.1	3139.9
		Gamma(2,3)	1496.0	472.5
		Gamma(4,3)	29.1	12.6
		Gamma(10,3)	**0.0019**	**0.0020**
	Laplace	Laplace	2.5	13.6
	Jeffreys	Jeffreys	**0.7**	**4.0**
30	Gamma(1,2)	Gamma(3,3)	357.1	125.6
	Gamma(2,2)		15626.6	5500.1
	Gamma(3,2)		428190.9	151300.7
	Gamma(2,2)	Gamma(1,3)	1936147.3	597773.2
		Gamma(2,3)	167656.2	54433.1
		Gamma(4,3)	1539.0	603.0
		Gamma(10,3)	**0.0113**	**0.0088**
	Laplace	Laplace	1.7	11.5
	Jeffreys	Jeffreys	**0.4**	**2.6**
50	Gamma(1,2)	Gamma(3,3)	2161996.0	796727.8
	Gamma(2,2)		156358014.0	58652977.0
	Gamma(3,2)		7218780790.0	2682940095.0
	Gamma(2,2)	Gamma(1,3)	53119836572.0	18718219507.0
		Gamma(2,3)	2776659384.0	993973186.0
		Gamma(4,3)	9361043.0	3659097.0
		Gamma(10,3)	3.8	2.4
	Laplace	Laplace	**0.7**	**4.7**
	Jeffreys	Jeffreys	**0.1**	**0.8**

Table 1: Mean and standard deviations for Bayes Factor (BF) when different prior distributions for the hypothesis H_1 and H_2 are considered. The BFs that provide support to H_2 are shown in bold.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

It is evident that the BF changes considerably when prior distributions and sample sizes change. In many cases, H_1 is more strongly supported by the data than H_2, which is not correct since $\lambda = 1.5$. And yet, the posterior distribution over λ is largely insensitive to these choices. For $\lambda = 1.5$ and a sample size of 50, the expected sum over all observations for is 75. Since a Gamma prior is conjugate for a Poisson likelihood, we can compute the expected posterior directly: for prior shape α and rate β, the expected posterior shape and rate are $\alpha + 75$ and $\beta + 50$, giving a posterior mean of

$$\frac{(\alpha + 75)}{(\beta + 50)}.$$

For Gamma(1,3) and Gamma(2,3) priors, we have posterior means of 1.43 and 1.45, respectively—a negligible difference. And yet, the Bayes factors resulting from these priors differ by an order of magnitude.

Additionally, Bayes factors may exhibit strange behavior when used to test point-nulls for continuous models (e.g. testing that a mean difference is exactly zero). Aitkin, Boys, and Chadwick (2005) study an example in which a hypothesis test for a binomial probability, conducted via Bayes factor, returns strong support for a point null $H_0: p = p_0$ when, in fact, the posterior distribution (and a data itself) overwhelmingly support a value $p \neq p_0$. The authors note that this is an example where two competing approaches—estimation and hypothesis testing—are in clear conflict. In general, researchers should be aware that the question they are asking—that is, do the data support the null value, or is *some other value* better supported by the data—is not necessarily the question being answered by the Bayes factor, nor is a Bayes factor used to test a point null on a parameter consistent in general with the posterior distribution over that same parameter. For these reasons, we do not feel that the Bayes factor is a satisfactory substitute for traditional hypothesis testing, nor does it address the fundamental problems associated with such approaches: namely, that they ignore uncertainty in favor of binary decision making. Using a threshold for the BF will result in a similar dilemma as with a threshold for the *p*-value. As Konijn et al. (2015) suggested, "God would love a Bayes Factor of 3.01 nearly as much as a BF of 2.99". Moreover, the Bayes factor does not provide a good measure of statistical evidence, as it fails the coherence desideratum (see Lavine & Schervish, 1999).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017

References

Aitkin, M., Boys, R. J., & Chadwick, T. (2005). Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing* **15**, 217–230.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis, Third Edition*. Boca Raton, FL: CRC press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Konijn, E. A., van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015) Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures* **9**, 280–302.

Lavine, M. and Schervish, M. J. (1999). Bayes factors: what they are and what they are not. *The American Statistician* **53**, 119–122.

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology* **72**, 19–32.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review* **14**, 779–804.

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review* **19**, 1057–1064.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3411v1 | CC BY 4.0 Open Access | rec: 14 Nov 2017, publ: 14 Nov 2017