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., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6-10. https://doi.org/10.1038/s41562-017-0189-z

Peer reviewed version

Link to published version (if available):
10.1038/s41562-017-0189-z

Title: Redefine Statistical Significance

Authors: Daniel J. Benjamin[1]*, James O. Berger[2], Magnus Johannesson[3]*, Brian A. Nosek[4,5], E.-J. Wagenmakers[6], Richard Berk[7, 10], Kenneth A. Bollen[8], Björn Brembs[9], Lawrence Brown[10], Colin Camerer[11], David Cesarini[12, 13], Christopher D. Chambers[14], Merlise Clyde[2], Thomas D. Cook[15,16], Paul De Boeck[17], Zoltan Dienes[18], Anna Dreber[3], Kenny Easwaran[19], Charles Efferson[20], Ernst Fehr[21], Fiona Fidler[22], Andy P. Field[18], Malcolm Forster[23], Edward I. George[10], Richard Gonzalez[24], Steven Goodman[25], Edwin Green[26], Donald P. Green[27], Anthony Greenwald[28], Jarrod D. Hadfield[29], Larry V. Hedges[30], Leonhard Held[31], Teck Hua Ho[32], Herbert Hoijtink[33], James Holland Jones[39,40], Daniel J. Hruschka[34], Kosuke Imai[35], Guido Imbens[36], John P.A. Ioannidis[37], Minjeong Jeon[38], Michael Kirchler[41], David Laibson[42], John List[43], Roderick Little[44], Arthur Lupia[45], Edouard Machery[46], Scott E. Maxwell[47], Michael McCarthy[48], Don A. Moore[49], Stephen L. Morgan[50], Marcus Munafó[51, 52], Shinichi Nakagawa[53], Brendan Nyhan[54], Timothy H. Parker[55], Luis Pericchi[56], Marco Perugini[57], Jeff Rouder[58], Judith Rousseau[59], Victoria Savalei[60], Felix D. Schönbrodt[61], Thomas Sellke[62], Betsy Sinclair[63], Dustin Tingley[64], Trisha Van Zandt[65], Simine Vazire[66], Duncan J. Watts[67], Christopher Winship[68], Robert L. Wolpert[2], Yu Xie[69], Cristobal Young[70], Jonathan Zinman[71], Valen E. Johnson[72]*

Affiliations:

[1]Center for Economic and Social Research and Department of Economics, University of Southern California, Los Angeles, CA 90089-3332, USA.

[2]Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA.

[3]Department of Economics, Stockholm School of Economics, SE-113 83 Stockholm, Sweden.

[4]University of Virginia, Charlottesville, VA 22908, USA.

[5]Center for Open Science, Charlottesville, VA 22903, USA.

[6]University of Amsterdam, Department of Psychology, 1018 VZ Amsterdam, The Netherlands.

[7]University of Pennsylvania, School of Arts and Sciences and Department of Criminology, Philadelphia, PA 19104-6286, USA.

[8]University of North Carolina Chapel Hill, Department of Psychology and Neuroscience, Department of Sociology, Chapel Hill, NC 27599-3270, USA.

[9]Institute of Zoology - Neurogenetics, Universität Regensburg, Universitätsstrasse 31 93040 Regensburg, Germany.

[10]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

[11]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA.

[12]Department of Economics, New York University, New York, NY 10012, USA.

[13]The Research Institute of Industrial Economics (IFN), SE- 102 15 Stockholm, Sweden.

[14]Cardiff University Brain Research Imaging Centre (CUBRIC), CF24 4HQ, UK.

[15]Northwestern University, Evanston, IL 60208, USA.

[16]Mathematica Policy Research, Washington, DC, 20002-4221, USA.

[17]Department of Psychology, Quantitative Program, Ohio State University, Columbus, OH 43210, USA.

[18]School of Psychology, University of Sussex, Brighton BN1 9QH, UK.

[19]Department of Philosophy, Texas A&M University, College Station, TX 77843-4237, USA.

[20]Department of Psychology, Royal Holloway University of London, Egham Surrey TW20 0EX, UK.

[21]Department of Economics, University of Zurich, 8006 Zurich, Switzerland.

[22]School of BioSciences and School of Historical & Philosophical Studies, University of Melbourne, Vic 3010, Australia.

[23]Department of Philosophy, University of Wisconsin - Madison, Madison, WI 53706, USA.

[24]Department of Psychology, University of Michigan, Ann Arbor, MI 48109-1043, USA.

[25]Stanford University, General Medical Disciplines, Stanford, CA 94305, USA.

[26]Department of Ecology, Evolution and Natural Resources SEBS, Rutgers University, New Brunswick, NJ 08901-8551, USA.

[27]Department of Political Science, Columbia University in the City of New York, New York, NY 10027, USA.

[28]Department of Psychology, University of Washington, Seattle, WA 98195-1525, USA.

[29]Institute of Evolutionary Biology School of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3JT, UK.

[30]Weinberg College of Arts & Sciences Department of Statistics, Northwestern University, Evanston, IL 60208, USA.

[31]Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, 8001 Zurich, Switzerland.

[32]National University of Singapore, Singapore 119077.

[33]Department of Methods and Statistics, Universiteit Utrecht, 3584 CH Utrecht, The Netherlands.

[34]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287-2402, USA.

[35]Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton NJ 08544, USA.

[36]Stanford University, Stanford, CA 94305-5015, USA.

[37]Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA.

[38] Advanced Quantitative Methods, Social Research Methodology, Department of Education, Graduate School of Education & Information Studies, University of California, Los Angeles, CA 90095-1521, USA.

[39]Department of Life Sciences, Imperial College London, Ascot SL5 7PY, UK.

[40]Department of Earth System Science, Stanford, CA 94305- 4216, USA.

[41]Department of Banking and Finance, University of Innsbruck and University of Gothenburg, A-6020 Innsbruck, Austria.

[42]Department of Economics, Harvard University, Cambridge, MA 02138, USA.

[43]Department of Economics, University of Chicago, Chicago, IL 60637, USA.

[44]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA.

[45]Department of Political Science, University of Michigan, Ann Arbor, MI 48109-1045, USA.

[46]Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh PA 15260, USA.

[47]Department of Psychology, University of Notre Dame, Notre Dame, IN 46556, USA.

[48]School of BioSciences, University of Melbourne, Vic 3010, Australia.

[49]Haas School of Business, University of California at Berkeley, Berkeley, CA 94720-1900A, USA.

[50]Johns Hopkins University, Baltimore, MD 21218, USA.

[51]MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 1TU, UK.

[52]UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol BS8 1TU, UK.

[53]Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia.

[54]Department of Government, Dartmouth College, Hanover, NH 03755, USA.

[55]Department of Biology, Whitman College, Walla Walla, WA 99362, USA.

[56]Department of Mathematics, University of Puerto Rico, Rio Piedras Campus, San Juan, PR 00936-8377.

[57]Department of Psychology, University of Milan - Bicocca, 20126 Milan, Italy.

[58]Department of Cognitive Sciences, University of California, Irvine, CA 92617, USA.

[59]Université Paris Dauphine, 75016 Paris, France.

[60]Department of Psychology, The University of British Columbia, Vancouver, BC Canada V6T 1Z4.

[61]Department Psychology, Ludwig-Maximilians-University Munich, Leopoldstraße 13, 80802 Munich, Germany.

[62]Department of Statistics, Purdue University, West Lafayette, IN 47907-2067, USA.

[63]Department of Political Science, Washington University in St. Louis, St. Louis, MO 63130-4899, USA.

[64]Government Department, Harvard University, Cambridge, MA 02138, USA.

[65]Department of Psychology, Ohio State University, Columbus, OH 43210, USA.

[66]Department of Psychology, University of California, Davis, CA, 95616, USA.

[67]Microsoft Research. 641 Avenue of the Americas, 7th Floor, New York, NY 10011, USA.

[68]Department of Sociology, Harvard University, Cambridge, MA 02138, USA.

[69]Department of Sociology, Princeton University, Princeton NJ 08544, USA.

[70]Department of Sociology, Stanford University, Stanford, CA 94305-2047, USA.

[71]Department of Economics, Dartmouth College, Hanover, NH 03755-3514, USA.

[72]Department of Statistics, Texas A&M University, College Station, TX 77843, USA.

*Correspondence to: Daniel J. Benjamin, daniel.benjamin@gmail.com; Magnus Johannesson, magnus.johannesson@hhs.se; Valen E. Johnson, vejohnson@exchange.tamu.edu.

One Sentence Summary: We propose to change the default *P*-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

Main Text:

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on "statistically significant" findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (e.g., multiple testing, P-hacking, publication bias, and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: Statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating "statistically significant" findings with $P < 0.05$ results in a high rate of false positives *even in the absence of other experimental, procedural and reporting problems*.

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called "significant" but do not meet the new threshold should instead be called "suggestive." While statisticians have known the relative weakness of using $P \approx 0.05$ as a threshold for discovery and the proposal to lower it to 0.005 is not new (*1, 2*), a critical mass of researchers now endorse this change.

We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (e.g., genomics and high-energy physics research; see Potential Objections below).

We also restrict our recommendation to studies that conduct null hypothesis significance tests. We have diverse views about how best to improve reproducibility, and many of us believe that other ways of summarizing the data, such as Bayes factors or other posterior summaries based on clearly articulated model assumptions, are preferable to *P*-values. However, changing the *P*-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.

Strength of evidence from *P*-values

In testing a point null hypothesis H_0 against an alternative hypothesis H_1 based on data x_{obs}, the *P*-value is defined as the probability, calculated under the null hypothesis, that a test statistic is as extreme or more extreme than its observed value. The null hypothesis is typically rejected—and the finding is declared "statistically significant"—if the *P*-value falls below the (current) Type I error threshold $\alpha = 0.05$.

From a Bayesian perspective, a more direct measure of the strength of evidence for H_1 relative to H_0 is the ratio of their probabilities. By Bayes' rule, this ratio may be written as:

$$\frac{\Pr(H_1|x_{\text{obs}})}{\Pr(H_0|x_{\text{obs}})} = \frac{f(x_{\text{obs}}|H_1)}{f(x_{\text{obs}}|H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)} \equiv BF \times (\text{prior odds}), \quad (1)$$

where BF is the Bayes factor that represents the evidence from the data, and the prior odds can be informed by researchers' beliefs, scientific consensus, and validated evidence from similar research questions in the same field. Multiple hypothesis testing, P-hacking, and publication bias all reduce the credibility of evidence. Some of these practices reduce the prior odds of H_1 relative to H_0 by changing the population of hypothesis tests that are reported. Prediction markets (*3*) and analyses of replication results (*4*) both suggest that for psychology experiments, the prior odds of H_1 relative to H_0 may be only about 1:10. A similar number has been suggested in cancer clinical trials, and the number is likely to be much lower in preclinical biomedical research (*5*).

There is no unique mapping between the *P*-value and the Bayes factor since the Bayes factor depends on H_1. However, the connection between the two quantities can be evaluated for particular test statistics under certain classes of plausible alternatives (Fig. 1).

[Figure 1 here]

A two-sided *P*-value of 0.05 corresponds to Bayes factors in favor of H_1 that range from about 2.5 to 3.4 under reasonable assumptions about H_1 (Fig. 1). This is weak evidence from at least three perspectives. First, conventional Bayes factor categorizations (*6*) characterize this range as "weak" or "very weak." Second, we suspect many scientists would guess that $P \approx 0.05$ implies stronger support for H_1 than a Bayes factor of 2.5 to 3.4. Third, using equation (1) and prior odds of 1:10, a *P*-value of 0.05 corresponds to ***at least*** *3:1 odds* (i.e., the reciprocal of the product $\frac{1}{10} \times$ 3.4) *in favor of the null hypothesis*!

Why 0.005?

The choice of any particular threshold is arbitrary and involves a trade-off between Type I and II errors. We propose 0.005 for two reasons. First, a two-sided *P*-value of 0.005 corresponds to Bayes factors between approximately 14 and 26 in favor of H_1. This range represents "substantial" to "strong" evidence according to conventional Bayes factor classifications (*6*).

Second, in many fields the $P < 0.005$ standard would reduce the false positive rate to levels we judge to be reasonable. If we let ϕ denote the proportion of null hypotheses that are true, $(1-\beta)$ the power of tests in rejecting false null hypotheses, and α the Type I error/significance threshold, then as the population of tested hypotheses becomes large, the false positive rate (i.e., the proportion of true null effects among the total number of statistically significant findings) can be approximated by

$$\text{false positive rate} \approx \frac{\alpha\phi}{\alpha\phi + (1-\beta)(1-\phi)}. \quad (2)$$

For different levels of the prior odds that there is a true effect, $\frac{1-\phi}{\phi}$, and for significance thresholds $\alpha = 0.05$ and $\alpha = 0.005$, Figure 2 shows the false positive rate as a function of power $1 - \beta$.

[Figure 2 here]

In many studies, statistical power is low (e.g., ref. *7*). Fig. 2 demonstrates that low statistical power and $\alpha = 0.05$ combine to produce high false positive rates.

For many, the calculations illustrated by Fig. 2 may be unsettling. For example, the false positive rate is greater than 33% with prior odds of 1:10 and a *P*-value threshold of 0.05, *regardless of the level of statistical power*. Reducing the threshold to 0.005 would reduce this minimum false positive rate to 5%. Similar reductions in false positive rates would occur over a wide range of statistical powers.

Empirical evidence from recent replication projects in psychology and experimental economics provide insights into the prior odds in favor of H_1. In both projects, the rate of replication (i.e., significance at $P < 0.05$ in the replication in a consistent direction) was roughly double for initial studies with $P < 0.005$ relative to initial studies with $0.005 < P < 0.05$: 50% versus 24% for psychology (*8*), and 85% versus 44% for experimental economics (*9*). Although based on relatively small samples of studies (93 in psychology, 16 in experimental economics, after excluding initial studies with $P > 0.05$), these numbers are suggestive of the potential gains in reproducibility that would accrue from the new threshold of $P < 0.005$ in these fields. In biomedical research, 96% of a sample of recent papers claim statistically significant results with the $P < 0.05$ threshold (*10*). However, replication rates were very low (*5*) for these studies, suggesting a potential for gains by adopting this new standard in these fields as well.

Potential Objections

We now address the most compelling arguments against adopting this higher standard of evidence.

The false negative rate would become unacceptably high. Evidence that does not reach the new significance threshold should be treated as suggestive, and where possible further evidence should be accumulated; indeed, the combined results from several studies may be compelling even if any particular study is not. Failing to reject the null hypothesis does *not* mean accepting the null hypothesis. Moreover, the false negative rate will not increase if sample sizes are increased so that statistical power is held constant.

For a wide range of common statistical tests, transitioning from a *P*-value threshold of $\alpha = 0.05$ to $\alpha = 0.005$ while maintaining 80% power would require an increase in sample sizes of about 70%. Such an increase means that fewer studies can be conducted using current experimental designs and budgets. But Figure 2 shows the benefit: false positive rates would typically fall by factors greater than two. Hence, considerable resources would be saved by not performing future studies based on false premises. Increasing sample sizes is also desirable

because studies with small sample sizes tend to yield inflated effect size estimates (*11*), and publication and other biases may be more likely in an environment of small studies (*12*). We believe that efficiency gains would far outweigh losses.

The proposal does not address multiple hypothesis testing, P-hacking, publication bias, low power, or other biases (e.g., confounding, selective reporting, measurement error), which are arguably the bigger problems. We agree. Reducing the *P*-value threshold complements—but does not substitute for—solutions to these other problems, which include good study design, ex ante power calculations, pre-registration of planned analyses, replications, and transparent reporting of procedures and all statistical analyses conducted.

The appropriate threshold for statistical significance should be different for different research communities. We agree that the significance threshold selected for claiming a new discovery should depend on the prior odds that the null hypothesis is true, the number of hypotheses tested, the study design, the relative cost of Type I versus Type II errors, and other factors that vary by research topic. For exploratory research with very low prior odds (well outside the range in Figure 2), even lower significance thresholds than 0.005 are needed. Recognition of this issue led the genetics research community to move to a "genome-wide significance threshold" of 5×10^{-8} over a decade ago. And in high-energy physics, the tradition has long been to define significance by a "5-sigma" rule (roughly a *P*-value threshold of 3×10^{-7}). We are essentially suggesting a move from a 2-sigma rule to a 3-sigma rule.

Our recommendation applies to disciplines with prior odds broadly in the range depicted in Figure 2, where use of $P < 0.05$ as a default is widespread. Within those disciplines, it is helpful for consumers of research to have a consistent benchmark. We feel the default should be shifted.

Changing the significance threshold is a distraction from the real solution, which is to replace null hypothesis significance testing (and bright-line thresholds) with more focus on effect sizes and confidence intervals, treating the P-value as a continuous measure, and/or a Bayesian method. Many of us agree that there are better approaches to statistical analyses than null hypothesis significance testing, but as yet there is no consensus regarding the appropriate choice of replacement. For example, a recent statement by the American Statistical Association addressed numerous issues regarding the misinterpretation and misuse of P-values (as well as the related concept of statistical significance), but failed to make explicit policy recommendations to address these shortcomings (*13*). Even after the significance threshold is changed, many of us will continue to advocate for alternatives to null hypothesis significance testing.

Concluding remarks

Ronald Fisher understood that the choice of 0.05 was arbitrary when he introduced it (*14*). Since then, theory and empirical evidence have demonstrated that a lower threshold is needed. A much larger pool of scientists are now asking a much larger number of questions, possibly with much lower prior odds of success.

For research communities that continue to rely on null hypothesis significance testing, reducing the *P*-value threshold for claims of new discoveries to 0.005 is an actionable step that will immediately improve reproducibility. We emphasize that this proposal is about standards of evidence, not standards for policy action nor standards for publication. Results that do not reach

the threshold for statistical significance (whatever it is) can still be important and merit publication in leading journals if they address important research questions with rigorous methods. This proposal should not be used to reject publications of novel findings with $0.005 < P < 0.05$ properly labeled as suggestive evidence. We should reward quality and transparency of research as we impose these more stringent standards, and we should monitor how researchers' behaviors are affected by this change. Otherwise, science runs the risk that the more demanding threshold for statistical significance will be met to the detriment of quality and transparency.

Journals can help transition to the new statistical significance threshold. 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." The new significance threshold will help researchers and readers to understand and communicate evidence more accurately.

References and Notes:

1. A. G. Greenwald *et al*., Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology* **33**, 175-183 (1996).
2. V. E. Johnson, Revised standards for statistical evidence. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19313-19317 (2013).
3. A. Dreber *et al*., Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15343-15347 (2015).
4. V. E. Johnson *et al*., On the reproducibility of psychological science. *J. Am. Stat. Assoc.* **112,** 1-10 (2016).
5. G. C. Begley, J. P. A. Ioannidis, Reproducibility in science: Improving the standard for basic and preclinical research. *Circ. Res.* **116,** 116-126 (2015).
6. R. E. Kass, A. E. Raftery, Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773-795 (1995).
7. D. Szucs, J. P. A. Ioannidis, Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **15**, (2017**)**.
8. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349,** (2015).
9. C. Camerer *et al*., Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433-1436 (2016).
10. D. Chavalarias *et al.,* Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA* **315**, 1141-1148 (2016).
11. A. Gelman, J. Carlin, Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641-651 (2014).
12. D. Fanelli, R. Costas, J. P. A. Ioannidis, Meta-assessment of bias in science. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3714-3719 (2017).
13. R. L. Wasserstein, N. A. Lazar, The ASA's statement on p-values: Context, process, and purpose. *Am. Stat.* **70** (and online comments), 129-133 (2016).
14. R. A. Fisher, *Statistical Methods for Research Workers* (Oliver & Boyd, Edinburgh, 1925).
15. T. Sellke, M. J. Bayarri, J. O. Berger, Calibration of p-values for testing precise null hypotheses. *Am. Stat.* **55**, 62-71 (2001).

Acknowledgements: We thank Deanna L. Lormand, Rebecca Royer and Anh Tuan Nguyen Viet for excellent research assistance.