Article:
Lakens, D, Adolfi, FG, Albers, CJ et al. (85 more authors) (2018) Justify your alpha. Nature Human Behaviour, 2 (3). pp. 168-171.

https://doi.org/10.1038/s41562-018-0311-x

Justify Your Alpha

In Press, Nature Human Behavior

Daniel Lakens*[1], Federico G. Adolfi[2], Casper J. Albers[3], Farid Anvari[4], Matthew A. J. Apps[5], Shlomo E. Argamon[6], Thom Baguley[7], Raymond B. Becker[8], Stephen D. Benning[9], Daniel E. Bradford[10], Erin M. Buchanan[11], Aaron R. Caldwell[12], Ben van Calster[13], Rickard Carlsson[14], Sau-Chin Chen[15], Bryan Chung[16], Lincoln J Colling[17], Gary S. Collins[18], Zander Crook[19], Emily S. Cross[20], Sameera Daniels[21], Henrik Danielsson[22], Lisa DeBruine[23], Daniel J. Dunleavy[24], Brian D. Earp[25], Michele I. Feist[26,] Jason D. Ferrell[27], James G. Field[28], Nicholas W. Fox[29], Amanda Friesen[30], Caio Gomes[31], Monica Gonzalez-Marquez[32], James A. Grange[33], Andrew P. Grieve[34], Robert Guggenberger[35], James Grist[36], Anne-Laura van Harmelen[37], Fred Hasselman[38], Kevin D. Hochard[39], Mark R. Hoffarth[40], Nicholas P. Holmes[41], Michael Ingre[42], Peder M. Isager[43], Hanna K. Isotalus[44], Christer Johansson[45], Konrad Juszczyk[46], David A. Kenny[47], Ahmed A. Khalil[48], Barbara Konat[49], Junpeng Lao[50], Erik Gahner Larsen[51], Gerine M. A. Lodder[52], Jiří Lukavský[53], Christopher R. Madan[54], David Manheim[55], Stephen R. Martin[56], Andrea E. Martin[57], Deborah G. Mayo[58], Randy J. McCarthy[59], Kevin McConway[60], Colin McFarland[61], Amanda Q. X. Nio[62], Gustav Nilsonne[63], Cilene Lino de Oliveira[64], Jean-Jacques Orban de Xivry[65], Sam Parsons[66], Gerit Pfuhl[67], Kimberly A. Quinn[68], John J. Sakon[69], S. Adil Saribay[70], Iris K. Schneider[71], Manojkumar Selvaraju[72], Zsuzsika Sjoerds[73], Samuel G. Smith[74], Tim Smits[75], Jeffrey R. Spies[76], Vishnu Sreekumar[77], Crystal N. Steltenpohl[78], Neil Stenhouse[79], Wojciech Świątkowski[80], Miguel A. Vadillo[81], Marcel A. L. M. Van Assen[82], Matt N. Williams[83], Samantha E. Williams[84], Donald R. Williams[85], Tal Yarkoni[86], Ignazio Ziano[87], Rolf A. Zwaan[88]

Affiliations

*[1]Human-Technology Interaction, Eindhoven University of Technology, Den Dolech, 5600MB, Eindhoven, The Netherlands

[2]Laboratory of Experimental Psychology and Neuroscience (LPEN), Institute of Cognitive and Translational Neuroscience (INCYT), INECO Foundation, Favaloro University, Pacheco de Melo 1860, Buenos Aires, Argentina
[2]National Scientific and Technical Research Council (CONICET), Godoy Cruz 2290, Buenos Aires, Argentina
[3]Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands
[4]College of Education, Psychology & Social Work, Flinders University, Adelaide, GPO Box 2100, Adelaide, SA, 5001, Australia
[5]Department of Experimental Psychology, University of Oxford, New Radcliffe House, Oxford, OX2 6GG, UK
[6]Department of Computer Science, Illinois Institute of Technology, Chicago, IL, 10 W. 31st Street, Chicago, IL 60645, USA
[7]Department of Psychology, Nottingham Trent University, Nottingham, 50 Shakespeare Street, Nottingham, NG1 4FQ, UK
[8]Faculty of Linguistics and Literature, Bielefeld University, Bielefeld, Universitätsstraße 25, 33615 Bielefeld, Germany
[9]Psychology, University of Nevada, Las Vegas, Las Vegas, 4505 S. Maryland Pkwy., Box 455030, Las Vegas, NV 89154-5030, USA
[10]Psychology, University of Wisconsin-Madison, Madison, 1202 West Johnson St. Madison WI. 53706, USA
[11]Psychology, Missouri State University, 901 S. National Ave, Springfield, MO, 65897, USA
[12]Health, Human Performance, and Recreation, University of Arkansas, Fayetteville, 155 Stadium Drive, HPER 321, Fayetteville, AR, 72701, USA
[13]Department of Development and Regeneration, KU Leuven, Leuven, Herestraat 49 box 805, 3000 Leuven, Belgium, Belgium
[13]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Postbus 9600, 2300 RC, Leiden, The Netherlands

[14]Department of Psychology, Linnaeus University, Kalmar, Stagneliusgatan 14, 392 34, Kalmar, Sweden
[15]Department of Human Development and Psychology, Tzu-Chi University, No. 67, Jieren St., Hualien City, Hualien County, 97074, Taiwan
[16]Department of Surgery, University of British Columbia, Victoria, #301 - 1625 Oak Bay Ave, Victoria BC Canada, V8R 1B1 , Canada
[17]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, UK
[18]Centre for Statistics in Medicine, University of Oxford, Windmill Road, Oxford, OX3 7LD, UK
[19]Department of Psychology, The University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, UK
[20]School of Psychology, Bangor University, Bangor, Adeilad Brigantia, Bangor, Gwynedd, LL57 2AS, UK
[21]Ramsey Decision Theoretics, 4849 Connecticut Ave. NW #132, Washington, DC 20008, USA
[22]Department of Behavioural Sciences and Learning, Linköping University, SE-581 83, Linköping, Sweden
[23]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, 58 Hillhead Street, UK
[24]College of Social Work, Florida State University, 296 Champions Way, University Center C, Tallahassee, FL, 32304, USA
[25]Departments of Psychology and Philosophy, Yale University, 2 Hillhouse Ave, New Haven CT 06511, USA
[26]Department of English, University of Louisiana at Lafayette, P. O. Box 43719, Lafayette LA 70504, USA
[27]Department of Psychology, St. Edward's University, 3001 S. Congress, Austin, TX 78704, USA

[27]Department of Psychology, University of Texas at Austin, 108 E. Dean Keeton Stop A8000, Austin, TX 78712-1043, USA
[28]Department of Management, West Virginia University, 1602 University Avenue, Morgantown, WV 26506, USA
[29]Department of Psychology, Rutgers University, New Brunswick, 53 Avenue E, Piscataway NJ 08854, USA
[30]Department of Political Science, Indiana University Purdue University, Indianapolis, Indianapolis, 425 University Blvd CA417, Indianapolis, IN 46202, USA
[31]Booking.com, Herengracht 597, 1017 CE Amsterdam, The Nederlands
[32]Department of English, American and Romance Studies, RWTH - Aachen University, Aachen, Kármánstraße 17/19, 52062 Aachen, Germany
[33]School of Psychology, Keele University, Keele, Staffordshire, ST5 5BG, UK
[34]Centre of Excellence for Statistical Innovation, UCB Celltech, 208 Bath Road, Slough, Berkshire SL1 3WE, UK
[35]Translational Neurosurgery, Eberhard Karls University Tübingen, Tübingen, Germany
[35]University Tübingen, International Centre for Ethics in Sciences and Humanities, Germany
[36]Department of Radiology, University of Cambridge, Box 218, Cambridge Biomedical Campus, CB2 0QQ, UK
[37]Department of Psychiatry, University of Cambridge, Cambridge, 18b Trumpington Road, CB2 8AH, UK
[38]Behavioural Science Institute, Radboud University Nijmegen, Montessorilaan 3, 6525 HR, Nijmegen, The Netherlands
[39]Department of Psychology, University of Chester, Chester, Department of Psychology, University of Chester, Chester, CH1 4BJ, UK
[40]Department of Psychology, New York University, 4 Washington Place, New York, NY 10003, USA
[41]School of Psychology, University of Nottingham, Nottingham, University Park, NG7 2RD, UK

[42]None, Independent, Stockholm, Skåpvägen 5, 12245 ENSKEDE, Sweden
[43]Department of Clinical and Experimental Medicine, University of Linköping, 581 83 Linköping,, Sweden
[44]School of Clinical Sciences, University of Bristol, Bristol, Level 2 academic offices, L&R Building, Southmead Hospital, BS10 5NB, UK
[45]Occupational Orthopaedics and Research, Sahlgrenska University Hospital, 413 45 Gothenburg, Sweden
[46]The Faculty of Modern Languages and Literatures, Institute of Linguistics, Psycholinguistics Department, Adam Mickiewicz University, Al. Niepodległości 4, 61-874, Poznań, Poland
[47]Department of Psychological Sciences, University of Connecticut, Storrs, CT, Department of Psychological Sciences, U-1020, Storrs, CT 06269-1020, USA
[48]Center for Stroke Research Berlin, Charité - Universitätsmedizin Berlin, Hindenburgdamm 30, 12200 Berlin, Germany
[48]Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, 04103 Leipzig, Germany
[48]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Luisenstraße 56, 10115 Berlin, Germany
[40]Social Sciences, Adam Mickiewicz University, Poznań, Szamarzewskiego 89, 60-568 Poznan, Poland
[50]Department of Psychology, University of Fribourg, Faucigny 2, 1700 Fribourg, Switzerland
[51]School of Politics and International Relations, University of Kent, Canterbury CT2 7NX, UK
[52] Department of Sociology / ICS, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands
[53]Institute of Psychology, Czech Academy of Sciences, Hybernská 8, 11000 Prague, Czech Republic
[54]School of Psychology, University of Nottingham, Nottingham, NG7 2RD, UK
[55]Pardee RAND Graduate School, RAND Corporation, 1200 S Hayes St, Arlington, VA 22202, USA

[56]Psychology and Neuroscience, Baylor University, Waco, One Bear Place 97310, Waco TX, USA
[57]Psychology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, Wundtlaan 1, 6525XD, The Netherlands
[57]Department of Psychology, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, 7 George Square, EH8 9JZ Edinburgh, UK
[58]Dept of Philosophy, Major Williams Hall, Virginia Tech, Blacksburg, VA, US
[59]Center for the Study of Family Violence and Sexual Assault, Northern Illinois University, DeKalb, IL, 125 President's BLVD., DeKalb, IL 60115, USA
[60]School of Mathematics and Statistics, The Open University, Milton Keynes, Walton Hall, Milton Keynes MK7 6AA, UK
[61]Skyscanner, 15 Laurison Place, Edinburgh, EH3 9EN, UK
[62]School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
[63]Stress Research Institute, Stockholm University, Stockholm, Frescati Hagväg 16A, SE-10691 Stockholm, Sweden
[63]Department of Clinical Neuroscience, Karolinska Institutet, Nobels väg 9, SE-17177 Stockholm, Sweden
[63]Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA
[64]Laboratory of Behavioral Neurobiology, Department of Physiological Sciences, Federal University of Santa Catarina, Florianópolis, Campus Universitário Trindade, 88040900, Brazil
[65]Department of Kinesiology, KU Leuven, Leuven, Tervuursevest 101 box 1501, B-3001 Leuven, Belgium
[66]Department of Experimental Psychology, University of Oxford, Oxford, UK
[67]Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway
[68]Department of Psychology, DePaul University, Chicago, 2219 N Kenmore Ave, Chicago, IL 60657, USA

[69]Center for Neural Science, New York University, 4 Washington Pl Room 809 New York, NY 10003, USA
[70]Department of Psychology, Boğaziçi University, Bebek, 34342, Istanbul, Turkey
[71]Psychology, University of Cologne, Cologne,Herbert-Lewin-St. 2, 50931, Cologne, Germany
[72]Saudi Human Genome Program, King Abdulaziz City for Science and Technology (KACST); Integrated Gulf Biosystems, Riyadh, Saudi Arabia
[73]Cognitive Psychology Unit, Institute of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands
[73]Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands
[74]Leeds Institute of Health Sciences, University of Leeds, Leeds, LS2 9NL, UK
[75]Institute for Media Studies, KU Leuven, Leuven, Belgium
[76]Center for Open Science, 210 Ridge McIntire Rd Suite 500, Charlottesville, VA 22903, USA
[76]Department of Engineering and Society, University of Virginia, Thornton Hall, P.O. Box 400259, Charlottesville, VA 22904, USA
[77]Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA
[78]Department of Psychology, University of Southern Indiana, 8600 University Boulevard, Evansville, Indiana, USA
[79]Life Sciences Communication, University of Wisconsin-Madison, Madison, Wisconsin, 1545 Observatory Drive, Madison, WI 53706, USA
[80]Department of Social Psychology, Institute of Psychology, University of Lausanne, Quartier UNIL-Mouline, Bâtiment Géopolis, CH-1015 Lausanne, Switzerland
[81]Departamento de Psicología Básica, Universidad Autónoma de Madrid, c/ Ivan Pavlov 6, 28049 Madrid, Spain
[82]Department of Methodology and Statistics, Tilburg University, Warandelaan 2, 5000 LE Tilburg, The Netherlands

[82]Department of Sociology, Utrecht University, Padualaan 14, 3584 CH, Utrecht, The Netherlands
[83]School of Psychology, Massey University, Auckland, Private Bag 102904, North Shore, Auckland, 0745, New Zealand
[84]Psychology, Saint Louis University, St. Louis, MO, 3700 Lindell Blvd, St. Louis, MO 63108, USA
[85]Psychology, University of California, Davis, Davis, One Shields Ave, Davis, CA 95616, USA
[86]Department of Psychology, University of Texas at Austin, 108 E. Dean Keeton Stop A8000, Austin, TX 78712-1043, USA
[87]Marketing Department, Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium
[88]Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, Burgemeester Oudlaan 50, 3000 DR, Rotterdam, The Netherlands

Author Contributions. Daniel Lakens, Nicholas W. Fox, Monica Gonzalez-Marquez, James A. Grange, Nicholas P. Holmes, Ahmed A. Khalil, Stephen R. Martin, Vishnu Sreekumar, and Crystal N. Steltenpohl participated in brainstorming, drafting the commentary, and data-analysis. Casper J. Albers, Shlomo E. Argamon, Thom Baguley, Erin M. Buchanan, Ben van Calster, Zander Crook, Sameera Daniels, Daniel J. Dunleavy, Brian D. Earp, Jason D. Ferrell, James G. Field, Anne-Laura van Harmelen, Michael Ingre, Peder M. Isager, Hanna K. Isotalus, Junpeng Lao, Gerine M. A. Lodder, David Manheim, Andrea E. Martin, Kevin McConway, Amanda Q. X. Nio, Gustav Nilsonne, Cilene Lino de Oliveira, Jean-Jacques Orban de Xivry, Gerit Pfuhl, Kimberly A. Quinn, Iris K. Schneider, Zsuzsika Sjoerds, Samuel G. Smith, Jeffrey R. Spies, Marcel A. L. M. Van Assen, Matt N. Williams, Donald R. Williams, Tal Yarkoni, and Rolf A. Zwaan participated in brainstorming and drafting the commentary. Federico G. Adolfi, Raymond B. Becker, Michele I. Feist, and Sam Parsons participated in drafting the commentary, and data-analysis. Matthew A. J. Apps, Stephen D. Benning, Daniel E. Bradford, Sau-Chin Chen, Bryan Chung, Lincoln J Colling, Henrik Danielsson, Lisa DeBruine, Mark R. Hoffarth, Erik Gahner Larsen, Randy J. McCarthy, John J. Sakon, S. Adil

Saribay, Tim Smits, Neil Stenhouse, Wojciech Świątkowski, and Miguel A. Vadillo participated in brainstorming. Farid Anvari, Aaron R. Caldwell, Rickard Carlsson, Emily S. Cross, Amanda Friesen, Caio Gomes, Andrew P. Grieve, Robert Guggenberger, James Grist, Kevin D. Hochard, Christer Johansson, Konrad Juszczyk, David A. Kenny, Barbara Konat, Jiří Lukavský, Christopher R. Madan, Deborah G. Mayo, Colin McFarland, Manojkumar Selvaraju, Samantha E. Williams, and Ignazio Ziano did not participate in drafting the commentary because the points that they would have raised had already been incorporated into the commentary, or endorse a sufficiently large part of the contents as if participation had occurred. Except for the first author, authorship order is alphabetical.

Acknowledgements: We'd like to thank Dale Barr, Felix Cheung, David Colquhoun, Hans IJzerman, Harvey Motulsky, and Richard Morey for helpful discussions while drafting this commentary. Daniel Lakens was supported by NWO VIDI 452-17-013. Federico G. Adolfi was supported by CONICET. Matthew Apps was funded by a Biotechnology and Biological Sciences Research Council AFL Fellowship (BB/M013596/1). Gary Collins was supported by the NIHR Biomedical Research Centre, Oxford. Zander Crook was supported by the Economic and Social Research Council [grant number C106891X]. Emily S. Cross was supported by the European Research Council (ERC-2015-StG-677270). Lisa DeBruine is supported by the European Research Council (ERC-2014-CoG-647910 KINSHIP). Anne-Laura van Harmelen is funded by a Royal Society Dorothy Hodgkin Fellowship (DH150176). Mark R. Hoffarth was supported by the National Science Foundation under grant SBE SPRF-FR 1714446. Junpeng Lao was supported by the SNSF grant 100014_156490/1. Cilene Lino de Oliveira was supported by AvH, Capes, CNPq. Andrea E. Martin was supported by the Economic and Social Research Council of the United Kingdom [grant number ES/K009095/1]. Jean-Jacques Orban de Xivry is supported by an internal grant from the KU Leuven (STG/14/054) and by the Fonds voor Wetenschappelijk Onderzoek (1519916N). Sam Parsons was supported by the European Research Council (FP7/2007–2013; ERC grant agreement no; 324176). Gerine Lodder was funded by NWO VICI 453-14-

016. Samuel Smith is supported by a Cancer Research UK Fellowship (C42785/A17965). Vishnu Sreekumar was supported by the NINDS Intramural Research Program (IRP). Miguel A. Vadillo was supported by Grant 2016-T1/SOC-1395 from Comunidad de Madrid. Tal Yarkoni was supported by NIH award R01MH109682.

Competing Interests: The authors declare no competing interests.

Abstract: In response to recommendations to redefine statistical significance to $p \leq .005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level.

Justify Your Alpha

Benjamin et al.[1] proposed changing the conventional "statistical significance" threshold (i.e., the alpha level) from $p \leq .05$ to $p \leq .005$ for all novel claims with relatively low prior odds. They provided two arguments for why lowering the significance threshold would "immediately improve the reproducibility of scientific research." First, a p-value near .05 provides weak evidence for the alternative hypothesis. Second, under certain assumptions, an alpha of .05 leads to high false positive report probabilities (FPRP[2]; the probability that a significant finding is a false positive).

We share their concerns regarding the apparent non-replicability of many scientific studies, and agree that a universal alpha of .05 is undesirable. However, redefining "statistical significance" to a lower, but equally arbitrary threshold, is inadvisable for three reasons: (1) there is insufficient evidence that the current standard is a "leading cause of non-reproducibility"[1]; (2) the arguments in favor of a blanket default of $p \leq .005$ do not warrant the immediate and widespread implementation of such a policy; and (3) a lower significance threshold will likely have negative consequences not discussed by Benjamin and colleagues. We conclude that the term "statistically significant" should no longer be used and suggest that researchers employing null hypothesis significance testing justify their choice for an alpha level before collecting the data, instead of adopting a new uniform standard.

Lack of evidence that $p \leq .005$ improves replicability

Benjamin et al.[1] claimed that the expected proportion of replicable studies should be considerably higher for studies observing $p \leq .005$ than for studies observing $.005 < p \leq .05$, due to a lower FPRP. Theoretically, replicability is related to the FPRP, and lower alpha levels will reduce false positive results in the literature. However, in practice, the impact of lowering alpha levels depends on several unknowns, such as the prior odds that the

examined hypotheses are true, the statistical power of studies, and the (change in) behavior of researchers in response to any modified standards.

An analysis of the results of the Reproducibility Project: Psychology[3] showed that 49% (23/47) of the original findings with p-values below .005 yielded $p \leq .05$ in the replication study, whereas only 24% (11/45) of the original studies with $.005 < p \leq .05$ yielded $p \leq .05$ ($\chi^2(1) = 5.92$, $p = .015$, $BF_{10} = 6.84$). Benjamin and colleagues presented this as evidence of "potential gains in reproducibility that would accrue from the new threshold." According to their own proposal, however, this evidence is only "suggestive" of such a conclusion, and there is considerable variation in replication rates across p-values (see Figure 1). Importantly, lower replication rates for p-values just below .05 are likely confounded by p-hacking (the practice of flexibly analyzing data until the p-value passes the "significance" threshold). Thus, the differences in replication rates between studies with $.005 < p \leq .05$ compared to those with $p \leq .005$ may not be entirely due to the level of evidence. Further analyses are needed to explain the low (49%) replication rate of studies with $p \leq .005$, before this alpha level is recommended as a new significance threshold for novel discoveries across scientific disciplines.

Weak justifications for the α = .005 threshold

We agree with Benjamin et al. that single p-values close to .05 never provide strong "evidence" against the null hypothesis. Nonetheless, the argument that p-values provide weak evidence based on Bayes factors has been questioned[4]. Given that the marginal likelihood is sensitive to different choices for the models being compared, redefining alpha levels as a function of the Bayes factor is undesirable. For instance, Benjamin and colleagues stated that p-values of .005 imply Bayes factors between 14 and 26. However, these upper bounds only hold for a Bayes factor based on a point null model and when the p-value is calculated for a two-sided test, whereas one-sided tests or Bayes factors for non-

point null models would imply different alpha thresholds. When a test yields BF = 25 the data are interpreted as strong relative evidence for a specific alternative (e.g., μ = 2.81), while a p ≤ .005 only warrants the more modest rejection of a null effect without allowing one to reject even small positive effects with a reasonable error rate[5]. Benjamin et al. provided no rationale for why the new p-value threshold should align with equally arbitrary Bayes factor thresholds. We question the idea that the alpha level at which an error rate is controlled should be based on the amount of relative evidence indicated by Bayes factors.

The second argument for α = .005 is that the FPRP can be high with α = .05. Calculating the FPRP requires a definition of the alpha level, the power of the tests examining true effects, and the ratio of true to false hypotheses tested (the prior odds). Figure 2 in Benjamin et al. displays FPRPs for scenarios where most hypotheses are false, with prior odds of 1:5, 1:10, and 1:40. The recommended p ≤ .005 threshold reduces the minimum FPRP to less than 5%, assuming 1:10 prior odds (the true FPRP might still be substantially higher in studies with very low power). This prior odds estimate is based on data from the Reproducibility Project: Psychology[3] using an analysis modelling publication bias for 73 studies[6]. Without stating the reference class for the "base-rate of true nulls" (e.g., does this refer to all hypotheses in science, in a discipline, or by a single researcher?), the concept of "prior odds that H1 is true" has little meaning. Furthermore, there is insufficient representative data to accurately estimate the prior odds that researchers examine a true hypothesis, and thus, there is currently no strong argument based on FPRP to redefine statistical significance.

How a threshold of p ≤ .005 might harm scientific practice

Benjamin et al. acknowledged that their proposal has strengths as well as weaknesses, but believe that its "efficacy gains would far outweigh losses." We are not convinced and see at least three likely negative consequences of adopting a lowered threshold.

Risk of fewer replication studies. All else being equal, lowering the alpha level requires larger sample sizes and creates an even greater strain on already limited resources. Achieving 80% power with $\alpha = .005$, compared to $\alpha = .05$, requires a 70% larger sample size for between-subjects designs with two-sided tests (88% for one-sided tests). While Benjamin et al. propose $\alpha = .005$ exclusively for "new effects" (and not replications), designing larger original studies would leave fewer resources (i.e., time, money, participants) for replication studies, assuming fixed resources overall. At a time when replications are already relatively rare and unrewarded, lowering alpha to .005 might therefore reduce resources spent on replicating the work of others. More generally, recommendations for evidence thresholds need to carefully balance statistical and non-statistical considerations (e.g., the value of evidence for a novel claim vs. the value of independent replications).

Risk of reduced generalisability and breadth. Requiring larger sample sizes across scientific disciplines may exacerbate over-reliance on convenience samples (e.g., undergraduate students, online samples). Specifically, without (1) increased funding, (2) a reward system that values large-scale collaboration, and (3) clear recommendations for how to evaluate research with sample size constraints, lowering the significance threshold could adversely affect the breadth of research questions examined. Compared to studies that use convenience samples, studies with unique populations (e.g., people with rare genetic variants, patients with post-traumatic stress disorder) or with time- or resource-intensive data collection (e.g., longitudinal studies) require considerably more research funds and effort to increase the sample size. Thus, researchers may become less motivated to study unique populations or collect difficult-to-obtain data, reducing the generalisability and breadth of findings.

Risk of exaggerating the focus on single p-values. Benjamin et al.'s proposal risks (1) reinforcing the idea that relying on p-values is a sufficient, if imperfect, way to evaluate findings, and (2) discouraging opportunities for more fruitful changes in scientific practice

and education. Even though Benjamin et al. do not propose p ≤ .005 as a publication threshold, some bias in favor of significant results will remain, in which case redefining p ≤ .005 as "statistically significant" would result in greater upward bias in effect size estimates. Furthermore, it diverts attention from the cumulative evaluation of findings, such as converging results of multiple (replication) studies.

No one alpha to rule them all

We have two key recommendations. First, we recommend that the label “statistically significant” should no longer be used. Instead, researchers should provide more meaningful interpretations of the theoretical or practical relevance of their results. Second, authors should transparently specify—and justify—their design choices. Depending on their choice of statistical approach, these may include the alpha level, the null and alternative models, assumed prior odds, statistical power for a specified effect size of interest, the sample size, and/or the desired accuracy of estimation. We do not endorse a single value for any design parameter, but instead propose that authors justify their choices before data are collected. Fellow researchers can then evaluate these decisions, ideally also prior to data collection, for example, by reviewing a Registered Report submission[7]. Providing researchers (and reviewers) with accessible information about ways to justify (and evaluate) design choices, tailored to specific research areas, will improve current research practices.

Benjamin et al. noted that some fields, such as genomics and physics, have lowered the “default” alpha level. However, in genomics the overall false positive rate is still controlled at 5%; the lower alpha level is only used to correct for multiple comparisons. In physics, researchers have argued against a blanket rule, and for an alpha level based on factors such as the surprisingness of the predicted result and its practical or theoretical impact[8]. In non-human animal research, minimizing the number of animals used needs to be directly balanced against the probability and cost of false positives. Depending on these and other

considerations, the optimal alpha level for a given research question could be higher or lower than the current convention of .05[9,10,11].

Benjamin et al. stated that a “critical mass of researchers” endorse the standard of a $p \leq .005$ threshold for “statistical significance.” However, the presence of a critical mass can only be identified after a norm has been widely adopted, not before. Even if a $p \leq .005$ threshold were widely accepted, this would only reinforce the misconception that a single alpha level is universally applicable. Ideally, the alpha level is determined by comparing costs and benefits against a utility function using decision theory[12]. This cost-benefit analysis (and thus the alpha level)[13] differs when analyzing large existing datasets compared to collecting data from hard-to-obtain samples.

Conclusion

Science is diverse, and it is up to scientists to justify the alpha level they decide to use. As Fisher noted[14]: "...no scientific worker has a fixed level of significance at which, from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." Research should be guided by principles of rigorous science[15], not by heuristics and arbitrary blanket thresholds. These principles include not only sound statistical analyses, but also experimental redundancy (e.g., replication, validation, and generalisation), avoidance of logical traps, intellectual honesty, research workflow transparency, and accounting for potential sources of error. Single studies, regardless of their p-value, are never enough to conclude that there is strong evidence for a substantive claim. We need to train researchers to assess cumulative evidence and work towards an unbiased scientific literature. We call for a broader mandate beyond p-value thresholds whereby all justifications of key choices in research design and statistical practice are transparently evaluated, fully accessible, and pre-registered whenever feasible.

References

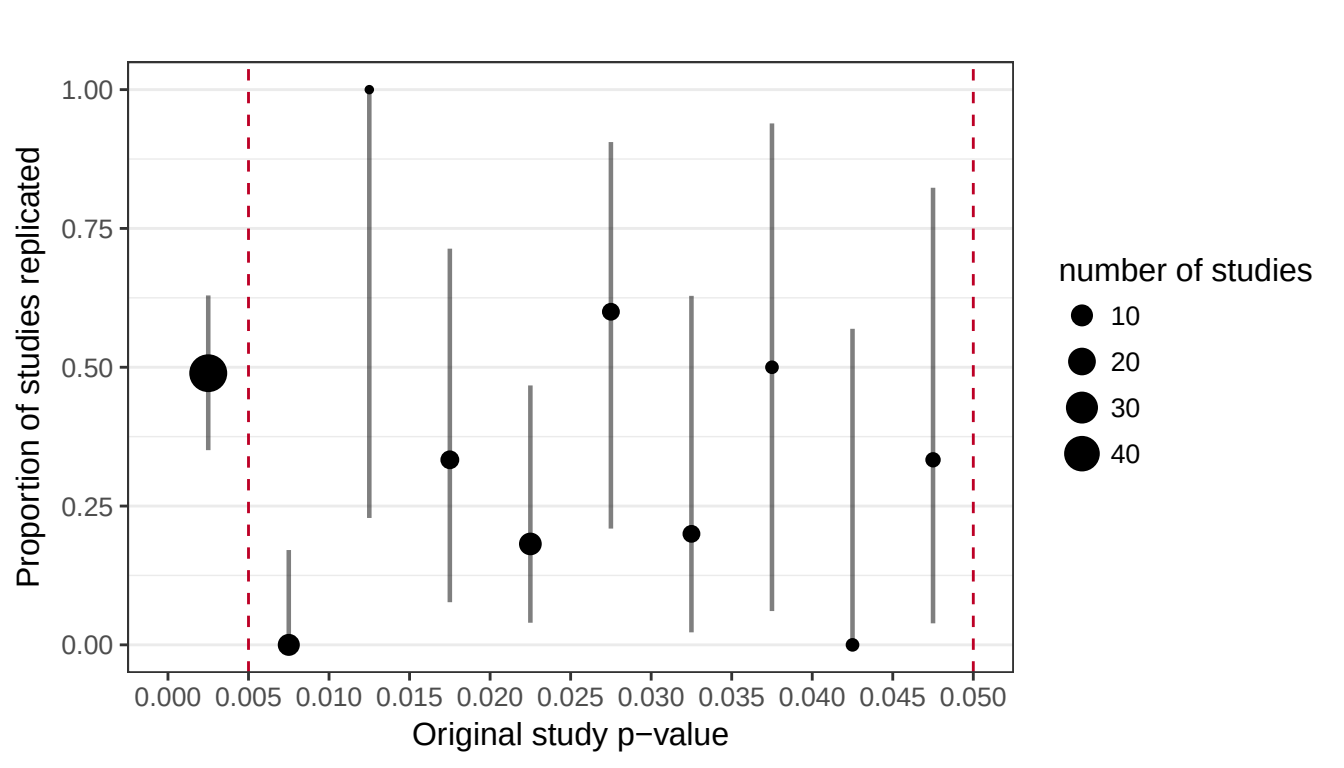
1. Benjamin, D. J., et al. Nature Human Behaviour 2, 6-10 https://doi.org/10.1038/s41562-017-0189-z (2017).
2. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. Journal of the National Cancer Institute 96, 434-442 https://doi.org/10.1093/jnci/djh075 (2004).
3. Open Science Collaboration. (2015). Science 349 (6251), 1-8 https://doi.org/10.1126/science.aac4716 (2015).
4. Senn, S. Statistical issues in drug development (2nd ed). (John Wiley & Sons, 2007).
5. Mayo, D. Statistical inference as severe testing: How to get beyond the statistics wars. (Cambridge University Press, 2018).
6. Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. Journal of the American Statistical Association 112(517), 1–10 https://doi.org/10.1080/01621459.2016.1240079 (2017).
7. Chambers, C.D., Dienes, Z., McIntosh, R.D., Rotshtein, P., & Willmes, K. Cortex 66, A1-2 https://doi.org/10.1016/j.cortex.2015.03.022 (2015).
8. Lyons, L. Discovering the Significance of 5 sigma. Preprint at http://arxiv.org/abs/1310.1284 (2013).
9. Field, S. A., Tyre, A. J., Jonzen, N., Rhodes, J. R., & Possingham, H. P. Ecology Letters 7(8), 669-675 https://doi.org/10.1111/j.1461-0248.2004.00625.x (2004).
10. Grieve, A. P. Pharmaceutical Statistics 14(2), 139–150 https://doi.org/10.1002/pst.1667 (2015).
11. Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. PLOS ONE 7(2), e32734 https://doi.org/10.1371/journal.pone.0032734 (2012).
12. Skipper, J. K., Guenther, A. L., & Nass, G. The American Sociologist 2(1), 16–18 (1967).
13. Neyman, J., & Pearson, E. S. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 231 694–706 https://doi.org/10.1098/rsta.1933.0009 (1933).

14. Fisher R. A. Statistical methods and scientific inferences. (Hafner, 1956).

15. Casadevall, A., & Fang, F. C. mBio 7(6), e01902-16. https://doi.org/10.1128/mbio.01902-16 (2016).

Figure Caption

Figure 1. The proportion of studies[3] replicated at α = .05 (with a bin width of .005). Window start and end positions are plotted on the horizontal axis. The error bars denote 95% Jeffreys confidence intervals. R code to reproduce Figure 1 is available from https://osf.io/by2kc/.


Proportion of studies replicated
1.00
0.75
0.50
0.25
0.00
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
0.050
Original study p−value
number of studies
10
20
30
40