

1 **Justify Your Alpha**

2 In Press, Nature Human Behavior

3
4 Daniel Lakens*¹, Federico G. Adolfini², Casper J. Albers³, Farid Anvari⁴, Matthew A. J. Apps⁵,
5 Shlomo E. Argamon⁶, Thom Baguley⁷, Raymond B. Becker⁸, Stephen D. Benning⁹, Daniel E.
6 Bradford¹⁰, Erin M. Buchanan¹¹, Aaron R. Caldwell¹², Ben van Calster¹³, Rickard Carlsson¹⁴,
7 Sau-Chin Chen¹⁵, Bryan Chung¹⁶, Lincoln J Colling¹⁷, Gary S. Collins¹⁸, Zander Crook¹⁹,
8 Emily S. Cross²⁰, Sameera Daniels²¹, Henrik Danielsson²², Lisa DeBruine²³, Daniel J.
9 Dunleavy²⁴, Brian D. Earp²⁵, Michele I. Feist²⁶, Jason D. Ferrell²⁷, James G. Field²⁸, Nicholas
10 W. Fox²⁹, Amanda Friesen³⁰, Caio Gomes³¹, Monica Gonzalez-Marquez³², James A.
11 Grange³³, Andrew P. Grieve³⁴, Robert Guggenberger³⁵, James Grist³⁶, Anne-Laura van
12 Harmelen³⁷, Fred Hasselman³⁸, Kevin D. Hochard³⁹, Mark R. Hoffarth⁴⁰, Nicholas P.
13 Holmes⁴¹, Michael Ingre⁴², Peder M. Isager⁴³, Hanna K. Isotalus⁴⁴, Christer Johansson⁴⁵,
14 Konrad Juszczyk⁴⁶, David A. Kenny⁴⁷, Ahmed A. Khalil⁴⁸, Barbara Konat⁴⁹, Junpeng Lao⁵⁰,
15 Erik Gahner Larsen⁵¹, Gerine M. A. Lodder⁵², Jiří Lukavský⁵³, Christopher R. Madan⁵⁴, David
16 Manheim⁵⁵, Stephen R. Martin⁵⁶, Andrea E. Martin⁵⁷, Deborah G. Mayo⁵⁸, Randy J.
17 McCarthy⁵⁹, Kevin McConway⁶⁰, Colin McFarland⁶¹, Amanda Q. X. Nio⁶², Gustav Nilsson⁶³,
18 Cilene Lino de Oliveira⁶⁴, Jean-Jacques Orban de Xivry⁶⁵, Sam Parsons⁶⁶, Gerit Pfuhl⁶⁷,
19 Kimberly A. Quinn⁶⁸, John J. Sakon⁶⁹, S. Adil Saribay⁷⁰, Iris K. Schneider⁷¹, Manojkumar
20 Selvaraju⁷², Zsuzsika Sjoerds⁷³, Samuel G. Smith⁷⁴, Tim Smits⁷⁵, Jeffrey R. Spies⁷⁶, Vishnu
21 Sreekumar⁷⁷, Crystal N. Steltenpohl⁷⁸, Neil Stenhouse⁷⁹, Wojciech Świątkowski⁸⁰, Miguel A.
22 Vadillo⁸¹, Marcel A. L. M. Van Assen⁸², Matt N. Williams⁸³, Samantha E. Williams⁸⁴, Donald
23 R. Williams⁸⁵, Tal Yarkoni⁸⁶, Ignazio Ziano⁸⁷, Rolf A. Zwaan⁸⁸

24
25 **Affiliations**

26
27 *¹Human-Technology Interaction, Eindhoven University of Technology, Den Dolech,
28 5600MB, Eindhoven, The Netherlands

1 ²Laboratory of Experimental Psychology and Neuroscience (LPEN), Institute of Cognitive
2 and Translational Neuroscience (INCYT), INECO Foundation, Favaloro University,
3 Pacheco de Melo 1860, Buenos Aires, Argentina

4 ²National Scientific and Technical Research Council (CONICET), Godoy Cruz 2290, Buenos
5 Aires, Argentina

6 ³Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat
7 2/1, 9712TS Groningen, The Netherlands

8 ⁴College of Education, Psychology & Social Work, Flinders University, Adelaide, GPO Box
9 2100, Adelaide, SA, 5001, Australia

10 ⁵Department of Experimental Psychology, University of Oxford, New Radcliffe House,
11 Oxford, OX2 6GG, UK

12 ⁶Department of Computer Science, Illinois Institute of Technology, Chicago, IL, 10 W. 31st
13 Street, Chicago, IL 60645, USA

14 ⁷Department of Psychology, Nottingham Trent University, Nottingham, 50 Shakespeare
15 Street, Nottingham, NG1 4FQ, UK

16 ⁸Faculty of Linguistics and Literature, Bielefeld University, Bielefeld, Universitätsstraße 25,
17 33615 Bielefeld, Germany

18 ⁹Psychology, University of Nevada, Las Vegas, Las Vegas, 4505 S. Maryland Pkwy., Box
19 455030, Las Vegas, NV 89154-5030, USA

20 ¹⁰Psychology, University of Wisconsin-Madison, Madison, 1202 West Johnson St. Madison
21 WI. 53706, USA

22 ¹¹Psychology, Missouri State University, 901 S. National Ave, Springfield, MO, 65897, USA

23 ¹²Health, Human Performance, and Recreation, University of Arkansas, Fayetteville, 155
24 Stadium Drive, HPER 321, Fayetteville, AR, 72701, USA

25 ¹³Department of Development and Regeneration, KU Leuven, Leuven, Herestraat 49 box
26 805, 3000 Leuven, Belgium, Belgium

27 ¹³Department of Medical Statistics and Bioinformatics, Leiden University Medical Center,
28 Postbus 9600, 2300 RC, Leiden, The Netherlands

- 1 ¹⁴Department of Psychology, Linnaeus University, Kalmar, Stagneliusgatan 14, 392 34,
2 Kalmar, Sweden
- 3 ¹⁵Department of Human Development and Psychology, Tzu-Chi University, No. 67, Jieren
4 St., Hualien City, Hualien County, 97074, Taiwan
- 5 ¹⁶Department of Surgery, University of British Columbia, Victoria, #301 - 1625 Oak Bay Ave,
6 Victoria BC Canada, V8R 1B1 , Canada
- 7 ¹⁷Department of Psychology, University of Cambridge, Cambridge CB2 3EB, UK
- 8 ¹⁸Centre for Statistics in Medicine, University of Oxford, Windmill Road, Oxford, OX3 7LD,
9 UK
- 10 ¹⁹Department of Psychology, The University of Edinburgh, 7 George Square, Edinburgh, EH8
11 9JZ, UK
- 12 ²⁰School of Psychology, Bangor University, Bangor, Adeilad Brigantia, Bangor, Gwynedd,
13 LL57 2AS, UK
- 14 ²¹Ramsey Decision Theoretics, 4849 Connecticut Ave. NW #132, Washington, DC 20008,
15 USA
- 16 ²²Department of Behavioural Sciences and Learning, Linköping University, SE-581 83,
17 Linköping, Sweden
- 18 ²³Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, 58 Hillhead
19 Street, UK
- 20 ²⁴College of Social Work, Florida State University, 296 Champions Way, University Center C,
21 Tallahassee, FL, 32304, USA
- 22 ²⁵Departments of Psychology and Philosophy, Yale University, 2 Hillhouse Ave, New Haven
23 CT 06511, USA
- 24 ²⁶Department of English, University of Louisiana at Lafayette, P. O. Box 43719, Lafayette LA
25 70504, USA
- 26 ²⁷Department of Psychology, St. Edward's University, 3001 S. Congress, Austin, TX 78704,
27 USA

1 ²⁷Department of Psychology, University of Texas at Austin, 108 E. Dean Keeton Stop A8000,
2 Austin, TX 78712-1043, USA

3 ²⁸Department of Management, West Virginia University, 1602 University Avenue,
4 Morgantown, WV 26506, USA

5 ²⁹Department of Psychology, Rutgers University, New Brunswick, 53 Avenue E, Piscataway
6 NJ 08854, USA

7 ³⁰Department of Political Science, Indiana University Purdue University, Indianapolis,
8 Indianapolis, 425 University Blvd CA417, Indianapolis, IN 46202, USA

9 ³¹Booking.com, Herengracht 597, 1017 CE Amsterdam, The Netherlands

10 ³²Department of English, American and Romance Studies, RWTH - Aachen University,
11 Aachen, Kármánstraße 17/19, 52062 Aachen, Germany

12 ³³School of Psychology, Keele University, Keele, Staffordshire, ST5 5BG, UK

13 ³⁴Centre of Excellence for Statistical Innovation, UCB Celltech, 208 Bath Road, Slough,
14 Berkshire SL1 3WE, UK

15 ³⁵Translational Neurosurgery, Eberhard Karls University Tübingen, Tübingen, Germany

16 ³⁵University Tübingen, International Centre for Ethics in Sciences and Humanities, Germany

17 ³⁶Department of Radiology, University of Cambridge, Box 218, Cambridge Biomedical
18 Campus, CB2 0QQ, UK

19 ³⁷Department of Psychiatry, University of Cambridge, Cambridge, 18b Trumpington Road,
20 CB2 8AH, UK

21 ³⁸Behavioural Science Institute, Radboud University Nijmegen, Montessorilaan 3, 6525 HR,
22 Nijmegen, The Netherlands

23 ³⁹Department of Psychology, University of Chester, Chester, Department of Psychology,
24 University of Chester, Chester, CH1 4BJ, UK

25 ⁴⁰Department of Psychology, New York University, 4 Washington Place, New York, NY
26 10003, USA

27 ⁴¹School of Psychology, University of Nottingham, Nottingham, University Park, NG7 2RD,
28 UK

1 ⁴²None, Independent, Stockholm, Skåpvägen 5, 12245 ENSKEDE, Sweden

2 ⁴³Department of Clinical and Experimental Medicine, University of Linköping, 581 83

3 Linköping,, Sweden

4 ⁴⁴School of Clinical Sciences, University of Bristol, Bristol, Level 2 academic offices, L&R

5 Building, Southmead Hospital, BS10 5NB, UK

6 ⁴⁵Occupational Orthopaedics and Research, Sahlgrenska University Hospital, 413 45

7 Gothenburg, Sweden

8 ⁴⁶The Faculty of Modern Languages and Literatures, Institute of Linguistics, Psycholinguistics

9 Department, Adam Mickiewicz University, Al. Niepodległości 4, 61-874, Poznań, Poland

10 ⁴⁷Department of Psychological Sciences, University of Connecticut, Storrs, CT, Department

11 of Psychological Sciences, U-1020, Storrs, CT 06269-1020, USA

12 ⁴⁸Center for Stroke Research Berlin, Charité - Universitätsmedizin Berlin, Hindenburgdamm

13 30, 12200 Berlin, Germany

14 ⁴⁸Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, 04103

15 Leipzig, Germany

16 ⁴⁸Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Luisenstraße 56, 10115

17 Berlin, Germany

18 ⁴⁰Social Sciences, Adam Mickiewicz University, Poznań, Szamarzewskiego 89, 60-568

19 Poznan, Poland

20 ⁵⁰Department of Psychology, University of Fribourg, Faucigny 2, 1700 Fribourg, Switzerland

21 ⁵¹School of Politics and International Relations, University of Kent, Canterbury CT2 7NX, UK

22 ⁵² Department of Sociology / ICS, University of Groningen, Grote Rozenstraat 31, 9712 TG

23 Groningen, The Netherlands

24 ⁵³Institute of Psychology, Czech Academy of Sciences, Hybernská 8, 11000 Prague, Czech

25 Republic

26 ⁵⁴School of Psychology, University of Nottingham, Nottingham, NG7 2RD, UK

27 ⁵⁵Pardee RAND Graduate School, RAND Corporation, 1200 S Hayes St, Arlington, VA

28 22202, USA

1 ⁵⁶Psychology and Neuroscience, Baylor University, Waco, One Bear Place 97310, Waco TX,
2 USA

3 ⁵⁷Psychology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen,
4 Wundtlaan 1, 6525XD, The Netherlands

5 ⁵⁷Department of Psychology, School of Philosophy, Psychology, and Language Sciences,
6 University of Edinburgh, 7 George Square, EH8 9JZ Edinburgh, UK

7 ⁵⁸Dept of Philosophy, Major Williams Hall, Virginia Tech, Blacksburg, VA, US

8 ⁵⁹Center for the Study of Family Violence and Sexual Assault, Northern Illinois University,
9 DeKalb, IL, 125 President's BLVD., DeKalb, IL 60115, USA

10 ⁶⁰School of Mathematics and Statistics, The Open University, Milton Keynes, Walton Hall,
11 Milton Keynes MK7 6AA, UK

12 ⁶¹Skyscanner, 15 Laurison Place, Edinburgh, EH3 9EN, UK

13 ⁶²School of Biomedical Engineering and Imaging Sciences, King's College London, London,
14 UK

15 ⁶³Stress Research Institute, Stockholm University, Stockholm, Frescati Hagväg 16A, SE-
16 10691 Stockholm, Sweden

17 ⁶³Department of Clinical Neuroscience, Karolinska Institutet, Nobels väg 9, SE-17177
18 Stockholm, Sweden

19 ⁶³Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

20 ⁶⁴Laboratory of Behavioral Neurobiology, Department of Physiological Sciences, Federal
21 University of Santa Catarina, Florianópolis, Campus Universitário Trindade, 88040900,
22 Brazil

23 ⁶⁵Department of Kinesiology, KU Leuven, Leuven, Tervuursevest 101 box 1501, B-3001
24 Leuven, Belgium

25 ⁶⁶Department of Experimental Psychology, University of Oxford, Oxford, UK

26 ⁶⁷Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway

27 ⁶⁸Department of Psychology, DePaul University, Chicago, 2219 N Kenmore Ave, Chicago, IL
28 60657, USA

1 ⁶⁹Center for Neural Science, New York University, 4 Washington PI Room 809 New York, NY
2 10003, USA

3 ⁷⁰Department of Psychology, Boğaziçi University, Bebek, 34342, Istanbul, Turkey

4 ⁷¹Psychology, University of Cologne, Cologne, Herbert-Lewin-St. 2, 50931, Cologne,
5 Germany

6 ⁷²Saudi Human Genome Program, King Abdulaziz City for Science and Technology
7 (KACST); Integrated Gulf Biosystems, Riyadh, Saudi Arabia

8 ⁷³Cognitive Psychology Unit, Institute of Psychology, Leiden University, Wassenaarseweg
9 52, 2333 AK Leiden, The Netherlands

10 ⁷³Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands

11 ⁷⁴Leeds Institute of Health Sciences, University of Leeds, Leeds, LS2 9NL, UK

12 ⁷⁵Institute for Media Studies, KU Leuven, Leuven, Belgium

13 ⁷⁶Center for Open Science, 210 Ridge McIntire Rd Suite 500, Charlottesville, VA 22903, USA

14 ⁷⁶Department of Engineering and Society, University of Virginia, Thornton Hall, P.O. Box
15 400259, Charlottesville, VA 22904, USA

16 ⁷⁷Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke,
17 National Institutes of Health, Bethesda, MD 20892, USA

18 ⁷⁸Department of Psychology, University of Southern Indiana, 8600 University Boulevard,
19 Evansville, Indiana, USA

20 ⁷⁹Life Sciences Communication, University of Wisconsin-Madison, Madison, Wisconsin, 1545
21 Observatory Drive, Madison, WI 53706, USA

22 ⁸⁰Department of Social Psychology, Institute of Psychology, University of Lausanne, Quartier
23 UNIL-Mouline, Bâtiment Géopolis, CH-1015 Lausanne, Switzerland

24 ⁸¹Departamento de Psicología Básica, Universidad Autónoma de Madrid, c/ Ivan Pavlov 6,
25 28049 Madrid, Spain

26 ⁸²Department of Methodology and Statistics, Tilburg University, Warandelaan 2, 5000 LE
27 Tilburg, The Netherlands

1 ⁸²Department of Sociology, Utrecht University, Padualaan 14, 3584 CH, Utrecht, The
2 Netherlands

3 ⁸³School of Psychology, Massey University, Auckland, Private Bag 102904, North Shore,
4 Auckland, 0745, New Zealand

5 ⁸⁴Psychology, Saint Louis University, St. Louis, MO, 3700 Lindell Blvd, St. Louis, MO 63108,
6 USA

7 ⁸⁵Psychology, University of California, Davis, Davis, One Shields Ave, Davis, CA 95616, USA

8 ⁸⁶Department of Psychology, University of Texas at Austin, 108 E. Dean Keeton Stop A8000,
9 Austin, TX 78712-1043, USA

10 ⁸⁷Marketing Department, Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium

11 ⁸⁸Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam,
12 Rotterdam, Burgemeester Oudlaan 50, 3000 DR, Rotterdam, The Netherlands

13

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

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

10

11 **Acknowledgements:** We'd like to thank Dale Barr, Felix Cheung, David Colquhoun, Hans
12 IJzerman, Harvey Motulsky, and Richard Morey for helpful discussions while drafting this
13 commentary. Daniel Lakens was supported by NWO VIDI 452-17-013. Federico G. Adolffi
14 was supported by CONICET. Matthew Apps was funded by a Biotechnology and Biological
15 Sciences Research Council AFL Fellowship (BB/M013596/1). Gary Collins was supported by
16 the NIHR Biomedical Research Centre, Oxford. Zander Crook was supported by the
17 Economic and Social Research Council [grant number C106891X]. Emily S. Cross was
18 supported by the European Research Council (ERC-2015-StG-677270). Lisa DeBruine is
19 supported by the European Research Council (ERC-2014-CoG-647910 KINSHIP). Anne-
20 Laura van Harmelen is funded by a Royal Society Dorothy Hodgkin Fellowship (DH150176).
21 Mark R. Hoffarth was supported by the National Science Foundation under grant SBE
22 SPRF-FR 1714446. Junpeng Lao was supported by the SNSF grant 100014_156490/1.
23 Cilene Lino de Oliveira was supported by AvH, Capes, CNPq. Andrea E. Martin was
24 supported by the Economic and Social Research Council of the United Kingdom [grant
25 number ES/K009095/1]. Jean-Jacques Orban de Xivry is supported by an internal grant from
26 the KU Leuven (STG/14/054) and by the Fonds voor Wetenschappelijk Onderzoek
27 (1519916N). Sam Parsons was supported by the European Research Council (FP7/2007–
28 2013; ERC grant agreement no; 324176). Gerine Lodder was funded by NWO VICI 453-14-

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

5

6 **Competing Interests:** The authors declare no competing interests.

7

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

11

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 (FPRP²; 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

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

3

4 An analysis of the results of the Reproducibility Project: Psychology³ showed that 49%
5 (23/47) of the original findings with p-values below .005 yielded $p \leq .05$ in the replication
6 study, whereas only 24% (11/45) of the original studies with $.005 < p \leq .05$ yielded $p \leq .05$
7 ($\chi^2(1) = 5.92, p = .015, BF_{10} = 6.84$). Benjamin and colleagues presented this as evidence of
8 “potential gains in reproducibility that would accrue from the new threshold.” According to
9 their own proposal, however, this evidence is only “suggestive” of such a conclusion, and
10 there is considerable variation in replication rates across p-values (see Figure 1).

11 Importantly, lower replication rates for p-values just below .05 are likely confounded by p-
12 hacking (the practice of flexibly analyzing data until the p-value passes the “significance”
13 threshold). Thus, the differences in replication rates between studies with $.005 < p \leq .05$
14 compared to those with $p \leq .005$ may not be entirely due to the level of evidence. Further
15 analyses are needed to explain the low (49%) replication rate of studies with $p \leq .005$, before
16 this alpha level is recommended as a new significance threshold for novel discoveries
17 across scientific disciplines.

18

19 ***Weak justifications for the $\alpha = .005$ threshold***

20

21 We agree with Benjamin et al. that single p-values close to .05 never provide strong
22 “evidence” against the null hypothesis. Nonetheless, the argument that p-values provide
23 weak evidence based on Bayes factors has been questioned⁴. Given that the marginal
24 likelihood is sensitive to different choices for the models being compared, redefining alpha
25 levels as a function of the Bayes factor is undesirable. For instance, Benjamin and
26 colleagues stated that p-values of .005 imply Bayes factors between 14 and 26. However,
27 these upper bounds only hold for a Bayes factor based on a point null model and when the
28 p-value is calculated for a two-sided test, whereas one-sided tests or Bayes factors for non-

1 point null models would imply different alpha thresholds. When a test yields $BF = 25$ the data
2 are interpreted as strong relative evidence for a specific alternative (e.g., $\mu = 2.81$), while a p
3 $\leq .005$ only warrants the more modest rejection of a null effect without allowing one to reject
4 even small positive effects with a reasonable error rate⁵. Benjamin et al. provided no
5 rationale for why the new p-value threshold should align with equally arbitrary Bayes factor
6 thresholds. We question the idea that the alpha level at which an error rate is controlled
7 should be based on the amount of relative evidence indicated by Bayes factors.

8

9 The second argument for $\alpha = .005$ is that the FPRP can be high with $\alpha = .05$. Calculating the
10 FPRP requires a definition of the alpha level, the power of the tests examining true effects,
11 and the ratio of true to false hypotheses tested (the prior odds). Figure 2 in Benjamin et al.
12 displays FPRPs for scenarios where most hypotheses are false, with prior odds of 1:5, 1:10,
13 and 1:40. The recommended $p \leq .005$ threshold reduces the minimum FPRP to less than
14 5%, assuming 1:10 prior odds (the true FPRP might still be substantially higher in studies
15 with very low power). This prior odds estimate is based on data from the Reproducibility
16 Project: Psychology³ using an analysis modelling publication bias for 73 studies⁶. Without
17 stating the reference class for the “base-rate of true nulls” (e.g., does this refer to all
18 hypotheses in science, in a discipline, or by a single researcher?), the concept of “prior odds
19 that H_1 is true” has little meaning. Furthermore, there is insufficient representative data to
20 accurately estimate the prior odds that researchers examine a true hypothesis, and thus,
21 there is currently no strong argument based on FPRP to redefine statistical significance.

22

23 **How a threshold of $p \leq .005$ might harm scientific practice**

24

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

28

1 Risk of fewer replication studies. All else being equal, lowering the alpha level requires larger
2 sample sizes and creates an even greater strain on already limited resources. Achieving
3 80% power with $\alpha = .005$, compared to $\alpha = .05$, requires a 70% larger sample size for
4 between-subjects designs with two-sided tests (88% for one-sided tests). While Benjamin et
5 al. propose $\alpha = .005$ exclusively for “new effects” (and not replications), designing larger
6 original studies would leave fewer resources (i.e., time, money, participants) for replication
7 studies, assuming fixed resources overall. At a time when replications are already relatively
8 rare and unrewarded, lowering alpha to .005 might therefore reduce resources spent on
9 replicating the work of others. More generally, recommendations for evidence thresholds
10 need to carefully balance statistical and non-statistical considerations (e.g., the value of
11 evidence for a novel claim vs. the value of independent replications).

12

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

25

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

1 and education. Even though Benjamin et al. do not propose $p \leq .005$ as a publication
2 threshold, some bias in favor of significant results will remain, in which case redefining $p \leq$
3 $.005$ as "statistically significant" would result in greater upward bias in effect size estimates.
4 Furthermore, it diverts attention from the cumulative evaluation of findings, such as
5 converging results of multiple (replication) studies.

6

7 **No one alpha to rule them all**

8

9 We have two key recommendations. First, we recommend that the label "statistically
10 significant" should no longer be used. Instead, researchers should provide more meaningful
11 interpretations of the theoretical or practical relevance of their results. Second, authors
12 should transparently specify—and justify—their design choices. Depending on their choice of
13 statistical approach, these may include the alpha level, the null and alternative models,
14 assumed prior odds, statistical power for a specified effect size of interest, the sample size,
15 and/or the desired accuracy of estimation. We do not endorse a single value for any design
16 parameter, but instead propose that authors justify their choices before data are collected.
17 Fellow researchers can then evaluate these decisions, ideally also prior to data collection,
18 for example, by reviewing a Registered Report submission⁷. Providing researchers (and
19 reviewers) with accessible information about ways to justify (and evaluate) design choices,
20 tailored to specific research areas, will improve current research practices.

21

22 Benjamin et al. noted that some fields, such as genomics and physics, have lowered the
23 "default" alpha level. However, in genomics the overall false positive rate is still controlled at
24 5%; the lower alpha level is only used to correct for multiple comparisons. In physics,
25 researchers have argued against a blanket rule, and for an alpha level based on factors
26 such as the surprisingness of the predicted result and its practical or theoretical impact⁸. In
27 non-human animal research, minimizing the number of animals used needs to be directly
28 balanced against the probability and cost of false positives. Depending on these and other

1 considerations, the optimal alpha level for a given research question could be higher or
2 lower than the current convention of .05^{9,10,11}.

3
4 Benjamin et al. stated that a “critical mass of researchers” endorse the standard of a $p \leq$
5 .005 threshold for “statistical significance.” However, the presence of a critical mass can only
6 be identified after a norm has been widely adopted, not before. Even if a $p \leq .005$ threshold
7 were widely accepted, this would only reinforce the misconception that a single alpha level is
8 universally applicable. Ideally, the alpha level is determined by comparing costs and benefits
9 against a utility function using decision theory¹². This cost-benefit analysis (and thus the
10 alpha level)¹³ differs when analyzing large existing datasets compared to collecting data from
11 hard-to-obtain samples.

12
13 **Conclusion**

14
15 Science is diverse, and it is up to scientists to justify the alpha level they decide to use. As
16 Fisher noted¹⁴: “...no scientific worker has a fixed level of significance at which, from year to
17 year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each
18 particular case in the light of his evidence and his ideas.” Research should be guided by
19 principles of rigorous science¹⁵, not by heuristics and arbitrary blanket thresholds. These
20 principles include not only sound statistical analyses, but also experimental redundancy
21 (e.g., replication, validation, and generalisation), avoidance of logical traps, intellectual
22 honesty, research workflow transparency, and accounting for potential sources of error.
23 Single studies, regardless of their p-value, are never enough to conclude that there is strong
24 evidence for a substantive claim. We need to train researchers to assess cumulative
25 evidence and work towards an unbiased scientific literature. We call for a broader mandate
26 beyond p-value thresholds whereby all justifications of key choices in research design and
27 statistical practice are transparently evaluated, fully accessible, and pre-registered whenever
28 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
- 3 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).
- 4
- 5 3. Open Science Collaboration. (2015). *Science* 349 (6251), 1-8
- 6 <https://doi.org/10.1126/science.aac4716> (2015).
- 7
- 8 4. Senn, S. *Statistical issues in drug development* (2nd ed). (John Wiley & Sons, 2007).
- 9
- 10 5. Mayo, D. *Statistical inference as severe testing: How to get beyond the statistics wars*. (Cambridge University Press, 2018).
- 11
- 12 6. Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. *Journal of the American Statistical Association* 112(517), 1–10
- 13 <https://doi.org/10.1080/01621459.2016.1240079> (2017).
- 14
- 15 7. Chambers, C.D., Dienes, Z., McIntosh, R.D., Rotshtein, P., & Willmes, K. *Cortex* 66, A1-2
- 16 <https://doi.org/10.1016/j.cortex.2015.03.022> (2015).
- 17
- 18 8. Lyons, L. *Discovering the Significance of 5 sigma*. Preprint at <http://arxiv.org/abs/1310.1284> (2013).
- 19
- 20 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).
- 21
- 22 10. Grieve, A. P. *Pharmaceutical Statistics* 14(2), 139–150 <https://doi.org/10.1002/pst.1667> (2015).
- 23
- 24 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).
- 25
- 26 12. Skipper, J. K., Guenther, A. L., & Nass, G. *The American Sociologist* 2(1), 16–18 (1967).
- 27
- 28 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).

- 1 14. Fisher R. A. Statistical methods and scientific inferences. (Hafner, 1956).
- 2 15. Casadevall, A., & Fang, F. C. mBio 7(6), e01902-16. [https://doi.org/10.1128/mbio.01902-](https://doi.org/10.1128/mbio.01902-16)
- 3 16 (2016).
- 4

1 **Figure Caption**

2

3 Figure 1. The proportion of studies³ replicated at $\alpha = .05$ (with a bin width of .005). Window

4 start and end positions are plotted on the horizontal axis. The error bars denote 95%

5 Jeffreys confidence intervals. R code to reproduce Figure 1 is available from

6 <https://osf.io/by2kc/>.

