

Justify your alpha

In response to recommendations to redefine statistical significance to $P \leq 0.005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level.

Daniel Lakens, Federico G. Adolphi, Casper J. Albers, Farid Anvari, Matthew A. J. Apps, Shlomo E. Argamon, Thom Baguley, Raymond B. Becker, Stephen D. Benning, Daniel E. Bradford, Erin M. Buchanan, Aaron R. Caldwell, Ben Van Calster, Rickard Carlsson, Sau-Chin Chen, Bryan Chung, Lincoln J. Colling, Gary S. Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa DeBruine, Daniel J. Dunleavy, Brian D. Earp, Michele I. Feist, Jason D. Ferrell, James G. Field, Nicholas W. Fox, Amanda Friesen, Caio Gomes, Monica Gonzalez-Marquez, James A. Grange, Andrew P. Grieve, Robert Guggenberger, James Grist, Anne-Laura van Harmelen, Fred Hasselman, Kevin D. Hochard, Mark R. Hoffarth, Nicholas P. Holmes, Michael Ingre, Peder M. Isager, Hanna K. Isotalus, Christer Johansson, Konrad Juszczyk, David A. Kenny, Ahmed A. Khalil, Barbara Konat, Junpeng Lao, Erik Gahner Larsen, Gerine M. A. Lodder, Jiří Lukavský, Christopher R. Madan, David Manheim, Stephen R. Martin, Andrea E. Martin, Deborah G. Mayo, Randy J. McCarthy, Kevin McConway, Colin McFarland, Amanda Q. X. Nio, Gustav Nilsson, Cilene Lino de Oliveira, Jean-Jacques Orban de Xivry, Sam Parsons, Gerit Pfuhl, Kimberly A. Quinn, John J. Sakon, S. Adil Saribay, Iris K. Schneider, Manojkumar Selvaraju, Zsuzsika Sjoerds, Samuel G. Smith, Tim Smits, Jeffrey R. Spies, Vishnu Sreekumar, Crystal N. Steltenpohl, Neil Stenhouse, Wojciech Świątkowski, Miguel A. Vadillo, Marcel A. L. M. Van Assen, Matt N. Williams, Samantha E. Williams, Donald R. Williams, Tal Yarkoni, Ignazio Ziano and Rolf A. Zwaan

Benjamin et al.¹ proposed changing the conventional “statistical significance” threshold (that is, the alpha level) from $P \leq 0.05$ to $P \leq 0.005$ 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 0.05 provides weak evidence for the alternative hypothesis. Second, under certain assumptions, an alpha level of 0.05 leads to high false positive report probabilities (FPRPs²; 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 0.05 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 favour of a blanket default of $P \leq 0.005$ 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.

$P \leq 0.005$ does not improve replicability

Benjamin et al.¹ claimed that the expected proportion of replicable studies should be considerably higher for studies observing $P \leq 0.005$ than for studies observing $0.005 < P \leq 0.05$, 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) behaviour of researchers in response to any modified standards.

An analysis of the results of the Reproducibility Project: Psychology³ showed that 49% (23/47) of the original findings with P values below 0.005 yielded $P \leq 0.05$ in the replication study, whereas only 24% (11/45) of the original studies with $0.005 < P \leq 0.05$ yielded $P \leq 0.05$ ($\chi^2(1) = 5.92$, $P = 0.015$, Bayes factor $BF_{10} = 6.84$). 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 Fig. 1). Importantly, lower replication rates for P values just below 0.05 are likely confounded by P -hacking (the practice of flexibly analysing data until the P value passes the “significance” threshold). Thus, the differences in replication rates between studies with $0.005 < P \leq 0.05$ compared with those with $P \leq 0.005$ may not be entirely due to the level of evidence. Further analyses are needed to explain the low (49%) replication rate of studies with $P \leq 0.005$, before this alpha level is recommended as a new significance threshold for novel discoveries across scientific disciplines.

Weak justifications for $\alpha = 0.005$

We agree with Benjamin et al. that single P values close to 0.05 never provide strong “evidence” against the null hypothesis. Nonetheless, the argument that P values provide weak evidence based on Bayes factors has been questioned⁴. 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 0.005 imply Bayes factors between 14 and 26. However, these upper bounds only hold for a Bayes factor based on a point null

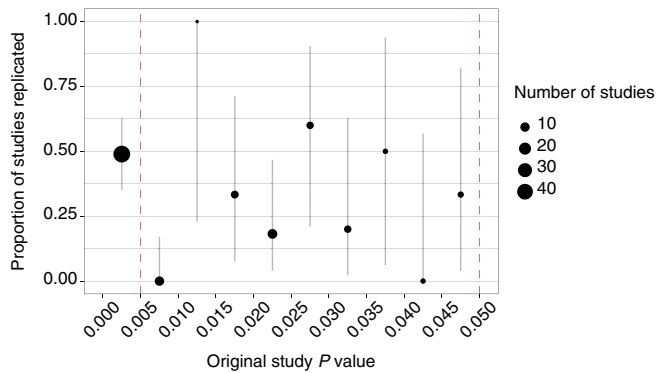


Fig. 1 | The proportion of studies³ replicated at $\alpha = 0.05$ (with a bin width of 0.005). Window start and end positions are plotted on the horizontal axis. Red dashed lines indicate 0.005 and 0.05 thresholds. The error bars denote 95% Jeffreys confidence intervals. R code to reproduce Fig. 1 is available from <https://osf.io/by2kc/>.

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 (for example, mean = 2.81), while a $P \leq 0.005$ only warrants the more modest rejection of a null effect without allowing one to reject even small positive effects with a reasonable error rate⁵. 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 $\alpha = 0.005$ is that the FPRP can be high with $\alpha = 0.05$. 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. shows FPRPs for scenarios where most hypotheses are false, with prior odds of 1:5, 1:10 and 1:40. The recommended $P \leq 0.005$ 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: Psychology³ using an analysis modelling publication bias for 73 studies⁶. Without stating the reference class for the “base rate of true nulls” (for example, does this refer to all hypotheses in science, in a discipline or by a single researcher?), the concept of “prior odds that the alternative hypothesis 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.

$P \leq 0.005$ might harm scientific practice

Benjamin et al. acknowledged that their proposal has strengths as well as weaknesses, but believe that its “efficiency 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 $\alpha = 0.005$, compared with $\alpha = 0.05$, requires a 70% larger sample size for between-subjects designs with two-sided tests (88% for one-sided tests). While Benjamin et al. propose $\alpha = 0.005$ exclusively for “new effects” (and not replications), designing larger original studies would leave fewer resources (that is, time, money, participants) for replication studies, assuming fixed resources overall. At a time when replications are already relatively rare and unrewarded, lowering alpha to 0.005 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 (for example, the value of evidence for a novel claim versus the value of independent replications).

Risk of reduced generalizability and breadth. Requiring larger sample sizes across scientific disciplines may exacerbate over-reliance on convenience samples (for

example, 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 with studies that use convenience samples, studies with unique populations (for example, people with rare genetic variants, patients with post-traumatic stress disorder) or with time- or resource-intensive data collection (for example, 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 generalizability and breadth of findings.

Risk of exaggerating the focus on single P values.

The proposal of Benjamin et al. 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 $P \leq 0.005$ as a publication threshold, some bias in favour of significant results will remain, in which case redefining $P \leq 0.005$ 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 before data collection, for example, by reviewing a registered report submission⁷. 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 impact⁸. 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 0.05 (refs^{9–11}).

Benjamin et al. stated that a “critical mass of researchers” endorse the standard of a $P \leq 0.005$ 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 $P \leq 0.005$ 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 theory¹². This cost–benefit analysis (and thus the alpha level)¹³ differs when analysing large existing datasets compared with 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 noted¹⁴: “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 science¹⁵, not by heuristics and arbitrary blanket thresholds. These principles include not only sound statistical analyses, but also experimental redundancy (for example, replication, validation and generalization), 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. □

Daniel Lakens^{1*}, Federico G. Adolfini^{2,3}, Casper J. Albers⁴, Farid Anvari⁵, Matthew A. J. Apps⁶, Shlomo E. Argamon⁷, Thom Baguley⁸, Raymond B. Becker⁹, Stephen D. Benning¹⁰, Daniel E. Bradford¹¹, Erin M. Buchanan¹², Aaron R. Caldwell¹³, Ben Van Calster^{14,15}, Rickard Carlsson¹⁶, Sau-Chin Chen¹⁷, Bryan Chung¹⁸, Lincoln J. Colling¹⁹, Gary S. Collins²⁰, Zander Crook²¹, Emily S. Cross^{22,23}, Sameera Daniels²⁴, Henrik Danielsson²⁵, Lisa DeBruine²³, Daniel J. Dunleavy²⁶, Brian D. Earp²⁷, Michele I. Feist²⁸, Jason D. Ferrell^{29,30}, James G. Field³¹, Nicholas W. Fox³², Amanda Friesen³³, Caio Gomes³⁴, Monica Gonzalez-Marquez³⁵, James A. Grange³⁶, Andrew P. Grieve³⁷, Robert Guggenberger^{38,39}, James Grist⁴⁰, Anne-Laura van Harmelen⁴¹, Fred Hasselman⁴², Kevin D. Hochard⁴³, Mark R. Hoffarth⁴⁴, Nicholas P. Holmes⁴⁵, Michael Ingre⁹⁷, Peder M. Isager⁴⁷, Hanna K. Isotalus⁴⁸, Christer Johansson⁴⁹, Konrad Juszczyk⁵⁰, David A. Kenny⁵¹, Ahmed A. Khalil^{52,53,54}, Barbara Konat⁵⁵, Junpeng Lao⁵⁶, Erik Gahner Larsen⁵⁷, Gerine M. A. Lodder⁵⁸, Jiří Lukavský⁵⁹, Christopher R. Madan⁴⁵, David Manheim⁶⁰, Stephen R. Martin⁶¹, Andrea E. Martin^{21,62}, Deborah G. Mayo⁶³, Randy J. McCarthy⁶⁴, Kevin McConway⁶⁵, Colin McFarland⁶⁶, Amanda Q. X. Nio⁶⁷, Gustav Nilsson^{68,69,70}, Cilene Lino de Oliveira⁷¹, Jean-Jacques Orban de Xivry⁷², Sam Parsons⁶, Gerit Pfuhl⁷³, Kimberly A. Quinn⁷⁴, John J. Sakon⁷⁵, S. Adil Saribay⁷⁶, Iris K. Schneider⁷⁷, Manojkumar Selvaraju^{78,79}, Zsuzsika Sjoerds^{80,81}, Samuel G. Smith⁸², Tim Smits⁸³, Jeffrey R. Spies^{84,85}, Vishnu Sreekumar⁸⁶, Crystal N. Steltenpohl⁸⁷, Neil Stenhouse⁸⁸, Wojciech Świątkowski⁸⁹, Miguel A. Vadillo⁹⁰, Marcel A. L. M. Van Assen^{91,92}, Matt N. Williams⁹³, Samantha E. Williams⁹⁴, Donald R. Williams⁹⁵, Tal Yarkoni³⁰, Ignazio Ziano⁹⁶ and Rolf A. Zwaan⁴⁶

¹Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, The Netherlands. ²National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina. ³Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany. ⁴Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands. ⁵College of Education, Psychology and Social Work, Flinders University, Adelaide, South Australia, Australia. ⁶Department of Experimental Psychology, University of Oxford, Oxford, UK. ⁷Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA. ⁸Department of Psychology, Nottingham Trent

University, Nottingham, UK. ⁹Faculty of Linguistics and Literature, Bielefeld University, Bielefeld, Germany. ¹⁰Psychology, University of Nevada, Las Vegas, Las Vegas, NV, USA. ¹¹Psychology, University of Wisconsin-Madison, Madison, WI, USA. ¹²Psychology, Missouri State University, Springfield, MO, USA. ¹³Health, Human Performance, and Recreation, University of Arkansas, Fayetteville, AR, USA. ¹⁴Department of Development and Regeneration, KU Leuven, Leuven, Belgium. ¹⁵Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands. ¹⁶Department of Psychology, Linnaeus University, Kalmar, Sweden. ¹⁷Department of Human Development and Psychology, Tzu-Chi University, Hualien City, Taiwan. ¹⁸Department of Surgery, University of British Columbia, Victoria, British Columbia, Canada. ¹⁹Department of Psychology, University of Cambridge, Cambridge, UK. ²⁰Centre for Statistics in Medicine, University of Oxford, Oxford, UK. ²¹Department of Psychology, School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, UK. ²²School of Psychology, Bangor University, Bangor, UK. ²³Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. ²⁴Ramsey Decision Theoretics, Washington, DC, USA. ²⁵Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden. ²⁶College of Social Work, Florida State University, Tallahassee, FL, USA. ²⁷Departments of Psychology and Philosophy, Yale University, New Haven, CT, USA. ²⁸Department of English, University of Louisiana at Lafayette, Lafayette, LA, USA. ²⁹Department of Psychology, St. Edward's University, Austin, TX, USA. ³⁰Department of Psychology, University of Texas at Austin, Austin, TX, USA. ³¹Department of Management, West Virginia University, Morgantown, WV, USA. ³²Department of Psychology, Rutgers University, New Brunswick, NJ, USA. ³³Department of Political Science, Indiana University Purdue University, Indianapolis, IN, USA. ³⁴Booking.com, Amsterdam, The Netherlands. ³⁵Department of English, American and Romance Studies, RWTH Aachen University, Aachen, Germany. ³⁶School of Psychology, Keele University, Keele, UK. ³⁷Centre of Excellence for Statistical Innovation, UCB Celltech, Slough, UK. ³⁸Translational Neurosurgery, Eberhard Karls University Tübingen, Tübingen, Germany. ³⁹International Centre for Ethics in Sciences and Humanities, Eberhard Karls University Tübingen, Tübingen, Germany. ⁴⁰Department of Radiology, University of Cambridge, Cambridge, UK. ⁴¹Department of Psychiatry, University of Cambridge, Cambridge, UK. ⁴²Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, The Netherlands. ⁴³Department of Psychology, University of Chester, Chester, UK. ⁴⁴Department of Psychology, New York University, New York, NY, USA. ⁴⁵School of Psychology, University of Nottingham, Nottingham, UK. ⁴⁶Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands. ⁴⁷Department of Clinical and Experimental Medicine,

University of Linköping, Linköping, Sweden. ⁴⁸School of Clinical Sciences, University of Bristol, Bristol, UK. ⁴⁹Occupational Orthopaedics and Research, Sahlgrenska University Hospital, Gothenburg, Sweden. ⁵⁰The Faculty of Modern Languages and Literatures, Institute of Linguistics, Psycholinguistics Department, Adam Mickiewicz University, Poznań, Poland. ⁵¹Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA. ⁵²Center for Stroke Research Berlin, Charité Universitätsmedizin Berlin, Berlin, Germany. ⁵³Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. ⁵⁴Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany. ⁵⁵Social Sciences, Adam Mickiewicz University, Poznań, Poland. ⁵⁶Department of Psychology, University of Fribourg, Fribourg, Switzerland. ⁵⁷School of Politics and International Relations, University of Kent, Canterbury, UK. ⁵⁸Department of Sociology/ICS, University of Groningen, Groningen, The Netherlands. ⁵⁹Institute of Psychology, Czech Academy of Sciences, Prague, Czech Republic. ⁶⁰Pardee RAND Graduate School, RAND Corporation, Arlington, VA, USA. ⁶¹Psychology and Neuroscience, Baylor University, Waco, TX, USA. ⁶²Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. ⁶³Department of Philosophy, Virginia Tech, Blacksburg, VA, USA. ⁶⁴Center for the Study of Family Violence and Sexual Assault, Northern Illinois University, DeKalb, IL, USA. ⁶⁵School of Mathematics and Statistics, The Open University, Milton Keynes, UK. ⁶⁶Skyscanner, Edinburgh, UK. ⁶⁷School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK. ⁶⁸Stress Research Institute, Stockholm University, Stockholm, Sweden. ⁶⁹Department of Clinical Neuroscience, Stockholm, Sweden. ⁷⁰Department of Psychology, Stanford University, Stanford, CA, USA. ⁷¹Laboratory of Behavioral Neurobiology, Department of Physiological Sciences, Federal University of Santa Catarina, Florianópolis, Brazil. ⁷²Department of Kinesiology, KU Leuven, Leuven, Belgium. ⁷³Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway. ⁷⁴Department of Psychology, DePaul University, Chicago, IL, USA. ⁷⁵Center for Neural Science, New York University, New York, NY, USA. ⁷⁶Department of Psychology, Boğaziçi University, Istanbul, Turkey. ⁷⁷Psychology, University of Cologne, Cologne, Germany. ⁷⁸Saudi Human Genome Program, King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia. ⁷⁹Integrated Gulf Biosystems, Riyadh, Saudi Arabia. ⁸⁰Cognitive Psychology Unit, Institute

of Psychology, Leiden University, Leiden, The Netherlands. ⁸¹Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands. ⁸²Leeds Institute of Health Sciences, University of Leeds, Leeds, UK. ⁸³Institute for Media Studies, KU Leuven, Leuven, Belgium. ⁸⁴Center for Open Science, Charlottesville, VA, USA. ⁸⁵Department of Engineering and Society, University of Virginia, Charlottesville, VA, USA. ⁸⁶Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. ⁸⁷Department of Psychology, University of Southern Indiana, Evansville, IN, USA. ⁸⁸Life Sciences Communication, University of Wisconsin-Madison, Madison, WI, USA. ⁸⁹Department of Social Psychology, Institute of Psychology, University of Lausanne, Lausanne, Switzerland. ⁹⁰Departamento de Psicología Básica, Universidad Autónoma de Madrid, Madrid, Spain. ⁹¹Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands. ⁹²Department of Sociology, Utrecht University, Utrecht, The Netherlands. ⁹³School of Psychology, Massey University, Auckland, New Zealand. ⁹⁴Psychology, Saint Louis University, St. Louis, MO, USA. ⁹⁵Psychology, University of California, Davis, Davis, CA, USA. ⁹⁶Marketing Department, Ghent University, Ghent, Belgium. ⁹⁷Unaffiliated: michael.ingre@gmail.com
*e-mail: D.Lakens@tue.nl

Published online: 26 February 2018

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

References

- Benjamin, D. J. et al. *Nat. Hum. Behav.* **2**, 6–10 (2018).
- Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. *J. Natl Cancer Inst.* **96**, 434–442 (2004).
- Open Science Collaboration *Science* **349**, aac4716 (2015).
- Senn, S. *Statistical Issues in Drug Development* 2nd edn (John Wiley & Sons, Chichester, 2007).
- Mayo, D. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (Cambridge Univ. Press, Cambridge, 2018).
- Johnson, V. E., Payne, R. D., Wang, T., Asher, A. & Mandal, S. *J. Am. Stat. Assoc.* **112**, 1–10 (2017).
- Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P. & Willmes, K. *Cortex* **66**, A1–A2 (2015).
- Lyons, L. Preprint at <http://arxiv.org/abs/1310.1284> (2013).
- Field, S. A., Tyre, A. J., Jonzen, N., Rhodes, J. R. & Possingham, H. *P. Ecol. Lett.* **7**, 669–675 (2004).
- Grieve, A. P. *Pharm. Stat.* **14**, 139–150 (2015).
- Mudge, J. F., Baker, L. F., Edge, C. B. & Houlihan, J. E. *PLoS ONE* **7**, e32734 (2012).
- Skipper, J. K., Guenther, A. L. & Nass, G. *Am. Sociol.* **2**, 16–18 (1967).
- Neyman, J. & Pearson, E. S. *Phil. Trans. R. Soc. Lond. Ser. A* **231**, 694–706 (1933).
- Fisher, R. A. *Statistical Methods and Scientific Inferences* (Hafner, Oxford, 1956).
- Casadevall, A. & Fang, F. C. *mBio* **7**, e01902–16 (2016).

Acknowledgements

We thank D. Barr, F. Cheung, D. Colquhoun, H. IJzerman, H. Motulsky and R. Morey for helpful discussions while drafting this Comment. D.L. was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek 452-17-013. F.G.A. was supported by CONICET. M.A.J.A. was funded by a Biotechnology and Biological Sciences Research Council AFL Fellowship (BB/M013596/1). G.S.C. was supported by the National Institute for Health Research Biomedical Research Centre, Oxford. Z.C. was supported by the Economic and Social Research Council (grant number C106891X). E.S.C. was supported by the European Research Council (ERC-2015-StG-677270). L.D. is supported by the European Research Council (ERC-2014-CoG-647910 KINSHIP). A.-L.v.H. is funded by a Royal Society Dorothy Hodgkin Fellowship (DH150176). M.R.H. was supported by the National Science Foundation under grant SBE SPRF-FR 1714446. J.Lao was supported by the Swiss National Science Foundation grant 100014_156490/1. C.L.d.O. was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico. A.E.M. was supported by the Economic and Social Research Council of the United Kingdom (grant number ES/K009095/1). J.-J.O.d.X. is supported by an internal grant from the KU Leuven (STG/14/054) and by the Fonds voor Wetenschappelijk Onderzoek (1519916N). S.P. was supported by the European Research Council (FP7/2007–2013; ERC grant agreement no. 324176). G.M.A.L. was funded by Nederlandse Organisatie voor Wetenschappelijk Onderzoek 453-14-016. S.G.S. is supported by a Cancer Research UK Fellowship (C42785/A17965). V.S. was supported by the National Institute for Neurological Disorders and Stroke Intramural Research Program (IRP). M.A.V. was supported by grant 2016-T1/SOC-1395 from Comunidad de Madrid. T.Y. was supported by National Institutes of Health award R01MH109682.

Author contributions

D.L., N.W.F., M.G.-M., J.A.G., N.P.H., A.A.K., S.R.M., V.S., S.D.B. and C.N.S. participated in brainstorming, drafting the Comment and data analysis. C.J.A., S.E.A., T.B., E.M.B., B.V.C., Z.C., G.S.C., S.D., D.J.D., B.D.E., J.D.F., J.G.F., A.-L.v.H., M.I., P.M.I., H.K.I., J.Lao, G.M.A.L., D.M., A.E.M., K.M., A.Q.X.N., G.N., C.L.d.O., J.-J.O.d.X., G.P., K.A.Q., I.K.S., Z.S., S.G.S., J.R.S., M.A.L.M.V.A., M.N.W., D.R.W., T.Y. and R.A.Z. participated in brainstorming and drafting the Comment. F.G.A., R.B.B., M.I.F., F.H. and S.P. participated in drafting the Comment and data analysis. M.A.J.A., D.E.B., S.-C.C., B.C., L.J.C., H.D., L.D., M.R.H., E.G.L., R.J.M., J.J.S., S.A.S., T.S., N.S., W.S. and M.A.V. participated in brainstorming. F.A., A.R.C., R.C., E.S.C., A.F., C.G., A.P.G., R.G., J.G., K.D.H., C.J., K.J., D.A.K., B.K., J.Lukavský, C.R.M., D.G.M., C.M., M.S., S.E.W. I.Z. did not participate in drafting the Comment because the points that they would have raised had already been incorporated, or they endorse a sufficiently large part of the contents as if participation had occurred. Except for the first author, authorship order is alphabetical.

Competing interests

The authors declare no competing interests.