Justify Your Alpha: A Response to "Redefine Statistical Significance"

Daniel Lakens[abc1], Federico G. Adolfi[bc2], Casper J. Albers[ab3], Farid Anvari[d4], Matthew A. J. Apps[a5], Shlomo E. Argamon[ab6], Thom Baguley[ab7], Raymond B. Becker[ac8], Stephen D. Benning[a9], Daniel E. Bradford[a10], Erin M. Buchanan[ab11], Aaron R. Caldwell[d12], Ben van Calster[ab13], Rickard Carlsson[d14], Sau-Chin Chen[a15], Bryan Chung[a16], Lincoln J Colling[a17], Gary S. Collins[b18], Zander Crook[ab19], Emily S. Cross[d20], Sameera Daniels[ab21], Henrik Danielsson[a22], Lisa DeBruine[a23], Daniel J. Dunleavy[ab24], Brian D. Earp[ab25], Michele I. Feist[bc26], Jason D. Ferrell[ab27], James G. Field[ab28], Nicholas W. Fox[abc29], Amanda Friesen[d30], Caio Gomes[d31], Monica Gonzalez-Marquez[abc32], James A. Grange[abc33], Andrew P. Grieve[d34], Robert Guggenberger[d35], James Grist[d36], Anne-Laura van Harmelen[ab37], Fred Hasselman[bc38], Kevin D. Hochard[d39], Mark R. Hoffarth[a40], Nicholas P. Holmes[abc41], Michael Ingre[ab42], Peder M. Isager[b43], Hanna K. Isotalus[ab44], Christer Johansson[d45], Konrad Juszczyk[d46], David A. Kenny[d47], Ahmed A. Khalil[abc48], Barbara Konat[d49], Junpeng Lao[ab50], Erik Gahner Larsen[a51], Gerine M. A. Lodder[ab52], Jiří Lukavský,[d53], Christopher R. Madan[d54], David Manheim[ab55], Stephen R. Martin[abc56], Andrea E. Martin[ab57], Deborah G. Mayo[d58], Randy J. McCarthy[a59], Kevin McConway[ab60], Colin McFarland[61], Amanda Q. X. Nio[ab62], Gustav Nilsonne[ab63], Cilene Lino de Oliveira[b64], Jean-Jacques Orban de Xivry[ab65],Sam Parsons[bc66], Gerit Pfuhl[ab67], Kimberly A. Quinn[b68], John J. Sakon[a69], S. Adil Saribay[a70], Iris K. Schneider[ab71], Manojkumar Selvaraju[d72], Zsuzsika Sjoerds[b73], Samuel G. Smith[b74], Tim Smits[a75], Jeffrey R. Spies[b76], Vishnu Sreekumar[abc77], Crystal N. Steltenpohl[abc78], Neil Stenhouse[a79], Wojciech Świątkowski[a80], Miguel A. Vadillo[a81], Marcel A. L. M. Van Assen[ab82], Matt N. Williams[ab83], Samantha E. Williams[d84], Donald R. Williams[ab85], Tal Yarkoni[b86], Ignazio Ziano[d87], Rolf A. Zwaan[ab88]

a) Participated in brainstorming. b) Participated in drafting the commentary. c) Conducted statistical analyses/data preparation. d) Did not participate in a, b, or c, 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.

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 Note: 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.

Funding Statement: 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.

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: A Response to "Redefine Statistical Significance"

"Tests should only be regarded as tools which must be used with discretion and understanding, and not as instruments which in themselves give the final verdict."
Neyman & Pearson, 1928, p. 58.

Renewed concerns about the non-replication of scientific findings have prompted widespread debates about its underlying causes and possible solutions. As an actionable step toward improving standards of evidence for new discoveries, 72 researchers proposed changing the conventional threshold that defines "statistical significance" (i.e., the alpha level) from $p \leq .05$ to $p \leq .005$ for all novel claims with relatively low prior odds (Benjamin et al., 2017). They argued that this change will "immediately improve the reproducibility of scientific research in many fields" (Benjamin et al., 2017, p. 5).

Benjamin et al. (2017) provided two arguments against the current threshold for statistical significance of .05. First, a *p*-value of .05 provides only weak evidence for the alternative hypothesis. Second, under certain assumptions, a *p*-value threshold of .05 leads to a high false positive report probability (FPRP; the probability that a significant finding is a false positive, Wacholder et al., 2004; also referred to as the false positive rate, or false positive risk, Benjamin et al., 2017; Colquhoun, 2017). The authors claim that lowering the threshold for statistical significance to .005 will increase evidential strength for novel discoveries and reduce the FPRP.

We share the concerns raised by Benjamin et al. (2017) regarding the apparent non-replicability[1] of many scientific studies and appreciate their attempt to provide a concrete, easy-to-implement suggestion to improve science. We further agree that the current default alpha level of .05 is arbitrary and may result in weak evidence for the alternative hypothesis. However, we do not think that redefining the threshold for statistical significance to the lower, but equally arbitrary threshold of $p \leq .005$ is advisable. In this commentary, we argue that (1) there is insufficient evidence that the current standard for statistical significance is in fact a "leading cause of non-reproducibility" (Benjamin et al., 2017, p. 5), (2) the arguments in favor of a blanket default of $p \leq .005$ are not strong enough to warrant the immediate and widespread implementation of such a policy, and (3) a lower significance threshold will likely have positive and negative consequences, both of which should be carefully evaluated

[1] We use 'replicability' to refer to the question of whether a conclusion that is sufficiently similar to an earlier study could be drawn from data obtained from a new study, and 'reproducibility' to refer to getting the same results when re-analysing the same data (Peng, 2009).

before any large-scale changes are proposed. We conclude with an alternative suggestion, whereby researchers justify their choice for an alpha level before collecting the data, instead of adopting a new uniform standard.

Lack of evidence that p ≤ .005 improves replicability

One of the main claims made by Benjamin et al. (2017) is that the expected proportion of studies that can be replicated will be considerably higher for studies that observe $p \leq .005$ than for studies that observe $.005 < p \leq .05$, due to a lower FPRP. All else being equal, we agree with Benjamin et al. (2017) that improvement in replicability is *in theory* related to the FPRP, and that lower alpha levels will reduce false positive results in the literature. However, it is difficult to predict how much the FPRP will change *in practice*, because quantifying the FPRP requires accurate estimates of several unknowns, such as the prior odds that the examined hypotheses are true, the true power of any performed experiments, and the (change in) actual behavior of researchers should the newly proposed threshold be put in place.

An analysis of the results of the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015) shows that 49% (23 out of 47) of the original findings with *p*-values below .005 yielded $p \leq .05$ in the replication study, whereas only 24% (11 out of 45) of the original studies with $.005 < p \leq .05$ yielded $p \leq .05$ in the replication study ($\chi^2(1) = 5.92$, $p = .015$, $BF_{10} = 6.84$). Benjamin et al. (2017, p. 9) presented this analysis as empirical evidence of the "potential gains in reproducibility that would accrue from the new threshold." However, as they acknowledged, their obtained *p*-value of .015 is only "suggestive" of such a conclusion, according to their own proposal. Moreover, there is considerable variation in replication rates across *p*-values (see Figure 1), with few observations in bins of size .005 for $.005 < p \leq .05$. In addition, the lower replication rate for *p*-values just below .05 is likely confounded by *p*-hacking (the practice of flexibly analysing data until the *p*-value passes the 'significance' threshold) in the original study. This implies that at least some of the differences in replication rates between studies with $.005 < p \leq .05$ compared to studies with $p \leq .005$ are not due to the level of evidence *per se,* but rather due to other mechanisms (e.g., flexibility during data analysis). Indeed, depending on the degree of flexibility exploited by researchers, such *p*-hacking can be used to overcome any inferential threshold.

Even with a $p \leq .005$ threshold, only 49% of studies replicated successfully. Furthermore, only 11 out of 30 studies (37%) with $.0001 < p \leq .005$ replicated at $\alpha = .05$. By contrast, a *prima facie* more satisfactory replication success rate of 71% was obtained only for $p <$

.0001 (12 out of 17 studies). This suggests that a relatively small number of studies with *p*-values much lower than .005 were largely responsible for the 49% replication rate for studies with $p \leq .005$. Further analysis is needed, therefore, to explain the low 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.

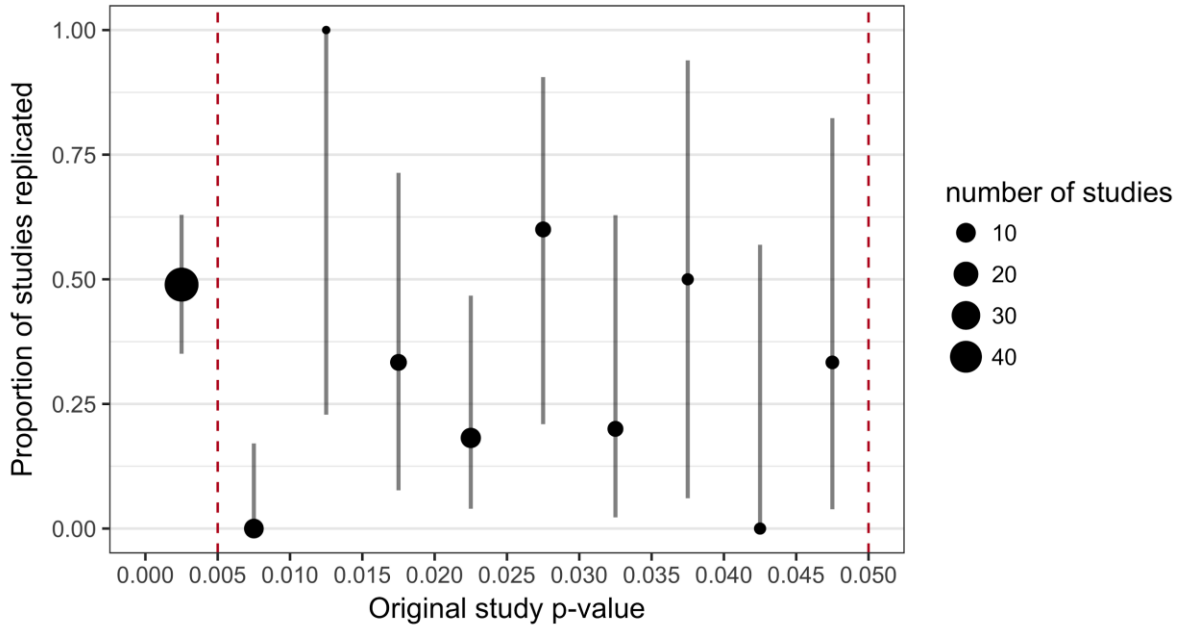

Figure 1. The proportion of studies (Open Science Collaboration, 2015) that replicated at α = .05 (with a bin width of 0.005). Window start and end positions are plotted on the horizontal axis. The error bars denote 95% Jeffreys confidence intervals. R code to reproduce Figure 1 is available from https://github.com/VishnuSreekumar/Alpha005

Weak justifications for the new p ≤ .005 threshold

Even though *p*-values close to .05 never provide strong 'evidence' against the null hypothesis on their own (Wasserstein & Lazar, 2016), the argument that *p*-values provide weak evidence based on Bayes factors has been called into question (Casella & Berger, 1987; Greenland et al., 2016; Senn, 2001). Redefining the alpha level as a function of the strength of relative evidence measured by the Bayes factor is undesirable, given that the marginal likelihood is very sensitive to different (somewhat arbitrary) choices for the models that are compared (Gelman et al., 2013). Benjamin et al. (2017) stated that *p*-values of .005 imply Bayes factors between 14 and 26, but the level of evidence depends on the model priors and the choice of hypotheses tested, and different modelling assumptions would imply a different *p*-value threshold. The Bayesian analysis that underlies the recommendation actually overstates the evidence against the null from the perspective of error statistics. It

would, with high probability, deem an alternative highly probable, even if it's false (Mayo, 1997, 2018). Finally, Benjamin et al. (2017) provided no rationale for why the new *p*-value threshold *should* align with equally arbitrary Bayes factor thresholds representing 'substantial' or 'strong' evidence. Indeed, it has been argued that such classifications of Bayes factors themselves introduce arbitrary meaning to a continuous measure (e.g., Morey, 2015). We (even those of us prepared to use likelihoods and Bayesian approaches in lieu of *p*-values when interpreting results) caution against the idea that the alpha level at which an error rate is controlled should be based on the amount of relative evidence indicated by a Bayes factor. Extending Morey, Wagenmakers, and Rouder (2016), who argued against the frequentist calibration of Bayes factors, we argue against the necessity of a Bayesian calibration of error rates.

The second argument Benjamin et al. (2017) provided for $p \leq .005$ is that the FPRP can be high with $\alpha = .05$. To calculate the FPRP one needs to define the alpha level, the power of the tests that examine true effects, and the ratio of true to false hypotheses tested (the prior odds). The FPRP is only problematic when a high proportion of examined hypotheses are false, and thus Benjamin et al. (2017, p. 10) stated that their "recommendation applies to disciplines with prior odds broadly in the range depicted in Figure 2." Their Figure 2 displays FPRPs for scenarios where many examined hypotheses are false, with ratios of true to false hypotheses (i.e., prior odds) of 1 to 5, 1 to 10, and 1 to 40. Benjamin et al. (2017) recommended $p \leq .005$ because this threshold reduces the *minimum* FPRP to less than 5%, assuming 1 to 10 prior odds of examining a true hypothesis (the true FPRP might still be substantially higher in studies with very low power). This estimate of prior odds is based on data from the RP:P (Open Science Collaboration, 2015) using an analysis that modelled publication bias for 73 studies (Johnson et al., 2017; see also Ingre, 2016, for a more conservative estimate). Without stating the reference class for the 'base-rate of true nulls' (i.e., 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 in practice. The modelling effort by Johnson et al. (2017) ignored practices that inflate error rates (e.g., *p*-hacking) and thus likely does not provide an accurate estimate of bias, given the prevalence of such practices (Fiedler & Schwarz, 2016; John et al., 2012). An estimate of the prior probability that a hypothesis is true, similar to that of Johnson et al. (2017), was derived from 92 participants' subjective ratings of the prior probability that the alternative hypothesis was true for 44 studies included in the RP:P (Dreber et al., 2015). As Dreber et al. (2015, p. 15345) noted, "This relatively low average prior may reflect [the fact] that top psychology journals focus on publishing surprising findings, i.e., positive findings on relatively unlikely hypotheses." These observations imply that there are not sufficient 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 to $p \leq .005$.

Ways in which a threshold of p ≤ .005 might harm scientific practice

Benjamin et al. (2017) acknowledged that lowering the *p*-value threshold will not ameliorate other practices that negatively impact the replicability of research findings (such as *p*-hacking, publication bias, and low power). Yet, they did not address ways in which a $p \leq .005$ threshold might harm scientific practice. Chief among our concerns are (1) a reduction in the number of replication studies that can be conducted if such a threshold is adopted, (2) a concomitant reduction in generalisability and breadth of research findings due to a likely increased reliance on convenience samples, and (3) exacerbation of an already exaggerated focus on single *p*-values.

Risk of fewer replication studies. Replication studies are central to generating reliable scientific knowledge, especially when conclusions are largely based on *p*-values. As Fisher (1926, p. 85) noted: “A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance.” Replication studies are at the heart of scientific progress. In the field of medicine, for example, the FDA requires two independent pre-registered clinical trials, both significant with $p \leq .05$, before issuing marketing approval for new drugs (for a discussion, see Senn, 2007, p. 188). Researchers have limited resources, and when studies require larger sample sizes scientists will have to decide what research they will invest in. Achieving 80% power with $\alpha = .005$, compared to $\alpha = .05$, will require a 70% larger sample size in a between-subjects design with a two-sided test (and an 88% larger sample size for one-sided tests). This means that researchers can complete almost two studies each powered at $\alpha = .05$ (e.g., one novel study and one replication study), or only one study powered at $\alpha = .005$. Therefore, at a time when replication studies are rare, lowering the alpha level to .005 might reduce the number of replication studies. Indeed, general recommendations for evidence thresholds need to carefully balance statistical and non-statistical considerations (e.g., the value of evidence per novel study vs. the value of independent replications).

Risk of reduced generalisability and breadth. All things equal, larger sample sizes increase the informational value of studies, but requiring larger sample sizes across all scientific disciplines would potentially compound problems with over-reliance on convenience samples (such as undergraduate students or Mechanical Turk workers). Lowering the significance threshold could adversely affect the type of breadth of research questions examined if it is

done without (1) increased funding, (2) a reward system that values large-scale collaboration, or (3) clear recommendations for how to evaluate research with lower evidential value due to sample size constraints. Achieving a lower *p*-value in studies with unique populations (e.g., people with rare genetic variants, people diagnosed with post-traumatic stress disorder) or in studies with time- or otherwise resource-intensive data collection (e.g., longitudinal studies) requires exponentially more effort than increasing the amount of evidence in studies that use undergraduate students or Mechanical Turk workers. Thus, researchers may become less motivated, or even tacitly discouraged, to study the former populations or collect those types of data. Hence, lowering the alpha threshold may indirectly reduce the generalisability and breadth of findings (Peterson & Merunka, 2014).

Risk of exaggerating the focus on single p-values. If anything, an excessive focus on *p*-value thresholds has the potential to mask or even discourage opportunities for more fruitful changes in scientific practice and education. Many researchers have come to recognise *p*-hacking, low power, and publication bias as more important reasons for non-replication. Benjamin et al. (2017) acknowledged that changing the threshold could be considered a distraction from other solutions, and yet their proposal risks reinforcing the idea that relying only on *p*-values is a sufficient, if imperfect, way to evaluate findings. The proposed $p \leq .005$ threshold is not intended as a publication threshold. However, given the long history of misuse of statistical recommendations, there is a substantial risk that redefining $p \leq .005$ as 'statistically significant' will increase publication bias, which, in turn, would bias effect size estimates upwards to an even greater extent (Lane & Dunlap, 1978). As such, Benjamin et al.'s recommendation could divert attention from the burgeoning movement towards a more cumulative evaluation of findings, where the converging results of multiple studies are taken into account when addressing specific research questions. Examples of such approaches are: multiple replications (both registered and multi-lab; see, e.g., Hagger et al., 2016), continuously updating meta-analyses (Braver et al., 2014), *p*-curve analysis (Simonsohn et al., 2014), and pre-registration of studies.

No one alpha to rule them all

Benjamin et al. (2017) recommended that only *p*-values lower than .005 should be called 'statistically significant' and that studies should generally be designed with $\alpha = .005$. Our recommendation is similarly twofold. First, when describing results, we recommend that the label 'statistically significant' simply no longer be used. Instead, researchers should provide a more meaningful interpretation (Eysenck, 1960). While *p*-values can inspire statements about the probability of data (e.g., 'the observed difference in the data was surprisingly large,

assuming the null hypothesis is true'), they should not be treated as indices that, on their own, signify evidence for a theory.

Second, when designing studies, we propose that authors transparently specify their design choices. These include (where applicable) 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. Without imposing a single value on any of these design parameters, we ask authors to *justify their choices* before the data are collected. Fellow researchers can evaluate these decisions on their merits and discuss how appropriate they are for a specific research question, and whether the conclusions follow from the study design. Ideally, this evaluation process occurs prior to data collection when reviewing a Registered Report submission (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015). Providing researchers (and reviewers) with accessible information on ways to justify (and evaluate) these design choices, tailored to specific research areas, would improve current research practices.

The optimal alpha level will sometimes be lower and sometimes be higher than the current convention of .05 (see Field, Tyre, Jonzén, Rhodes, & Possingham, 2004; Grieves, 2015; Mudge, Baker, Edge, & Houlahan, 2012; Pericchi & Pereira, 2016). Some fields, such as genomics and physics, have lowered the 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 (Storey & Tibshirani, 2003). In physics, a five sigma threshold ($p \leq 2.87 \times 10^{-7}$) is required to publish an article with 'discovery of' in the title, with less stringent alpha levels being used for article titles with 'evidence for' or 'measurement of' (Franklin, 2014). In physics researchers have also argued against a blanket rule, and instead setting the alpha level based on factors such as how surprising the result would be and how much practical or theoretical impact the discovery would have (Lyons, 2013). In non-human animal research, minimising the number of animals used needs to be directly balanced against the probability of false positives; other trade-offs may be relevant in other areas. Thus, a broadly applied $p \leq .005$ threshold will rarely be optimal.

Benjamin et al. (2017, p. 5) stated that a "critical mass of researchers" now 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 or practice has been widely adopted, not *before*. Even if a $p \leq .005$ threshold was widely endorsed, this would only reinforce the flawed idea that a single alpha level is universally applicable. Ideally, the decision of where to set the alpha level for a study should be based on statistical decision theory, where costs

and benefits are compared against a utility function (Neyman & Pearson, 1933; Skipper, Guenther, & Nass, 1967). Such an analysis can be expected to differ based on the type of study being conducted: for example, analysis of a large existing dataset versus primary data collection relying on hard-to-obtain samples. Science is necessarily diverse, and it is up to scientists within specific fields to justify the alpha level they decide to use. To quote Fisher (1956, p. 42): "...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."

Conclusion

It is laudable that Benjamin et al. (2017) suggested a concrete step designed to immediately improve science. However, it is not clear that lowering the significance threshold to $p \leq .005$ will in practice amount to an improvement in replicability that is worth the potential costs. Instead of simple heuristics and an arbitrary blanket threshold, research should be guided by principles of rigorous science (Casadevall & Fang, 2016; LeBel, Vanpaemel, McCarthy, Earp, & Elson, 2017; Meehl, 1990). 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 full 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 *theory*. We need to train researchers to recognise what cumulative evidence looks like and work towards an unbiased scientific literature.

Although we agree with Benjamin et al. (2017) that the relatively high rate of non-replication in the scientific literature is a cause for concern, we do not believe that redefining statistical significance is a desirable solution: (1) there is not enough evidence that a blanket threshold of $p \leq .005$ will improve replication sufficiently to be worth the additional cost in data collection, (2) the justifications given for the new threshold are not strong enough to warrant the widespread implementation of such a policy, and (3) there are realistic concerns that a $p \leq .005$ threshold will have negative consequences for science, which should be carefully examined before a change in practice is instituted. Instead of a narrower focus on *p*-value thresholds, we call for a broader mandate whereby all *justifications* of key choices in research design and statistical practice are pre-registered whenever possible, fully accessible, and transparently evaluated.

References

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. (2017, July 22). Redefine statistical significance. https://doi.org/10.17605/OSF.IO/MKY9J

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333-342. https://doi.org/10.1177/1745691614529796

Casella, G., & Berger, R. L. (1987). Testing Precise Hypotheses: Comment. *Statistical Science*, *2*(3), 344–347. https://doi.org/10.1214/ss/1177013243

Casadevall, A., & Fang, F. C. (2016). Rigorous Science: a How-To Guide. *mBio, 7*(6), e01902-16. https://doi.org/10.1128/mBio.01902-16

Chambers, C.D., Dienes, Z., McIntosh, R.D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1-2. https://doi.org/10.1016/j.cortex.2015.03.022

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of *p*-values. *bioRxiv*, 144337. https://doi.org/10.1101/144337

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., … Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*(50)*,* 15343–15347. https://doi.org/10.1073/pnas.1516179112

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review*, *67*(4), 269-271. http://dx.doi.org/10.1037/h0048412

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. http://doi.org/10.1177/1948550615612150

Field, S. A., Tyre, A. J., Jonzen, N., Rhodes, J. R., & Possingham, H. P. (2004). Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters*, *7*(8), 669-675. https://doi.org/10.1111/j.1461-0248.2004.00625.x

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503-513.

Fisher R. A. (1956). Statistical Methods and Scientific Inferences. Hafner: New York.

Franklin, A. (2014). *Shifting standards: Experiments in particle physics in the twentieth century.* University of Pittsburgh Press.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a

guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grieve, A. P. (2015). How to test hypotheses if you must. *Pharmaceutical Statistics*, *14*(2), 139–150. https://doi.org/10.1002/pst.1667

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11,* 546–573. https://doi.org/10.1177/1745691616652873

Ingre, M. (2016). Recent reproducibility estimates indicate that negative evidence is observed over 30 times before publication. *arXiv preprint arXiv:1605.06414.* https://arxiv.org/abs/1605.06414

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.2139/ssrn.1996631

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association, 112*(517)*,* 1–10. https://doi.org/10.1080/01621459.2016.1240079

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608-614. https://doi.org/10.1177/1745691612462586

Lakens, D. (2015). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ, 3*, e1142. https://doi.org/10.7717/peerj.1142

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31(2)*, 107-112. https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

LeBel, E. P., Vanpaemel, W, McCarthy, R. J., Earp, B. D., & Elson, M. (2017). A Unified Framework to Quantify the Trustworthiness of Empirical Research. https://doi.org/10.17605/OSF.IO/UWMR8

Lyons, L. (2013). Discovering the Significance of 5 sigma. *arXiv preprint arXiv:1310.1284*.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Mayo, D. (1997). Error statistics and learning from error: Making a virtue of necessity. *Philosophy of Science, Vol. 64,* Part II: Symposia Papers, S195-S212.

Mayo, D. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

Morey, R. (2015). On verbal categories for the interpretation of Bayes Factors. *BayesFactor.* https://bayesfactor.blogspot.nl/2015/01/on-verbal-categories-for-interpretation.html

Morey, R. D., Wagenmakers, E.-J., & Rouder, J. N. (2016). Calibrated Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*, *51*(1), 11–19. http://dx.doi.org/10.1080/00273171.2015.1052710

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests. *PLOS ONE*, *7*(2), e32734. https://doi.org/10.1371/journal.pone.0032734

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 175-240. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia II: Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science,* 7(6), 615-631. http://doi.org/10.1177/1745691612459058

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

Peng, R. D. (2009). Reproducible research and biostatistics. *Biostatistics*, *10*(3), 405–408. https://doi.org/10.1093/biostatistics/kxp014

Pericchi, L., & Pereira, C. (2016). Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics, 30(1)*, 70–90. https://doi.org/10.1214/14-BJPS257

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, *67*(5), 1035-1041. https://doi.org/10.1016/j.jbusres.2013.08.010

Senn, S. (2001) Two cheers for p-values? *Journal of Epidemiology and Biostatistics, 6*, 193-204. https://doi.org/10.1080/135952201753172953

Senn, S. (2007). *Statistical issues in drug development* (2nd ed). Chichester, England; Hoboken, NJ: John Wiley & Sons.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666-681. https://doi.org/10.1177/1745691614553988

Skipper, J. K., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, *2*(1), 16–18.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences, 100*(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, *96*, 434-442. https://doi.org/10.1093/jnci/djh075

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108