Abandon Statistical Inference

Valentin Amrhein[1,2], David Trafimow[3] and Sander Greenland[4]

16 April 2018

[1]Zoological Institute, University of Basel, Basel, Switzerland. [2]Swiss Ornithological Institute, Sempach, Switzerland. [3]Department of Psychology, New Mexico State University, USA. [4]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, USA.
E-mail: v.amrhein@unibas.ch; dtrafimo@nmsu.edu; lesdomes@g.ucla.edu

Abstract. There is a massive crisis of confidence in statistical inference, which has largely been attributed to overemphasis on and abuse of hypothesis testing. Much of the abuse stems from failure to recognize that statistical tests not only test hypotheses, but countless assumptions and the entire environment in which research takes place. Unedited and unselected results *must* vary from replication to replication because of varying assumption violations and random variation; excessive agreement itself would suggest deeper problems, such as failure to publish results in conflict with group expectations or desires. Considerable non-replication is thus to be expected even with honest and complete reporting practices, and generalizations from single studies are rarely if ever warranted. Because of all the uncertain and unknown assumptions that underpin statistical inferences, we should treat inferential statistics as highly unstable local descriptions of relations between model predictions and data, rather than as generalizable inferences about hypotheses or models. And that means we should treat statistical results as being much more incomplete and uncertain than is currently the norm. Rather than focusing our study reports on uncertain conclusions, we should thus focus on describing accurately how the study was conducted, what problems occurred, and what analysis methods were used.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

The crisis of unreplicable research is not only about so-called replication failures. It is also about scientists and other people perceiving non-replication of scientific results as a sign of bad science (Baker 2016). Yes, there is an epidemic of misinterpretation of statistics and of more or less subtle forms of scientific misconduct (such as selectively reporting studies that "worked"; Martinson, Anderson, and de Vries 2005; John, Loewenstein, and Prelec 2012). But all results are uncertain and highly variable, even those from the most honest and rigorous studies.

Indeed, Fisher (1937) wrote that "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon." And Boring (1919) said a century ago, "scientific generalization is a broader question than mathematical description." Yet today we still indoctrinate students with methods that claim to produce scientific generalizations from mathematical descriptions of isolated studies. Naturally, such generalizations will often fail to agree with those from other studies – and thus statistical inference will fail to replicate. Because our current academic reward system is built on publishing inferences from single studies, it should come as no surprise that many conflicting generalizations are published, and hence that a high proportion of generalizations must be wrong.

A core problem is that both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct. Statistical results must eventually mislead us when they are used and communicated as if they present this complex reality, rather than a model for it. This is not a problem of our statistical methods. It is a problem of interpretation and communication of results.

In the following, we argue that the crisis of unreplicable research is mainly a crisis of overconfidence in statistical results. We recommend that we should use, communicate, and teach our statistical methods as being descriptive of logical relations between assumptions and data, rather than as allowing specific generalized inferences about universal populations.

Inferences are not about hypotheses

A statistical model is a set of assumptions (and thus a compound hypothesis) about how the data could have been generated. The model matches reality to the degree that assumptions are met, starting from the assumption that we measured what we think we measured, and that measurement errors were either absent or adequately accounted for. Such model assumptions are

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

part of what Meehl (1990) calls "auxiliary theories," or what McShane et al. (2018) call "neglected factors."

Thus statistical models imply countless assumptions about the underlying reality. A null hypothesis such as "the means of these two populations do not differ" is an explicit assumption. But the validity of statistical inferences depends on the entire set of assumptions needed for the statistical test to be valid. Some of these assumptions may not even be recognized or mentioned in research reports, such as that there was no selection of results for presentation.

For example, we should think of a p-value as referring not only to the hypothesis it claims to test, such as a null hypothesis. A p-value refers to the entire model including other usually explicit assumptions like randomization of treatment and linearity of effects, and also including usually implicit procedural assumptions such as that the equipment for taking measurements was in perfect working order (Greenland et al. 2016; Greenland 2017; Amrhein 2018). Whether recognized or not, these assumptions underpin the usual inferences from a test (Greenland 2018). A small p-value is either a result of random variation, or it indicates that at least one model assumption is violated. But it does not indicate which assumption is violated.

Yes, a small p-value may mean that the null hypothesis is false. But it can also mean that some mathematical aspect of the model was not correctly specified, that sampling was not a hundred percent random, that we accidentally switched the names of some factor levels, that we unintentionally, or intentionally, selected analyses that led to a small p-value (downward "p-hacking"), or that a cable in our measuring device was loose (Amrhein 2018). And symmetrically, a large p-value may arise from mistakes and procedural errors, such as selecting analyses that led to a large p-value (upward p-hacking).

Even the best single studies will be imperfect. Because of varying assumption violations, whether recognized or hidden, and because of random variation, their results will vary from replication to replication. Some degree of non-replication is the norm. This is why generalizations about a particular hypothesis from single studies should be avoided. Having confidence in generalizations from single studies means having overconfidence. Inference that could deserve the description "relatively reliable" would require merging information from multiple studies and lines of evidence.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

Overconfidence triggers selection bias

Unfortunately, even a combination of studies does not guarantee that assumptions will be valid. Published results tend to be biased, for example because they may be selected from unpublished results on the basis of some statistical procedure. Such bad yet common scientific practice introduces bias by accumulating statistical inferences that go into a certain direction, typically emphasizing results that cross some p-value threshold (Amrhein, Korner-Nievergelt, and Roth 2017; Locascio 2017).

We think a major reason for result-selection bias is overconfidence in statistical inference. For decades, scientists were taught to judge which results are reliable and which are not, and which results are thus worth being published or not, based on statistics obtained from single studies. Statistics was misused as an automated decision machine, both for statements about hypotheses and for selection of scientific studies for publication. And this made interpretation, publication, and advertising much easier, because everybody assumed that statistical inferences based on p-value thresholds or other rigid criteria would be reliable.

But any selection criterion will introduce bias. If there is a tendency to publish results because the estimates are yellow, because confidence intervals are short, and because p-values are small, then the published literature will become biased towards yellow results with underestimated variances and overestimated effect sizes. The latter effect is the "winner's curse" that is reflected in the findings of the Open Science Collaboration (2015) project: the average effect size in the original studies was about twice as large as in the replication studies that apparently reported all results and thus did not suffer from selection bias.

Even if authors report all study outcomes, but then select what to discuss and to highlight based on p-value thresholds or other aids to judgement, their conclusions and what is reported in subsequent news and reviews will be biased (Amrhein, Korner-Nievergelt, and Roth 2017). Selective attention based on study outcomes will therefore not only distort the literature but will slant published descriptions of study results – biasing the summary descriptions reported to practicing professionals and the general public.

One way to reduce this selection bias and related misreporting is to maintain that all results are uncertain. If we obtain a small p-value, a large effect estimate, or a narrow confidence interval – or even all three – we should not be confident about textbook inferences from these results. In one of the next replications, p will be large, the effect estimate will be small, or the confidence

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

interval wide, and the textbook inferences will shift dramatically as a consequence. This caution is of course even more in force if there is any selective reporting of, or attention to, results.

We should trust in uncertainty and instead focus on describing accurately how the study was conducted, what problems occurred (e.g., nonresponse of some subjects, missing data), and what analysis methods were used, with detailed data tabulation and graphs, and complete reporting of results. The advent of online supplements eliminates the common excuse of space limitations preventing such detail.

Don't blame the p-value

A clear sign that overconfidence ruled the era of hypothesis testing is that many people still are surprised by the "dance of the p-values" (Cumming 2014), that is, the way a valid p-value bounces around its range even in the largest of samples. This variability means that $p<0.05$ is no guarantee for $p<0.05$ in a replication (Gelman and Stern 2006). Even if our alternative hypothesis and all other assumptions are correct, the p-value in the next sample will probably differ widely from our current sample: "The fickle P value generates irreproducible results" (Halsey et al. 2015).

But the p-value itself is not supposed to be reliable in the sense of staying put. Its fickleness reliably indicates variation in the data from sample to sample. If sample averages vary among samples, then p-values will vary as well, because they are calculated from sample averages. And we don't usually take a single sample average and announce it to be the truth. So if instead of simply reporting the p-value, we engage in "dichotomania" (Greenland 2017) and use it to decide which hypothesis is wrong and which is right, such scientifically destructive behavior is our fault, even if socially encouraged; it is not the fault of the p-value.

Further, if we overlook the sensitivity of p-values to possible violations of background assumptions, by assuming that p-values are only about deciding whether to reject "null hypotheses," we are privileging what may be a scientifically irrelevant hypothesis and are engaging in "nullism," a compulsion to test only one hypothesis among many of importance. But again, such bad behavior is our fault, even if socially encouraged. And if we interpret a small p-value as providing support for some alternative hypothesis (which currently seems to be a standard interpretation), this too is our fault, not the fault of the p-value.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

Ban your p-value

It may help to ban some practices within specific contexts. We ban alcohol drinking[1] before or during driving, despite its general acceptance in relaxed settings. There are even studies showing positive effects of small doses. But use easily becomes abuse, and it is essential to abstain from alcohol in many settings.

By analogy, we think it may do good to ban p-values and confidence intervals from certain settings. Reaching for p-values is, like drinking alcohol, a culturally ingrained habit that is easily implemented. P-values and alcohol often give the wrong impression that complex decisions can be oversimplified without negative consequences, for example by relying on significance testing. And many people are addicted psychologically. These habits and addictions are worth breaking.

At the very least, partial or temporal bans are one way to force researchers to learn how to analyze data in alternative ways (Trafimow and Marks 2015). Hopefully, thinking about advantages and disadvantages of alternatives will lead to more sober interpretation of statistics. One concern, however, is that complete prohibition could lead to misuse and abuse of other methods, such as Bayesian techniques – which have an additional source of nonreplicability insofar as what are acceptable priors can vary dramatically across research groups.

Long live no king

Hypothesis testing was king for nearly a century. We propose not to banish the king – the Neyman-Pearson decision procedure may be useful, for example, in industrial or laboratory quality control, or "sampling tests laid down in commercial specifications" (Neyman and Pearson 1933), in which automated decisions to stop a production line or to recalibrate equipment may be necessary. For scientific inference, however, we hope that dichotomania from which the Neyman-Pearson procedure suffers can be cured by abandoning hypothesis testing based on p-value thresholds (Hurlbert and Lombardi 2009; Amrhein, Korner-Nievergelt, and Roth 2017; Greenland 2017; Amrhein and Greenland 2018; McShane et al. 2018; Trafimow et al. 2018).

Since statistical models include not only our hypotheses but countless explicit and implicit assumptions, traditional hypothesis tests hardly test the hypotheses that we think they test.

[1] We prefer this analogy above comparing p-values with guns, as we have heard or read sometimes.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

Yet hypothesis tests still precipitate decisions between significant/nonsignificant or reject/accept. They have been destructive agents by allowing overconfident claims from isolated studies to become the prevailing currency in the academic system.

Yes, sometimes we need to make decisions in science, for example whether to further pursue a study or not. For such a decision, we will usually weigh scientific and personal costs and benefits of our decision, applying informed personal judgment (Gigerenzer 1993). But when it comes to weigh evidence against, or in favor of, a scientific hypothesis, statistical tests cannot suffice and may even be destructive if degraded into a binary form as in reporting tests as significant/non-significant, or in basing conclusions on whether the null value was included in or excluded from an interval. This is especially true when (as almost always) these results are sensitive to doubtful assumptions, such as absence of measurement-error dependencies. And even in the unlikely case that all model assumptions are met, we would still need to consider costs and benefits, as well as the published and unpublished literature, for making a decision whether or not to judge a scientific hypothesis as being largely correct (subject to further evidence to the contrary).
We hope that classical hypothesis testing will be retired quickly, so that regicide is not necessary.

Empire of diversity

But what comes next? There are countless possibilities. If we want to report a continuous measure of refutational evidence against a model, we could use the traditional p-value in a continuous fashion, or better still the Shannon information or S-value (surprisal) of the test, $-\log_2(p)$, which is unbounded above and thus difficult to misinterpret as a hypothesis probability (Greenland 2017, 2018). If we want to compare the relative support for different models we could use likelihood ratios or Bayesian methods. But we should not lapse back into dichotomous thinking by using some p-value threshold, or by making binary inferences based on confidence intervals or Bayes factors. And we should not pledge uncritical loyalty to posterior probabilities, especially since they rely on the same fundamental assumptions about the data-generating model that hypothesis tests and confidence intervals use.

A 95% confidence interval encompasses a range of hypotheses that have a p-value exceeding 0.05. Instead of talking about hypothetical coverage of this interval (which will fail under various assumption violations), we can instead think of it as a compatibility interval (Greenland 2018), showing the effect sizes that meet the 0.05 predictive criterion for the data under the model used

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

to compute the interval. Again, whether the interval includes or excludes zero should play no role in its interpretation, because even with only random variation present, the intervals from different data sets will be very different.

With additional (and inevitable) nonrandom variation, the true effect size will frequently be outside the 95% interval. In reality, it will not happen that every assumption is met, nor will we be aware of every assumption. Stating that our data "support" any value in the compatibility interval (e.g., a zero effect), or that the interval covers the true value at some rate, or that the interval measures uncertainty by indicating the range of possible effect sizes, makes the compatibility interval into an overconfidence interval.

The empire of "statistical significance" has its roots in the 19th century writings of Edgeworth (1885) and reached full dominance with the spread of cutoffs for testing, formalized by Jerzy Neyman and Egon Pearson as Type-I error rates. Like the political empires of their period, such significance testing for scientific (as opposed to mechanical) inference is a relic of a bygone era, whose destructive effects reverberates to this today. We hope this era is over. As for what comes next, there is no substitute for accepting methodologic diversity (Good 1957; Cox 1978; Box 1980; Barnard 1996; Little 2006; Senn 2011; Efron and Hastie 2016; Crane 2017; Trafimow and Earp 2017), with careful assessment of uncertainty as the core motivation for statistical practice (Gelman 2016).

If we are researchers ...

... we should continue doing our studies, be they isolated experiments or replications of other studies. Single studies are the life blood of science. If we think we did a good study, we should report it, but avoid being overconfident about our results. If we believe we have strong evidence, we are putting too much faith on statistical inference. But we should not be overconfident about our weak evidence either. Almost never will we have found absolutely no effect. Let us free our negative results by allowing them to be potentially positive (Amrhein, Korner-Nievergelt, and Roth 2017). And let us free our positive results by allowing them to be potentially negative. The compatibility interval will usually show that any practical conclusion is uncertain anyway.

In particular, if we obtain a large p-value, our interval estimate will show that the null hypothesis is only one of the many practically different hypotheses that are compatible with the data (Rothman, Greenland, and Lash 2008; Amrhein, Korner-Nievergelt, and Roth 2017). So we

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

cannot claim statistics demonstrate there is no effect whatsoever, because even if the data remain consistent with a zero effect, they remain consistent with many other effects as well. And we should remember there are lots of additional hypotheses outside the compatibility interval that will also be compatible with our data, due to methodologic limitations that we have not modeled.

Because of all the known and unknown assumptions about our model, we should treat every inferential statistic, like the point estimate, the compatibility interval, or the p-value, as being descriptive of the relation of statistical models to the data. It is hard enough to describe the known assumptions about our model. We should not draw inference and generalize based on assumptions we cannot be certain about or we do not even think about.

P is merely the probability of one particular test statistic being as or more extreme than observed in our particular study, given that the current model is true. There is no inferential meaning that must be attached to that. For the next set of data, the p-value will be different. A small p-value is just a warning signal that the current model could have a problem, so we should check our model assumptions. And this assumption checking does not only mean inspecting residuals, but also checking the extent of deviations of our study from a perfect randomized experiment or random survey, whether from failures of protocol, measurement, equipment, or any of the innumerable details that real research must confront.

Science is learning about assumption violations, then addressing those violations and improving the performance of our models about reality.

We should thus communicate our local conclusions about our study system, not our generalized inferences about some ill-defined universal population. And presentation decisions should not be based on p-values, nor on any other statistic. This will make our own conclusions more valid and published scientific knowledge more reliable. If we think we did a good study, we should be modest about our conclusions, but be proud about our painfully honest and thorough description and discussion of our methods and of our data.

If we are science writers and journalists ...

... we should continue writing about isolated experiments and replications. Single studies are the life blood of science. If we think we found a good study, or a bad study, we may report it. But let us try not to be impressed by what researchers say is surprising about their study – surprising results are often products of data dredging or random error, and are thus less reproducible (Open

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

Science Collaboration 2015). So surprising results may not point to general scientific discoveries, although they may still be valuable because they lead to new insights about study problems and violations of assumptions.

Then too, we should not overemphasize what researchers say was unsurprising, since that may largely reflect their conformity to group expectations rather than what the data would actually show under close scrutiny. Indeed, we might consider not to ask researchers about surprising or unsurprising results at all, but rather ask which results appeared most boring because they were shown several times before and thus seem to be reliable. More generally, we should not fall for overconfident claims by researchers, science writers, or journalists. Rather, we should try to uncover overconfident claims and the bad incentives leading to overconfident claims.

Science is about accumulating information. If replications do not find the same results, this is not necessarily a crisis, but a natural process by which science evolves. The goal of scientific methodology should be to direct this evolution toward ever more accurate descriptions of the world and how it works, not toward ever more publication of alleged stories, inferences, or conclusions.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

Amrhein, V. (2018), "Inferential statistics is not inferential," *sci five, University of Basel*, http://bit.ly/2oLY7t9.

Amrhein, V., and Greenland, S. (2018), "Remove, rather than redefine, statistical significance," *Nature Human Behaviour*, 2, 4.

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research," *PeerJ*, 5, e3544. doi: 10.7717/peerj.3544.

Baker, M. (2016), "Is there a reproducibility crisis?" *Nature*, 533, 452–454.

Barnard, G.A. (1996), "Fragments of a statistical autobiography," *Student*, 1, 257–268.

Boring, E.G. (1919), "Mathematical vs. scientific significance," *Psychological Bulletin*, 16, 335–338. doi: 10.1037/h0074554.

Box, G.E.P. (1980), "Sampling and Bayes' inference in scientific modeling and robustness," *Journal of the Royal Statistical Society, Series A*, 143, 383–430. doi: 10.2307/2982063.

Cox, D.R. (1978), "Foundations of statistical inference: the case for eclecticism," *Australian Journal of Statistics*, 20, 43–59.

Crane, H. (2017), "Why 'Redefining statistical significance' will not improve reproducibility and could make the replication crisis worse," arXiv:1711.07801.

Cumming, G. (2014), "The new statistics: why and how," *Psychological Science*, 25, 7–29. doi: 10.1177/0956797613504966.

Edgeworth, F.Y. (1885), "Methods of Statistics," *Journal of the Statistical Society of London*, Jubilee Volume, 181–217.

Efron, B., and Hastie, T. (2016), *Computer age statistical inference: algorithms, evidence, and data science*, New York: Cambridge University Press.

Fisher, R.A. (1937), *The design of experiments*, second ed, Edinburgh: Oliver and Boyd.

Gelman, A. (2016), "The problems with p-values are not just with p-values," *The American Statistician, supplemental material to the ASA statement on p-values and statistical significance*. doi: 10.1080/00031305.2016.1154108.

Gelman, A., and Stern, H. (2006), "The difference between 'significant' and 'not significant' is not itself statistically significant," *The American Statistician*, 60, 328–331. doi: 10.1198/000313006x152649.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

Gigerenzer, G. (1993), "The superego, the ego, and the id in statistical reasoning," in *A handbook for data analysis in the behavioral sciences*, edited by G. Keren and C. Lewis, 311–339, Hillsdale: Lawrence Erlbaum Associates.

Good, I.J. (1957), "Some logic and history of hypothesis testing," in *Philosophical foundations of economics*, edited by J.C. Pitt, 149–174, Dordrecht, Holland: D. Reidel. Reprinted as Ch. 14 in Good, I.J. (1983), *Good Thinking*, 129–148, Minneapolis: U. Minnesota Press.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.C., Poole, C., Goodman, S.N., and Altman, D.G. (2016), "Statistical tests, confidence intervals, and power: A guide to misinterpretations," *European Journal of Epidemiology*, 31, 337–350. doi: 10.1007/s10654-016-0149-3.

Greenland, S. (2017), "Invited commentary: The need for cognitive science in methodology," *American Journal of Epidemiology*, 186, 639–645. doi: 10.1093/aje/kwx259.

Greenland, S. (2018), "The unconditional information in *P*-values, and its refutational interpretation via *S*-values," *under submission*.

Halsey, L.G., Curran-Everett, D., Vowler, S.L., and Drummond, G.B. (2015), "The fickle *P* value generates irreproducible results," *Nature Methods*, 12, 179–185. doi: 10.1038/nmeth.3288.

Hurlbert, S.H. and Lombardi, C.M. (2009), "Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, 46, 311–349. DOI 10.5735/086.046.0501.

John, L.K., Loewenstein, G., and Prelec, D. (2012), "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychological Science*, 23, 524–532. doi: 10.1177/0956797611430953.

Little, R.J. (2006), "Calibrated Bayes: A Bayes/frequentist roadmap," *American Statistician*, 60, 213–223. doi: 10.1198/000313006x117837.

Locascio, J. (2017), "Results blind science publishing," *Basic and Applied Social Psychology*, 39, 239–246. https://doi.org/10.1080/01973533.2017.1336093.

Martinson, B.C., Anderson, M.S., and de Vries, R. (2005), "Scientists behaving badly," *Nature*, 435, 737–738. doi: 10.1038/435737a.

McShane, B.B., Gal, D., Gelman, A., Robert, C., and Tackett, J.L. (2018), "Abandon statistical significance," *arXiv*:1709.07588.

Meehl, P.E. (1990), "Why summaries of research on psychological theories are often uninterpretable," *Psychological Reports*, 66, 195–244. doi: 10.2466/pr0.66.1.195-244.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018

Neyman, J., and Pearson, E.S. (1933), "The testing of statistical hypotheses in relation to probabilities a priori," *Proceedings of the Cambridge Philosophical Society*, 29, 492–510.

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

Rothman, K., Greenland, S., and Lash, T.L. (2008), *Modern Epidemiology*, 3rd Edition, Ch. 10, Philadelphia, PA: Lippincott Williams & Wilkins.

Senn, S.J. (2011), "You may believe you are a Bayesian but you are probably wrong," *Rational Markets and Morals*, 2, 48–66.

Trafimow, D., Amrhein, V., Areshenkoff, C.N., Barrera-Causil, C., Beh, E.J., Bilgiç, Y., Bono, R., Bradley, M.T., Briggs, W.M., Cepeda-Freyre, H.A., Chaigneau, S.E., Ciocca, D.R., Carlos Correa, J., Cousineau, D., de Boer, M.R., Dhar, S.S., Dolgov, I., Gómez-Benito, J., Grendar, M., Grice, J., Guerrero-Gimenez, M.E., Gutiérrez, A., Huedo-Medina, T.B., Jaffe, K., Janyan, A., Karimnezhad, A., Korner-Nievergelt, F., Kosugi, K., Lachmair, M., Ledesma, R., Limongi, R., Liuzza, M.T., Lombardo, R., Marks, M., Meinlschmidt, G., Nalborczyk, L., Nguyen, H.T., Ospina, R., Perezgonzalez, J.D., Pfister, R., Rahona, J.J., Rodríguez-Medina, D.A., Romão, X., Ruiz-Fernández, S., Suarez, I., Tegethoff, M., Tejo, M., van de Schoot, R., Vankov, I., Velasco-Forero, S., Wang, T., Yamada, Y., Zoppino, F.C., and Marmolejo-Ramos, F. (2018), "Manipulating the alpha level cannot cure significance testing," *PeerJ Preprints*, 6, e3411v2.

Trafimow, D., and Earp, B.D. (2017), "Null hypothesis significance testing and Type I error: The domain problem," *New Ideas in Psychology*, 45, 19–27. http://dx.doi.org/10.1016/j.newideapsych.2017.01.002.

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. doi: 10.1080/01973533.2015.1012991.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v1 | CC BY 4.0 Open Access | rec: 16 Apr 2018, publ: 16 Apr 2018