Inferential Statistics as Descriptive Statistics: There is No Replication Crisis if We Don't Expect Replication

Forthcoming in *The American Statistician*

Valentin Amrhein[1,2], David Trafimow[3] and Sander Greenland[4]
24 October 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@ucla.edu

Abstract. Statistical inference often fails to replicate. One reason is that many results may be selected for drawing inference because some threshold of a statistic like the P-value was crossed, leading to biased reported effect sizes. Nonetheless, considerable non-replication is to be expected even without selective reporting, and generalizations from single studies are rarely if ever warranted. Honestly reported 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. A general perception of a "replication crisis" may thus reflect failure to recognize that statistical tests not only test hypotheses, but countless assumptions and the entire environment in which research takes place. 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 assumptions 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. Acknowledging this uncertainty could help reduce the allure of selective reporting: Since a small P-value could be large in a replication study, and a large P-value could be small, there is simply no need to selectively report studies based on statistical results. Rather than focusing our study reports on uncertain conclusions, we should thus focus on describing accurately how the study was conducted, what problems occurred, what data were obtained, what analysis methods were used and why, and what output those methods produced.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

The "crisis of unreplicable research" is not only about alleged replication failures. It is also about perceived non-replication of scientific results being interpreted as a sign of bad science (Baker 2016). Yes, there is an epidemic of misinterpretation of statistics and what amounts to scientific misconduct, even though it is common practice (such as selectively reporting studies that "worked" or that were "significant"; Martinson, Anderson, and de Vries 2005; John, Loewenstein, and Prelec 2012). But all results are uncertain and highly variable, even those from the most rigorous studies.

Because a small P-value could result from random variation alone, 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 single publications (usually reporting the results of one or a few similar 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 and of the reality being studied. 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 inferential statistical methods as describing logical relations between assumptions and data (as detailed in the Appendix), rather than as providing generalized inferences about universal populations.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

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 assumptions that we measured what we think we measured and that measurement errors were either absent or adequately accounted for. Such model assumptions are part of what are called "auxiliary hypotheses" (Popper 1968), "auxiliary theories" (Meehl 1990), or "currently subordinate factors" (McShane et al. 2018).

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. Further assumptions that are often explicitly addressed in research reports are that sampling was random or that residuals are independent and identically distributed. Other assumptions may not even be recognized or mentioned in research reports, such as that there was no selection of particular results for presentation, or that the population from which we drew our random sample is equivalent to the population we have targeted for inference. Whether it is assumptions that are reviewed by inspecting residuals, or further assumptions that link statistics to reality, the validity of statistical inferences depends on the entire set of assumptions.

For example, we should think of a P-value[1] 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, plus 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 2018b). A small P-value is the net result of some combination of random variation and violations of model assumptions, but does not indicate which (if any) assumption is violated.

Yes, a small P-value may arise because 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-

[1] We focus on P-values and confidence intervals not because they are better or worse than other methods, but because they are probably the most often used, and most often misused, inferential statistics. Throughout the text, P is the random variable and p is its observed value (realization) in a given sample.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

hacking"), that we did not measure what we think we measured, or that a cable in our measuring device was loose (Amrhein 2018). And 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), or using a measurement so noisy that the relation of the measured construct to anything else is hopelessly obscured.

Replication studies have a false-negative problem

Even the best single studies will be imperfect. In addition to random variation, their results will usually vary from replication to replication because of varying assumption violations, whether recognized or hidden, and thus the observed effect sizes can easily differ across settings. Consider the replication project by the Open Science Collaboration (2015): Of 97 psychological studies with "significant" results ($p \leq 0.05$) out of 100 that were subject to replication, 35 had $p \leq 0.05$ in the replication. This is much less than would have been expected if all original null hypotheses were false and only random variation caused differences between the original and replication – under these circumstances, with an average power of 92% in the replications, 89 of the 97 replicates were expected to have $p \leq 0.05$. One explanation by the authors is that in the original studies, results were selected for reporting based on having $p \leq 0.05$, which led to inflated effect sizes (see next subsection) that could not be replicated.

The reports of such replication projects are often misinterpreted as showing that the original "significant" studies were mostly or entirely false positives. To see the error in such interpretations, consider that the Open Science Collaboration (2015) observed $97 - 35 = 62$ replications with $p > 0.05$ for which the original study had $p \leq 0.05$, which is 64% of the 97 replications. This emphatically does *not* mean that 64% of the 97 original null hypotheses with "significant" P-values were correct, or that 62 of the 97 "significant" results were false positives. Why not? If as suggested (and indicated by the replications) the original reported effect sizes were largely inflated due to selective reporting based on $p \leq 0.05$, then the actual effect sizes (both in the original and replication studies) could easily be too small to provide high power for replication (Camerer et al. 2018).

Suppose for example that among the 97 "significant" original studies, 70% of the null hypotheses were false, for a total of $0.70(97) = 68$ false nulls, and the average power for the effects in these non-null cases was 50% in the replication studies (and not 92% as calculated

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

based on the probably inflated original effect sizes). Then the expected number of "true positive" replications (those with $p \leq 0.05$ for a false null) would be $0.50(68) = 34$, while the number of "false positive" replications (those with $p \leq 0.05$ for a true null) would be $0.05(97 - 68) = 1.45$, resulting in a total of about 35 out of 97 replications having $p \leq 0.05$, as observed. That means that, given selective reporting in the original studies, the observed 64% of the 97 replication attempts with $p > 0.05$ could have been expected even if only $97 - 68 = 29$ or 30% of the null hypotheses were correct!

Thus, with selective reporting in the original studies, it may be not surprising to get "non-significant" results in about two-thirds of replications. And even with no selective reporting and only random variation present, replication studies remain subject to what may be severe *false-negative* errors. Consequently, "non-significant" replications (those with $p > 0.05$) do not reliably flag original studies as being false positives. For example, with a statistical power of 80%, two studies will be traditionally judged as "conflicting," meaning that one is "significant" and the other is not, in one third of the cases (Greenland et al. 2016; Amrhein, Korner-Nievergelt, and Roth 2017). This means that, unless statistical power of the replication is nearly 100%, interpretations of replication attempts must allow for false-negative errors as well as false-positive errors, and that "significance" and "non-significance" cannot be used to reliably judge success or failure of replication (Goodman 1992; Senn 2001, 2002). As importantly, valid interpretation must take into account any background research (prior information) on the hypothesis under study, especially if there are conflicts between the original and replication studies.

Another often overlooked point about hacking and replication failures is that both random and selection artefacts will arise even if we ban all tests and focus instead on estimates. Some degree of variation, and hence non-replication, is the norm across honestly reported studies, even when all assumptions are met, and estimates can be selected for reporting based directly on their size rather than their P-value. Add to these problems other assumption violations and we will find that error rates are usually far from the nominal rates given in the report – e.g., 95% confidence intervals cannot realistically be claimed to have as much as 95% coverage of the true effect when study imperfections exist. Consequently, having confidence in generalizations from single studies means having overconfidence in most cases. Inference that could be called trustworthy would require merging information from multiple studies and lines of evidence.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

Overconfidence triggers selection bias

Unfortunately, even a combination of studies does not guarantee that inferences will be valid. As noted above, published results tend to be biased, for example because they may be selected from unpublished results based on some statistical criterion. 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 suspect that a major driver of result-selection bias is overconfidence in statistical inference. For decades, scientists were taught to judge which results are trustworthy 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 scientific decision machine, both for statements about hypotheses and for selection of studies for publication. And this made interpretation, publication, and public advertising much easier, because everybody assumed that statistical inferences based on P-value thresholds or other rigid criteria would be "reliable," in the sense that a replication would probably meet the same thresholds or criteria again. So if researchers expect that a small P-value or a short interval estimate indicate "reliable" results, while all other results are "unreliable," they may be "prepared to ignore all results which fail to reach this standard" (Fisher 1937, p. 15; one of many published pleas by various authors encouraging selective reporting).

But any selection criterion will introduce bias. If there is a tendency to publish results because the estimates are yellow, because interval estimates are short, and because P-values are small or the point estimates are far from null, 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," or inflation of effect sizes, that is reflected in the findings of the Open Science Collaboration (2015): the average effect size in the original studies was about twice as large as in the replication studies that 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). Such *selective attention* based on study outcomes will therefore not only distort the literature but will

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

slant published descriptions of study results – biasing the summary descriptions reported to practicing professionals and the general public.

One way to reduce selective reporting and attention is to maintain that all results are uncertain. If we obtain a small P-value, a large effect estimate, or a narrow interval estimate – 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 interval estimate wide, and thus the textbook inferences will shift dramatically due to random variation or to assumptions we have not modeled. Because of this uncertainty, there is simply no need to selectively report studies based on statistical results.

We should thus "move toward a greater acceptance of uncertainty and embracing of variation" (Gelman 2016) and focus on describing accurately how the study was conducted, what problems occurred (e.g., non-response 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 and preprint servers eliminate the common excuse that space limitations prevent reporting such detail.

Don't blame the P-value

A clear sign that overconfidence has ruled the era of hypothesis testing is that many people still are surprised by the "dance of the P-values" (Cumming 2014), that is, by 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 (Goodman 1992; Senn 2001, 2002; Gelman and Stern 2006); after all, if the (null) hypothesis tested is correct and experimental conditions are ideal, the P-value will vary uniformly between 0 and 1. And even if our alternative hypothesis is correct, the P-value in the next sample will typically differ widely from our current sample: "The fickle P value generates irreproducible results" (Halsey et al. 2015), at least if reproducibility is defined by whether P is above or below a threshold and the power is not very high.

But the P-value itself is not supposed to be "reliable" in the sense of staying put (Greenland 2018a). Its fickleness 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. But if instead of simply reporting the P-value, we engage in "dichotomania" (Greenland 2017)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

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. We could instead provide P-values for relevant alternatives, and arguably *should* do so if we compute any P-value at all (Poole 1987a,b). And if we interpret a small null P-value as providing support for some alternative hypothesis (which currently seems to be a standard interpretation) without testing the alternative as well, this too is our fault, not the fault of the P-value.

Ban statistical tests?

It may help to ban some practices, at least temporary and in specific contexts. Despite the general acceptance of alcohol[2] in relaxed settings and possible beneficial effects from light use, we ban its drinking before or during driving, and recognize that its use easily becomes abuse. Reaching for statistical tests to force out "inferences" (whether traditional "$p \leq \alpha$" testing or substitutes like tests using Bayes-factor criteria) is, like drinking alcohol, a culturally ingrained habit. Statistical testing (like alcohol) often gives the wrong impression that complex decisions can be oversimplified without negative consequences, for example by making decisions solely because p was above or below some cutoff like 0.05. And many researchers are addicted to such oversimplification. These addictions are worth breaking.

At the very least, partial or temporary 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 non-replicability insofar as what are acceptable priors can vary dramatically across research groups.

[2] We prefer this analogy above comparing P-values or hypothesis tests with guns, as we have heard or read sometimes.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

Long live no king

Fixed-cutoff (α-level) hypothesis testing has been king for over 80 years. We propose not to banish the king – a fixed-cutoff 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 such procedures suffer can be cured by abandoning them in favor of data description and direct presentation of precise P-values – including P-values for alternative hypotheses (Poole 1987a,b; Cohen 1994; Ziliak and McCloskey 2008; Hurlbert and Lombardi 2009; Amrhein, Korner-Nievergelt, and Roth 2017; Greenland 2017; Greenland 2018a,b; Amrhein and Greenland 2018; McShane et al. 2018; Trafimow et al. 2018). Parallel criticisms and remedies apply to tests or decisions (such as whether to report or highlight results) based on Bayes factors, posterior odds, or any other statistical criterion.

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, 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 from research reporting, so that regicide is not necessary.

Empire of diversity

But what comes next? There are countless possibilities. The most common proposal is to replace hypothesis tests with interval estimates. While doing so is helpful for sophisticated researchers, it has not reduced what we see as the core psychological problems – which is unsurprising, because

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

the classical confidence interval is nothing more than a summary of dichotomized hypothesis tests. Consider that a 95% confidence interval encompasses a range of hypotheses (effect sizes) that have a P-value exceeding 0.05. Instead of talking about hypothetical coverage of the true value by such intervals, which will fail under various assumption violations, we can think of the confidence interval as a "compatibility interval" (Greenland 2018a,b), showing effect sizes most compatible with the data according to their P-values, under the model used to compute the interval. Likewise, we can think of a posterior probability interval, or Bayesian "credible interval," as a compatibility interval showing effect sizes most compatible with the data, under the model *and prior distribution* used to compute the interval (Greenland 2018a). Again, whether such intervals include or exclude zero should play no role in their interpretation, because even with only random variation the intervals from different studies can easily be very different (Cumming 2014).

With additional (and inevitable) nonrandom variation, the true effect size will frequently be outside the 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, upon unlimited replication, the intervals would cover the true value at some rate, or that the interval "measures uncertainty" by indicating the range of *possible* effect sizes (as opposed to *compatible* effect sizes, given the model), makes the compatibility interval into an *overconfidence* interval.

To avoid the dichotomization retained by interval estimates, one could report a measure of refutational evidence such as a traditional P-value in a continuous fashion (as recommended by classic texts on testing such as Lehmann 1986, p. 71), reporting an observed P-value as a measure of the degree of compatibility between the hypothesis or model it tests and the data (Greenland 2018a). Better still, we could report the Shannon information or S-value (surprisal) of the test, $-\log_2(p)$, which is a measure of the evidence against the model supplied by the test, expressed in units of *bits* (binary digits) of information. Among its advantages, the S-value is unbounded above and thus difficult to misinterpret as a hypothesis probability (Greenland 2017, 2018a,b). Considering that the 95% confidence interval is the range in which the parameter values have $p > 0.05$, the values in the interval are those for which the S-value s is less than $-\log_2(0.05) = 4.3$. This means that, under the model used to construct the interval (e.g., a regression model), the values in a 95% confidence interval have only about 4 bits or less information against them; that

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

is very little information indeed (4 bits is the same as the evidence against "fairness" of coin tosses provided by obtaining 4 heads in a row).

If we want to compare the relative support for different parameter values or 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 expect posterior probabilities to solve the problems, especially since they rely on the same often questionable assumptions about the data-generating model that both hypothesis tests and confidence intervals use.

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 hypothesis testing for scientific (as opposed to mechanical) inference is a relic of a bygone era, whose destructive effects reverberate to this day. 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), with careful assessment of uncertainty as the core motivation for statistical practice (e.g., by discussing the effect sizes compatible with the data, given the model, as outlined above).

The replacement for hypothesis tests

We "don't look for a magic alternative to NHST [null hypothesis significance testing], some other objective mechanical ritual to replace it. It doesn't exist" (Cohen 1994). And if it existed, we would probably not recommend it for scientific inference. What needs to change is not necessarily the statistical methods we use, but how we select our results for interpretation and publication, and what conclusions we draw. Why would we want a mechanical decision procedure for single studies, if not for selecting results for publication or interpretation? As we described above, every selection criterion would introduce bias. We therefore join others who have advised that we should, to the extent feasible:

(a) Target results for publication and interpretation *before* data are collected, i.e., state our hypotheses and predictions in a defined protocol or a binding research proposal.

(b) Before analyzing data (and preferably before collecting them), make an analysis plan (i.e., a pre-analysis protocol), setting out how data will be analyzed; and, in the publication,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

show what results the protocol produced before displaying the results of any analyses deviating from the predefined protocol.

(c) Emphasize and interpret our estimates rather than tests, explicitly discussing both the lower and upper limits of our interval estimates.

(d) When reporting statistics, give their precise values rather than mere inequalities; for example, if we are reporting a P-value and it is 0.03, report "p = 0.03," not "p < 0.05".[3]

(e) Not use the word "significant" to describe scientific results, as it implies an inappropriate level of certainty based on an arbitrary criterion, and has produced far too much confusion between statistical, scientific, and policy meanings.

(f) Acknowledge that our statistical results describe relations between assumptions and the data *in our study*, and that scientific generalization from a single study is unwarranted.

(g) Openly and fully report our detailed methods, materials, procedures, data, and analysis scripts.

As an example, consider a study by Brown et al. (2017), who reported that "in utero serotonergic antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child," based on an estimated hazard-rate ratio (HR) of 1.61 (a 61% rate increase in the exposed relative to the unexposed) and a 95% confidence interval of 0.997 to 2.59. As is often the case, the authors misused the confidence interval as a hypothesis test, and they claim to have demonstrated no association because the lower limit of the interval was slightly below no association (which corresponds to a hazard-rate ratio of HR = 1), ignoring that the upper limit exceeded 2.50. A more correct summary of the results would have been: "Our estimate of the hazard-rate ratio was 1.61, and thus exposure could be associated with autism; however, possible hazard-rate ratios that are highly compatible with our data, given our model, ranged from 0.997 (essentially no association) to 2.59 (a relatively strong association)." If applicable, this could then be followed by a discussion of why the authors seem to think the exposure effect might be negligible despite the association, and how strong they judge their evidence not only based on the width of an interval estimate, but also in view of possible shortcomings of their study, of their prior knowledge about other studies on autism, and of possible costs of their interpretation for the health of the patients.

[3] An exception: If a P-value is below the limit of numerical accuracy of the data, an inequality would be called for, but the precision would be context dependent: e.g., $p < 10^{-8}$ is typical in genomics, $p < 0.001$ is common in medicine and ecology.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

Had the authors found an interval of 1.003 to 2.59 rather than 0.997 to 2.59, the reporting should have been the same. Even with an interval of 0.900 to 2.59, the description of the results should be largely similar – the point estimate would still be a HR well above 1, indicating a possible positive association. What would need to change with the latter interval is the description that not only relatively large positive, but also small negative associations would fall within the interval.

Anything goes?

So what do we conclude from a study like Brown et al. (2017)? If we would interpret an interval of 1.003 to 2.59 and of 0.997 to 2.59 in the same way, does that mean that the floodgates of "anything goes" are wide open? Yes, the floodgates should be open – for reporting our results. Everything should be published in some form if whatever we measured made sense *before we obtained the data* because it was connected in a potentially useful way to some research question. If after doing the study, it appears the measure did not make sense or the methods were faulty, at least other researchers can learn that lesson without repeating the error – provided our report contains enough detail to allow such critical evaluation.

However, the floodgates should be closed for drawing conclusions from virtually any single study. For example, because they found a confidence interval that barely included the null value, Brown et al. (2017) reported conflict with previously observed associations that were nearly the same size (hazard ratios around 1.6) but had confidence intervals that did not include the null value. We think that, at most, a conclusion could be drawn that the new study was largely consistent with previous studies, but that the null value was also compatible with their data, given the model.

In view of all the unmodeled uncertainties, it would be good to plan and publish single studies with the goal of their easy entry into meta-analyses. In the words of Trafimow et al. (2018): "It is desirable to obtain precise estimates in those studies, but a more important goal is to eliminate publication bias by including wide confidence intervals and small effects in the literature, without which the cumulative evidence will be distorted."

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

Abandon statistical inference?

We do not suggest to completely abandon inference from our data to a larger population (although the title of a preprint of this paper was "Abandon statistical inference"; Amrhein, Trafimow, and Greenland 2018). But we say this inference must be scientific rather than statistical, even if we use inferential statistics. Because all statistical methods require subjective choices (Gelman and Hennig 2017), there is no objective decision machine for automatic scientific inference; it must be we who make the inference, and claims about a larger population will always be uncertain.

So when can we be confident that we know something? This is the topic of the vast domains epistemology, scientific inference, and philosophy of science, and thus far beyond the scope of the present paper (and its authors). Nonetheless, a successful theory is one that survives decades of scrutiny. If every study claims to provide decisive results (whether from inferential statistics or narrative impressions – or a confusion of the two), there will be ever more replication failures, which in turn will further undermine public confidence in science. We thus believe that decision makers must act based on cumulative knowledge – which means they should preferably not rely solely on single studies or even single lines of research (even if such contributions may determine a decision when all other evidence appears ambiguous or unreliable).

If we are researchers ...

... and we obtained a large P-value for the null hypothesis or an interval estimate that includes a null effect, our interval will show that the null hypothesis is only one of many different hypotheses that are compatible with the data (Rothman, Greenland, and Lash 2008; Greenland et al. 2016). Unlike what Brown et al. (2017) suggested with their hazard-rate ratios, we cannot claim our statistics indicate there is no effect, because even if the data remain consistent with a zero effect, they remain consistent with many other effects as well. A "proof of the null hypothesis" such as "the earth is flat ($p > 0.05$)" is therefore not possible (Greenland 2011; Amrhein, Korner-Nievergelt, and Roth 2017). And we should remember there are lots of additional hypotheses outside the interval estimate that will also be compatible with our data, due to methodologic limitations that we have not modeled.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

Thus, we should not be overconfident about our "weak evidence" from a large P-value or an interval estimate that includes the null. Almost never will we have found absolutely no effect. Let us free our "negative" results by allowing them to be positive contributions to knowledge (Amrhein, Korner-Nievergelt, and Roth 2017). This means that, unless the interval estimate is entirely within an explicit interval of "practical equivalence" to the null,[4] we should first and foremost ban the following statements from our thinking, from our papers, and from our talks: "there was no effect," "there was no difference," "there was no interaction," or "we deleted this term from the model because it had no influence."

Such statements will usually be wrong even if our point estimate is exactly null (and thus p = 1), because our interval estimate will usually show there are many important non-null effects that are highly compatible with our data. This means that an outcome of a study can only be "negative" in a falsificationist sense of finding little incompatibility between the data and the predictions of a model that includes a hypothesis of no effect. A P-value of 1 for the test of the null only means our data are perfectly compatible with our model (including the null hypothesis); but "perfectly compatible" with one hypothesis, or model, does not mean that all other hypotheses, or models, are refuted. Indeed, typical interval estimates will reveal a large number of non-null hypotheses that we would call highly compatible with our data (e.g., because their P-values exceed 0.05), given our model is correct.

Thus, most studies with large P-values or interval estimates that include the null should be considered "positive" in the sense that they leave open the possibility of important effects, even if they also leave open the possibility of no effect. The best we can do is describe the values covered by our interval estimates – and if those values have qualitatively different practical consequences, we should admit that our current set of data could not settle the matter even if we knew that all the auxiliary assumptions were correct.

Conversely, if we believe we have "strong evidence" because our P-value is small and our point estimate is large, or because our interval estimate is not near the null, we are placing too much faith in our inferential statistics. Keep in mind that these statistics do not "measure" uncertainty. At best, the interval estimate may give a rough idea of uncertainty, given that all the assumptions used to create it are correct. And even then, we should remember the "dance of the

[4]In terms of α-level equivalence testing (Wellek 2010; Lakens et al. 2018), for example, this means that all effects with symmetric two-sided $p > 2\alpha$ are inside the interval of equivalence; for Bayesian analysis this would mean the posterior probabilities of being below the interval and being above the interval both exceed α.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

confidence intervals" (Cumming 2014) shows a valid interval will bounce around from sample to sample due to random variation.

Because every inferential statistic (including P-values, interval estimates, and posterior probabilities) is derived from the multiple implicit as well as explicit assumptions that compose the model, we should treat these statistics as descriptions of the relation of the model to the data rather than as statements about the correctness of the model. This discipline may be more difficult for Bayesians to accept, since Bayesian methods do produce hypothetical probabilities of models by using assumptions hidden in restrictions on the models considered. Regardless, 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.

For example, a P-value is merely the probability of one particular test statistic being as or more extreme than observed in our particular study, given that the model it is computed from is correct. No inferential meaning need 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 (including an assumption such as "the means of these two populations do not differ", i.e., our tested hypothesis). 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 (Greenland 2017; Stark and Saltelli 2018).

Science includes learning about assumption violations, then addressing those violations and improving the performance of our models about reality. Statistics can help by formalizing parts of the models, and by assisting in careful assessment of uncertainty. We should thus communicate our limited conclusions about our data, not our generalized inferences about some ill-defined universal population. And decisions to communicate and interpret a result should not be based on P-values, nor on any other statistic. Presentations that start with analysis plans that were formulated before the analysis (pre-analysis protocols) can help strengthen both the validity and credibility of our inferences. The reported description of our results will be a good description if it is complete and honest. If we think we did a good study, we should thus be modest about our conclusions, but be proud about our painfully honest and thorough description and discussion of our methods and of our data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

And if we are working as journal editors, we should be proud about our exhaustive methods sections and consider "results blind evaluation" of manuscripts (Locascio 2017), i.e., basing our decisions about the suitability of a study for publication on the quality of its materials and methods rather than on results and conclusions; the quality of the presentation of the latter is only judged after it is determined that the study is valuable based on its materials and methods.

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 Science Collaboration 2015). So surprising results will often 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 asking researchers about surprising or unsurprising results, but instead ask which results appeared most boring because they were shown several times before and thus seem to be trustworthy. More generally, we should not fall for overconfident claims by researchers or by other science writers or journalists. Rather, we should try to uncover overconfident claims and the bad incentives that lead to those claims.

Clear signs of overconfidence are formulations like "we proved" or "we disproved" or "we rejected" a hypothesis, or "we have shown" or "demonstrated" a relation exists or is explained in some manner. So are "there was *no* effect / *no* association / *no* difference" (which almost always would be an impossible proof of the null hypothesis), and "our study confirms / validates / invalidates / refutes previous results" (because a single study has nothing definitive, it can only add one further data point to the larger picture; at most it can be "consistent / inconsistent with previous results"). If we find any of those or related phrases, we should question the interpretations being offered in the paper and search for arguments provided by the authors. If the main argument for a conclusion is that the results were "significant" or "not significant," this does

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

not automatically mean that the study is bad. But it does flag the paper as likely providing an unreliable interpretation of the reported results.

The hard truth is that journalists cannot decide whether a result from a single study can be generalized – and the same is usually true for the authors of the study, and for editors and external experts. An important role for statistics in research is the summary and accumulation of information. If replications do not find the same results, this is not necessarily a crisis, but is part of 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 inferences, conclusions, or decisions.

Acknowledgements: We thank three referees, the associate editor, and the editor Allen Schirm for exceptionally helpful critical comments. For additional comments, we thank Natalie Jeanneret, Fränzi Korner-Nievergelt, Timothy L. Lash, Lilla Lovász, Michel Montagner, Freya Pappert, and Tobias Roth.

Appendix: A Descriptive View of P-values and Posterior Probabilities

We here provide a more technical explanation of why *inferential* statistics like P-values and posterior probabilities can be interpreted as being *descriptive* of logical relations between assumptions and the observed data.

Consider a data-generating model M we wish to test and a test statistic (function of the data) T that has been selected as a summary measure of the absolute deviation of the data from the model predictions, with an observed value for T of t.[5] Here, M is the assumption set or constraints used to derive the distribution of T. To say that this test gave back $p = 0.04$ is to say that under the model M, the probability that $T \geq t$ is 0.04, or symbolically that $\Pr(T \geq t|M) = 0.04$. This Fisherian[6] P-value of 0.04 is thus a *logical* inference from the model M and the observation that $T = t$, the final deduction "p = 0.04" from a derivation that begins with the premises "the data

[5] There are of course many technical details we must gloss over such as optimization of T to detect a scientifically relevant type of deviation, and how to proceed when as usual M does not fully specify the distribution of T; see Bayarri and Berger (2000) for further discussion. For simplicity we here assume discrete data and model spaces.
[6] That is, a tail probability of an observed statistic. This device actually precedes Fisher, for example in the goodness-of-fit tests of Karl Pearson.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

were generated from M" and "those data produced T = t." Thus the model and data are in a logical relation in which their conjunction implies a probability p for the tail event $T \geq t$.

Without further elements (such as an α-level cutoff) this observed P-value p implies absolutely no decision, inference, bet, or behavior. What people make of p = 0.04 in these or other practical terms requires additional contextual detail such as a loss function, acceptable error rates, or whatever else the analysis team can bring to bear (although usually it is just social conventions like α = 0.05 and program defaults that determine what gets claimed).

The same comments apply to posterior probabilities. Now however "the model" (the set of assumptions used for deducing posterior probabilities) must include or entail a prior distribution Pr(M) over some restricted model family for generating a probability Pr(x|M) for the observed data x so that the full model provides a formula for Pr(x,M) = Pr(x|M)Pr(M) (Box 1980). Again the deduction of Pr(M|x) follows by conditioning on the observed data x (or equivalently on a sufficient statistic t, using T = t as a premise); that is, the posterior Pr(M|x) becomes a deduction from the observed data x and the full model Pr(x,M). Nonetheless, because one cannot construct a prior distribution over every conceivable model, this deduction is limited by the assumption that the model family used can approximate (within statistical error) the true data-generating process. In contrast, Fisherian P-values can be used to test this assumption before using it in the Bayesian deduction (Bayarri and Berger 2000; Robins et al. 2000).

References

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

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, available at https://doi.org/10.7717/peerj.3544.

Amrhein, V., Trafimow, D., and Greenland, S. (2018), "Abandon statistical inference," *PeerJ Preprints*, 6, e26857v1, available at https://doi.org/10.7287/peerj.preprints.26857v1.

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

Bayarri, M.J., and Berger, J.O. (2000), "P-values for composite null models," *Journal of the American Statistical Association*, 95, 1127–1142.

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

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.

Brown, H.K., Ray, J.G., Wilton, A.S., Lunsky, Y., Gomes, T., and Vigod, S.N. (2017), "Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children," *Jama-Journal of the American Medical Association*, 317, 1544–1552.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T.Z., Chen, Y.L., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.J. & Wu, H. (2018), "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2, 637–644.

Cohen, J. (1994), "The earth is round ($p < .05$)," *American Psychologist*, 49, 997–1003.

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," available at https://arxiv.org/abs/1711.07801.

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

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* (2nd 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*.

Gelman, A., and Hennig, C. (2017), "Beyond subjective and objective in statistics," *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 180, 967–1033.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

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

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.

Goodman, S.N. (1992), "A comment on replication, P-values and evidence," *Statistics in Medicine*, 11, 875–879.

Greenland, S. (2011), "Null misinterpretation in statistical testing and its impact on health risk assessment," *Preventive Medicine*, 53, 225–228.

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

Greenland, S. (2018a), "Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values," *The American Statistician*, in press.

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

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," *The American Statistician*, 70, online supplement 1 at http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf; reprinted in *European Journal of Epidemiology*, 31, 337–350.

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.

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.

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

Lakens, D., Scheel, A.M., and Isager, P.M. (2018), "Equivalence testing for psychological research: A tutorial," *Advances in Methods and Practices in Psychological Science*, 1, 259–269, available at https://doi.org/10.1177/2515245918770963.

Lehmann, E.L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Springer.

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

Locascio, J. (2017), "Results blind science publishing," *Basic and Applied Social Psychology*, 39, 239–246.

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

McShane, B.B., Gal, D., Gelman, A., Robert, C., and Tackett, J.L. (2018), "Abandon statistical significance," *The American Statistician*, in press.

Meehl, P.E. (1990), "Why summaries of research on psychological theories are often uninterpretable," *Psychological Reports*, 66, 195–244.

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

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

Poole, C. (1987a), "Beyond the confidence interval," *American Journal of Public Health*, 77, 195–199.

Poole, C. (1987b), "Confidence intervals exclude nothing," *American Journal of Public Health*, 77, 492–493.

Popper, K.R. (1968), *The Logic of Scientific Discovery* (2nd English ed.), London: Routledge.

Robins, J.M., van der Vaart, A., and Ventura, V. (2000), "Asymptotic distribution of P values in composite null models," *Journal of the American Statistical Association*, 95, 1143–1156.

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

Senn, S.J. (2001), "Two cheers for P-values?" *Journal of Epidemiology and Biostatistics*, 6, 193–204.

Senn, S.J. (2002), "Letter to the Editor" re: Goodman 1992, *Statistics in Medicine*, 21, 2437–2444.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018

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

Stark, P.B., and Saltelli, A. (2018), "Cargo-cult statistics and scientific crisis," *Significance*, 15, 40–43.

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," *Frontiers in Psychology*, 9, 699, available at https://doi.org/10.3389/fpsyg.2018.00699.

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

Wellek, S. (2010), *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2nd ed.), New York: Chapman & Hall.

Ziliak, S.T., and McCloskey, D.N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v4 | CC BY 4.0 Open Access | rec: 25 Oct 2018, publ: 25 Oct 2018