A peer-reviewed version of this preprint was published in PeerJ on 7 July 2017.

View the peer-reviewed version (peerj.com/articles/3544), which is the preferred citable publication unless you specifically need to cite this preprint.

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

The earth is flat (p>0.05): Significance thresholds and the crisis of unreplicable research

Valentin Amrhein[1,2,3], Fränzi Korner-Nievergelt[3,4], Tobias Roth[1,2]

[1]Zoological Institute, University of Basel, Basel, Switzerland
[2]Research Station Petite Camargue Alsacienne, Saint-Louis, France
[3]Swiss Ornithological Institute, Sempach, Switzerland
[4]Oikostat GmbH, Ettiswil, Switzerland
Email address: v.amrhein@unibas.ch

Abstract

The widespread use of 'statistical significance' as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process (according to the American Statistical Association). We review why degrading p-values into 'significant' and 'nonsignificant' contributes to making studies irreproducible, or to making them seem irreproducible. A major problem is that we tend to take small p-values at face value, but mistrust results with larger p-values. In either case, p-values tell little about reliability of research, because they are hardly replicable even if an alternative hypothesis is true. Also significance ($p \leq 0.05$) is hardly replicable: at a good statistical power of 80%, two studies will be 'conflicting', meaning that one is significant and the other is not, in one third of the cases if there is a true effect. A replication can therefore not be interpreted as having failed only because it is nonsignificant. Many apparent replication failures may thus reflect faulty judgment based on significance thresholds rather than a crisis of unreplicable research. Reliable conclusions on replicability and practical importance of a finding can only be drawn using cumulative evidence from multiple independent studies. However, applying significance thresholds makes cumulative knowledge unreliable. One reason is that with anything but ideal statistical power, significant effect sizes will be biased upwards. Interpreting inflated significant results while ignoring nonsignificant results will thus lead to wrong conclusions. But current incentives to hunt for significance lead to selective reporting and to publication bias against nonsignificant findings. Data dredging, p-hacking, and publication bias should be addressed by removing fixed significance thresholds. Consistent with the recommendations of the late Ronald Fisher, p-values should be interpreted as graded measures of the strength of evidence against the null hypothesis. Also larger p-values offer some evidence against the null hypothesis, and they cannot be interpreted as supporting the null hypothesis, falsely concluding that 'there is no effect'. Information on possible true effect sizes that are compatible with the data must be obtained from the point estimate, e.g., from a sample average, and from the interval estimate, such as a confidence interval. We review how confusion about interpretation of larger p-values can be traced back to historical disputes among the founders of modern statistics. We further discuss potential arguments against removing significance thresholds, for example that decision rules should rather be more stringent, that sample sizes could decrease, or that p-values should better be completely abandoned. We conclude that whatever method of statistical inference we use, dichotomous threshold thinking must give way to non-automated informed judgment.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Introduction

"It seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an 'uncertainty laundering' that begins with data and concludes with success as measured by statistical significance. (...) The solution is not to reform p-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation."
Andrew Gelman (2016)

Scientific results can be irreproducible for at least six major reasons (Academy of Medical Sciences 2015). There may be (1) technical problems that are specific to the particular study. There may be more general problems like (2) weak experimental design or (3) methods that are not precisely described so that results cannot be reproduced. And there may be statistical issues affecting replicability that are largely the same in many fields of research. Such issues are (4) low statistical power, and (5) 'data dredging' or 'p-hacking' by trying alternative analyses until a significant result is found, which then is selectively reported without mentioning the nonsignificant outcomes. Related to that, (6) publication bias occurs when papers are more likely to be published if they report significant results (Bishop & Thompson 2016).

Is a major part of an apparent crisis of unreplicable research caused by the way we use statistics for analyzing, interpreting, and communicating our data? Significance testing has been severely criticized for about a century (e.g., Boring 1919; Berkson 1938; Rozeboom 1960; Oakes 1986; Cohen 1994; Ziliak & McCloskey 2008; Kline 2013), but the prevalence of p-values in the biomedical literature is still increasing (Chavalarias et al. 2016). For this review, we assume that a revolution in applied statistics with the aim of banning p-values is not to be expected nor necessarily useful, and that the main problem is not p-values but how they are used (Gelman 2013b; Gelman 2016). We argue that one of the smallest incremental steps to address statistical issues of replicability, and at the same time a most urgent step, is to remove thresholds of statistical significance like p=0.05 (see Box 1). This may still sound fairly radical to some, but for the following reasons it is actually not.

First, p-values can be traditionally employed and interpreted as evidence against null hypotheses also without using a significance threshold. However, what needs to change for reducing data dredging and publication bias is our overconfidence in what significant p-values can tell, and, as the other side of the coin, our bad attitude towards p-values that do not pass a threshold of significance. As long as we treat our larger p-values as unwanted children, they will continue disappearing in our file drawers, causing publication bias, which has been identified as the possibly most prevalent threat to reliability and replicability of research already a long time ago (Sterling 1959; Wolf 1961; Rosenthal 1979). Still today, in an online survey of 1576 researchers, selective reporting was considered the most important factor contributing to irreproducible research (Baker 2016).

Second, the claim to remove fixed significance thresholds is widely shared among statisticians. In 2016, the American Statistical Association (ASA) published a statement on p-values, produced by a group of more than two dozen experts (Wasserstein & Lazar 2016). While there were controversial discussions about many topics, the consensus report of the ASA features the following statement: "The widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process" (Wasserstein & Lazar 2016). And a subgroup of seven ASA statisticians published an extensive review of 25 misinterpretations of p-values, confidence intervals, and power, closing with the words: "We join others in singling out the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

degradation of p-values into 'significant' and 'nonsignificant' as an especially pernicious statistical practice" (Greenland et al. 2016).

The idea of using p-values not as part of a binary decision rule but as a continuous measure of evidence against the null hypothesis has had many advocates, among them the late Ronald Fisher (Fisher 1956; Fisher 1958; Eysenck 1960; Skipper, Guenther & Nass 1967; Labovitz 1968; Edgington 1970; Oakes 1986; Rosnow & Rosenthal 1989; Stoehr 1999; Sterne & Smith 2001; Gelman 2013a; Greenland & Poole 2013; Higgs 2013; Savitz 2013; Madden, Shah & Esker 2015; Drummond 2016; Lemoine et al. 2016; van Helden 2016). Removing significance thresholds was also suggested by authors sincerely defending p-values against their critics (Weinberg 2001; Hurlbert & Lombardi 2009; Murtaugh 2014a).

In the following, we start with reviewing what p-values can tell about replicability and reliability of results. That this will not be very encouraging should not be taken as another advice to stop using p-values. Rather, we want to stress that reliable information about reliability of results cannot be obtained from p-values nor from any other statistic calculated in individual studies. Instead, we should design, execute, and interpret our research as a 'prospective meta-analysis' (Ioannidis 2010), to allow combining knowledge from multiple independent studies, each producing results that are as unbiased as possible. Our aim is to show that not p-values, but significance thresholds are a serious obstacle in this regard.

We therefore do not focus on general misconceptions about p-values, but on problems with, history of, and solutions for applying significance thresholds. After discussing why significance cannot be used to reliably judge the credibility of results, we review why applying significance thresholds reduces replicability. We then describe how the switch in interpretation that often follows once a significance threshold is crossed leads to proofs of the null hypothesis like 'the earth is flat ($p>0.05$)'. We continue by summarizing opposing recommendations by Ronald Fisher versus Jerzy Neyman and Egon Pearson that led to the unclear status of nonsignificant results, contributing to publication bias. Finally, we outline how to use graded evidence and discuss potential arguments against removing significance thresholds. We conclude that we side with a neoFisherian paradigm of treating p-values as graded evidence against the null hypothesis. We think that little would need to change, but much could be gained by respectfully discharging significance, and by cautiously interpreting p-values as continuous measures of evidence.

Box 1: **Significance thresholds and two sorts of reproducibility**

Inferential reproducibility might be the most important dimension of reproducibility and "refers to the drawing of qualitatively similar conclusions" from an independent replication of a study (Goodman, Fanelli & Ioannidis 2016). Some people erroneously conclude that a nonsignificant replication automatically contradicts a significant original study. Others will look at the observed effect, which might hint into the same direction as in the original study, and therefore come to the opposite conclusion. Since judgment based on significance is faulty, judgment based on effect sizes will increase inferential reproducibility. Further, it is current practice to interpret p-values >0.05 as evidence either against the null hypothesis, or (falsely) in favor of a null effect, or as no evidence at all. Researchers will increase inferential reproducibility if they refrain from turning their conclusion upside down once a significance threshold is crossed, but instead take the p-value as providing graded evidence against the null hypothesis.

Results reproducibility, or **replicability**, "refers to obtaining the same results from the conduct of an independent study" (Goodman, Fanelli & Ioannidis 2016). How results should look like to be

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

considered 'the same', however, remains operationally elusive. What matters, according to Goodman, Fanelli & Ioannidis (2016), "is not replication defined by the presence or absence of statistical significance, but the evaluation of the cumulative evidence and assessment of whether it is susceptible to major biases." Unfortunately, adhering to significance thresholds brings considerable bias to the published record of cumulative evidence. If results are selected for publication and interpretation because they are significant, conclusions will be invalid. One reason is that the lens of statistical significance usually sees only inflated effects and results that are "too good to be true" (Gelman 2015). Researchers will increase replicability if they report and discuss all results, irrespective of the sizes of their p-values.

P-values are hardly replicable

In most cases, null hypothesis significance testing is used to examine how compatible some data are with the null hypothesis that the true effect size is zero. The statistical test result is a p-value informing on the probability of the observed data, or data more extreme, given that the null hypothesis is true (and given that all other assumptions about the model are correct; Greenland et al. 2016). If $p \leq 0.05$, we have learned in our statistics courses to call this significant, to reject the null hypothesis, and to accept an alternative hypothesis about some non-zero effect in the larger population.

However, we do not know nor can we infer whether the null hypothesis or an alternative hypothesis is true. On the basis of one single study, it is logically impossible to draw a firm conclusion (Oakes 1986, p. 128), for example because a small p-value either means the null hypothesis is not true, or else it is true but we happened to find relatively unlikely data. It is for those "possible effects of chance coincidence" that Ronald Fisher wrote: "No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon" (Fisher 1937, p. 16).

Unlike a widespread belief, the p-value itself does not indicate how replicable our results are (Miller 2009; Greenland et al. 2016). We hope that a small p-value means our results are reliable and a replication study would have a good chance to find a small p-value again. Indeed, an extensive research project replicating 100 psychological studies reported that the chance to find a significant result in a replication was higher if the p-value in the original study was smaller; but of 63 original studies with $p<0.02$, only 26 (41%) had $p<0.05$ in the replication (Open Science Collaboration 2015).

Apparently, p-values are hardly replicable. This is most evident if the null hypothesis is true, because then p-values are uniformly distributed and thus all values are equally likely to occur (Hung et al. 1997; Boos & Stefanski 2011; Colquhoun 2014). However, a null hypothesis of an effect of exactly zero is often unlikely to be true (Loftus 1993; Cohen 1994; Stahel 2016). After all, in most cases we did our study because we had some *a priori* reason to believe that the true effect is not zero.

Unfortunately, p-values are highly variable and thus are hardly replicable also if an alternative hypothesis is true, or if the observed effect size is used for calculating the distribution of p (Goodman 1992; Senn 2002; Cumming 2008; Halsey et al. 2015). Cumming (2008) showed that if we observe an effect with $p=0.05$ in a first study, a replication study will find a p-value between 0.00008 and 0.44 with 80% probability (given by an 80% 'prediction interval'), and of >0.44 with 10% probability. For $p=0.5$ in a first study, Lazzeroni, Lu & Belitskaya-Levy (2014) found that 95% of replication p-values will have sizes between 0.003 and 0.997 (given by a 95% prediction interval).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

This enormous variability from sample to sample was called the 'dance of the p-values' (Cumming 2012; Cumming 2014). Because the p-value is based upon analysis of random variables, it is a random variable itself, and it behaves as such (Hung et al. 1997; Sackrowitz & Samuel-Cahn 1999; Murdoch, Tsai & Adcock 2008). For some reason, however, the stochastic aspect of p-values is usually neglected, and p is reported as a fixed value without a measure of vagueness or unreliability (Sackrowitz & Samuel-Cahn 1999; Cumming 2008; Barber & Ogle 2014). Indeed, we cannot use standard errors or confidence intervals for p, because they would estimate unobservable population parameters; and because the p-value is a property of the sample, there is no unobservable 'true p-value' in the larger population (Cumming 2012, p. 133). But as shown, e.g., by Cumming (2008), we could present our p-values with prediction intervals, which are intervals with a specified chance of including the p-value given by a replication.

If we would make vagueness of p-values visible by using prediction intervals, it would become immediately apparent that the information content of p=0.04 and of p=0.06 is essentially the same (Dixon 2003; Halsey et al. 2015; Giner-Sorolla 2016), and that "the difference between 'significant' and 'not significant' is not itself statistically significant" (Gelman & Stern 2006). It is a good habit to publish exact p-values rather than uninformative statements like 'p>0.05'; but additional decimal places and an equal sign should not mislead us to give p-values an aura of exactitude (Boos & Stefanski 2011; Halsey et al. 2015).

P-values are only as reliable as the samples from which they are obtained (Halsey et al. 2015). They inherit their vagueness from the uncertainty of point estimates like the sample average from which they are calculated. But they clearly give less information on uncertainty, reliability or repeatability of the point estimate than is evident from a 95% confidence interval (which is an 83% prediction interval for the point estimate of a replication; Cumming 2014). While the confidence interval measures precision and, therefore, reliability of the point estimate, the p-value mixes information on the size of the effect and how precisely it was measured. Thus, two point estimates can be equally reliable but may have different effect sizes and therefore different p-values (Fig. 1A, D). And a small p-value can arise because a point estimate is far off the null value or because the sample size is large; but data may still show considerable variation around the point estimate that therefore would not be very reliable (Fig. 1A versus E).

So by definition, the p-value reflects our observed evidence against a null hypothesis, but it does not directly measure reliability of the effect that we found in our sample. And we saw that p-values are much more variable than most people think (Lai, Fidler & Cumming 2012). We therefore must learn to treat p-values like any other descriptive statistic and refrain from taking them at face value when we want to draw inference beyond our particular sample data (Miller 2009). Using observed p-values to make a binary decision whether or not to reject a hypothesis is as risky as placing our bets on a sample average without considering that there might be error attached to it. If p-values are hardly replicable, so too are decisions based on them (Kline 2013, p. 13).

It seems that the only way to know how replicable our results are is to actually replicate our results. Science will proceed by combining cumulative knowledge from several studies on a particular topic, summarized for example in meta-analyses (Schmidt 1992; Schmidt 1996; Goodman, Fanelli & Ioannidis 2016). And one reason why replication studies are rarely done (Kelly 2006) may be that between 37% and 60% of academic professionals seem to think the p-value informs on the probability of replication (Gigerenzer, Krauss & Vitouch 2004). After all, why actually replicate a study when the p-value gives us virtual replications (Ziliak & McCloskey 2008, p. 127)?

If we do a true replication study, however, our observed p-value will be a realization of a random variable again and will be as unreliable as in the first study. A single replication thus can neither

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

validate nor invalidate the original study (Maxwell, Lau & Howard 2015; Leek & Jager 2016; Nosek & Errington 2017). It simply adds a second data point to the larger picture.

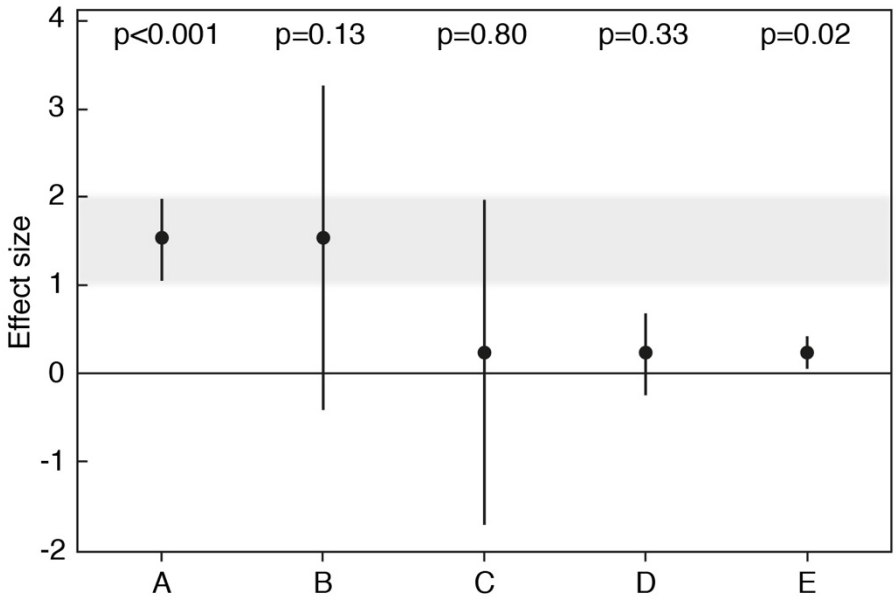

Figure 1 **Averages and 95% confidence intervals from five simulated studies.** P-values are from one sample t-tests, and sample sizes are n=30 each (adapted from Korner-Nievergelt & Hüppop 2016). Results A and E are relatively incompatible with the null hypothesis that the true effect size (the population average) is zero. Note that p-values in A versus D, or B versus C, are very different, although the estimates have the same precision and are thus equally reliable. Note also that the p-value in A is smaller than in E although variation is larger, because the point estimate in A is farther off the null value. If we define effect sizes between 1 and 2 as scientifically or practically important, result A is strong evidence that the effect is important, and result E is clear evidence that the effect is not important, because the small effect size was estimated with high precision. Result B is relatively clear evidence that the effect is not strongly negative and could be important, given that a value close to the center of a 95% confidence interval is about seven times as likely to be the true population parameter as is a value near a limit of the interval (Cumming 2014). Result C is only very weak evidence against the null hypothesis, and because plausibility for the parameter is greatest near the point estimate, we may say that the true population average could be relatively close to zero. However, result C also shows why a large p-value cannot be used to 'confirm' or 'support' the null hypothesis: first, the point estimate is larger than zero, thus the null hypothesis of a zero effect is not the hypothesis most compatible with the data. Second, the confidence interval shows population averages that would be consistent with the data and that could be strongly negative, or positive and even practically important. Because of this large uncertainty covering qualitatively very different parameter values, we should refrain from drawing conclusions about practical consequences based on result C. In contrast, result D is only weak evidence against the null hypothesis, but precision is sufficient to infer that possible parameter values are not far off the null and that the effect is practically not important. Result C is thus a case in which the large p-value and the wide confidence interval roughly say the same, which is that inference is difficult. Results B and D can be meaningfully interpreted even though p-values are relatively large.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Significance is hardly replicable

There is currently no consensus how replicability or results reproducibility should be measured (see Box 1; Open Science Collaboration 2015; Baker 2016; Goodman, Fanelli & Ioannidis 2016). Whether 'the same results' were obtained may be judged by comparing effect sizes and interval estimates, or by relying on subjective assessment by the scientists performing the replication (Open Science Collaboration 2015). What counts in the end will be the cumulative evidential weight from multiple independent studies; and those studies will probably show considerable variation in effect sizes (Goodman, Fanelli & Ioannidis 2016; Patil, Peng & Leek 2016).

Traditionally, the success versus failure of a replication is defined in terms of whether an effect in the same direction as in the original study has reached statistical significance again (Miller 2009; Open Science Collaboration 2015; Simonsohn 2015; Fabrigar & Wegener 2016; Nosek & Errington 2017). However, p-values are difficult to compare because they are sensitive to many differences among studies that are irrelevant to whether results are in agreement (Greenland et al. 2016, p. 343). For example, even if the point estimates are exactly the same, p-values of two studies may be on opposite sides of 0.05 because estimation precision or sample sizes differ (Simonsohn 2015). Further, if the p-value itself is hardly replicable, we would be surprised if $p \leq 0.05$ were replicable.

So how likely will two studies both turn out to be significant? It is sometimes suggested that replicability of significance is given by the statistical power of the test used in the replication study. Power is defined as the probability that a test will be significant given that the alternative hypothesis is true. However, because in real life we do not know whether the alternative hypothesis is true, also power does not help in judging replicability of empirical results. But if we theoretically assume that the alternative hypothesis is true, we can use power to make simple calculations about the probability that p-values cross a significance threshold such as $p=0.05$.

If two studies on a true alternative hypothesis have a reasonable sample size and thus the recommended statistical power of 80%, the probability that both studies are significant is 80% * 80% = 64%. As exemplified by Greenland et al. (2016), this means that under the luckiest circumstances, e.g. when the alternative hypothesis is true, when there is no publication bias, and when statistical power is good, then two studies will both be significant in only 64% of cases. The probability that one study is significant and the other is not is (80% * 20%)+(20% * 80%) = 32% (Greenland et al. 2016; 20% is the beta error of accepting the null hypothesis although it is false, which equals 1 - power). In one third of fairly ideal replications, results will traditionally be interpreted as conflicting, and replication as having failed.

However, the above replicabilities of significance are probably overestimated for two reasons. First, replication studies often report smaller effect sizes than the original studies due to publication bias in the original studies (see below). Second, most studies end up with a power much smaller than 80%. For example, average power to detect medium sized effects was about 40-47% in 10 journals on animal behavior (Jennions & Møller 2003), and median power in neuroscience was reported to be about 21% (Button et al. 2013b). In the Journal of Abnormal Psychology, in which Jacob Cohen (1962) found a median power of 46% for a medium effect in 1960, power dropped to 37% in 1984 (Sedlmeier & Gigerenzer 1989). As summarized from 44 reviews on statistical power in the social, behavioral and biological sciences, average power to detect small effects was 24% (Smaldino & McElreath 2016).

If we repeat the exercise by Greenland et al. (2016) with a more realistic power of 40%, we obtain a probability that both studies are significant of 40% * 40% = 16%, and a probability that there are conflicting results of (40% * 60%)+(60% * 40%) = 48%. This means that even if we did everything right, except for having only about average power, and if there is a true effect in the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

larger population, about half of our replications will fail by traditional significance standards (i.e., one study is significant and the other is not). And only about one in six studies will significantly replicate the significant result of another study.

This is of course not the fault of the p-value. It is the fault of us defining replication success as the event of crossing a significance threshold. As Greenland et al. (2016) put it, "one could anticipate a 'replication crisis' even if there were no publication or reporting bias, simply because current design and testing conventions treat individual study results as dichotomous outputs of significant/nonsignificant or reject/accept". Even in ideal replication studies, significance as defined by classical thresholds is not to be expected, and nonsignificance cannot be used as a criterion to undermine the credibility of a preceding study (Goodman, Fanelli & Ioannidis 2016).

Significance thresholds reduce replicability

In fact, it would be highly dubious if replication success in terms of statistical significance were larger than just described. This would indicate that researchers suppress nonsignificant replications and selectively report significant outcomes, and that there is publication bias against nonsignificant studies (Francis 2013). However, when nonsignificant results on a particular hypothesis remain unpublished, any significant support for the same hypothesis is rendered essentially uninterpretable (ASA statement; Wasserstein & Lazar 2016). If white swans remain unpublished, reports of black swans cannot be used to infer on general swan color. In the worst case, publication bias means according to Rosenthal (1979) that the 95% of studies that correctly yield nonsignificant results may be vanishing in file drawers, while journals may be filled with the 5% of studies committing the alpha error by claiming to have found a significant effect when in reality the null hypothesis is true.

However, selective reporting was encouraged since the early days of significance testing. As Ronald Fisher (1937, p. 15) wrote, "it is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard", an advice he gave at least since 1926 (Fisher 1926). As can be read in guidelines on writing papers and theses, students should "expect that you will produce many more figures and perform many more statistical tests than will be included in the final written product" (Lertzman 1995). Based on a survey of over 2000 psychologists, John, Loewenstein & Prelec (2012) estimated that among researchers, the prevalence of having engaged in selective reporting of studies that 'worked' or of only a subset of measured variables is 100%.

No wonder that a bias towards publishing significant results was observed for a long time and in many research areas (Sterling 1959; Csada, James & Espie 1996; Gerber & Malhotra 2008; Song et al. 2010; Dwan et al. 2013). Today, so-called negative results may disappear from many disciplines and countries (Fanelli 2012; but see de Winter & Dodou 2015), threatening the reliability of our scientific conclusions and contributing to the crisis of unreplicable research (Gelman 2015). No outright fraud, no technical fault or bad experimental design are necessary to render a study irreproducible; it is sufficient that we report results preferentially if they cross a threshold of significance.

The problem with selective reporting becomes even worse because significant results are not a random sample from all possible results – significant results are biased results. If a number of studies are done on a population with a fixed effect size, studies that due to sampling variation find a larger effect are more likely to be significant than those that happen to find smaller effects (Schmidt 1992). Using statistical significance as a guideline thus typically selects for large effects that are "too good to be true" (Gelman 2015). The consequence is that "most discovered true associations are inflated" (Ioannidis 2008). This effect was called 'truth inflation' (Reinhart 2015),

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

or 'winner's curse' (Zöllner & Pritchard 2007; Young, Ioannidis & Al-Ubaydi 2008; Button et al. 2013b) after how the term is used in economics: in high-risk situations with competitive bidding, the winner tends to be the bidder who most strongly overestimates the value of an object being auctioned (Capen, Clapp & Campbell 1971; Foreman & Murnighan 1996).

The inflation of effect sizes in significant results declines as statistical power increases (for example because sample sizes are large), and inflation becomes negligible as power approaches 100% (Colquhoun 2014; Gelman & Carlin 2014; Lemoine et al. 2016). One way to see how this works is to imagine that with a power of 100%, every test becomes significant given that the alternative hypothesis is true; thus, decisions for reporting based on significance would no longer select for large effects, because every effect from every random sample would be significant. However, with a more realistic power of, e.g., 40%, only inflated effects that on average are about 1.5 times larger than the true effect size may cross the significance threshold (Colquhoun 2014, Fig. 7).

In the ecological and neurological studies summarized in Lemoine et al. (2016) and Button et al. (2013b), sample sizes required to minimize inflation of effect sizes were n>100. However, as Lemoine et al. (2016) note, increased sample sizes can only partially offset the problem of inflated effect sizes, because a power near 100% will usually not be obtained.

Selective reporting of inflated significant effects while ignoring smaller and nonsignificant effects will lead to wrong conclusions in meta-analyses synthesizing effect sizes from a larger number of studies (Ferguson & Heene 2012; van Assen et al. 2014). This is one of the reasons why reliance on significance testing has been accused of systematically retarding the development of cumulative knowledge (Schmidt 1996; Branch 2014).

Of course, selective reporting of significant results leads to inflated effects not only in meta-analyses but in every single study. Even in cases in which authors report all conducted tests regardless of their p-values, but then select what to interpret and to discuss based on significance thresholds, the effects from which the authors draw their conclusions will be biased upwards.

The problem arises not only by consciously discarding nonsignificant findings. Also largely automated selection procedures may produce inflated effects, for example if genome-wide association studies select findings based on significance thresholds (Göring, Terwilliger & Blangero 2001; Garner 2007). In statistical model simplification, or model selection, significant predictors will have inflated point estimates (Whittingham et al. 2006; Ioannidis 2008; Forstmeier & Schielzeth 2011), and defining the importance of a predictor variable based on statistical significance will thus lead to distorted results.

Truth inflation

The p-value can be seen as a measure of surprise (Greenwald et al. 1996): the smaller it is, the more surprising the results are if the null hypothesis is true (Reinhart 2015, p. 9). If one wants to determine which patterns are unusual enough to warrant further investigation, p-values are thus perfectly suitable as explorative tools for selecting the largest effects. Whoever is interested in describing the average state of a system, however, should not "choose what to present based on statistical results", because "valid interpretation of those results is severely compromised" unless all tests that were done are disclosed (ASA statement, Wasserstein & Lazar 2016). And, we might add, unless all results are used for interpretation and for drawing conclusions, irrespective of their p-values.

Of course, current incentives lead to 'significance chasing' (Ioannidis 2010) rather than to publishing nonsignificant results. To put it more bluntly, "research is perverted to a hunt for statistically significant results" (Stahel 2016). The road to success is nicely summarized in the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

online author guidelines of the journal 'Nature' (accessed 2017): "The criteria for a paper to be sent for peer-review are that the results seem novel, arresting (illuminating, unexpected or surprising)." And the p-value, as a measure of surprise, seems to be a great selection tool for that purpose. However, the urge for large effect sizes in novel fields with little prior research is a "perfect combination for chronic truth inflation" (Reinhart 2015, p. 25). As wrote Ioannidis (2008), "at the time of first postulated discovery, we usually cannot tell whether an association exists at all, let alone judge its effect size."

Indeed, the strength of evidence for a particular hypothesis usually declines over time, with replication studies presenting smaller effects than original studies (Jennions & Møller 2002; Brembs, Button & Munafo 2013; Open Science Collaboration 2015). The reproducibility project on 100 psychological studies showed that larger original effect sizes were associated with greater effect size differences between original and replication, and that "surprising effects were less reproducible" (Open Science Collaboration 2015).

Small, early, and highly cited studies tend to overestimate effects (Fanelli, Costas & Ioannidis 2017). Pioneer studies with inflated effects often appear in higher-impact journals, while studies in lower-impact journals apparently tend to report more accurate estimates of effect sizes (Ioannidis 2005; Munafo, Stothart & Flint 2009; Munafo & Flint 2010; Siontis, Evangelou & Ioannidis 2011; Brembs, Button & Munafo 2013). The problem is likely publication bias towards significant and inflated effects particularly in the early stages of a potential discovery. At a later time, authors of replication studies might then want, or be allowed by editors, to report results also if they found only negligible or contradictory effects, because such results find a receptive audience in a critical scientific discussion (Jennions & Møller 2002). Replications therefore tend to suffer less from publication bias than original studies (Open Science Collaboration 2015).

So far, academic reward mechanisms often focus on statistical significance and newsworthiness of results rather than on reproducibility (Ioannidis et al. 2014). Also journalists and media consumers and, therefore, all of us ask for the novel, unexpected and surprising. Thus the average truth often does not make it to the paper and the public, and much of our attention is attracted by exaggerated results.

The earth is flat (p>0.05)

The average truth might be nonsignificant and non-surprising. But this does not mean the truth equals zero. In the last decades, many authors have compiled lists with misinterpretations regarding the meaning of the p-value (e.g., Greenland et al. 2016), and surveys showed that such false beliefs are widely shared among researchers (Oakes 1986; Lecoutre, Poitevineau & Lecoutre 2003; Gigerenzer, Krauss & Vitouch 2004; Badenes-Ribera et al. 2016). The "most devastating" of all false beliefs is probably that "if a difference or relation is not statistically significant, then it is zero, or at least so small that it can safely be considered to be zero" (Schmidt 1996). For example, if two studies are called conflicting or inconsistent because one is significant and the other is not, it may be implicitly assumed that the nonsignificant effect size was zero (Cumming 2012, p. 31).

Jacob Cohen (1994) published his classic critique of the use of significance tests under the title "The earth is round ($p < .05$)". What if this test happens to be nonsignificant? In 38% to 63% of articles sampled from five journals of psychology, neuropsychology and conservation biology, nonsignificant results were interpreted as 'there is no effect', which means that a null hypothesis was accepted or 'proven' (Finch, Cumming & Thomason 2001; Schatz et al. 2005; Fidler et al. 2006; Hoekstra et al. 2006).

In Cohen's example, 'no effect' would probably mean 'the earth is flat ($p>0.05$).' And this is not far from reality. Similar cases abound in the published literature, such as "lamb kill was not

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

correlated to trapper hours (r_{12} = 0.50, P = 0.095)" (cited in Johnson 1999). It may be completely obvious that the null hypothesis of 'no effect' cannot be true, as judging from a large but nonsignificant correlation coefficient, from clear but nonsignificant differences between averages in a figure, or from common sense; but still we do not hesitate to apply our "binary thinking, in which effects and comparisons are either treated as zero or are treated as real" (Gelman 2013b). How is this possible, since "of course, everyone knows that one can't actually prove null hypotheses" (Cohen 1990)?

We probably tend to misinterpret p-values because significance testing "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does" (Cohen 1994). Null hypothesis significance testing is not about estimating the probability that the null hypothesis or the alternative hypothesis is true – such a claim would be reserved to Bayesian testers, and even they would not be able to 'prove' any hypothesis. Null hypothesis testing is about the probability of our data *given* that the null hypothesis is true.

The problem is "the researcher's 'Bayesian' desire for probabilities of hypotheses" (Gigerenzer 1993). It may be hopeless to temper this desire, since also Fisher himself held a "quasi-Bayesian view that the exact level of significance somehow measures the confidence we should have that the null hypothesis is false" (Gigerenzer 1993). Yet, Fisher seemed to be clear about proving the null: a hypothesis cannot be "proved to be true, merely because it is not contradicted by the available facts" (Fisher 1935). And, therefore, "it is a fallacy, so well known as to be a *standard* example, to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened" (Fisher 1955; italics in original).

The last sentence, however, shows that also Fisher vacillated (Gigerenzer et al. 1989, p. 97). In fact, the null hypothesis cannot be confirmed nor strengthened, because very likely there are many better hypotheses: "Any p-value less than 1 implies that the test [null] hypothesis is not the hypothesis most compatible with the data, because any other hypothesis with a larger p-value would be even more compatible with the data" (Greenland et al. 2016). This can be seen when looking at a 'nonsignificant' 95% confidence interval that encompasses not only zero but also many other null hypotheses that would be compatible with the data, or, in other words, that would not be rejected using a 5% threshold (Fig. 1C; Tukey 1991; Tryon 2001; Hoekstra, Johnson & Kiers 2012). Within the confidence interval, zero is usually not the value that is closest to the observed point estimate. And even if the point estimate is exactly zero and thus "p=1, there will be many other hypotheses [i.e., the values covered by the confidence interval] that are highly consistent with the data, so that a definitive conclusion of 'no association' cannot be deduced from a p-value, no matter how large" (Greenland et al. 2016).

"Limbo of suspended disbelief"

It is easy to imagine research in which falsely claiming a true null effect causes great harm. To give a drastic example, Ziliak & McCloskey (2008, p. 28) cite a clinical trial on the painkiller Vioxx that reports data on heart attacks and other adverse events (Lisse et al. 2003). Lisse and colleagues found several 'p>0.2', among them that "the rofecoxib ['Vioxx'] and naproxen [generic drug] groups did not differ significantly in the number of thrombotic cardiovascular events (...) (10 vs. 7; P>0.2)". The conclusion was that "the results demonstrated no difference between rofecoxib and naproxen." Later, the unjustified proof of the null caused more suffering, and the manufacturer Merck took Vioxx off the market and faced more than 4200 lawsuits by August 20, 2005 (Ziliak & McCloskey 2008).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Unfortunately, if we finally accept that we cannot accept a null hypothesis, obtaining a nonsignificant result becomes downright annoying. If the null hypothesis cannot be rejected because $p>0.05$, it is often recommended to 'suspend judgment' (Tryon 2001; Hurlbert & Lombardi 2009), which leaves the null hypothesis "in a kind of limbo of suspended disbelief" (Edwards, Lindman & Savage 1963). As Cohen (1990) put it, "all you could conclude is that you couldn't conclude that the null was false. In other words, you could hardly conclude anything."

This unfortunate state becomes even worse because usually the researchers are blamed for not having collected more data. Indeed, in correlational research, most null hypotheses of an effect of exactly zero are likely wrong at least to a small degree (Edwards, Lindman & Savage 1963; Meehl 1967; Delong & Lang 1992; Lecoutre & Poitevineau 2014 – but see Hagen 1997; Hagen 1998; Thompson 1998; Krueger 2001 for a critical discussion). Therefore, most tests would probably produce significant results if only one had large enough sample sizes (Oakes 1986; Cohen 1990; Cohen 1994; Gill 1999; Stahel 2016). In other words, a significance test often does not make a clear statement about an effect, but instead it "examines if the sample size is large enough to detect the effect" (Stahel 2016). And because "you can pretty much always get statistical significance if you look hard enough" (Gelman 2015), you were probably "not trying hard enough to find significant results" (Ferguson & Heene 2012). Nonsignificance therefore seems to be regarded as "the sign of a badly conducted experiment" (Gigerenzer et al. 1989, p. 107).

As an outgoing editor of a major psychological journal wrote, "the [false] decision not to reject the null hypothesis can be a function of lack of power, lack of validity for the measures, unreliable measurement, lack of experiment control, and so on" (Campbell 1982). Surely all of those influences could also lead to falsely claiming a significant outcome, but the editor concludes: "it is true that there is an evaluation asymmetry between significant and nonsignificant results."

Naturally, we develop an aversion to 'null results' and fail to publish them (Ferguson & Heene 2012). Instead, we engage in significance chasing that may promote harmful practices like excluding variables or data with unwanted effects on p-values, stopping to collect data after looking at preliminary tests, rounding down p-values that are slightly larger than 0.05, or even falsifying data (non-exhaustive list after the survey by John, Loewenstein & Prelec 2012).

Although it is now widely discussed that significant results may be much less reliable than we used to believe, mistrust in nonsignificant results still seems larger than mistrust in significant results, leading to publication bias, which in turn causes significant results to be less reliable. What could be done?

Accepting history

First, we should briefly review the rise of modern statistics that started in the 1920s and 1930s. We think that problems like the unclear status of nonsignificant p-values have their roots in history, and that understanding those roots will help finding a future role for p-values (for more details, see Gigerenzer et al. 1989; Gill 1999; Salsburg 2001; Lenhard 2006; Hurlbert & Lombardi 2009; Lehmann 2011).

The three most influential founding fathers, Ronald A. Fisher, Jerzy Neyman and Egon S. Pearson, disagreed on many things, but they agreed that scientific inference should not be made mechanically (Gigerenzer & Marewski 2015). In their 'hypothesis tests', Neyman and Pearson confronted a point null hypothesis with a point alternative hypothesis. Based on this scenario they discovered alpha and beta errors as well as statistical power. Neyman and Pearson did not request to report p-values, but to make decisions based on predefined alpha and beta errors. They never recommended a fixed significance threshold (Lehmann 2011, p. 55), but rather held that defining error rates "must be left to the investigator" (Neyman & Pearson 1933a, p. 296).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Some years earlier, Fisher had introduced 'significance tests' using p-values on single null hypotheses, and he generally opposed the consideration of alternative hypotheses and of power (Lehmann 2011, p. 51). Fisher did not invent the p-value, which he called 'level of significance', but he was the first to outline formally the logic of its use (Goodman 1993). In 1925, he defined a threshold of p=0.05, based on the proportion of the area under a normal distribution that falls outside of roughly two standard deviations from the mean: "The value for which P =·.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant" (Fisher 1925). The choice of 0.05 was acknowledged to be "arbitrary" by Fisher (1929, p. 191) and was influenced by earlier definitions of significance by William Gosset, Karl Pearson (the father of Egon) and others (discussed in Cowles & Davis 1982b; Sauley & Bedeian 1989; Hurlbert & Lombardi 2009). A main reason why 0.05 was selected and still persists today may be that it fits our subjective feeling that events that happen at least as rarely as 10% or 1% of the time are suspiciously unlikely (Cowles & Davis 1982a; Weiss 2011).

Throughout his life, Fisher used the p-value mainly to determine whether a result was statistically significant (Lehmann 2011, p. 53). In his last new book, however, Fisher famously wrote that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (Fisher 1956, p. 42). In the thirteenth edition of "Statistical methods for research workers" (Fisher 1958), he then stated that "the actual value of P (...) indicates the strength of the evidence against the hypothesis" (p. 80) and that "tests of significance are used as an aid to judgment, and should not be confused with automatic acceptance tests, or 'decision functions' " (p. 128).

A main point of controversy was 'inductive inference' that was central to Fisher's thinking (Lehmann 2011, p. 90). Fisher believed that significance tests allow drawing inference from observations to hypotheses, or from the sample to the population (although deducing the probability of data given that a null hypothesis is true may actually look more like *deductive* inference, from the population to the sample; Thompson 1999).

In contrast, Neyman and Pearson thought that inductive inference is not possible in statistical analyses on single studies: "As far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis" (Neyman & Pearson 1933a, p. 291). Their hypothesis test was therefore not meant to give a measure of evidence (Goodman 1993), but to provide "rules to govern our behavior" with regard to our hypotheses, to insure that "in the long run of experience, we shall not be too often wrong" (Neyman & Pearson 1933a, p. 291). The Neyman-Pearson decision procedure was particularly suitable for industrial quality control, or "sampling tests laid down in commercial specifications" (Neyman & Pearson 1933b). Here, the quality of production may be tested very often over long periods of time, so that a frequentist 'long run of experience' is indeed possible, and manufacturers may indeed require thresholds to decide when to stop the production and to look for the cause of a quality problem (Gigerenzer et al. 1989, p. 100).

Applying Neyman-Pearson decision rules means accepting the null hypothesis while rejecting the alternative hypothesis, or vice versa. And Neyman and Pearson did indeed use the word 'accept' (e.g., Neyman & Pearson 1933b). Today, we may still speak about accepting a null hypothesis when it is false as committing a beta error, or error of the second kind. But was it not said that null hypotheses cannot be accepted? Fisher wanted 'inductive inference', and for drawing scientific conclusions, a nonsignificant result can never mean to accept a null hypothesis: "errors of the second kind are committed only by those who misunderstand the nature and application of tests of significance" (Fisher 1935). But Neyman and Pearson invented 'inductive behavior' explicitly for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

avoiding inductive inference. And for a behavioral decision, it is of course possible to accept any kind of premises: "As Neyman emphasized, to accept a hypothesis is not to regard it as true, or to believe it. At most it means to act as if it were true" (Gigerenzer et al. 1989, p. 101).

Thus accepting null hypotheses was encouraged by half of the founding schools of modern statistics. No wonder that "in approximately half of all cases, authors interpret their nonsignificant results as evidence that the null hypothesis is true" (McCarthy 2007, p. 43).

Our current null hypothesis significance tests are anonymous hybrids mixing elements both from the Fisherian and the Neyman-Pearson concepts. But mixing two essentially incompatible approaches, one for measuring evidence and the other for making behavioral decisions, of course creates all sorts of problems (Gigerenzer 1993; Goodman 1993; Gill 1999; Hubbard & Bayarri 2003; Schneider 2015). We often use Neyman-Pearson to refer to statistical power and to the two kinds of error. But then in practice we follow Fisher by refusing to specify a concrete point alternative hypothesis, and by interpreting exact p-values as graded measures of the strength of evidence against the null hypothesis (Mundry 2011; Cumming 2012, p. 25). However, we only consistently interpret p-values as strength of evidence as long as they are smaller than 0.05. For p-values between 0.05 and 0.1, some authors are willing to acknowledge a statistical 'trend', while others are not. For even larger p-values, we often switch back to a kind of Neyman-Pearson decision making, which offers no positive inference but seems to allow at least accepting the null hypothesis.

It looks like our mistrust in nonsignificant results that leads to publication bias is caused by confusion about interpretation of larger p-values that goes back to historical disputes among the founders of modern statistics.

Removing significance thresholds

What could be done? We could again define what a p-value means: the probability of the observed data, or data more extreme, given that the null hypothesis is true. According to the interpretation by the ASA (Wasserstein & Lazar 2016), smaller p-values cast more, and larger p-values cast less doubt on the null hypothesis. If we apply those definitions, it falls naturally to take p-values as graded evidence. We should try and forget our black-and-white thresholds (Tukey 1991) and instead consider the p-value as a continuous measure of the compatibility between the data and the null hypothesis, "ranging from 0 for complete incompatibility to 1 for perfect compatibility" (Greenland et al. 2016).

We need to move away from Fisher's early recommendation to ignore nonsignificant results, because following this rule leads to publication bias and to reported effects that are biased upwards. We need to move away from the Neyman-Pearson reject/accept procedure, because it leads to proofs of the null like 'not correlated (p=0.095)'. Instead, we should listen to the ASA-statisticians who say that if the "p-value is less than 1, some association must be present in the data, and one must look at the point estimate to determine the effect size most compatible with the data under the assumed model" (Greenland et al. 2016).

We are thus encouraged to interpret our point estimate as "our best bet for what we want to know" (Cumming 2007). According to the central limit theorem, sample averages are approximately normally distributed, so with repeated sampling most of them would cluster around the true population average. Within a 95% confidence interval, the sample average (the point estimate) is therefore about seven times as plausible, or seven times as good a bet for the true population average, as are the limits of the confidence interval (Cumming 2012, p. 99; Cumming 2014).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

After looking at the point estimate, we should then interpret the upper and lower limits of the confidence interval, which indicate values that are still plausible for the true population parameter. Those values should not appear completely unrealistic or be qualitatively very different; otherwise the width of our confidence interval suggests our estimate is so noisy that we should refrain from drawing firm conclusions about practical consequences (Fig. 1C).

If necessary, we should then focus on the p-value as a continuous measure of compatibility (Greenland et al. 2016), and interpret larger p-values as perhaps less convincing but generally 'positive' evidence against the null hypothesis, instead of evidence that is either 'negative' or uninterpretable or that only shows we did not collect enough data. In short, we should develop a critical but positive attitude towards larger p-values. This alone could lead to less proofs of the null hypothesis, to less significance chasing, less data dredging, less p-hacking, and ultimately to less publication bias, less inflated effect sizes and more reliable research.

And removing significance thresholds is one of the smallest steps that we could imagine to address issues of replicability. Using p-values as graded evidence would not require a change in statistical methods. It would require a slight change in interpretation of results that would be consistent with the recommendations by the late Ronald Fisher and thus with a neoFisherian paradigm described by Hurlbert & Lombardi (2009). A difference to Hurlbert & Lombardi (2009) is that for larger p-values we do not propose 'suspending judgment', which we believe would contribute to selective reporting and publication bias because we usually do not want to report results without evaluating possible conclusions. Instead, we recommend "suspending firm decisions (i.e., interpreting results with extra caution)" (Greenland & Poole 2013). As we saw, some results can be meaningfully interpreted although their p-values are relatively large (Fig. 1B, D); and even if p-values were small, firm decisions would not be possible based on an isolated experiment (Fisher 1937, p. 16).

For our next scientific enterprise using frequentist statistics, we suggest that we

a) do our study and our analysis as planned
b) report our point estimate, interpret our effect size
c) report and interpret an interval estimate, e.g. a 95% confidence interval
d) report the exact p-value
e) do not use the word 'significant' and do not deny our observed effect if the p-value is relatively large
f) discuss how strong we judge the evidence, and how practically important the effect is.

Do we need a scale for interpreting the strength of evidence against the null hypothesis? Graded evidence means there are no thresholds to switch from 'strong' to 'moderate' to 'weak' evidence (Sterne & Smith 2001). It means that "similar data should provide similar evidence" (Dixon 2003), because "surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal 1989).

There are no 'few-sizes-fit-all' grades of evidence. Instead of following the same decision rules, no matter how large the sample size or how serious we judge measurement error, we should "move toward a greater acceptance of uncertainty and embracing of variation" (Gelman 2016). If we have obtained a small p-value, we must be aware of the large variability of p-values, keep in mind that our evidence against the null hypothesis might not be as strong as it seems, and acknowledge that our point estimate is probably biased upwards. If we have obtained a large p-value, we must be even more aware that many other hypotheses are compatible with our data, including the null hypothesis. Looking at the values covered by the confidence interval will help identifying those competing hypotheses. Very likely there are hypotheses compatible with the data that would cause

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

even greater concern than a zero effect, e.g. if the effect would be in the opposite direction (Fig. 1C) or be much smaller or larger than what we observed in our point estimate. Note that the true effect can also be much smaller or larger than our point estimate if the p-value is small.

When discussing our results, we should "bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis" (ASA statement; Wasserstein & Lazar 2016). For example, results from exploratory studies are usually less reliable than from confirmatory (replication) studies also if their p-values were the same, because exploratory research offers more degrees of freedom in data collection and analysis (Gelman & Loken 2014; Higginson & Munafo 2016; Lew 2016). Already half a century ago, Labovitz (1968) compiled a list of criteria for evaluating evidence from p-values that include the practical consequences (costs) of a conclusion, the plausibility of alternative hypotheses, or the robustness of the test.

And we must keep in mind that support for our hypothesis will require not just one, but many more independent replications. If those replications do not find the same results, this is not necessarily a crisis, but a natural process by which science proceeds. In most research disciplines in which results are subject to a substantial random error, "the paradigm of accumulating evidence might be more appropriate than any binary criteria for successful or unsuccessful replication" (Goodman, Fanelli & Ioannidis 2016). Thus, "the replication crisis perhaps exists only for those who do not view research through the lens of meta-analysis" (Stanley & Spence 2014).

We therefore need to publish as many of our results as possible, as long as we judge them sufficiently reliable to be reported to the scientific community. And we should publish our results also if we find them neither novel nor arresting, illuminating, unexpected or surprising. Those results that we find trustworthy are the best approximates of reality, especially if they look familiar and expected and replicate what we already think we know.

Increasing inferential reproducibility

Of course, people and journals who want to publish only small p-values will probably continue publishing only small p-values. Perhaps, "forced to focus their conclusions on continuous measures of evidence, scientists would selectively report continuous measures of evidence" (Simonsohn 2014). We hope, however, that bidding farewell to significance thresholds will free at least some of our negative results by allowing them to be positive. And we think the chances that this will happen are good: because it is already happening.

People seem prepared to interpret p-values as graded evidence. When asked about the degree of confidence that an experimental treatment really has an effect, the average researcher's confidence was found to drop quite sharply as p-values rose to about 0.1, but then confidence levelled off until 0.9, essentially showing a graded response (Poitevineau & Lecoutre 2001). Nuzzo (2015) cites Matthew Hankins, who has collected "more than 500 creative phrases that researchers use to convince readers that their nonsignificant results are worthy of attention (see go.nature.com/pwctoq). These include 'flirting with conventional levels of significance ($p>0.1$)', 'on the very fringes of significance ($p=0.099$)' and 'not absolutely significant but very probably so ($p>0.05$)'."

As Hurlbert & Lombardi (2009) note, even Jerzy Neyman (1977, p. 112) labelled p-values of 0.09, 0.03, and <0.01 reported in an earlier paper (Lovasich et al. 1971, table 1, right column) as 'approximately significant', 'significant', and 'highly significant', respectively. And creativity is increasing. Pritschet, Powell & Horne (2016) searched over 1500 papers in three journals of psychology for terms like 'marginally significant' or 'approaching significance' that were used for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

p-values between 0.05 to 0.1 and up to 0.18. They found that "the odds of an article describing a p-value as marginally significant in 2010 were 3.6 times those of an article published in 1970." In 2000 and 2010, the proportion of articles describing at least one p-value as marginally significant were 59% and 54%.

The rules how to label results if p>0.05 are usually unwritten (Pritschet, Powell & Horne 2016), so it is current practice to interpret p-values between 0.05 and 0.1, or even larger p-values, as evidence either against the null hypothesis, or (falsely) in favor of a null effect, or as no evidence at all. This anarchical state of affairs undermines 'inferential reproducibility', which might be the most important dimension of reproducibility and "refers to the drawing of *qualitatively* similar conclusions from either an independent replication of a study or a reanalysis of the original study" (Goodman, Fanelli & Ioannidis 2016; italics supplied). Interpreting larger p-values as graded evidence, but as evidence that can only speak against the null hypothesis, would clearly help increasing inferential reproducibility by reducing the choices for qualitative interpretation. For example, a large p-value would not be taken to support the null hypothesis, and only rarely a result would be interpreted as providing no evidence at all.

As listed in the beginning of this paper, many authors have argued for removing fixed thresholds. In 2016, Lemoine et al. (2016) wrote: "we join others in calling for a shift in the statistical paradigm away from thresholds for statistical significance", and similar words were used in the closing remarks of Greenland et al. (2016). So far, however, we see few researchers performing this shift. Although black-and-white seems to give way to a more flexible approach allowing for trends and marginal significance, dichotomous threshold thinking is still the rule among applied researchers and even among many statisticians (McShane & Gal 2016; McShane & Gal 2017). Perhaps this is because of serious issues with interpreting p-values as graded evidence?

Arguments against removing thresholds

In the following, we discuss potential problems. We start each paragraph with a possible argument that could be raised against removing fixed significance thresholds.

'We need more stringent decision rules'

Evidently, removing significance thresholds "may lead to an increased prevalence of findings that provide weak evidence, at best, against the null hypothesis" (Pritschet, Powell & Horne 2016). This will not cause any problem, as long as the authors of a study acknowledge that they found only weak evidence. Quite the contrary, publishing weak evidence is necessary to reduce publication bias and truth inflation. However, p-values just below 0.05 are currently interpreted as generally allowing a decision to reject the null hypothesis, which is one of the reasons why scientific claims may often be unreliable (Oakes 1986; Sterne & Smith 2001; Johnson 2013; Colquhoun 2014). One alternative proposition to enhance replicability of research was therefore not to remove thresholds, but rather to apply more stringent thresholds (Johnson 2013; Ioannidis 2014; Academy of Medical Sciences 2015). For example, it may be said that you should "not regard anything greater than p<0.001 as a demonstration that you have discovered something" (Colquhoun 2014).

We agree it is currently too "easy for researchers to find large and statistically significant effects that could arise from noise alone" (Gelman & Carlin 2014). Interpreting p-values as graded evidence would mean to mistrust p-values around 0.05, or smaller, depending on the circumstances and the scientific discipline. To announce a discovery in particle physics, a p-value as small as 0.0000003 may be needed (Johnson 2014). Also in other disciplines, we should often judge our evidence as strong only if the p-value is much smaller than 0.05: if we want to demonstrate a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

surprising, counterintuitive effect; if we know that our null hypothesis has a high prior probability (Bayarri et al. 2016); if our sample size is large (Anderson, Burnham & Thompson 2000; Pericchi, Pereira & Perez 2014); or if postulating an effect that in reality is negligible would have serious practical consequences.

However, even a p-value of 0.1 or larger may be interpreted as sufficiently strong evidence: if we collected our data truly randomized and by minimizing bias and measurement error; if we stuck to our pre-planned protocol for data analysis without trying multiple alternative analyses; if our effect size is small; or if claiming only weak evidence for an effect that in reality is practically important would have serious consequences (Gaudart et al. 2014).

A large effect with a large p-value could potentially have much more impact in the real world than a small effect with a small p-value (Lemoine et al. 2016), although in the first case, the evidence against the null hypothesis is weaker (Fig. 1B, E). And in exploratory studies screening data for possible effects, results may be considered interesting even if their p-values are large (Madden, Shah & Esker 2015). Since p-values from exploratory research should be taken as descriptive with little inferential content (Berry 2016), such studies are perhaps the clearest cases in which there is no need for significance thresholds (Lew 2016). When exploring "what the data say" (Lew 2016), it simply makes no sense to listen to them only if $p<0.05$. We should, however, clearly state that the research was exploratory (Simmons, Nelson & Simonsohn 2011; Berry 2016). And we should keep in mind that if we recommend effects for further investigation because their p-values are small, our effect sizes are likely inflated and will probably be smaller in a follow-up study.

Another argument against stringent thresholds like $p=0.001$ is that sample sizes in fields like biomedicine or animal behavior are often bound to be small for practical and ethical reasons (Nakagawa 2004; Gaudart et al. 2014; Academy of Medical Sciences 2015). With small or moderate sample sizes, the chances that small but important effects will not become significant is very high, and this applies both to more stringent thresholds and to a possible scenario in which the conventional threshold of 0.05 were reinforced.

Because smaller p-values usually come with larger sample sizes, the question whether we should aim for smaller p-values boils down to whether we should conduct fewer but larger or more but smaller studies. Confidence and precision clearly increase with sample size (Button et al. 2013a). Based on simulations, however, IntHout, Ioannidis & Borm (2016) recommend doing rather several studies with a moderate power of 30% to 50% than one larger study that would need a sample size four times to twice as large, respectively, to obtain 80% power. The reason is that every study will suffer from some sort of bias in the way it is conducted or reported, and that "in a series of studies, it is less likely that all studies will suffer from the same type of bias; consequently, their composite picture may be more informative than the result of a single large trial" (IntHout, Ioannidis & Borm 2016).

And of course, to be of any use, all of those studies should be published. Cumulative evidence often builds up from several studies with larger p-values that only when combined show clear evidence against the null hypothesis (Greenland et al. 2016, p. 343). Very possibly, more stringent thresholds would lead to even more results being left unpublished, enhancing publication bias (Gaudart et al. 2014; Gelman & Robert 2014). What we call winner's curse, truth inflation or inflated effect sizes will become even more severe with more stringent thresholds (Button et al. 2013b). And one reason why 'false-positives' are vastly more likely than we think, or as we prefer to say, why the evidence is often vastly overestimated, is that we often flexibly try different approaches to analysis (Simmons, Nelson & Simonsohn 2011; Gelman & Loken 2014). Exploiting those researcher degrees of freedom, and related phenomena termed multiplicity, data dredging, or

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

p-hacking, would probably become more severe if obtaining significant results were harder due to more stringent thresholds.

We think that while aiming at making our published claims more reliable, requesting more stringent fixed thresholds would achieve quite the opposite.

'Sample sizes will decrease'

Ideally, significance thresholds should force researchers to think about the sizes of their samples – this is perhaps the only advantage thresholds could potentially offer. If significance is no longer required, it could happen that researchers more often content themselves with smaller sample sizes.

However, the argument could as well be reversed: fixed significance thresholds may lead to unreasonably small sample sizes. We have had significance thresholds for decades, and average power of studies was constantly low (Sedlmeier & Gigerenzer 1989; Button et al. 2013b). Researchers seem often to rely on rules of thumb by selecting sample sizes they think will yield p-values just below 0.05, although it is clear that p-values in this order of magnitude are hardly replicable (Vankov, Bowers & Munafo 2014). Oakes (1986, p. 85-86) argues that the major reason for the abundance of low-power studies is the belief that "experimental findings suddenly assume the mantle of reality at the arbitrary 0.05 level of significance" – if a result becomes true once a threshold is crossed, there is simply no need to strive for larger sample sizes.

In their summary of 44 reviews on statistical power, Smaldino & McElreath (2016) found not only that average power to detect small effects was 24%, but that there was no sign of increase over six decades. The authors blame an academic environment that only rewards significant findings, because in such a system, "an efficient way to succeed is to conduct low power studies. Why? Such studies are cheap and can be farmed for significant results" (Smaldino & McElreath 2016). Similarly, Higginson & Munafo (2016) suggest that "researchers acting to maximize their fitness should spend most of their effort seeking novel results and conduct small studies that have only 10%–40% statistical power."

Quite often, significance testing appears like a sort of gambling. Even a study with minimized investment into sample sizes will yield significant results, if only enough variables are measured and the right buttons in the statistical software are pushed. Small sample sizes further have it that significant effects will probably be inflated, so novel and surprising results are almost guaranteed. And we may have become somewhat addicted to this game – it is satisfying to feed data into the machine and find out whether the right variables have turned significant.

Indeed, "Fisher offered the idea of p-values as a means of protecting researchers from declaring truth based on patterns in noise. In an ironic twist, p-values are now often manipulated to lend credence to noisy claims based on small samples" (Gelman & Loken 2014). And the manipulation can happen completely unintentionally, "without the researcher performing any conscious procedure of fishing through the data" (Gelman & Loken 2014), just as a side-effect of how the game of significance testing is usually applied.

We should discourage significance farming by insisting that a p-value of 0.02 is not automatically convincing, and that there is no need to throw away a study with p=0.2. We hope that removing significance thresholds will allow researchers to take the risk and put more effort into larger studies, without the fear that all the effort could be wasted if there were nonsignificance in the end.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

'We need objective decisions'

Graded evidence means no strict criterion to decide whether there is evidence or not, no algorithm that makes the decision for us. Can graded evidence be objectively interpreted? Fisher's honorable aim was to develop a "rigorous and objective test" (Fisher 1922, p. 314), and Neyman and Pearson made the test even more rigorous by introducing automatic decisions between two competing hypotheses.

Unfortunately, it seems like "the original Neyman-Pearson framework has no utility outside quality control type applications" (Hurlbert & Lombardi 2009). In most studies, we do not formally specify alternative hypotheses, and we tend to use small sample sizes, so that beta errors are usually both high and unknown (Fidler et al. 2004). This is one reason why our Fisher-Neyman-Pearson hybrid tests offer only an "illusion of objectivity" (Berger & Berry 1988; Gigerenzer 1993). Another reason is that all statistical methods require subjective choices (Gelman & Hennig 2017). Prior to calculating p-values, we make all kinds of personal decisions: in formulating our research question, in selecting the variables to be measured, in determining the data sampling scheme, the statistical model, the test statistic, how to verify whether model assumptions are met, how to handle outliers, how to transform the data, which software to use. We do all of that and are used to justifying our choices; but interestingly, when it comes to interpreting test results, "one can feel widespread anxiety surrounding the exercise of informed personal judgment" (Gigerenzer 1993).

For statistical tests, our *a priori* assumptions about the true model may be, for example, that residuals are independent and identically distributed. Since "it is impossible logically to distinguish between model assumptions and the prior distribution of the parameter", using prior information is not a feature peculiar to Bayesian inference, but a necessity for all scientific inference (Box 1980). As explained by Oakes (1986, p. 114), the statistical schools differ in the manner in which they employ this prior information.

Pearson (1962) wrote about his work with Neyman: "We were certainly aware that inferences must make use of prior information", thus "we left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment in such matters – to use our terminology – as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities." But these judgments had to be made before data collection, and "every detail of the design, including responses to all possible surprises in the incoming data, must be planned in advance" (Berger & Berry 1988). To change the plan after data collection is to violate the model (Oakes 1986).

We agree this is the way to go in confirmatory studies that replicate procedures from earlier research. Sticking to the rules is one reason why results from replication studies are usually more reliable than from exploratory studies (Gelman & Loken 2014; Higginson & Munafo 2016); another reason is that confirmatory studies may suffer less from publication bias than exploratory studies (Open Science Collaboration 2015). We remind, however, that even in ideal confirmatory studies, significance is not to be expected (see above), thus significance thresholds should not be applied to judge replication success. And in exploratory studies, we see even less utility for the original Neyman-Pearson framework and their predefined significance thresholds. Such studies have all the rights to interrogate the data repeatedly and intensively (Lew 2016), as long as they are acknowledged to be exploratory (Gelman & Loken 2014; Berry 2016), and, of course, as long as they report all results irrespective of their p-values.

We thus side with Ronald Fisher that the decision how to deal with the null hypothesis should reside with the investigator rather than be taken for him in the Neyman-Pearson manner (Oakes 1986, p. 114). It should be remembered, however, that the decision whether there is an important effect cannot be answered in a single study, and that it is often more interesting to discuss the size

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

of the effect than to speculate on its mere existence (e.g., Cumming 2012; Gelman 2013a). In many cases, "research requires no firm decision: it contributes incrementally to an existing body of knowledge" (Sterne & Smith 2001). Or in the words of Rozeboom (1960): "The primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested."

'Null hypotheses should be acceptable'

As stated earlier, most null hypotheses of an effect of exactly zero are likely wrong at least to a small degree. This claim seems to be widely accepted for correlational research (Edwards, Lindman & Savage 1963; Meehl 1967; Cohen 1990; Tukey 1991; Cohen 1994; Lecoutre & Poitevineau 2014; Stahel 2016), although its general applicability and philosophical background are debated (Hagen 1997; Hagen 1998; Thompson 1998; Krueger 2001). However, the claim may not apply to experimental studies (Meehl 1990). A null hypothesis of an effect of exactly zero may be valid "for true experiments involving randomization (e.g., controlled clinical trials) or when any departure from pure chance is meaningful (as in laboratory experiments on clairvoyance)" (Cohen 1994).

Under such circumstances, we would like to be able to show that the null hypothesis is true. But how should this be done, if any p-value that we calculate can only speak against the null hypothesis (Delong & Lang 1992)? Interpreting p-values as graded evidence will mean that we can never support the null hypothesis, and that alternative hypotheses may be "virtually unkillable", buried at "a vast graveyard of undead theories" (Ferguson & Heene 2012, although the authors coined those words decrying publication bias against 'null' results and not the use of graded evidence).

So "how to accept the null hypothesis gracefully" (Greenwald 1975)? The hard truth is that p-values from classical null hypothesis tests are not made for this task – whether or not we use them as graded evidence. As we discussed at length, also nonsignificant results cannot not support the null hypothesis, and high power is of no help (Greenland 2012). There is only one way to attach a number to the probability of a null hypothesis: "use a range, rather than a point, null hypothesis", and "compute the posterior probability of the null (range) hypothesis" (Greenwald 1975), or use similar Bayesian methods (Gallistel 2009; Morey & Rouder 2011; Dienes 2014). Alternative procedures in a frequentist framework include equivalence tests (Levine et al. 2008) or the careful interpretation of effect sizes and confidence intervals, to determine whether an effect is so precisely small as to be practically unimportant (Fig. 1D, E; Cumming 2014, p. 17; Simonsohn 2015).

'We need to get rid of p-values'

So p-values vary enormously from sample to sample even if there is a true effect. Unless they are very small, p-values therefore give only unreliable evidence against the null hypothesis (Cumming 2012, p. 152). For this and other reasons, p-values should probably be viewed as descriptive statistics, not as formal quantifications of evidence (Lavine 2014). To make it worse, we usually translate the observed evidence against the null hypothesis into evidence in favor of an alternative hypothesis, without even mentioning the null hypothesis ('females were twice as large as males, $p<0.001$'). However, knowing that the data are unlikely under the null hypothesis is of little use unless we consider whether or not they are also unlikely under the alternative hypothesis (Sellke, Bayarri & Berger 2001; Barber & Ogle 2014). Since very often we are not honestly interested in describing evidence *against* a hypothesis, null hypothesis testing is usually a "parody of falsificationism in which straw-man null hypothesis A is rejected and this is taken as evidence in favor of preferred alternative B" (Gelman 2016).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Why do we not join others and discourage using p-values and null hypothesis tests (e.g., Carver 1978; Schmidt 1996; Ziliak & McCloskey 2008; Orlitzky 2012; Cumming 2014; Trafimow & Marks 2015; Gorard 2016)?

We agree to (Cumming 2012, p. 33) that "thinking of p as strength of evidence may be the least bad approach". We thus propose to describe our observed gradual evidence against the null hypothesis rather than to 'reject' the null. A first step would be to stop using the word 'significant' (Higgs 2013; Colquhoun 2014). Indeed, often null hypotheses about zero effects are automatically chosen only to be rejected (Gelman 2013b), and null hypotheses on effect sizes other than zero, or on ranges of effect sizes, would be more appropriate (Cohen 1994; Greenland et al. 2016). The way to become aware of our zero effect automatism is to "replace the word significant with a concise and defensible description of what we actually mean to evoke by using it" (Higgs 2013).

These are small changes that may look insufficient to some. But the beauty in such simple measures to help address the crisis of unreplicable research is that they do not hurt. We do not need to publicly register our study protocol before we collect data, we do not need to learn new statistics, we do not even need to use confidence intervals if we prefer to use standard errors. Of course all of those measures would be helpful (e.g., Cumming 2014; Academy of Medical Sciences 2015), but they usually require an active change of research practices. And many researchers seem to hesitate changing statistical procedures that have been standard for nearly a century (Thompson 1999; Sharpe 2013).

Although there are hundreds of papers arguing against null hypothesis significance testing, we see more and more p-values (Chavalarias et al. 2016) and the ASA feeling obliged to tell us how to use them properly (Goodman 2016; Wasserstein & Lazar 2016). Apparently, bashing or banning p-values does not work. We need a smaller incremental step that at the same time is highly efficient. There are not many easy ways to improve scientific inference, but together with Higgs (2013) and others, we believe that removing significance thresholds is one of them.

Also, we do not deny that some thresholds are necessary when interpreting statistics. If we want to draw error bars around a mean, the lines of the bars must end at some point that is defined by a threshold. By convention, we use 95% confidence intervals, cutting off parameter values that would be rejected at the p=0.05 threshold; of course we could also use 90%, 80% or 75% confidence intervals with 0.10, 0.20 or 0.25 thresholds (Hurlbert & Lombardi 2009). Geoff Cumming's (2012, p. 76) recommendation is to concentrate on 95% confidence intervals, because "it's challenging enough to build up good intuitions about the standard 95% level of confidence".

Unfortunately, many researchers seem to use confidence intervals mostly to decide whether the null value is excluded, thus converting them to significance tests (Hoekstra, Johnson & Kiers 2012; McCormack, Vandermeer & Allan 2013; Rothman 2014; Savalei & Dunn 2015; van Helden 2016). Potentially, using confidence intervals could encourage estimation thinking and meta-analytic thinking (Cumming 2014), but for full benefit, researchers would need to interpret confidence intervals without recourse to significance (Coulson et al. 2010). 'Cat's-eye' pictures of 95% confidence intervals show how the plausibility that a value is the true population average is greatest for values near our point estimate, in the center of the interval; plausibility then drops smoothly to either end of the confidence interval, then continues to drop further outside the interval (Cumming 2012, p. 95). This means that "we should not lapse back into dichotomous thinking by attaching any particular importance to whether a value of interest lies just inside or just outside our confidence interval" (Cumming 2014).

Further, we recommend choosing not only from the available null hypothesis tests but also from the toolbox provided by Bayesian statistics (Korner-Nievergelt et al. 2015). But we agree that we should not "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).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Whatever method of statistical inference we use, biased effect sizes will be a problem with criteria that are used to select results for publication or interpretation. If effect sizes are reported because they are large, this will of course create an upwards bias, similarly to selecting p-values because they are small. If effect sizes are reported because they are small, or p-values because they are large, for example when people are interested to show that some treatment may not have an (adverse) effect, this will create a downwards bias (Greenland et al. 2016). If results are reported because Bayesian posterior probabilities are at least three times larger for an alternative than for a null hypothesis, this is equivalent to selective reporting based on significance thresholds (Simonsohn 2014).

Inflated effects will occur when a discovery is claimed because a Bayes factor is better than a given value or a false discovery rate is below a given value (Button et al. 2013b). Selecting a model because the difference in the Akaike information criterion (ΔAIC) passes some threshold is equivalent to model selection based on significance and will generate inflated effects (Murtaugh 2014a; Murtaugh 2014b; Parker et al. 2016). Whenever we use effect sizes, confidence intervals, AIC, posteriors, Bayes factors, likelihood ratios, or false discovery rates in threshold tests and decision heuristics for reporting or interpreting selected results rather than all results, we create biases and invalidate the answers we give to our questions (Simonsohn 2014; Yu et al. 2014; Parker et al. 2016; Wasserstein & Lazar 2016).

However, the greatest source of bias probably comes from selectively reporting small p-values, simply because of the dominant role of null hypothesis significance testing. If we learn to judge the strength of evidence based not on the event of passing a threshold, but on graded summaries of data like the p-value, we will become more aware of uncertainty. And statistical methods are not simply applied to a discipline but change the discipline itself (Gigerenzer & Marewski 2015). A greater acceptance of uncertainty and embracing of variation (Gelman 2016) could shift our focus back to core values like discussing practical importance of effect sizes, minimizing measurement error, performing replication studies, and using informed personal judgment (Cumming 2014; Gigerenzer & Marewski 2015; Lemoine et al. 2016).

Conclusions

Part of an apparent crisis of unreplicable research is caused by the way we use statistics for analyzing, interpreting and communicating our data. Applying significance thresholds leads to overconfidence in small but highly variable p-values, to publication bias against larger p-values, and to reported effects that are biased upwards. But larger p-values are to be expected also if there is a true effect, and they must be published because otherwise smaller p-values are uninterpretable. Thus, smaller p-values need to lose reputation, and larger p-values need to gain reputation. This is best accomplished by removing fixed significance thresholds, by cautiously interpreting p-values as graded evidence against the null hypothesis, and by putting more emphasis on interpreting effect sizes and interval estimates, using non-automated informed judgment. We give the last word to Edwin G. Boring (1919), who one century ago wrote in his paper "Mathematical vs. scientific significance": "Conclusions must ultimately be left to the scientific intuition of the experimenter and his public."

Acknowledgements

For helpful comments on the manuscript we thank Daniel Berner, Lilla Lovász, the students and colleagues from our journal clubs, and the three referees.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

References

Academy of Medical Sciences 2015. *Reproducibility and reliability of biomedical research: improving research practice. Symposium report*. Academy of Medical Sciences, BBSRC, MRC, Wellcome Trust.

Anderson DR, Burnham KP, Thompson WL. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* 64:912-923. 10.2307/3803199

Badenes-Ribera L, Frias-Navarro D, Iotti B, Bonilla-Campos A, Longobardi C. 2016. Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology* 7:1247. 10.3389/fpsyg.2016.01247

Baker M. 2016. Is there a reproducibility crisis? *Nature* 533:452-454.

Barber JJ, Ogle K. 2014. To P or not to P? *Ecology* 95:621-626. 10.1890/13-1402.1

Bayarri MJ, Benjamin DJ, Berger JO, Sellke TM. 2016. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology* 72:90-103. 10.1016/j.jmp.2015.12.007

Berger JO, Berry DA. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76:159-165.

Berkson J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33:526-536. 10.2307/2279690

Berry DA. 2016. P-values are not what theyre cracked up to be. *The American Statistician, supplemental material to the ASA statement on p-values and statistical significance*. 10.1080/00031305.2016.1154108

Bishop DVM, Thompson PA. 2016. Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ* 4:e1715. 10.7717/peerj.1715

Boos DD, Stefanski LA. 2011. P-value precision and reproducibility. *American Statistician* 65:213-221. 10.1198/tas.2011.10129

Boring EG. 1919. Mathematical vs. scientific significance. *Psychological Bulletin* 16:335-338. 10.1037/h0074554

Box GEP. 1980. Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A* 143:383-430. 10.2307/2982063

Branch M. 2014. Malignant side effects of null-hypothesis significance testing. *Theory & Psychology* 24:256-277. 10.1177/0959354314525282

Brembs B, Button K, Munafo M. 2013. Deep impact: unintended consequences of journal rank. *Frontiers in Human Neuroscience* 7:291. 10.3389/fnhum.2013.00291

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafo MR. 2013a. Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience* 14. 10.1038/nrn3475-c4

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafo MR. 2013b. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14:365-376. 10.1038/nrn3475

Campbell JP. 1982. Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology* 67:691-700.

Capen EC, Clapp RV, Campbell WM. 1971. Competitive bidding in high-risk situations. *Journal of Petroleum Technology* 23:641-653.

Carver RP. 1978. Case against statistical significance testing. *Harvard Educational Review* 48:378-399.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Chavalarias D, Wallach JD, Li AH, Ioannidis JP. 2016. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA-Journal of the American Medical Association* 315:1141-1148. 10.1001/jama.2016.1952

Cohen J. 1962. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Psychology* 65:145-153. 10.1037/h0045186

Cohen J. 1990. Things I have learned (so far). *American Psychologist* 45:1304-1312. 10.1037//0003-066x.45.12.1304

Cohen J. 1994. The earth is round (p < .05). *American Psychologist* 49:997-1003. 10.1037/0003-066x.50.12.1103

Colquhoun D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1:140216. 10.1098/rsos.140216

Coulson M, Healey M, Fidler F, Cumming G. 2010. Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology* 1:26. 10.3389/fpsyg.2010.00026

Cowles M, Davis C. 1982a. Is the .05 level subjectively reasonable? *Canadian Journal of Behavioural Science* 14:248-252. 10.1037/h0081256

Cowles M, Davis C. 1982b. On the origins of the .05 level of statistical significance. *American Psychologist* 37:553-558. 10.1037/0003-066x.37.5.553

Csada RD, James PC, Espie RHM. 1996. The "file drawer problem" of non-significant results: Does it apply to biological research? *Oikos* 76:591-593. 10.2307/3546355

Cumming G. 2007. Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics* 29:89-93.

Cumming G. 2008. Replication and p intervals. *Perspectives on Psychological Science* 3:286-300. 10.1111/j.1745-6924.2008.00079.x

Cumming G. 2012. *Understanding the new statistics*. New York: Routledge.

Cumming G. 2014. The new statistics: why and how. *Psychological Science* 25:7-29. 10.1177/0956797613504966

de Winter JCF, Dodou D. 2015. A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ* 3:e733. 10.7717/peerj.733

Delong JB, Lang K. 1992. Are all economic hypotheses false? *Journal of Political Economy* 100:1257-1272.

Dienes Z. 2014. Using Bayes to get the most out of non-significant results. *Frontiers in Psychology* 5:781. 10.3389/fpsyg.2014.00781

Dixon P. 2003. The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology* 57:189-202. 10.1037/h0087425

Drummond GB. 2016. Most of the time, P is an unreliable marker, so we need no exact cut-off. *British Journal of Anaesthesia* 116:893-893. 10.1093/bja/aew146

Dwan K, Gamble C, Williamson PR, Kirkham JJ, the Reporting Bias Group. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. *PLoS One* 8:e66844. 10.1371/journal.pone.0066844

Edgington ES. 1970. Hypothesis testing without fixed levels of significance. *Journal of Psychology* 76:109-115.

Edwards W, Lindman H, Savage LJ. 1963. Bayesian statistical inference for psychological research. *Psychological Review* 70:193-242. 10.1037/h0044139

Eysenck HJ. 1960. The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review* 67:269-271. 10.1037/h0048412

Fabrigar LR, Wegener DT. 2016. Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology* 66:68-80. 10.1016/j.jesp.2015.07.009

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Fanelli D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90:891-904. 10.1007/s11192-011-0494-7

Fanelli D, Costas R, Ioannidis JPA. 2017. Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America* 114:3714-3719. 10.1073/pnas.1618569114

Ferguson CJ, Heene M. 2012. A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science* 7:555-561. 10.1177/1745691612459059

Fidler F, Burgman MA, Cumming G, Buttrose R, Thomason N. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology* 20:1539-1544. 10.1111/j.1523-1739.2006.00525.x

Fidler F, Geoff C, Mark B, Neil T. 2004. Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics* 33:615-630. 10.1016/j.socec.2004.09.035

Finch S, Cumming G, Thomason N. 2001. Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement* 61:181-210. 10.1177/00131640121971167

Fisher RA. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* 222:309-368.

Fisher RA. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher RA. 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* 33:503-513.

Fisher RA. 1929. The statistical method in psychical research. *Proceedings of the Society for Psychical Research* 39:189-192.

Fisher RA. 1935. Statistical tests. *Nature* 136:474-474.

Fisher RA. 1937. *The design of experiments*. 2nd edition. Edinburgh: Oliver and Boyd.

Fisher R. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 17:69-78.

Fisher RA. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Fisher RA. 1958. *Statistical methods for research workers*. 13th edition. Edinburgh: Oliver and Boyd.

Foreman P, Murnighan JK. 1996. Learning to avoid the winner's curse. *Organizational Behavior and Human Decision Processes* 67:170-180. 10.1006/obhd.1996.0072

Forstmeier W, Schielzeth H. 2011. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* 65:47-55. 10.1007/s00265-010-1038-5

Francis G. 2013. Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology* 57:153-169. 10.1016/j.jmp.2013.02.003

Gallistel CR. 2009. The importance of proving the null. *Psychological Review* 116:439-453. 10.1037/a0015251

Garner C. 2007. Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology* 31:288-295. 10.1002/gepi.20209

Gaudart J, Huiart L, Milligan PJ, Thiebaut R, Giorgi R. 2014. Reproducibility issues in science, is P value really the only answer? *Proceedings of the National Academy of Sciences of the United States of America* 111:E1934-E1934. 10.1073/pnas.1323051111

Gelman A. 2013a. Interrogating p-values. *Journal of Mathematical Psychology* 57:188-189. 10.1016/j.jmp.2013.03.005

Gelman A. 2013b. The problem with p-values is how they're used. Available at http://www.stat.columbia.edu/~gelman (accessed 6 June 2017)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Gelman A. 2015. The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management* 41:632-643. 10.1177/0149206314525208

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*. 10.1080/00031305.2016.1154108

Gelman A, Carlin J. 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9:641-651. 10.1177/1745691614551642

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

Gelman A, Loken E. 2014. The statistical crisis in science. *American Scientist* 102:460-465.

Gelman A, Robert CP. 2014. Revised evidence for statistical standards. *Proceedings of the National Academy of Sciences of the United States of America* 111:E1933-E1933. 10.1073/pnas.1322995111

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

Gerber AS, Malhotra N. 2008. Publication bias in empirical sociological research – Do arbitrary significance levels distort published results? *Sociological Methods & Research* 37:3-30. 10.1177/0049124108318973

Gigerenzer G. 1993. The superego, the ego, and the id in statistical reasoning. In: Hillsdale NJ, ed. *A handbook for data analysis in the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates, 311-339.

Gigerenzer G, Krauss S, Vitouch O. 2004. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In: Kaplan D, ed. *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks: Sage Publications, 391-408.

Gigerenzer G, Marewski JN. 2015. Surrogate science: The idol of a universal method for scientific inference. *Journal of Management* 41:421-440. 10.1177/0149206314547522

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. 1989. *The empire of chance: How probability changed science and everyday life*. New York: Cambridge University Press.

Gill J. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52:647-674. 10.2307/449153

Giner-Sorolla R. 2016. Approaching a fair deal for significance and other concerns. *Journal of Experimental Social Psychology* 65:1-6. 10.1016/j.cjesp.2016.01.010

Goodman SN. 1992. A comment on replication, p-values and evidence. *Statistics in Medicine* 11:875-879. 10.1002/sim.4780110705

Goodman SN. 1993. P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 137:485-496.

Goodman SN. 2016. The next questions: who, what, when, where, and why? *The American Statistician, supplemental material to the ASA statement on p-values and statistical significance*. 10.1080/00031305.2016.1154108

Goodman SN, Fanelli D, Ioannidis JPA. 2016. What does research reproducibility mean? *Science Translational Medicine* 8. 10.1126/scitranslmed.aaf5027

Gorard S. 2016. Damaging real lives through obstinacy: re-emphasising why significance testing is wrong. *Sociological Research Online* 21. 10.5153/sro.3857

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Göring HHH, Terwilliger JD, Blangero J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics* 69:1357-1369. 10.1086/324471

Greenland S. 2012. Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology* 22:364-368. 10.1016/j.annepidem.2012.02.007

Greenland S, Poole C. 2013. Living with statistics in observational research. *Epidemiology* 24:73-78. 10.1097/EDE.0b013e3182785a49

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31:337-350. 10.1007/s10654-016-0149-3

Greenwald AG. 1975. Consequences of prejudice against the null hypothesis. *Psychological Bulletin* 82:1-19. 10.1037/h0076157

Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. 1996. Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology* 33:175-183. 10.1111/j.1469-8986.1996.tb02121.x

Hagen RL. 1997. In praise of the null hypothesis statistical test. *American Psychologist* 52:15-24. 10.1037/0003-066x.52.1.15

Hagen RL. 1998. A further look at wrong reasons to abandon statistical testing. *American Psychologist* 53:801-803. 10.1037/0003-066x.53.7.801

Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. 2015. The fickle P value generates irreproducible results. *Nature Methods* 12:179-185. 10.1038/nmeth.3288

Higginson AD, Munafo MR. 2016. Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology* 14:e2000995. 10.1371/journal.pbio.2000995

Higgs MD. 2013. Do we really need the s-word? *American Scientist* 101:6-9.

Hoekstra R, Finch S, Kiers HAL, Johnson A. 2006. Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review* 13:1033-1037. 10.3758/bf03213921

Hoekstra R, Johnson A, Kiers HAL. 2012. Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement* 72:1039-1052. 10.1177/0013164412450297

Hubbard R, Bayarri MJ. 2003. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *American Statistician* 57:171-178. 10.1198/0003130031856

Hung HMJ, Oneill RT, Bauer P, Kohne K. 1997. The behavior of the p-value when the alternative hypothesis is true. *Biometrics* 53:11-22. 10.2307/2533093

Hurlbert SH, Lombardi CM. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici* 46:311-349.

IntHout J, Ioannidis JPA, Borm GF. 2016. Obtaining evidence by a single well-powered trial or several modestly powered trials. *Statistical Methods in Medical Research* 25:538-552. 10.1177/0962280212461098

Ioannidis JPA. 2005. Contradicted and initially stronger effects in highly cited clinical research. *JAMA-Journal of the American Medical Association* 294:218-228. 10.1001/jama.294.2.218

Ioannidis JPA. 2008. Why most discovered true associations are inflated. *Epidemiology* 19:640-648. 10.1097/EDE.0b013e31818131e7

Ioannidis JPA. 2010. Meta-research: The art of getting it wrong. *Research Synthesis Methods* 1:169-184. 10.1002/jrsm.19

Ioannidis JPA. 2014. How to make more published research true. *PLoS Medicine* 11:e1001747. 10.1371/journal.pmed.1001747

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 383:166-175. 10.1016/s0140-6736(13)62227-8

Jennions MD, Møller AP. 2002. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society B-Biological Sciences* 269:43-48. 10.1098/rspb.2001.1832

Jennions MD, Møller AP. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* 14:438-445. 10.1093/beheco/14.3.438

John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23:524-532. 10.1177/0956797611430953

Johnson DH. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763-772. 10.2307/3802789

Johnson VE. 2013. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America* 110:19313-19317. 10.1073/pnas.1313476110

Johnson VE. 2014. Reply to Gelman, Gaudart, Pericchi: More reasons to revise standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America* 111:E1936-E1937. 10.1073/pnas.1400338111

Kelly CD. 2006. Replicating empirical research in behavioral ecology: How and why it should be done but rarely ever is. *Quarterly Review of Biology* 81:221-236. 10.1086/506236

Kline RB. 2013. *Beyond significance testing: Statistics reform in the behavioral sciences*. Washington: American Psychological Association.

Korner-Nievergelt F, Hüppop O. 2016. Kurze Einführung in Bayes-Statistik mit R für Ornithologen. *Vogelwarte* 54:181-194.

Korner-Nievergelt F, Roth T, von Felten S, Guélat J, Almasi B, Korner-Nievergelt P. 2015. *Bayesian data analysis in ecology using linear models with R, BUGS, and Stan*. London: Academic Press.

Krueger J. 2001. Null hypothesis significance testing – On the survival of a flawed method. *American Psychologist* 56:16-26. 10.1037//0003-066x.56.1.16

Labovitz S. 1968. Criteria for selecting a significance level: A note on the sacredness of .05. *American Sociologist* 3:220-222.

Lai J, Fidler F, Cumming G. 2012. Subjective p intervals – Researchers underestimate the variability of p values over replication. *Methodology* 8:51-62. 10.1027/1614-2241/a000037

Lavine M. 2014. Comment on Murtaugh. *Ecology* 95:642-645. 10.1890/13-1112.1

Lazzeroni LC, Lu Y, Belitskaya-Levy I. 2014. P-values in genomics: Apparent precision masks high uncertainty. *Molecular Psychiatry* 19:1336-1340. 10.1038/mp.2013.184

Lecoutre B, Poitevineau J. 2014. *The significance test controversy revisited*. Heidelberg: Springer.

Lecoutre M-P, Poitevineau J, Lecoutre B. 2003. Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology* 38:37-45. 10.1080/00207590244000250

Leek JT, Jager LR. 2017. Is most published research really false? *Annual Review of Statistics and Its Application* 4:109-122. 10.1146/annurev-statistics-060116-054104

Lehmann EL. 2011. *Fisher, Neyman, and the creation of classical statistics*. New York: Springer.

Lemoine NP, Hoffman A, Felton AJ, Baur L, Chaves F, Gray J, Yu Q, Smith MD. 2016. Underappreciated problems of low replication in ecological field studies. *Ecology* 97:2554-2561. 10.1002/ecy.1506

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Lenhard J. 2006. Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science* 57:69-91. 10.1093/bjps/axi152

Lertzman KP. 1995. Notes on writing papers and theses. *Bulletin of the Ecological Society of America* 76:86-90.

Levine TR, Weber R, Park HS, Hullett CR. 2008. A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research* 34:188-209. 10.1111/j.1468-2958.2008.00318.x

Lew MJ. 2016. Three inferential questions, two types of p-value. *The American Statistician, supplemental material to the ASA statement on p-values and statistical significance*. 10.1080/00031305.2016.1154108

Lisse JR, Perlman M, Johansson G, Shoemaker JR, Schechtman J, Skalky CS, Dixon ME, Polls AB, Mollen AJ, Geba GP. 2003. Gastrointestinal tolerability and effectiveness of rofecoxib versus naproxen in the treatment of osteoarthritis – A randomized, controlled trial. *Annals of Internal Medicine* 139:539-546.

Loftus GR. 1993. A picture is worth 1000 p-values: On the irrelevance of hypothesis-testing in the microcomputer age. *Behavior Research Methods Instruments & Computers* 25:250-256. 10.3758/bf03204506

Lovasich JL, Neyman J, Scott EL, Wells MA. 1971. Hypothetical explanations of negative apparent effects of cloud seeding in whitetop experiment. *Proceedings of the National Academy of Sciences of the United States of America* 68:2643-2646. 10.1073/pnas.68.11.2643

Madden LV, Shah DA, Esker PD. 2015. Does the P value have a future in plant pathology? *Phytopathology* 105:1400-1407. 10.1094/phyto-07-15-0165-le

Maxwell SE, Lau MY, Howard GS. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist* 70:487-498. 10.1037/a0039400

McCarthy MA. 2007. *Bayesian methods for ecology*. Cambridge: Cambridge University Press.

McCormack J, Vandermeer B, Allan GM. 2013. How confidence intervals become confusion intervals. *BMC Medical Research Methodology* 13:134. 10.1186/1471-2288-13-134

McShane BB, Gal D. 2016. Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science* 62:1707-1718. 10.1287/mnsc.2015.2212

McShane BB, Gal D. 2017. Statistical significance and the dichotomization of evidence: The relevance of the ASA statement on statistical significance and p-values for statisticians. *Journal of the American Statistical Association* 112.

Meehl PE. 1967. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34:103-115. 10.1086/288135

Meehl PE. 1990. Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports* 66:195-244. 10.2466/pr0.66.1.195-244

Miller J. 2009. What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review* 16:617-640. 10.3758/pbr.16.4.617

Morey RD, Rouder JN. 2011. Bayes factor approaches for testing interval null hypotheses. *Psychological Methods* 16:406-419. 10.1037/a0024377

Munafo MR, Flint J. 2010. How reliable are scientific studies? *British Journal of Psychiatry* 197:257-258. 10.1192/bjp.bp.109.069849

Munafo MR, Stothart G, Flint J. 2009. Bias in genetic association studies and impact factor. *Molecular Psychiatry* 14:119-120. 10.1038/mp.2008.77

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Mundry R. 2011. Issues in information theory-based statistical inference – A commentary from a frequentist's perspective. *Behavioral Ecology and Sociobiology* 65:57-68. 10.1007/s00265-010-1040-y

Murdoch DJ, Tsai Y-L, Adcock J. 2008. P-values are random variables. *American Statistician* 62:242-245. 10.1198/000313008x332421

Murtaugh PA. 2014a. In defense of P values. *Ecology* 95:611-617. 10.1890/13-0590.1

Murtaugh PA. 2014b. Rejoinder. *Ecology* 95:651-653. 10.1890/13-1858.1

Nakagawa S. 2004. A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology* 15:1044-1045. 10.1093/beheco/arh107

Neyman J. 1977. Frequentist probability and frequentist statistics. *Synthese* 36:97-131. 10.1007/bf00485695

Neyman J, Pearson ES. 1933a. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A* 231:289-337. 10.1098/rsta.1933.0009

Neyman J, Pearson ES. 1933b. The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society* 29:492-510.

Nosek BA, Errington TM. 2017. Making sense of replications. *eLife* 6:e23383. 10.7554/eLife.23383

Nuzzo R. 2015. Fooling ourselves. *Nature* 526:182-185.

Oakes M. 1986. *Statistical inference: Commentary for the social and behavioural sciences*. Chichester: Wiley.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349:aac4716. 10.1126/science.aac4716

Orlitzky M. 2012. How can significance tests be deinstitutionalized? *Organizational Research Methods* 15:199-228. 10.1177/1094428111428356

Parker TH, Forstmeier W, Koricheva J, Fidler F, Hadfield JD, Chee YE, Kelly CD, Gurevitch J, Nakagawa S. 2016. Transparency in ecology and evolution: Real problems, real solutions. *Trends in Ecology & Evolution* 31:711-719. 10.1016/j.tree.2016.07.002

Patil P, Peng RD, Leek JT. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science* 11:539-544. 10.1177/1745691616646366

Pearson ES. 1962. Some thoughts on statistical inference. *Annals of Mathematical Statistics* 33:394-403. 10.1214/aoms/1177704566

Pericchi L, Pereira CAB, Perez M-E. 2014. Adaptive revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America* 111:E1935-E1935. 10.1073/pnas.1322191111

Poitevineau J, Lecoutre B. 2001. Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review* 8:847-850. 10.3758/bf03196227

Pritschet L, Powell D, Horne Z. 2016. Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science* 27:1036-1042. 10.1177/0956797616645672

Reinhart A. 2015. *Statistics done wrong*. San Francisco: No Starch Press.

Rosenthal R. 1979. The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86:638-641. 10.1037//0033-2909.86.3.638

Rosnow RL, Rosenthal R. 1989. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* 44:1276-1284. 10.1037//0003-066x.44.10.1276

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Rothman KJ. 2014. Six persistent research misconceptions. *Journal of General Internal Medicine* 29:1060-1064. 10.1007/s11606-013-2755-z

Rozeboom WW. 1960. The fallacy of the null-hypothesis significance test. *Psychological Bulletin* 57:416-428. 10.1037/h0042040

Sackrowitz H, Samuel-Cahn E. 1999. P values as random variables – Expected P values. *American Statistician* 53:326-331. 10.2307/2686051

Salsburg D. 2001. *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Henry Holt & Co.

Sauley KS, Bedeian AG. 1989. .05: A case of the tail wagging the distribution. *Journal of Management* 15:335-344. 10.1177/014920638901500209

Savalei V, Dunn E. 2015. Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology* 6:245. 10.3389/fpsyg.2015.00245

Savitz DA. 2013. Reconciling theory and practice – What is to be done with P values? *Epidemiology* 24:212-214. 10.1097/EDE.0b013e318281e856

Schatz P, Jay KA, McComb J, McLaughlin JR. 2005. Misuse of statistical tests in Archives of Clinical Neuropsychology publications. *Archives of Clinical Neuropsychology* 20:1053-1059. 10.1016/j.acn.2005.06.006

Schmidt FL. 1992. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist* 47:1173-1181. 10.1037/0003-066x.47.10.1173

Schmidt FL. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1:115-129. 10.1037//1082-989x.1.2.115

Schneider JW. 2015. Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* 102:411-432. 10.1007/s11192-014-1251-5

Sedlmeier P, Gigerenzer G. 1989. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105:309-316. 10.1037//0033-2909.105.2.309

Sellke T, Bayarri MJ, Berger JO. 2001. Calibration of p values for testing precise null hypotheses. *American Statistician* 55:62-71. 10.1198/000313001300339950

Senn S. 2002. A comment on replication, p-values and evidence. *Statistics in Medicine* 21:2437-2444. 10.1002/sim.1072

Sharpe D. 2013. Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods* 18:572-582. 10.1037/a0034177

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22:1359-1366. 10.1177/0956797611417632

Simonsohn U. 2014. Posterior-hacking: Selective reporting invalidates Bayesian results also. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2374040 (accessed 6 June 2017)

Simonsohn U. 2015. Small telescopes: Detectability and the evaluation of replication results. *Psychological Science* 26:559-569. 10.1177/0956797614567341

Siontis KCM, Evangelou E, Ioannidis JPA. 2011. Magnitude of effects in clinical trials published in high-impact general medical journals. *International Journal of Epidemiology* 40:1280-1291. 10.1093/ije/dyr095

Skipper JK, Guenther AL, Nass G. 1967. The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *American Sociologist* 2:16-18.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Smaldino PE, McElreath R. 2016. The natural selection of bad science. *Royal Society Open Science* 3:160384. 10.1098/rsos.160384

Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I. 2010. Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment* 14. 10.3310/hta14080

Stahel WA. 2016. Statistical issues in reproducibility. In: Atmanspacher H, Maasen S, eds. *Reproducibility: Principles, problems, practices, and prospects*: Wiley, 87-114.

Stanley DJ, Spence JR. 2014. Expectations for replications: Are yours realistic? *Perspectives on Psychological Science* 9:305-318. 10.1177/1745691614528518

Sterling TD. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association* 54:30-34. 10.2307/2282137

Sterne JAC, Smith GD. 2001. Sifting the evidence – what's wrong with significance tests? *British Medical Journal* 322:226-231. 10.1136/bmj.322.7280.226

Stoehr AM. 1999. Are significance thresholds appropriate for the study of animal behaviour? *Animal Behaviour* 57:F22-F25. 10.1006/anbe.1998.1016

Thompson B. 1998. In praise of brilliance: Where that praise really belongs. *American Psychologist* 53:799-800. 10.1037//0003-066x.53.7.799

Thompson B. 1999. Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *Journal of Psychology* 133:133-140.

Trafimow D, Marks M. 2015. Editorial. *Basic and Applied Social Psychology* 37:1-2. 10.1080/01973533.2015.1012991

Tryon WW. 2001. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods* 6:371-386. 10.1037//1082-989x.6.4.371

Tukey JW. 1991. The philosophy of multiple comparisons. *Statistical Science* 6:100-116.

van Assen MALM, van Aert RCM, Nuijten MB, Wicherts JM. 2014. Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS One* 9:e84896. 10.1371/journal.pone.0084896

van Helden J. 2016. Confidence intervals are no salvation from the alleged fickleness of the P value. *Nature Methods* 13:605-606. 10.1038/nmeth.3932

Vankov I, Bowers J, Munafo MR. 2014. On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology* 67:1037-1040. 10.1080/17470218.2014.885986

Wasserstein RL, Lazar NA. 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* 70:129-133. 10.1080/00031305.2016.1154108

Weinberg CR. 2001. It's time to rehabilitate the p-value. *Epidemiology* 12:288-290. 10.1097/00001648-200105000-00004

Weiss KM. 2011. The 5% solution – How do we make decisions in science? *Evolutionary Anthropology* 20:81-84. 10.1002/evan20304

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* 75:1182-1189. 10.1111/j.1365-2656.2006.01141.x

Wolf IS. 1961. Perspectives in psychology – XVI. Negative findings. *Psychological Record* 11:91-95.

Young NS, Ioannidis JPA, Al-Ubaydi O. 2008. Why current publication practices may distort science. *PLoS Medicine* 5:e201. 10.1371/journal.pmed.0050201

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Yu EC, Sprenger AM, Thomas RP, Dougherty MR. 2014. When decision heuristics and science collide. *Psychonomic Bulletin & Review* 21:268-282. 10.3758/s13423-013-0495-z

Ziliak ST, McCloskey DN. 2008. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Zöllner S, Pritchard JK. 2007. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *American Journal of Human Genetics* 80:605-615. 10.1086/512821

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2921v2 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017