When null hypothesis significance testing is unsuitable for research: a reassessment

Denes Szucs*[1], John PA Ioannidis[2]

[1]Department of Psychology, University of Cambridge, UK

[2]Meta-Research Innovation Center at Stanford (METRICS) and Department of Medicine, Department of Health Research and Policy, and Department of Statistics, Stanford University, Stanford, USA

* Correspondence: Denes Szucs; ds377@cam.ac.uk

Abstract: 151

Null hypothesis significance testing (NHST) has several shortcomings that are likely contributing factors behind the widely debated replication crisis of psychology, cognitive neuroscience and biomedical science in general. We review these shortcomings and suggest that, after about 60 years of negative experience, NHST should no longer be the default, dominant statistical practice of all biomedical and psychological research. Different inferential methods (NHST, likelihood estimation, Bayesian methods, false-discovery rate control) may be most suitable for different types of research questions. Whenever researchers use NHST they should justify its use, and publish pre-study power calculations and effect sizes, including negative findings. Studies should optimally be pre-registered and raw data published. The current statistics lite educational approach for students that has sustained the widespread, spurious use of NHST should be phased out. Instead, we should encourage either more in-depth statistical training of more researchers and/or more widespread involvement of professional statisticians in all research.

Five Keywords: replication crisis, false positive findings, null hypothesis significance testing

'What used to be called judgment is now called prejudice and what used to be called prejudice is now called a null hypothesis. In the social sciences, particularly, it is dangerous nonsense (dressed up as the 'scientific method') and will cause much trouble before it is widely appreciated as such.' (Edwards, 1972; p.180.)

*'...the mathematical rules of probability theory are not merely rules for calculating frequencies of random variables; they are also the unique consistent rules for conducting inference (ie. plausible reasoning)' (*Jaynes, 2003; p.xxii)

1. The replication crisis and Null Hypothesis Significance Testing (NHST)

There is increasing discontent that many areas of psychological science, cognitive neuroscience, and biomedical research (Ioannidis 2005; Ioannidis et al. 2014) are in a crisis of producing too many false positive non-replicable results (Nosek et al. 2015). This wastes research funding, erodes credibility and slows down scientific progress. Since more than half a century many methodologists have claimed repeatedly that this crisis may at least in part be related to problems with Null Hypothesis Significance Testing (NHST) (Rozeboom 1960; Bakan 1966; Meehl 1978; Gigerezner 1998; Nickerson 2000). However, most scientists (and in particular psychologists, biomedical scientists, social scientists, cognitive scientists and neuroscientists) are still near exclusively educated in NHST, they tend to misunderstand and abuse NHST and the method is near fully dominant in scientific papers (Chavalarias, Wallach and Ioannidis, 2016). Here we provide an accessible critical reassessment of NHST and suggest that while it may have some legitimate uses NHST should be abandoned as the de factor *cornerstone* of research.

2. The origins of NHST as a weak heuristic and a decision rule

2.1 NHST as a weak heuristic based on the p value: Fisher

p values were widely popularized by Fisher (1925). In the context of the current NHST approach Fisher *only* relied on the concepts of the null hypothesis (H_0) and the *exact p value*

(hereafter p will refer to the p value and 'pr' to probability; see **Appendix 1** for terms). He thought that experiments should aim to reject (or 'nullify'; henceforth the name 'null hypothesis') H_0 which assumes that the data demonstrates random variability according to some distribution around a certain value. Discrepancy from H_0 is measured by a test statistic whose values can be paired with one or two-tailed p values which tell us how likely it is that we would have found our data *or* more extreme data if H_0 was really correct. Formally we will refer to the p value as: pr(data or more extreme data | H_0). It is important to realize that the p value represents the 'extremeness' of the data according to an imaginary data distribution assuming there is no bias in data sampling.

The late Fisher viewed the *exact* p value as a *heuristic piece of inductive evidence* which gives an indication of the plausibility of H_0 together with other available evidence, like effect sizes (see Gigerenzer et al. 2004; Hubbard and Bayarri, 2003). Fisher recommended that H_0 can usually be rejected if $p \leq 0.05$ but in his system there is no mathematical justification for selecting a particular p value for the rejection of H_0. Rather, this is up to the substantively informed judgment of the experimenter. Fisher thought that a hypothesis is demonstrable only when properly designed experiments *'rarely fail'* to give us statistically significant results (Gigerenzer et al. 1989, p96; Goodman, 2008). Hence, a single significant result should not represent a 'scientific fact' but should merely draw attention to a phenomenon which seems worthy of further investigation including replication (Goodman 2008). In contrast to the above, until recently replication studies have been very rare in many scientific fields; lack of replication efforts has been a particular problem in the psychological sciences (Makel, Plucker and Hegarty, 2012), but this may hopefully change with the wide attention that replication has received (Nosek et al. 2015).

2.2 Neyman and Pearson: a decision mechanism optimized for the long-run

The concepts of the alternative hypothesis (H_1), α, power, β, Type I and Type II errors

were introduced by Neyman and Pearson (Neyman and Pearson, 1933; Neyman 1950) who set up a formal decision procedure motivated by industrial quality control problems (Gigerenzer et al. 1989). Their approach aimed to minimize the false negative (Type II) error rate to an acceptable level (β) and consequently to maximize power (1-β) *subject* to a bound (α) on false positive (Type I) errors (Hubbard and Bayarri, 2003). α can be set by the experimenter to an arbitrary value and Type-II error can be controlled by setting the sample size so that the required effect size can be detected (see **Appendix 2** for illustration). In contrast to Fisher, this framework does not use the p value as a measure of evidence. We merely determine the critical value of the test statistic associated with α and reject H_0 whenever the test statistic is larger than the critical value. The exact p value is irrelevant because the sole objective of the decision framework is long-run error minimization and only the critical threshold but not the exact p value plays any role in achieving this goal (Hubbard and Bayarri, 2003). Neyman and Pearson rejected the idea of inductive reasoning and offered *a reasoning-free inductive behavioural rule* to choose between two behaviours, accepting or rejecting H_0, irrespective of the researcher's belief about whether H_0 and H_1 are true or not (Neyman and Pearson, 1933).

Crucially, the Neyman-Pearson approach is designed to work efficiently (Neyman and Pearson, 1933) in the context of long-run repeated testing (exact replication). Hence, there is a major difference between the p value which is computed for a *single* data set and α, β, power, Type I and Type II error which are so called *'frequentist'* concepts and they make sense in the context of *a long-run of many repeated experiments*. If we only run a single experiment all we can claim is that if we *had* run a long series of experiments we *would have had* 100α% false positives (Type I error) had H_0 been true and 100β% false negatives (Type II error) had H_1 been true *provided* we got the power calculations right. Note the conditionals.

In the Neyman-Pearson framework optimally setting α and β assures long-term decision-making efficiency in light of our costs and benefits by committing Type I and Type II

errors. However, optimizing α and β is much easier in industrial quality control than in research where often there is no reason to expect a specific effect size associated with H_1 (Gigerenzer et al. 1989). For example, if a factory has to produce screw heads with a diameter of 1±0.01 cm than we know that we have to be able to detect a deviation of 0.01 cm to produce acceptable quality output. In this setting we know exactly the smallest effect size we are interested in (0.01 cm) and we can also control the sample size very efficiently because we can easily take a sample of a large number of screws from a factory producing them by the million assuring ample power. On the one hand, failing to detect too large or too small screws (Type II error) will result in our customers cancelling their orders (or, in other industrial settings companies may deliver faulty cars or exploding laptops to customers exposing themselves to substantial litigation and compensation costs). On the other hand, throwing away false positives (Type I error), i.e. completely good batches of screws which we think are too small or too large, will also cost us a certain amount of money. Hence, we have a very clear scale (monetary value) to weigh the costs and benefits of both types of errors and we can settle on some rationally justified values of α and β so as to minimize our expenses and maximize our profit.

In contrast to such industrial settings, controlling the sample size and effect size and setting rational α and β levels is not that straightforward in most research settings where the effect sizes being pursued are largely unknown and deciding about the requested size of a good enough effect can be very subjective. For example, what is the smallest difference of interest between two participant groups in a measure of 'fMRI activity'? Or, what is the smallest difference of interest between two groups of participants when we measure their IQ or reaction time? And, even if we have some expectations about the 'true effect size', can we test enough participants to ensure a small enough β? Further, what is the cost of falsely claiming that a vaccine causes autism thereby generating press coverage that grossly misleads the public (Godlee, 2011; Deer 2011)? What is the cost of running too many underpowered studies

thereby wasting perhaps most research funding, boosting the number of false positive papers and complicating interpretation (Schmidt, 1992; Ioannidis 2005; Button et al. 2013)? More often than not researchers do not know the 'true' size of an effect they are interested in, so they cannot assure adequate sample size and it is also hard to estimate general costs and benefits of having particular α and β values. While some "rules of thumb" exist about what are small, modest, and large effects (e.g. Cohen, 1962; Cohen, 1988; Sedlmeier and Gigerenzer, 1989; Jaeschke et al. 1989), some large effects may not be actionable (e.g. a change in some biomarker that is a poor surrogate and thus bears little relationship to major, clinical outcomes), while some small effects may be important and may change our decision (e.g. most survival benefits with effective drugs are likely to be small, but still actionable).

Given the above ambiguity, researchers fall back to the default α=0.05 level with usually undefined power, these unjustified α and β levels completely discredit the originally intended *'efficiency'* rationale of the creators of the Neyman-Pearson decision mechanism (Neyman and Pearson, 1933).

2.4. NHST in its current form

The current NHST merged the above concepts and is often applied stereotypically as a 'mindless null ritual' (Gigerenzer, 2004). Researchers set H_0 nearly always 'predicting' zero effect but do not quantitatively define H_1. Hence, pre-experimental power cannot be calculated for most tests which is a crucial omission in the Neyman-Pearson framework. Researchers compute the *exact* p value as Fisher did but also *mechanistically* reject H_0 and accept the undefined H_1 if $p \leq (\alpha=0.05)$ without flexibility following the *behavioural decision rule* of Neyman and Pearson. As soon as $p \leq \alpha$, findings have the supposed right to become a scientific fact defying the exact replication demands of Fisher and the belief neutral approach of Neyman and Pearson. Researchers also interpret the *exact p* value and use it as a relative *measure of evidence* against H_0, as Fisher did. A *'highly significant'* result with a small p value is perceived

as much stronger evidence than a weakly significant one. However, while Fisher was conscious of the weak nature of the evidence provided by the p value (Wasserstein and Lazar, 2016), generations of scientists encouraged by incorrect editorial interpretations (Bakan 1966) started to exclusively rely on the p value in their decisions even if this meant neglecting their substantive knowledge: scientific conclusions *merged* with reading the p value (Goodman, 1999).

3. Neglecting the full context of NHST leads to confusions about the p value

Most textbooks illustrate NHST by partial 2×2 tables (see **Table 1**) which fail *t*o contextualize long-run conditional probabilities and fail to clearly distinguish between long-run probabilities and the p value which is computed for a single data set (Pollard and Richardson, 1987). This leads to major confusions about the meaning of the p value (see **Box 1**).

First, both H_0 and H_1 have some pre-study or 'prior' probabilities, pr(H_0) and pr(H_1). This means that before the study is run we may have some knowledge about the validity of H_0 and H_1. For example, we may know about a single published study claiming to demonstrate H_1 by showing a difference between appropriate experimental conditions. However, in conferences we may have also heard about 9 highly powered but failed replication attempts very similar to the original study. In this case we may assume that the odds of H_0:H_1 are 9:1, that is, pr(H_1) is 1/10. Of course, these pre-study odds are usually hard to judge unless we demand to see our colleagues' 'null results' hidden in their drawers because of the practice of not publishing negative findings. Current scientific practices appreciate the single published 'positive' study more than the 9 unpublished negative ones perhaps because NHST logic only allows for rejecting H_0 but does not allow for accepting it *and* because researchers *erroneously* often think that the single published positive study has a very small, acceptable error rate of providing false positive statistically significant results which equals α, or the p value. So, they

often spuriously assume that the negative studies somehow lacked the sensitivity to show an effect while the single positive study is perceived as a well-executed sensitive experiment delivering a 'conclusive' verdict rather than being a 'lucky' false positive (Bakan, 1966).

NHST completely neglects the above mentioned pre-study information and exclusively deals with rows 2-4 of **Table 1**. NHST computes the one or two-tailed p value for a particular data set assuming that H_0 is true. Additionally, NHST logic takes long-run error probabilities (α and β) into account conditional on H_0 and H_1 These long-run probabilities are represented in typical 2x2 NHST contingency tables but note that β is usually unknown in real studies.

As we have seen, NHST *never* computes the probability of H_0 and H_1 being true or false, all we have is a decision mechanism hoping for the best individual decision in view of long-run Type I and Type II error expectations. Nevertheless, following the repeated testing logic of the NHST framework, for many experiments we can denote the *long-run probability* of H_0 being true given a statistically significant result as False Report Probability (FRP), and the *long-run probability* of H_1 being true given a statistically significant result as True Report Probability (TRP). FRP and TRP are represented in **Row 5 of Table 1** and it is important to see that they refer to completely *different conditional probabilities than* the p value.

Simply put, the p value is pretty much the only thing that NHST computes but scientists usually would like to know the probability of their theory being true or false in light of their data (Jaynes, 2003; Pollard and Richardson, 1987; Goodman 1993). That is, researchers are interested in the post-experimental probability of H_0 and H_1. Most probably, for the reason that researchers do not get what they really want to see and the only parameter NHST computes is the p value it is well-documented (Oakes, 1986; Gliner et al. 2002; Wilkerson and Olson, 2010; Hoekstra et al. 2014; Castro-Sotos 2007; 2009) that many, if not most researchers confuse FRP with the p value or α and they also confuse the complement of p value (1-p) or α (1-α) with TRP (Pollard and Richardson, 1987; Cohen 1994). These confusions are of major portend because

the difference between these completely different parameters is not minor, they can differ by orders of magnitude, the long-run FRP being much larger than the p value under realistic conditions (Sellke et al. 2001; Ioannidis 2005). The complete misunderstanding of the probability of producing false positive findings is most probably a key factor behind vastly inflated confidence in research findings and we suggest that this inflated confidence is an important contributor to the current replication crisis in biomedical science and psychology.

3.1 Serious underestimation of the proportion of false positive findings in NHST

Ioannidis (2005) has shown that most published research findings relying on NHST are likely to be false. The modelling supporting this claim refers to the long-run FRP and TRP which we can compute by applying Bayes' theorem (see computational details and illustrations in **Appendix 3**). The calculations must consider α, the power (1-β) of the statistical test used, the pre-study probabilities of H_0 and H_1, and it is also insightful to consider bias (Berger 1985; Berger and Sellke, 1987; Berger and Delampady, 1987; Sellke, Bayari and Berger, 2001; Pollard and Richardson, 1987; Lindley 1993; Sterne and Smith, 2001; Ioannidis 2005).

While NHST neglects the pre-study odds of H_0 and H_1, these are crucial to take into account when calculating FRP and TRP. For example, let's assume that we run 200 experiments and in 100 studies our experimental ideas are wrong (that is, we test true H_0 situations) while in 100 studies our ideas are correct (that is, we test true H_1 situations). Let's also assume that the power (1-β) of our statistical test is 0.6 and α = 0.05. In this case in 100 studies (true H_0) we will have 5% of results significant by chance alone and in the other 100 studies (true H_1) 60% of studies will come up significant. FRP is the ratio of false positive studies to all studies which come up significant:

$$FRP = \frac{False\ positives}{All\ statistically\ significant\ results}$$

$$= \frac{5\%\ of\ 100\ studies}{5\%\ of\ 100\ studies + 60\%\ of\ 100\ studies} = \frac{5}{5\ +\ 60} = \frac{5}{65} = 0.0769$$

That is, we will have 5 false positives out of a total of 65 statistically significant outcomes which means that the proportion of false positive studies amongst all statistically significant results is 7.69%, higher than the usually assumed 5%. However, this example still assumes that we get every second hypothesis right. If we are not as lucky and only get every sixth hypothesis right then if we run 600 studies, 500 of them will have true H_0 true situations and 100 of them will have true H_1 situations. Hence, the computation will look like:

$$FRP = \frac{False\ positives}{All\ statistically\ significant\ results}$$

$$= \frac{5\%\ of\ 500\ studies}{5\%\ of\ 500\ studies + 60\%\ of\ 100\ studies} = \frac{25}{25\ +\ 60} = \frac{25}{85}$$

$$= 0.2941$$

Hence, nearly 1/3 of all statistically significant findings will be false positives irrespective of the p value. Of course, estimating pre-study odds is difficult, primarily due to the lack of publishing negative findings and to the lack of proper documentation of experimenter intentions before an experiment is run: We do not know what percent of the published statistically significant findings are lucky false positives explained post-hoc when in fact researchers could not detect the originally hypothesized effect. However, it is reasonable to assume that only the most risk avoidant studies have lower H_0:H_1 odds than 1, relatively conservative studies have low to moderate H_0:H_1 odds (1-10) while H_0:H_1 odds can be much higher in explorative research (50-100 or even higher) (Ioannidis, 2005).

Bias is another important determinant of FRP and TRP (Ioannidis 2005). Whenever H_0 is not rejected findings have far more difficulty to be published and the researcher may feel that she wasted her efforts. Further, positive findings are more likely to get cited than negative findings (Kivimäki et al. 2014; Jannot et al. 2013; Kjaergard and Gluud, 2002). Consequently, researchers may often be highly biased to reject H_0 and publish positive findings. Researcher bias affects FRP even if our NHST decision criteria, α and β, are formally unchanged.

Ioannidis (2005) introduced the *u* bias parameter. The impact of u is that after some data tweaking and selective reporting (see **4.6**) u fraction of otherwise non-significant true H_0 results will be reported as significant and u faction of otherwise non-significant true H_1 results will be reported as significant. If u increases, FRP increases and TRP decreases. For example, if $\alpha = 0.05$, power = 0.6 and H_0:H_1 odds = 1 then a 10% bias (u = 0.1) will raise FRP to 18.47%. A 20% bias will raise FRP to 26.09%. If H_0:H_1 odds = 6 then FRP will be 67.92%. Looking at these numbers the replication crisis does not seem surprising: using NHST very high FRP can be expected even with modestly high H_0:H_1 odds and moderate bias (Etz and Vanderckhove, 2016). Hence, under realistic conditions FRP not only *extremely rarely* equals α or the p value (and TRP extremely rarely equals 1-α and/or 1-p value) but also, FRP is *much* larger than the generally assumed 5% and TRP is much lower than the generally assumed 95%. Overall, α or the p value practically says nothing about the likelihood of our research findings being true or false.

3.2 The neglect of power reinterpreted

In contrast to the importance of power in determining FRP and TRP, NHST studies tend to ignore power and β and emphasize α and low p values. Often, finding a statistically significant effect erroneously seems to override the importance of power. However, statistical significance does not protect us from false positives. FRP can only be minimized by keeping H_0:H_1 odds and bias low and power high (Button et al. 2013; Pollard and Richardson, 1987; Bayarri et al. 2016). Hence, power is not only important so that we increase our chances to detect true effects but it is also crucial in keeping FRP low. While power in principle can be adjusted easily by increasing sample size, power in many/most fields of biomedical science and psychology has been notoriously low and the situation has not improved much during the past 50 years (Button et al. 2013; Cohen 1962; Sedlmeier and Gigerenzer, 1989; Rossi, 1990; Hallahan and Rosenthal, 1996). Clearly, besides making sure that research funding is not

wasted, minimizing FRP also provides very strong rationale for increasing the typically used sample sizes in studies.

4. NHST logic is incomplete

4.1 NHST misleads because it neglects pre-data probabilities

Besides creating conceptual confusion and generating misleading inferences especially in the setting of weak power, NHST has further serious problems. NHST logic is based on the so-called *modus tollens* (denying the consequent) argumentation: It sets up a H_0 model and assumes that if the data fits this model than the test statistic associated with the data should not take more extreme values than a certain threshold (Meehl, 1967; Pollard and Richardson, 1987). If the test statistic contradicts this expectation then NHST assumes that H_0 can be rejected and consequently its complement, H_1 can be accepted. While this logic may be able to minimize Type I error in well-powered high-quality well-controlled tests (**2.2**), it is inadequate if we use it to decide about the truth of H_1 in a single experiment, because there is always space for Type I and Type II error (Falk and Greenbaum, 1995). So, our conclusion is never certain and the only way to see how much error we have is to calculate the long-run FRP and TRP using appropriate α and power levels and prior H_0:H_1 odds. The outcome of the calculation can easily conflict with NHST decisions (see **Appendix 4**).

4.2 NHST neglects predictions under H_1 facilitating sloppy research

NHST does not require us to specify exactly what data H_1 would predict. Whereas the Neyman-Pearson approach requires researchers to specify an effect size associated with H_1 and compute power (1-β), in practice this is easy to *neglect* because the NHST machinery only computes the p value conditioned on H_0 and it is able to provide this result even if H_1 is not specified at all. A widespread *misconception* flowing from the fuzzy attitude of NHST to H_1 is that rejecting H_0 allows for accepting a *specific* H_1 (Nickerson 2000). This is what most practicing researchers do in practice when they reject H_0 and argue for their specific H_1 in turn.

However, NHST only computes probabilities conditional on H_0 and it does not allow for the acceptance of either H_0, a specific H_1 or a generic H_1. Rather, it only allows for the rejection of H_0. Hence, if we reject H_0 we will have no idea about how well our data fits a specific H_1. This cavalier attitude to H_1 can easily lead us astray even when contrasting H_0 just with a single alternative hypothesis as illustrated by the invalid inference based on NHST logic in **Table 2** (Pollard and Richardson, 1987).

Our model says that if H_0 is true, it is a *very rare* event that Harold is a member of congress. This rare event then happens which is equivalent to finding a small p value. Hence, we conclude that H_0 can be rejected and H_1 is accepted. However, if we carefully explicate all probabilities it is easy to see that we are being mislead by NHST logic. First, because we have absolutely no idea about Harold's nationality we can set pre-data probabilities of both H_1 and H_0 to 1/2, which means that H_0:H_1 odds are uninformative, 1:1. Then we can explicate the important conditional probabilities of the data (Harold *is* a member of congress) given the possible hypotheses. We can assign arbitrary but plausible probabilities:

$$pr(data|H_0) = pr(\text{Harold is member of congress} \mid \text{American}) = 10^{-7}$$

$$pr(data|H_1) = pr(\text{Harold is member of congress} \mid \text{not American}) = 0$$

That is, while the data is indeed rare under H_0, its probability is actually zero under H_1 (in other words, the data is very unlikely under both the null and the alternative models). So, even if $p \approx 0.0000001$, it does not make sense to reject H_0 and accept H_1 because this data just cannot happen if H_1 is true. If we only have these two hypotheses to choose from then it only makes sense to accept H_0 because the data is still possible under H_0 (Jaynes, 2003). In fact, using Bayes' theorem we can formally show that the probability of H_0 is actually 1 (**Appendix 5**).

In most real world problems multiple alternative hypotheses compete to explain the data. However, by using NHST we can only reject H_0 and argue for *some* H_1 without any formal justification of why we prefer a particular hypothesis whereas it can be argued that it

only makes sense to reject any hypothesis if another one better fits the data (Jaynes, 2003). We only have qualitative arguments to accept a specific H_1 and the exclusive focus on H_0 makes unjustified inference too easy. For example, if we assume that H_0 predicts normally distributed data with mean 0 and standard deviation 1 then we have endless options to pick H_1 (Hubbard and Bayarri, 2003): Does H_1 imply that the data have a mean other than zero, the standard deviation other than 1 and/or does it represent non-normally distributed data? NHST allows us to consider any of these options *implicitly* and then accept one of them post-hoc without any quantitative justification of why we chose that particular option. Further, merging all alternative hypotheses into a single H_1 is not only too simplistic for most real world problems but it also poses an 'inferential double standard' (Rozeboom, 1960): The procedure pits the well-defined H_0 against a potentially infinite number of alternatives.

Vague H_1 definitions (the lack of quantitative predictions) enable researchers to avoid the falsification of their favourite hypotheses by intricately redefining them (especially in fields such as psychology and cognitive neuroscience where theoretical constructs are often vaguely defined) and ever providing any definitive assessment of the plausibility of a favourite hypothesis in light of credible alternatives (Meehl, 1967). This problem is reflected in papers aiming at the mere demonstration of often little motivated significant differences between conditions (Giere, 1972) and post-hoc explanations of likely unexpected but statistically significant findings. For example, neuroimaging studies often attempt to explain why an fMRI BOLD signal ‘deactivation’ happened instead of a potentially more reasonable looking ‘activation’ (or, vice versa). Most such findings may be the consequence of the data randomly deviating into the wrong direction relative to zero between-condition difference. Even multiple testing correction will not help such studies as they still rely on standard NHST just with adjusted α thresholds. Similarly, patient studies often try to explain an unexpected difference between patient and control groups (e.g. the patient group is ‘better’ on a measure) by some

kind of 'compensatory mechanism'. In such cases what happens is that *'the burden of inference has been delegated to the statistical test'*, indeed, and simply because $p \leq \alpha$ odd looking observations and claims are to be trusted as scientific facts (Bakan, 1966, p423; Lykken 1968).

Finally, paradoxically, when we achieve our goal and successfully reject H_0 we will actually be left in complete existential vacuum because during the rejection of H_0 NHST *'saws off its own limb'* (Jaynes, 2003; p524): If we manage to reject H_0 then it follows that pr(data or more extreme data|H_0) is useless because H_0 is not true. Thus, we are left with nothing to characterize the probability of our data in the real world; we will not know pr(data|H_1) for example, because H_1 is formally undefined and NHST never tells us anything about it. In light of these problems Jaynes (2003) suggested that the NHST framework addresses an ill-posed problem and provides invalid responses to questions of statistical inference.

4.4 The p value may exaggerate evidence against H_0

The definition of the p value as pr(data or more extreme data | H_0) is only justified informally by claiming that the p value is a measure of the 'surprise' felt when a rare event happens (Berger and Delampady, 1987). However, as we have seen our surprise at a rare event does not guarantee that H_0 can be rejected: it may be surprising to find a member of Congress but this does not make him/her less likely to be American. Second, it can be shown formally that the definition of the p value does exaggerate the evidence against H_0 by about one order of magnitude which greatly biases NHST procedures towards the rejection of H_0 (see Berger and Sellke, 1987; Berger and Delamdapy, 1987; Goodman, 1993).

4.5 NHST is unsuitable for large datasets

In consequence of the recent 'big data' revolution access to large databases has increased dramatically potentially increasing power tremendously. However, NHST leads to worse inference with large databases than with smaller ones (Meehl, 1967; Khoury and Ioannidis, 2014). This is due to how NHST tests statistics are computed, the properties of real

data and to the lack of specifying data predicted by H_1 (Bruns and Ioannidis, 2016).

Most NHST studies rely on nil null hypothesis testing which means that H_0 expects a true mean difference of exactly zero between conditions with some variation around this true zero mean. Further, NHST machinery guarantees that we can detect any tiny irrelevant effect sizes if sample size is large enough. This is because test statistics are typically computed as the ratio of the relevant between condition differences and associated variability of the data weighted by some function of the sample size (difference/variability × f(sample size)). The p value is smaller if the test statistic is larger. Thus, the larger is the difference between conditions and/or the smaller is variability and/or the larger is the sample size the larger is the test statistic and the smaller is the p value. Consequently, by increasing sample size enough it is guaranteed that H_0 can be rejected even with miniature effect sizes (Ziliak et al. 2008).

Parameters of many real data sets are much more likely to differ than to be the same for reasons completely unrelated to our hypotheses (Edwards, 1972; Meehl, 1967; 1990). First, many psychological, social and biomedical phenomena are extremely complex reflecting the contribution of very large numbers of interacting (latent) factors, let it be at the level of society, personality or heavily networked brain function or other biological networks (Lykken 1968; Gelman, 2014). Hence, if we select any two variables related to these complex networks most probably there will be some kind of at least remote connection between them. This phenomenon is called ‘crud factor’ Meehl (1990) or ‘ambient correlational noise’ (Lykken, 1968) and it is unlikely to reflect a causal relationship. In fact some types of variables, such as intake of various nutrients and other environmental exposures are very frequently correlated among themselves and with various disease outcomes without this meaning that they have anything to do with causing disease outcomes (Patel and Ioannidis, 2014a,b). Second, unlike in physical sciences it is near impossible to control for the relationship of all irrelevant variables which are correlated with the variable(s) of interest (Rozeboom 1960; Lykken 1968).

Consequently, there can easily be a small effect linking two randomly picked variables even if their statistical connection merely communicates that they are part of a vast complex interconnected network of variables. Only a few of these tiny effects are likely to be causal and of any portend (Siontis and Ioannidis, 2011).

The above issues have been demonstrated empirically and by simulations. For example, Bakan (1966; see also Meehl, 1967; Nunally, 1960; Berkson, 1938) subdivided the data of 60,000 persons according to completely arbitrary criteria, like living east or west of the Mississippi river, living in the north or south of the USA, etc. and found all tests coming up statistically significant. Waller (2004) examined the personality questionnaire data of 81,000 individuals to see how many randomly chosen directional null hypotheses can be rejected. If sample size is large enough, 50% of directional hypothesis tests should be significant irrespective of the hypothesis. As expected, nearly half (46%) of Waller's (2004) results were significant. Simulations suggest that in the presence of even tiny residual confounding (e.g. some omitted variable bias) or other bias, large observational studies of null effects will generate results that may be mistaken as revealing thousands of true relationships (Bruns and Ioannidis, 2016). Experimental studies may also suffer the same problem, if they have even minimal biases.

Due to the combination of the above properties of some psychological data sets and statistical machinery theory testing radically differs in sciences with exact and non-exact quantitative predictions (Meehl, 1967). In physical sciences increased measurement precision and increased amounts of data increase the difficulties a theory must pass before it is accepted. This is because theoretical predictions are well defined, numerically precise and it is also easier to control measurements (Lykken, 1968). That is, a theory may predict that a quantity should exactly be let’s say 5 and the experimental setup can assure that really only very few factors

influence measurements - these factors can then be taken into account during analysis. Hence, increased measurement precision will make it easier to demonstrate a departure from numerically exact predictions. So, a 'five sigma' deviation rule may make good sense in physics where precise models are giving precise predictions about variables.

In sciences using NHST without clear numerical predictions the situation is the opposite of the above, because NHST does not demand the exact specification of H_1, so theories typically only predict a fairly vague *'difference'* between groups or experimental conditions rather than an exact numerical discrepancy between measures of groups or conditions. However, as noted, groups are actually likely to differ and if sample size increases and variability in data decreases it will become easier and easier to reject any kind of H_0 when following the NHST approach. In fact, with precise enough measurements and large enough sample size H_0 is guaranteed to be rejected on the long run even if the underlying processes generating the data in two experimental conditions are exactly the same. Hence, ultimately any H_1 can be accepted, claiming support for any kind of theory. For example, in an amusing demonstration Carver (1993) used Analysis of Variance to re-analyze the data of Michelson and Morley (1887) who suggested that the speed of light is constant (H_0) thereby providing the empirical basis for Einstein's theory of relativity. Carver (1993) found that that the speed of light was actually not constant at $p<0.001$. The catch? The effect size as measured by Eta^2 was 0.005. While some may feel that Einstein's theory has now been falsified, perhaps it is also worth considering that here the statistically significant result is essentially insignificant.

A typical defence of NHST may be that we actually may not want to increase power endlessly, just as much as we still think that it allows us to detect reasonable effect sizes (Giere, 1972). A more reasoned approach may be to consider explicitly what the consequences ("costs") are of a false-positive, true-positive, false-negative, and true-negative result. Explicit modelling can suggest that the optimal combination of Type 1 error and power may need to be

different depending on what these assumed costs are (Ioannidis, Hozo and Djulbegovic, 2014). Different fields may need to operate at different optimal ratios of false-positives to false-negatives (Ioannidis, Tarone and McLaughlin, 2011).

4.6 NHST may foster selective reporting

Because NHST never evaluates H_1 formally and it is fairly biased towards the rejection of H_0, reporting bias against H_0 can easily infiltrate the literature even if formal NHST parameters are fixed (see about the 'u' bias parameter in **3.1;**). Overall, a long series of exploratory tools and questionable research practices are utilized in search for statistical significance (Johns 2012, Ioannidis and Trikalinos, 2007). Researchers can influence their data during undocumented analysis and pre-processing steps and by the mere choice of structuring the data (constituting *researcher degrees of freedom*; Simmons et al. 2011). This is particularly a problem in neuroimaging where the complexity and idiosyncrasy of analyses is such that it is usually impossible to replicate exactly what happened and why during data analysis (Carp 2012; Vul et al. 2009; Kriegeskorte et al. 2009). Another term that has been used to describe the impact of diverse analytical choices is "vibration of effects" (Ioannidis, 2008). Different analytical options, e.g. choice of adjusting covariates in a regression model can result in a cloud of results, instead of a single result, and this may entice investigators to select a specific result that is formally significant, while most analytical options would give non-significant results or even results with effects in the opposite direction ('Janus effect'; Patel, Burford and Ioannidis, 2015). Another common mechanism that may generate biased results with NHST is when investigators continue data collection and re-analyse the accumulated data sequentially without accounting for the penalty induced by this repeated testing (DeMets and Lan, 1994; Goodman 1999). The unplanned testing is usually undocumented and researchers may not even be conscious that it exposes them to Type I error accumulation. Bias may be the key explanation why in most biomedical and social science disciplines, the vast majority of published papers

with empirical data report statistically significant results (Fanneli 2010; Kavvoura et al. PLoS Med 2007; Chavalarias et al. 2016).

4.7 The rejection of H_0 is guaranteed on the long-run

If H_0 is false, with $\alpha = 0.05$, 5% of our tests will be statistically significant on the long-run. The riskier experiments we run, the larger are H_0:H_1 odds and bias and the larger is the long-run FRP (**3.3**). Coupled with the fact that a large number of unplanned tests may be run in each study and that negative results and failed replications are often not published, this leads to *'unchallenged fallacies'* clogging up the research literature (Ioannidis, 2012; p1; Bakan 1966; Sterling 1959; Sterling et al. 1995). Moreover, such published false positive true H_0 studies will also inevitably overestimate the effect size of the non-existent effects or of existent, but unimportantly tiny, effects (Schmidt, 1992; 1996; Sterling et al. 1995; Ioannidis 2008). These effects may even be confirmed by meta-analyses, because meta-analyses typically are not able to incorporate unpublished negative results (Sterling et al. 1995) and they cannot correct many of the biases that have infiltrated the primary studies.

Given that the predictions of H_1 are rarely precise and that theoretical constructs in many scientific fields (including psychology and cognitive neuroscience) are often poorly defined, it is easy to claim support for a popular theory with many kinds of data falsifying H_0 even if the constructs measured in many papers are just very weakly linked to the original paper, or not linked at all. Overall, the literature may soon give the impression of a steady stream of replications throughout many years. Even when "negative" results appear, citation bias may still continue to distort the literature and the prevailing theory may continue to be based on the "positive" results. Hence, citation bias may maintain prevailing theories even when they are clearly false and unfounded (Greenberg, 2009).

4.8 NHST does not facilitate systematic knowledge integration

Due to high FRP the contemporary research literature provides statistically significant

'evidence' for nearly everything (Schoenfeld and Ioannidis, 2012). Because NHST emphasizes all or none p value based decisions rather than the magnitude of effects, often only p values are reported for critical tests, effect size reports are often missing and interval estimates and confidence intervals are not reported. In an assessment of the entire biomedical literature in 1990-2015, 96% of the papers that used abstracts reported at least some p value below 0.05, while only 4% of a random sample of papers presented consistently effect sizes with confidence intervals (Chalalarias, Wallach and Ioannidis, 2016). However, oddly enough, the main NHST 'measure of evidence', the p value cannot be compared across studies. It is a frequent *misconception* that a lower p value always means stronger evidence irrespective of the sample size and effect size (Oakes, 1986; Schmidt, 1996; Nickerson, 2000). Besides the non-comparable p values, NHST does not offer any *formal* mechanism for systematic knowledge accumulation and integration (Schmidt, 1996) unlike Bayesian methods which can take such pre-study information into account. Hence, we end up with many fragmented studies which are most often unable to say anything formal about their favourite H_1s (accepted in a qualitative manner). Methods do exist for the meta-analysis of p values (see e.g. Cooper, Hedges and Valentine, 2009) and these are still used in some fields. However, practically such meta-analyses still say nothing about the magnitude of the effect size of the phenomenon being addressed. These methods are potentially acceptable when the question is whether there is any non-null signal among multiple studies that have been performed, e.g. in some types of genetic associations where it is taken for granted that the effect sizes are likely to be small anyhow (Evangelou and Ioannidis, 2013).

5. The state of the art must change

5.1 NHST is unsuitable as the cornerstone of scientific inquiry in most fields

In summary, NHST provides *the illusion of certainty* through supposedly 'objective' binary accept/reject decisions (Cohen, 1997; Ioannidis 2012) based on practically meaningless

p values (Bakan 1966). However, researchers usually never give any formal assessment of how well their theory (a specific H_1) fits the facts and, instead of gradual model building (Gigerenzer 1988) and comparing the plausibility of theories, they can get away with destroying a strawman: they disprove an H_0 (which happens inevitably sooner or later) with a machinery biased to disproving it without ever going into much detail about the *exact* behaviour of variables under *exactly* specified hypotheses (Kranz 1999; Jaynes 2003). NHST also does not allow for systematic knowledge accumulation. In addition, both because of its shortcomings and because it is subject to major misunderstandings it facilitates the production of non-replicable false positive reports. Such reports ultimately erode scientific credibility and result in wasting perhaps most of the research funding in some areas (Nosek 2015; Ioannidis 2005; Macleod, Michie, Roberts, Dirnagl, Chalmers, Ioannidis, Al-Shahi Salman, Chan and Glasziou, 2014).

NHST seems to survive for various reasons. First, it allows for the easy production of a large number of publishable papers (irrespective of their truth value) providing a response to publication pressure. Second, NHST seems deceptively simple: because the burden of inference (Bakan, 1966) has been delegated to the significance test all too often researchers' statistical world view is narrowed to checking an inequality: is $p \leq 0.05$ (Cohen, 1994)? After passing this test, an observation can become a 'scientific fact' contradicting the random nature of statistical inference (Gelman, 2014). Third, in biomedical and social science NHST is often falsely perceived as the *single* objective approach to scientific inference (Gigerenzer et al. 1989) and alternatives are simply not taught and/or understood.

We have now decades of negative experience with NHST which gradually achieved dominance in biomedical and social science since the 1930s (Gigerenzer et al. 1989). Critique of NHST started not much later (Jeffreys, 1939) and has been forcefully present since then (Rozeboom, 1960; Nunnaly, 1960; Eysenck 1960; Clark 1963; Jeffreys, 1961; Bakan 1966;

Mehl 1967; Lykken 1968) and continues to-date (Wasserstein and Lazar, 2016). The problems are numerous, and as Edwards (1972, p179) concluded 44 years ago: *'any method which invites the contemplation of a null hypothesis is open to grave misuse, or even abuse'*. Time has proven this statement and that problems are unlikely to go away. We suggest that that it is *really* time for change now.

5.2 If theory is weak we need to focus on estimating effect sizes and their uncertainty

In basic biomedical and psychology research we often cannot provide very well worked out hypotheses and even a simple directional hypothesis may seem particularly enlightening. Such rudimentary state of knowledge can be respected. However, in such pre-hypothesis stage substantively blind all or nothing accept/reject decisions may be unhelpful and may maintain our ignorance rather than facilitate organizing new information into proper scientific models. It is much more meaningful to focus on assessing the magnitude of effects along with estimates of uncertainty, let these be error terms or Bayesian credible intervals (Luce, 1988; Edwards, 1972; Jaynes 2003; Schmidt, 1996; Gelman 2013; Lakens & Evers, 2014; see Morey et al. 2016 for why classical confidence intervals are not appropriate uncertainty measures). These provide more direct information on the actual 'empirical' behaviour of our variables. Gaining enough experience with interval estimates and assuring their robustness by building replication into design (Nosek et al. 2013) may then allow us to describe the behaviour of variables by more and more precise scientific models which may provide more clear predictions (Jaynes 2003; Schmidt, 1996; Gelman 2013).

The above problem does not only concern perceived 'soft areas' of science where measurement, predictions, control and quantification are thought to be less rigorous than in ‘hard’ areas (Meehl, 1978). In many fields, for example, in cognitive neuroscience, the measurement methods may be 'hard' but theoretical predictions and analysis often may be just

as 'soft' as in any area of 'soft' psychology: Using a state of the art fMRI scanner for data collection and extremely complicated and often not clearly understood black box software for data analysis will not make a badly defined theory well defined. In fact, in such *pretend-hard* areas the situation may even be worse because technology allows us to measure a huge number of variables and run an immense number of tests (many of them undocumented and hence, uncorrected for multiple testing) and analyze the data by highly complex obscure black box processes and non-replicable idiosyncratic approaches. All these problems will only boost the number of false positive unreplicable findings (Carp, 2012; Vul et al. 2009; Kriegeskorte et al. 2009).

5.3 Improved reporting and alternative statistical inference methods are needed

While we need substantial change, criticism of NHST should not *'lapse into methodological anarchy out of despair or confusion'* (Giere, 1972, p.171). For example, recently, the Journal of Basic and Applied Social Psychology banned NHST from their articles (Trafimov and Marks, 2015; see Hunter, 1997). The decision prompted critical responses from several high profile statisticians who objected to the approach of the journal editors and their negative view on the controversies of statistical inference in general (https://www.statslife.org.uk/news/2116-academic-journal-bans-p-value-significance-test), which sharply contrasts with the view of the renowned physicist ET Jaynes (2003; pp. xxii; see starting quotes) who plausibly defined probability theory as the *'logic of science'*. Let's make it clear that we are not arguing against statistical inference in general and we do not want to ban NHST. Quantitative and well justified statistical inference should be at the *core* of the scientific enterprise. We argue against the *default* and mindless application of NHST.

It may be reasonable to use NHST in some cases. One such case is when very precise quantitative theoretical predictions can be tested, hence, both power and effect size can be estimated well as intended by Neyman and Pearson (1933). Further, when theoretical

predictions are not precise, we powered NHST tests may be used as an initial heuristic look at the data as Fisher (1925) intended. In this second case NHST tests must be followed up by more robust procedures to estimate effect sizes and interval estimates (e.g. by the now widely available bootstrap and permutation procedures) and (if there are clear hypotheses) more robust likelihood estimation or Bayesian techniques to test hypotheses. As Fisher was well conscious, NHST procedures can only suggest that something is really important if they *'rarely fail'* to give us statistically significant results (Gigerenzer et al. 1989, p96; Goodman, 2008). Hence, strong claims require the replication of NHST tests optimally within the initial study. These replications must be well powered to keep FRP low. In all cases when NHST is used its use must be justified clearly rather than used as an automatic default and single cornerstone procedure. As discussed, NHST can only reject H_0 and can accept neither a generic or specific H_1. So, on its own NHST cannot provide evidence 'for' something even if findings are replicated.

Statistical reporting must improve substantially. Optimally, raw data should be published because data parameters of interest depend on the choice of models and analyses. Regarding the actually chosen analyses, the distribution (very rarely plotted at the moment) and important parameters of the data should be communicated (e.g. means and standard deviations in the original units of measurement as well as confidence and/or credible intervals). Regarding NHST procedures, researchers should report power calculations for each test including those with non-significant results and the number of cases should clearly be reported for each test. In the age of internet all important results can be communicated as online supplementary material, so there is not much excuse for not doing this. Improved access to raw data, algorithms and code would also be helpful and efficient ways to promote such reproducible research practices need to be found (Doshi, Goodman, Ioannidis, 2013; Diggle and Zeger, 2010; Keiding 2010; Laine, Goodman and Griswold, 2007; Peng, 2009; Peng,

2011). Incentives such as a badge system may help promote availability of more raw data (Nosek et al. 2015). Diverse stakeholders (journals, funders, institutions, and more) may contribute to align incentives with better research practices (Ioannidis 2014).

Hypotheses could be tested by either likelihood ratio testing, and/or Bayesian methods which usually view probability as characterizing the state of our beliefs about the world (Jaynes, 2003; Pearl 1998; MacKay, 2003; Gelman et al. 2014; Sivia and Skilling, 2006). The above alternative approaches require model specifications about alternative hypotheses, they can give probability statements about H_0 and alternative hypotheses, they allow for clear model comparison, are insensitive to data collection procedures and do not suffer from problems with large samples. In addition, Bayesian methods can also factor in pre-study (prior) information into model evaluations which may be important for integrating current and previous research findings. Hence, the above alternative approaches seem more suitable for the purpose of scientific inquiry than NHST and ample literature is available on both. The problem is that usually none of these alternative approaches are taught properly in statistics courses for students in psychology, neuroscience, biomedical science and social science. For example, across 1000 abstracts randomly selected from the biomedical literature of 1990-2015, none reported results in a Bayesian framework (Chavalarias, et al. 2016).

5.4 Better training and better use of statistical methods: from believers to thinkers

We argue that the practice of relying on (good-willed) editorial dictates rather than informed statistical thinking is a symptom of a core problem: the statistical subject knowledge of many researchers in biomedical and social science has been shown to be poor (Oakes, 1986; Gliner et al. 2002; Wilkerson and Olson, 2010; Hoekstra et al. 2014; Castro-Sotos 2007; 2009). NHST perfectly fits with poor understanding because of the perceived simplicity of interpreting its outcome: is $p \leq 0.05$ (Cohen, 1994)?

We suggest that the weak statistical understanding is probably due to inadequate

'statistics lite' education based on supposedly 'user friendly' dumbed down statistics cookbooks which may do more harm than good. This approach does not build up appropriate mathematical fundamentals and does not provide scientifically rigorous introduction into statistics. Hence, students' knowledge may remain imprecise, patchy and prone to serious misunderstandings. What this approach achieves, however, is providing students with false confidence of being able to use inferential tools whereas *they usually only interpret the p value provided by black box statistical software*. While this educational problem remains unaddressed, poor statistical practices will prevail regardless of what procedures and measures may be favoured and/or banned by editorials.

Understanding probability is difficult. Common sense is notoriously weak in understanding phenomena based on probabilities (Gigerenzer et al. 2005). We cannot assume that without proper training biomedical and social science graduates would get miraculously enlightened about probability. Some of the best symbolic thinking minds of humanity devoted hundreds of years to the proper understanding of probability and statisticians still do not agree on how best to draw statistical inference (Stigler 1986; Gigerenzer et al. 1989), e.g. the recent ASA statement on p values (Wasserstein and Lazar, 2016) was accompanied by 21 editorials from the statisticians and methodologists who participated in crafting it and who disagreed in different aspects among themselves.

One approach would be to phase out the 'statistics lite education approach for all research stream students and teach statistics rigorously, based on two years of calculus. Besides NHST, Bayesian and likelihood based approaches should also be taught, with explanation of the strengths and weaknesses of each inferential method. An alternative and/or complementary approach would be to enhance the training of professional applied statisticians and to ensure that all research involves knowledgeable statisticians or equivalent methodologists. At a minimum, all scientists should be well trained in understanding evidence and statistics and

being in a position to recognize that they may need help from a methodologist expert (Marusic A, Marusic, 2003; Moharar, Rahimi and Najafi, 2009; Vujaklija, Hren, Sambunjak, Vodopivec, Ivanis, Marusic and Marusic, 2010).

All too often statistical understanding is perceived as something external to the subject matter of substantive research. However, it is important to see that statistical understanding influences most decisions about substantive questions, because it underlies the *thinking* of researchers even if this remains *implicit*. While common sense 'statistics' may be able to cope with simple situations, common sense is not enough to decipher scientific puzzles involving dozens, hundreds, or even thousands of interrelated variables. In such cases well justified applications of probability theory are necessary (Jaynes, 2003). Hence, instead of delegating their judgment to 'automatized' but ultimately spurious decision mechanisms, researchers should have confidence in their own *informed judgment* when they make an inference.

5.4 There is no automatic inference: New-old dangers ahead?

Perhaps the most worrisome false belief about statistics is the belief in automatic statistical inference (Bakan 1966), the illusion that plugging in some numbers into some black box algorithm will give a number (perhaps the p value or some other metric) that conclusively proves or disproves hypotheses (Bakan 1966). There is no reason to assume that any kind of 'new statistics' (Cummings, 2008) will not suffer the fate of NHST if statistical understanding is inadequate. For example, it has been shown that confidence intervals are misinterpreted just as badly as p values by undergraduates, graduates and researchers alike and self-declared statistical experience even slightly positively correlates with the number of errors (Hoekstra et al. 2014). Similarly, the proper use of Bayesian methods may require use of advanced simulation methods and a clear understanding and justification of probability distribution models. In contrast to this, it is frequent to see a kind of 'automatic' determination of Bayes factors or posterior estimates, again, provided by black box statistical packages which again,

promise to take the load of thinking off the shoulders of researchers.

There is no reason to assume that understanding 21st and 22nd century science will require less mathematical and statistical understanding than before. If statistical understanding does not improve it will not matter whether editorials enforce bootstrapping, likelihood estimation or Bayesian approaches, they will all remain mystical to the untrained mind and open to abuse such as the NHST of the 20th century.

Acknowledgments. DS is supported by the James S McDonnell Foundation.

BOX 1: Major confusions about the p value

1. Many practicing researchers and even some statisticians confuse the roles of the p value and α (Hubbard and Bayarri 2003). These researchers set α = 0.05 before they run an experiment but once they compute the p value they falsely assume that the p value will now represent the actual data-dependent Type I error probability somehow replacing the Neyman-Pearson α level while also interpreting it as the strength of evidence against H_0 as used by Fisher (Goodman, 1993; 1999; Nickerson 2000). However, α is always fixed independently of what p value we find in an experiment whereas p values can be considered random variables, varying widely from experiment to experiment (Murdoch et al. 2008; Hung et al. 1997; Simonsohn et al. 2014a,b; Sterling 1959). Currently, the expression 'significance level' is used interchangeably for both the p value and α reflecting the confusion about them (Hubbard and Bayarri 2003).

2. Many practicing researchers falsely assume that if p = 0.01 then the probability of a false positive finding given the data (pr(H_0|data) is 0.01. Conversely, they also assume that if p = 0.01 then the probability of a truly positive finding given the data (pr(H_1|data) is 1 - p = 0.99. Yet, others confuse the p value with the 'updated' H_0:H_1 odds after a study was run, and/or with replication success (Bakan, 1966; Meehl, 1967; Pollard and Richardson, 1987; Cohen 1994; Hunter, 1997; Goodman, 1999; Oakes, 1986; Gliner et al. 2002; Wilkerson and Olson, 2010;

Hoekstra et al. 2014; Castro-Sotos 2007; 2009). These *false* assumptions are not only *thoroughly wrong*, they also deeply *underestimate* the probability of false positive findings and highly *overestimate* the probability of truly positive findings and replication success. The network of confusions outlined here constitute what Goodman (1999) termed the *'p value fallacy'* (see Goodman, 1999; Goodman 2008; Nickerson 2000 and Wagenmakers, 2007 for excellent reviews).

References

Ambaum, M.H.P. 2010. Significance tests in climate science. *Journal of Climate. 23*, 5927-5932.

Bakan, D. 1966. The test of significance in psychological research. *Psychological Bulletin*. *66*, 423-437.

Bakker, M., & Wicherts, J.M. 2001. The misreporting of statistical results in psychology journals. *Behav Res Methods*. *43*, 666-78.

Bayarri M.J., Benjamin, D.J., Berger, J.O., Sellke, T.M. 2016, Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*. In Press.

Berger, 1985. *Statistical decision theory and Bayesian analysis* 2nd edition. New York: Springer.

Berger, J. O., & Delampady, M. 1987. Testing precise hypothesis. *Statistical Science*. *2*, 317–352.

Berger, J. O., & Sellke, T. 1987. Testing a Point Null Hypothesis: the Irreconcilability of *p*-Values and Evidence. *Journal of the American Statistical Association*. *82*, 112–122.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*. *33*, 526-542.

Boutron, I., Altman, D.G., Hopewell, S., Vera-Badillo, F., Tannock, I., & Ravaud, P. 2014. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *J ClinOncol*. *32*, 4120-26.

Boutron, I., Dutton, S., Ravaud, P., & Altman, D.G. 2010. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*. *303*, 2058-64.

Bruns, S., & Ioannidis, J.P. 2016. p-Curve and p-Hacking in observational research. *PLoS ONE*. 112:e0149144

Button, K.S., Ioannidis, J., Mokrysz, C., & Nosek, B.A., 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*. *14*,

365-376.

Carp, J. 2012. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage. 63*, 289-300. http://dx.doi.org/10.1016/j.neuroimage.2012.07.004

Carver, R.P. 1993. The case against statistical significance testing, revisited. *Journal of Experimental Education*. *61*, 287-292.

Castro Sotos, A.E. Vanhoof, S. Van den Noortage, W., & Onghena, P. 2007. Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*. *2*, 98-113.

Castro Sotos, A.E., Vanhoof, S., Van den Noortage. W,. & Onghena, P. 2009. How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education*. *17*, No 2.

Chalalarias, D., Wallach, J., Li, A., & Ioannidis, J.P. 2016. Evolution of reporting P-values in the biomedical literature, 1990-2015. *JAMA.* in press

Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., Brooks, L.D., Cardon, L.R., Daly, M., Donnelly, P., Fraumeni, J.F. Jr., Freimer, N.B., Gerhard, D.S., Gunter, C., Guttmacher, A.E., Guyer, M.S., Harris, E.L., Hoh, J., Hoover, R., Kong, C.A., Merikangas, K.R., Morton, C.C., Palmer, L.J., Phimister, E.G., Rice, J.P., Roberts, J., Rotimi, C., Tucker, M.A., Vogan, K.J., Wacholder, S., Wijsman, E.M., Winn, D.M., & Collins, F.S. 2007. Replicating genotype-phenotype associations. *Nature. 447*, 655-60.

Chavalarias D, Wallach J, Li A, Ioannidis JP. Evolution of reporting P-values in the biomedical literature, 1990-2015. JAMA in press March 2016

Clark, C.A. 1963. Hypothesis testing in relation to statistical methodology. *Review of Educational Research*. *33*, 455-473.

Cohen, J. 1962. The statistical power of abnormal - social psychological research: A review.

Journal of Abnormal and Social Psychology. *65*, 145-153.

Cohen, J. 1988. *Statistical power analysis for the behavioural sciences*. Academic Press.

Cohen, J. 1994. The earth is roundp < 0.05. *American Psychologist*. *49*, 997-1003.

Cumming, G. 2014. The new statistics: Why and how? *Psychological Science*, *25*, 7-28. http://dx.doi.org/0956797613504966.

Deer, B. 2011. How the case against the MMR vaccine was fixed. *British Medical Journal*. *342*, c5347. http://dx.doi.org/10.1136/bmj.c5347

DeMets, D., & Lan, K.K.G. 1994. Interim analysis: The alpha spending function approach. *Statistics in Medicine*. *13*, 1341-1352.

Diggle, P.J., & Zeger, S.L. 2010. Embracing the concept of reproducible research. *Biostatistics*. 11, 375.

Doshi, P., Goodman, S.N., & Ioannidis, J.P. 2013. Raw data from clinical trials: within reach? *Trends Pharmacol Sci*. *34*, 645-7.

Edwards, A.W.F. 1972. *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge, UK: Cambridge University Press.

Etz A, Vandekerckhove J 2016, A Bayesian perspective on the reproducibility project: Psychology. *PLOS One*. 112: e0149794. DOI: 10.1371/journal/pone.0149794

Evangelou, E., & Ioannidis, J. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet*. *14*, 379-89.

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

Falk, R. Greenbaum, C.W. 1995. Significance tests die hard: The Amazing persistence of a probabilistic misconception. *Theory and Psychology*. *5*, 75-98.

Fanelli, D. 2010. *Do pressures to publish increase scientists' bias? An empirical support from US states data*. *PLoS ONE. 5*, 1-7.

Fisher, R. 1925. *Statistical methods for research workers*. First Edition. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1956. *Statistical Methods and Scientific Inference*. London: Oliver & Boyd; second revised edition, New York 1959: Hafner Publishing Co.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin D 2014, Bayesian data analysis. CRC Press.

Gelman, A. 2013. Commentary: P values and statistical practice. *Epidemiology*. *24*, 69-72.

Gelman, A. 2013. Interrogating p values. *Journal of Mathematical Psychology*. *57*, 188-189.

Giere, R.N. 1972. The significance test controversy. *The British journal for the philosophy of science. 23*, 170-181.

Gigerenzer, G. 1998. We need statistical thinking, not statistical rituals. *Behavioural and Brain Sciences*, *21*, 199-200.

Gigerenzer, G. 2004. Mindless statistics. *The Journal of Socio-economics. 33*, 587-606.

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K.V. 2005, 'A 30% chance tomorrow': How does the public understand probabilistic weather forecasts? *Risk Analysis*. *25*, 623-629.

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*. pp391-408. Thousand Oaks, CA: Sage.

Gigerenzer, G., Marewski, J.N. 1998. Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*. *41*, 421-400.

Gigerenzer, G., Swijtnik, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. 1989. *The empire of chance.* Camridge, UK: Cambridge University Press. Cambridge.

Gliner, J.A., Leech, N.L., Morgan, G.A. 2002. Problems with null hypothesis significance testing NHST: What do the textbooks say? *The Journal of Experimental Education. 7*, 83-92.

Godlee, F. 2011. Wakefield's article linking MMR vaccine and autism was fraudulent. *British Medical Journa*l. *342*.

Goodman, S.N. 1993. p values, hypothesis tests and likelihood: implications for epidemiology of a neglected historical debate. *Epidemiology*. *5*, 485-496.

Goodman, S.N. 1999. Toward evidence-based medical statistics 1: The p value fallacy. *Annals of Internal Medicine*. *130*, 995-1004.

Goodman, S.N. 2008. A dirty dozen: Twelve p value misconceptions. *Seminars in Hematology*. *45*, 135-140.

Greenberg, S.A. 2009. How citation distortions create unfounded authority: analysis of a citation network. *BMJ*. 1-14. Hallahan, M. Rosenthal, R. 1996. Statistical power: Concepts, procedures and applications. *Behavioural Research and Theory*. *34*, 489-499.

Hoekstra, R. Morey, R.D., Rouder, J.N., Wagenmakers, E.J. 2014. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review*. *21*, 1157-1164.

Hubbard, R. & Bayarri, M.J. 2003. Confusion over measures of evidence p's versus errors α's in classical statistical testing. *The American Statistician*. *57*, 171-182.

Hung, H.M.J., O'Neill, T., Bauer, P., & Kohne, K. 1997. The behavior of the p value when the alternative hypothesis is true. *Biometrics*. *53*, 11-22.

Hunter, J.E. 1997. Needed: A ban on the significance test. *Psychological Science*. *8*, 3-7.

Ioannidis JP, Hozo I, Djulbegovic D 2014, Improving the drug development process: More not less random trials. *Journal of Clinical Epidemiology*. 311, 355-6.

Ioannidis, J. 2008. Why most true discovered associations are inflated. *Epidemiology*. *19*, 640-8. http://dx.doi.org/

Ioannidis, J.P., Tarone, R., & McLaughlin, J. 2011. The false-positive to false-negative ratio in epidemiological studies. *Epidemiology*. *22*, 450-6.

Ioannidis, J.P., Trikalinos, T.A. 2007. An exploratory test for an excess of significant findings.

Clinical Trials. *4*, 245-53.

Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Medicine*. *2*, e124.

Ioannidis, J.P.A. 2008. Why most discovered true associations are inflated. *Epidemiology*. *19*, 640-648.

Ioannidis, J.P.A. 2012. Why science is not necessarily self-correcting. *Perspectives on Psychological Science*. *7*, 645-654.

Ioannidis, J.P.A. 2014. How to make more published research true. PLoS Medicine. 1110: e1001747.

Ioannidis, J.P.A., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R., Moher, D., Schulz, K.F., & Tibshirani, R. 2014. Increasing value and reducing waste and research design, conduct and analysis. *Lancet*, *383*, 166-175.

Jaeschke, R., Singer, J., & Guyatt, G.H. 1989. Measurement of health status: ascertaining the minimal clinically important difference. *Controlled clinical trials*. *10*4, 407-415. http://dx.doi.org/

Jannot, A.S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T.V. 2013. Citation bias favoring statistically significant studies was present in medical research. *J Clin Epidemiol*. *66*, 296-301.

Jaynes, E.T. 2003. *Probability theory: The logic of science.* Cambridge, UK: Cambridge University Press.

John L. K., Loewenstein G., Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*. 23, 524–532.

Kavvoura, F.K., Liberopoulos, G., & Ioannidis, J.P. 2007. Selection in Reported Epidemiological Risks: An Empirical Assessment. *PLoS Med*. *3*, 456-65.

Keiding, N. 2010. Reproducible research and the substantive context. *Biostatistics*. *11*, 376-8.

Khoury, M.J., & Ioannidis, J.P. 2014. Big data meets public health: Human well-being could

benefit from large-scale data if large-scale noise is minimized. *Science*. *346*, 1054-5.

Kivimäki, M., Batty, G.D., Kawachi, I., Virtanen, M., Singh-Manoux, A., & Brunner, E.J. 2014. Don't let the truth get in the way of a good story: an illustration of citation bias in epidemiologic research. *Am J Epidemiol*. *180*, 446-8.

Kjaergard, L.L., Gluud, C. 2002. Citation bias of hepato-biliary randomized clinical trials. *J Clin Epidemiol*. *55*, 407-10.

Kranz, D.H. 1999. The null hypothesis testing controversy in psychology. *Journal of American Statistical Association*. *94*, 1372-1381.

Kriegeskorte, N. Simmons, W.K., Bellgowan, P.S.F., & Baker, C.I. 2009. Circular analysis in systems neuroscience – the dangers of double dipping. *Nature Neuroscience 12*, 535-40.

Laine, C., Goodman, S.N., Griswold, M.E., & Sox, H.C. 2007. Reproducible research: moving toward research the public can really trust. *Ann Intern Med*. *146*, 450-3.

Lakens, D. & Evers, E.,R.,K. 2014, Sailing from the seas of chaos into the corridor of stability. Perspectives on psychological science. 9, 278-292.

Lazarus, C., Haneef, R., Ravaud, P., & Boutron, I. 2015. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol*. *15*, 1-8. http://dx.doi.org/10.1186/s12874-015-0079-x

Luce, R.D. 1988. The tools to theory hypothesis. Review of G. Gigerenzer and D.J. Murray, 'Cognition as intuitive statistics'. *Contemporary Psychology*. *33*. 582-583.

Lykken, D.T. 1968. Statistical significance in psychological research. *Psychological Bulletin*. *70*, 151-159.

MacKay, D.J.C. 2003. *Information theory, inference and learning algorithms*. Cambridge, UK: Cambridge University Press.

Macleod, M.R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J.P., Al-Shahi Salman, R., Chan, A.W., & Glasziou, P. 2014. Biomedical research: increasing value, reducing

waste. *Lancet*. *383*, 101-4.

Makel M., Plucker J., Hegarty B. 2012. Replications in psychology research: How often do they really occur? Perspectives in Psychological Science, 7, 537–542

Marusic, A., & Marusic, M. 2003. Teaching students how to read and write science: a mandatory course on scientific research and communication in medicine. *Acad Med*. *78*, 1235-9.

Meehl, P.E. 1967. Theory testing in psychology and physics: A methodological paradox. *Philosophy of science*. *34*, 103-115.

Meehl, P.E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of consulting and clinical psychology*. *46*, 806-834.

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

Moharari, R.S., Rahimi, E.R., & Najafi, A et al 2009. Teaching critical appraisal and statistics in anesthesia journal club. *Q J Med*. *102*, 139-41.

Moher, D., Dulberg, C.S., Wells, G.A. 1994, Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*. *272*, 122-4.

Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E-J 2016, The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review*. 23, 103-123.

Murdoch, D.J., Tsai, Y.L., & Adcock, J. 2008. P values are random variables. *The American Statistician*. *62*, 242-245.

Neyman, J. 1950. *Probability and statistics*. New York: Holt.

Neyman, J., & Pearson, E.S. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Ser. A, 231*, 289-337.

Nickerson, R.S. 2000. Null hypothesis significance testing: A review of an old and continuing

controversy. *Psychological Methods*. *5*, 241-301.

Nosek et al. 2015. Estimating the reproducibility of psychological science. *Science*. *349*, 943. http://dx.doi.org/10.1126/science.aac4716

Nosek, B.A., Spies, J.R., & Motyl, M. 2013. Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*. *7*, 615-631.

Nosek, BA, Alter, G, Banks, GC, Borsboom, D, Bowman, SD, Breckler, SJ, Buck, S, Chambers, CD, Chin, G, Christensen, G, Contestabile, M, Dafoe, A, Eich, E, Freese, J, Glennerster, R, Goroff, D, Green, DP, Hesse, B, Humphreys, M, Ishiyama, J, Karlan D, Kraut A, Lupia, A, Mabry, P, Madon, TA, Malhotra, N, Mayo-Wilson, E, McNutt, M, Miguel, E, Paluck, EL, Simonsohn, U, Soderberg, C, Spellman, BA, Turitto, J, VandenBos, G, Vazire, S, Wagenmakers, EJ, Wilson, R, Yarkoni, T. 2015, Promoting an open research culture. Science. Jun 26;3486242:1422-5

Nuijten, M.B., Hartgerink, C.H., van Assen, M.A., Epskamp, S., & Wicherts, J.M. 2015. The prevalence of statistical reporting errors in psychology 1985-2013. *Behav Res Methods*. [Epub ahead of print]

Nunally, J. 1960. The place of statistics in psychology. *Education and psychological measurement*. *20*, 641-650.

Oakes, M.L. 1986. *Statistical inference: A commentary for the social and behavioural sciences*. New York: Wiley.

Panagiotou, O.A., & Ioannidis, J.P. 2012. Genome-Wide Significance Project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol*. *41*, 273-86.

Patel, C.J., & Ioannidis, J.P. 2014. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J Epidemiol Community Health*. *68*, 1096-100.

Patel, C.J., & Ioannidis, J.P. 2014. Studying the elusive environment in large scale. *JAMA. 311*, 2173-4.

Patel, C.J., Burford, B., & Ioannidis, J.P. 2015. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. *68*, 1046-58.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan.

Peng, R.D. 2009. Reproducible research and Biostatistics. *Biostatistics. 10*, 405-8.

Peng, R.D. 2011. Reproducible research in computational science. *Science*. *334*, 1226-7.

Pollard P, Richardson JTE (1987), On the probability of making Type-I errors. Psychological Bulletin, 102, 159-163.

Rossi, J.S. 1990. Statistical power of psychological research: What have we gained in 20 years? *Journal of consulting and clinical psychology*. *58*, 646-656. http://dx.doi.org/

Rozeboom, W.W. 1960. The fallacy of the null hypothesis significance test. *Psychological Bulletin*. *57*, 416-428.

Schmidt, F.L. 1992. What do data really mean? Research findings, meta-analysis and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Schmidt, F.L. 1996. Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods*. *1*, 115-129.

Schoenfeld, J.D., Ioannidis, J.P.A. 2012, Is everything we eat is associated with cancer? A systematic cookbook review. *American Journal of Clinical Nutrition.* 97, 127-134.

Sedlmeyer, P., & Gigerenzer, G. 1989. Do studies of statistical power have an effect on the power of the studies? *Psychological Bulletin. 105*, 309-316.

Sellke, T., Bayarri, M.J., & Berger, J.O. 2001. Calibration of p values for testing precise null hypotheses. *The American Statistician*. *55*, 62-71.

Simmons, J., Nelson, L., & Simonsohn, U. 2011. False-positive psychology: Undisclosed

flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science*, 22, 1359-1366.

Simonsohn, U., Nelson, L.D., & Simmons, J.P. 2014a. P-Curve: A key to the file drawer. *Journal of Experimental Psychology: General. 143*2, 534-547.

Simonsohn, U., Nelson, L.D., & Simmons, J.P., 2014b. *p*-Curve and effect size: Correcting for publication bias using only significant results. *Psychological Science*. *9*6, 666-681.

Siontis, G.C., & Ioannidis, J.P. 2011. Risk factors and interventions with statistically significant tiny effects. *Int J Epidemiol. 40*, 1292-307.

Sivia DS, Skilling J 2006, Data Analysis: A Bayesian tutorial. Oxford University Press.

Statslife.org.uk 2015. http://www.statslife.org.uk/opinion/2114-journal-s-ban-on-null-hypothesis-significance-testing-reactions-from-the-statistical-arena. Retrieved: 27 Oct 2015.

Sterling, T.D. 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.

Sterne, J. A. C. 2002. Teaching hypothesis tests – time for significant change? *Statistics in Medicine*, 21, 985-994.

Sterne, J.A.C., & Smith, G.D. 2001. Sifting the evidence - what's wrong with significance tests? *British Medical Journal*, *322*, 226-231.

Storey, J.D., & Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A. 100*, 9440-5.

Trafimov, D., & Marks, M. 2015. Editorial, *Basic and Applied Social Psychology*, *37*, 1-2.

Veldkamp, C.L., Nuijten, M.B., Dominguez-Alvarez, L., van Assen, M.A., & Wicherts, J.M. 2014. Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One*. *9*, 1-19.

Vujaklija, A., Hren, D., Sambunjak, D., Vodopivec, I., Ivanis, A., Marusic, A., & Marusic, M.

2010. Can teaching research methodology influence students' attitude toward science? Cohort study and nonrandomized trial in a single medical school. *J Investig Med*. *58*, 282-6.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality and social cognition. *Perspectives on Psychological Science*. *4*, 274-324.

Wagenmakers, E-J. 2007, A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*. 14, 779-804.

Waller, N.G. 2004. The fallacy of the null hypothesis in soft psychology. *Applied and preventive psychology*. *11*, 83-86.

Wasserstein, R.L., & Lazar, N.A. The ASA statement on p values: context, process, and purpose. American Statistician. In Press. 2016

Wilkerson, M., & Olson, M.R. 2010. Misconceptions about sample size, statistical significance and treatment effect. *The Journal of Psychology: Interdisciplinary and Applied*. *131*, 627-631.

Yates, F. 1951. The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association, 46,* 19-34.

Ziliak, T., & McCloskey, N. 2008, *The Cult of Statistical Significance*. The University of Michigan Press.