Statistical Heaven Does not Exist: Why Non-Significant Results Matter

Hartgerink, Chris H.J.1

Van Assen, Marcel A.L.M.

Wicherts, Jelte M.1

1 Tilburg University, the Netherlands

**Abstract**

TITLE.

Statistical significance is over abundantly present in the (psychological) academic literature (Sterling, Rosenbaum, & Weinkam, 1995), while studies are vastly and systematically underpowered (i.e., 1-β < .8; Bakker, Van Dijk, & Wicherts, 2012). An underpowered study has a smaller probability of detecting an effect if there truly is one. Based on statistical principles in Null Hypothesis Significance Testing (NHST), high rates of significant values should *not* be possible in underpowered situations, because there is a probabilistic element to outcomes of a study. Even in a highly powered situation, one would expect to find non-significant effects at times, when there truly is an effect, purely due to sample fluctuation – in other words: chance (Francis, 2012a, 2012b).

Assume an underpowered situation, where power is .4. It is now expected that 40 out of 100 studies would show a significant effect, and 60 would show non-significant results. Power provides an expected value for tallied significant results. Extrapolating this and applying it to the significance rates 97% (Sterling, 1959) or 96% (Sterling, Rosenbaum, & Weinkam, 1995), these results would only be theoretically plausible under a power of approximately .97 or .96. A highly powered situation causing this many significant findings is unrealistic, as the sample sizes in psychology are small (). Power in psychological studies has been estimated to be as low as 35% (Bakker et al., 2012). Additionally, power values larger than .8 require exponentially large samples, due to decreasing efficiency of more responses when it crosses the .8 mark (). In other words, the combination of underpowered studies and an abundance of significant results indicates that the theoretical principles of NHST are violated, and other factors than sample fluctuation are at play.

These other factors are human factors, commonly seen as including *publication bias* (), *questionable research practices* (QRPs; Simmons, Nelson, & Simonsohn, 2011)*,* and, in the worst case scenario, *misconduct* (). Assuming these human factors intend to find *significant* effects most of the time, they cause an inflated significance rate. *Publication bias* entails the selective publication of studies, either by authors or by editors/reviewers, where non-significant effects are less likely to be published than significant effects. *Questionable research practices* entail researcher’s degrees of freedom that increase corroboration of the data with the hypotheses (e.g., p-fishing), or adjusting hypotheses to the data (e.g., HARKing; Kerr, 1998). *Misconduct* influences significance rates when data falsification or data fabrication occurs. Combined, but also separately, these human factors inflate significance rates in the literature.

We consider another human factor to be pervasive, as researchers are human beings, and human beings have been shown to have flawed intuition for probability (see Bar-Hillel & Wagenaar, 1991). Most clearly, this flawed intuition shows in the Monte Hall problem (). This flawed intuition for probability gives rise to the notion of *statistical heaven*: if an effect is found, this is a true-positive, and if we find no effect, there *is* no effect. More formally, statistical heaven implies no Type I or Type II errors (i.e., α = 0; β = 0). Put simply, this means that researchers fail to take into account sample fluctuations when evaluating findings, base conclusions on individual study results, and think about effects in a dichotomous manner.

The proposed new statistics (Cumming, 2013) finds a way around this problem, by depending more on meta-analytic strategies to estimate averages across studies. Meta-analysis is highly powerful, and *can* be unbiased if no publication bias is present (). This approach clearly removes the dichotomy of effect/no-effect and moves more towards direction and magnitude of effects. \*Talk about Schimmack and Francis that non-significant effects can increase credibility of results\* In the current paper, we intend to show that entire studies with non-significant effects can still

Gelman has provided a different way of thinking about statistical errors: Type S and Type M errors.

However, we consider this notion of statistical heaven to give rise to *both* publication bias and QRPs. We intend to show that this makes no sense, as non-significant results can still be indicative of an effect

The prevalence of these human factors is variable. Misconduct is estimated to occur in ‘only’ 2% of the published academic work (Fanelli, 2009). Publication bias and QRPs are experienced by researchers as widespread (John, Loewenstein, & Prelec, 2012). This perceived widespread intent to find significanteffects arises from the notion we call *statistical heaven*: This notion of statistical heaven is pervasive in research, because human intuition for probability is flawed. \*Talk about monte hall problem?\* As this intuition for probability is flawed, we consider it an unconscious notion.

Others have approached the abundance

\*Need to talk about statistical heaven

\*Need to talk about how we want to show that non-signif results are important because they give information

Francis (REF) has repeatedly indicated, in recent years, that papers are reporting too many significant effects and that, by means of interrogation, these p-values can be tortured into telling something about the plausibility of the study. When too many significant p-values are reported (i.e., < α), this could be indicative of unlikely effects. We note that, in the reverse situation (i.e., > α) this might be the exact opposite: researchers stating there is no effect, while in truth, there was an effect-but that they lacked the power to detect it in light of Null Hypothesis Significance Testing (NHST).

Such reporting of findings based on p-values alone, often gives rise to what we dub *statistical heaven*. In this situation, researchers interpret findings as if α- and β equal zero: an effect found can not be a Type I or a Type II error by default. This makes sense, in light of humans lacking the randomness of mind to even randomly generate numbers properly (e.g., Nickerson, 2002). We note however, that such a dichotomous accept/reject situation can make us reject much relevant information and artificially select results, even though that (hopefully) is not the researchers intention. In this paper, we intend to convince readers that, based on non-significant tests, we can still find an effect across the board.

To do this, we consider the distributional properties of p-values. Under the null hypothesis, these are uniformly distributed, meaning that a p-value of .5 is just as likely to occur as .4 or any other value in the state space [0;1]. However, just as some have proposed that NHST is redundant, because the probability of the null being correct is impossible (densities work in ranges, not in points), we propose that the uniform distribution of p-values is a theoretical ideal, which is never found in research. We therefore think that p-values, as reported in the reports, can be used to investigate what the effect size is.

In light of these ideas, we develop two studies to answer the research question: do non-significant p-values give indication for an effect size being present, even if researchers proclaim there is none? More specifically, what are the odds ratios at a paper level for the null being actually true, versus the study just having a lack of power under estimated effect size in that field. This question goes counter to what many people believe about publication bias: only significant effects are being reported, but we proclaim that *real* effects are being underreported, plainly because of a lack of power to detect them.

For example, if the odds ratio of *true* null versus lack of power were to be .5:1, this would indicate that in 1/3 of the papers, the null is estimated to be rightfully rejected. Conversely, this would immediately imply that 2/3 (!) of the null hypotheses are falsely *not* rejected. We try to estimate average odds ratios across fields and provide a tool that can estimate these odds ratios on a paper level. Subsequently, we might investigate whether papers are concerned with their power, and how that predicts such odds ratios.

**Method**

We use an automated tool to extract p-values from papers (statcheck; Epskamp & Nuijten, 2013). This program only extracts test statistics which are fully reported and up to APA norms. Critics might ask why we only include these, and why not the others (these might be just as important). Because we inspect a large array of papers, and individual inspection per paper is not possible with the scale we are working on, we assumed that if test statistics are being properly and fully reported, this indicates the results belong to the important effects which were studied. Sometimes non-significant tests are reported in a collective (e.g, *p*s > .10), which we do not pick up. Thus, we assume that these are non-vital to the report, an assumption that might be called into question sometimes.

\*Blablabla talk about fisher test and how we compute our p-values\*

\*blablabla talk about how we are going to run an awesome simulation study which will blow everyone’s mind\*