The world of research has gone berserk: modeling the consequences of requiring "greater statistical stringency" for scientific publication

Harlan Campbell and Paul Gustafson
Department of Statistics, University of British Columbia, Vancouver, Canada

December 15, 2018

Abstract

In response to growing concern about the reliability and reproducibility of published science, researchers have proposed adopting measures of 'greater statistical stringency,' including suggestions to require larger sample sizes and to lower the highly criticized '$p < 0.05$' significance threshold. While pros and cons are vigorously debated, there has been little to no modeling of how adopting these measures might affect what type of science is published. In this paper, we develop a novel optimality model that, given current incentives to publish, predicts a researcher's most rational use of resources in terms of the number of studies to undertake, the statistical power to devote to each study, and the desirable pre-study odds to pursue. We then develop a methodology that allows one to estimate the reliability of published research by considering a distribution of preferred research strategies. Using this approach, we investigate the merits of adopting measures of 'greater statistical stringency' with the goal of informing the ongoing debate.

Keywords: reliability, reproducibility, publication, meta-research, null hypothesis significance testing, statistical power

1 Introduction

> It is to be remarked that the theory here given rests on the supposition that the object of the investigation is the ascertainment of truth. When an investigation is made for the purpose of attaining personal distinction, the economics of the problem are entirely different. But that seems to be well enough understood by those engaged in that sort of investigation.
>
> *Note on the Theory of the Economy of Research,*
>
> Charles Sanders Peirce, 1879

In a highly cited essay, Ioannidis (2005) uses Bayes theorem to claim that more than half of published research findings are false. While not all agree with the extent of this conclusion (e.g. Goodman & Greenland (2007), Leek & Jager (2017)), recent large-scale efforts to reproduce published results in a number of different fields (economics, Camerer et al. (2016); psychology, Open-ScienceCollaboration (2015); oncology, Begley & Ellis (2012)), have also raised concerns about the reliability and reproducibility of published science. Unreliable research not only reduces the credibility of science, but is also very costly (Freedman et al. 2015) and as such, addressing the underlying issues is of "vital importance" (Spiegelhalter 2017). Many researchers have recently proposed adopting measures of "greater statistical stringency," including suggestions to require larger sample sizes and to lower the highly criticized "$p < 0.05$" significance threshold. In statistical terms, this represents selecting lower levels for acceptable type I and type II error rates. The main argument against "greater statistical stringency" is that these changes would increase the costs of a study and ultimately reduce the number of studies conducted.

Consider the debate about lowering the significance threshold in response to the work of Johnson (2013), who, based on the correspondence between uniformly most powerful Bayesian tests and classical significance tests, recommends lowering significance thresholds by a factor of 10 (e.g. from $p < 0.05$ to $p < 0.005$). Gaudart et al. (2014), voicing a common objection, contend that such a reduction in the allowable type I error rate will result in inevitable increases to the type II error rate. While larger sample sizes could compensate, this can be costly: "increasing the size of clinical trials will reduce their feasibility and increase their duration" (Gaudart et al. 2014). In the view of Johnson (2014), this may not necessarily be such a bad thing, pointing to the excess of false positives and the idea that (in the context of clinical trials) "too many ineffective drugs are subjected to phase III testing [...] wast[ing] enormous human and financial resources."

More recently, a highly publicized call by over seven dozen authors to "redefine statistical

significance" has made a similar suggestion: lower the threshold of what is considered "significant" for claims of new discoveries from $p \leq 0.05$ to $p \leq 0.005$ (Benjamin et al. 2018). This has prompted a familiar response (e.g. Wei & Chen (2018)). Amrhein et al. (2017) review the arguments for and against more stringent thresholds for significance and conclude that: "[v]ery possibly, more stringent thresholds would lead to even more results being left unpublished, enhancing publication bias. [...] [W]hile aiming at making our published claims more reliable, requesting more stringent fixed thresholds would achieve quite the opposite."

There is also substantial disagreement about suggestions to require larger sample sizes. In some fields, showing that a study has a sufficient sample size (i.e., high power) is common practice and an expected requirement for funding and/or publication, while in others it rarely occurs. For example, Charles et al. (2009) found that 95% of randomized controlled trials (RCTs) report sample size calculations. In contrast, only a tiny fraction of articles in some fields –about 3% for psychology, and about 2% for toxicology - justify their sample size (Fritz et al. 2013, Bosker et al. 2013). The number is only marginally higher in conservation biology at approximately 8% (Fidler et al. 2006).

One argument is that, once a significant finding is achieved, the size of a study is no longer relevant. Aycaguer & Galbán (2013) explain as follows: "If a study finds important information by blind luck instead of good planning, I still want to know the results." Another viewpoint is that, while far from ideal, underpowered studies should be published since cumulatively, they can contribute to useful findings (Walker 1995). Others disagree and contend that small sample sizes undermine the reliability of published science (Button et al. 2013*a*, Dumas-Mallet et al. 2017, Nord et al. 2017). In the context of clinical trials, IntHout et al. (2016) review the many conflicting opinions about whether trials with suboptimal power are justified and conclude that, in circumstances when evidence for efficacy can be effectively combined across a series of trials (e.g. via meta-analysis), small sample sizes might be justified.

Despite the long-running and ongoing debates on significance thresholds and sample size requirements, there has been little to no modeling of how changes to a publication policy might affect what type of studies are pursued, the incentive structures driving research, and ultimately, the reliability of published science. One example is Borm et al. (2009) who conclude, based on simulation studies, that the consequences of publication bias do not warrant the exclusion of trials with low power. Another recent example is Higginson & Munafò (2016), who, based on results from an optimality model of the "scientific ecosystem," conclude that in order to "improve the

scientific value of research" peer-reviewed publications should indeed require larger sample sizes, lower the α significance threshold, and give "less weight to strikingly novel findings." Our work here aims to build upon these modeling efforts to better inform the ongoing discussion on the reproducibility of published science.

The model and methodology we present seeks to add three features absent from the model of Higginson & Munafò (2016). While these authors consider how researchers balance available resources between exploratory and confirmatory studies, this simple dichotomy does not allow for a detailed assessment of the willingness of researchers to pursue high-risk studies (those studies that are, *a priori*, unlikely to result in a statistically significant finding). Our approach addresses this issue by considering a continuous spectrum of *a priori* risk, i.e. the "pre-study probability." Secondly, Higginson & Munafò (2016) define the "total fitness of a researcher" (i.e., the payoff for a given research strategy) with diminishing returns for confirmatory studies, but not for exploratory studies. This choice, however well intended, has problematic repercussions for their optimality model. (Under their framework, the optimal research strategy will depend on T, an arbitrary total budget parameter.) Finally, by failing to incorporate the number of correct studies that go unpublished within their metric for the value of scientific research, many potential downsides of adopting measures to increase statistical stringency are ignored. Other differences between our approach and previous ones will be made evident and include: considering outcomes in terms of distributional differences, and specific modeling of how sample size requirements are implemented.

The remainder of this paper is structured as follows. In Section 2, we describe the model and methodology proposed to evaluate different publication policies. We also list a number of metrics of interest and look to recent meta-research analyses to calibrate our model. In Section 3, we use the proposed methodology to evaluate the potential impact of lowering the significance threshold; and in Section 4, the impact of requiring larger sample sizes. Finally, in Section 5, we conclude with suggestions as to how publication policies can be defined to best improve the reliability of published research.

2 Methods

Recently, economic models have been rather useful for evaluating proposed research reforms (Gall et al. 2017). However, modeling of how resources ought to be allocated among research projects is not new. See for example the work of Greenwald (1975), Dasgupta & Maskin (1987), McLaughlin (2011), and Miller & Ulrich (2016). As noted in the introduction, our framework for modeling

the scientific ecosystem is closest in spirit to that of Higginson & Munafò (2016) who formulate a relationship between a researcher's strategy and his/her payoff, with the strategy involving a choice of mix between exploratory and confirmatory studies, and a choice of pursuing fewer studies with larger samples or more studies with smaller samples.

The publication process is complex and includes both objective and subjective considerations of efficacy and relevance. The title of this article was chosen specifically to emphasize this point (Gornitzki et al. 2015). A large, complicated human process like that of scientific publication cannot be entirely reduced to metrics and numbers: there are often financial, political and even cultural reasons for a paper being accepted or rejected for publication. With this in mind, the model presented here should not be seen as an attempt to precisely map out the peer-review process, but rather, as a useful tool for determining the consequences of implementing different publication policies.

Within our optimality model, many assumptions and simplifications are made. Most importantly, we assume that each researcher must make decisions consisting of only two choices: what statistical power (i.e., sample size) to adopt and what "pre-study probability" to pursue. Before elaborating further, let us briefly discuss these two concepts.

2.1 Statistical power

Increasing statistical power by conducting studies with larger sample sizes would undeniably result in more published research being true. However, these improvements may only prove modest, given current publication guidelines. When we consider the perspectives of both researchers and journal editors, it is not surprising that statistical power has not improved substantially (Smaldino & McElreath 2016, Lamberink et al. 2018) despite being highlighted as an issue over five decades ago (Cohen 1962).

In practice, multiple factors can influence sample size determination (Lenth 2001, Roos 2017, Hazra & Gogtay 2016). Strictly in terms of publication prospects however, there is little incentive to conduct high-powered studies: basic logic suggests that the likelihood of publication is only minimally affected by power. To illustrate, consider a large number of hypotheses tested, out of which 10% are truly non null. Under the assumption that only (and all) positive results are published with $\alpha = 0.05$ (which may in fact be realistic in certain fields, Fanelli (2011)), simple analytical calculation shows that increasing average power from an "unacceptably low" 55% to a "respectable" 85% (at the cost of more than doubling sample size), results in only a minimal

increase in the likelihood of publication: from 10% to 13%. Moreover, the proportion of true findings amongst those published is only increased modestly: from 55% to 64%. Indeed, a main finding of Higginson & Munafò (2016) is that the rational strategy of a researcher is to "carry out lots of underpowered small studies to maximize their number of publications, even though this means around half will be false positives." This result is in line with the views of many; see for example Bakker et al. (2012), Button et al. (2013*b*) and Gervais et al. (2015).

From a journal's perspective, there is also little incentive to require larger sample sizes as a requirement for publication. There are minimal consequences from publishing false claims. Fraley & Vazire (2014) review the publication history of six major journals in social-personality psychology and find that "journals that have the highest impact [factor] also tend to publish studies that have smaller samples." This finding is in agreement with Szucs & Ioannidis (2016) who conclude that, in the fields of cognitive neuroscience and psychology, journal impact factors are negatively correlated with statistical power; see also Brembs et al. (2013).

2.2 Pre-study probability

We use the term "pre-study probability" (*psp*) as shorthand for the *a priori* probability that a study's null hypothesis is false. In this sense, highly exploratory research will typically have very low *psp*, whereas confirmatory studies will have a relatively high *psp*. Studies with low *psp* are not problematic per se. To the contrary, there are undeniable benefits to pursuing "long-shot" novel ideas that are very unlikely to work out, see Cohen (2017). While replication studies (i.e., studies with higher *psp*) may be useful to a certain extent, there is little benefit in confirming a result that is already widely accepted. As Button et al. (2013*a*) note: "As R [the pre-study odds] increases [...] the incremental value of further research decreases." Most scientific journals no doubt take this into account in deciding what to publish, with more surprising results more likely to be published. This state of affairs persists, despite recent calls for more replication studies (e.g. Moonesinghe et al. (2007)). Indeed, replication studies are still often rejected on the grounds of "lack of novelty"; see Makel et al. (2012), Yeung (2017) and Martin & Clarke (2017). As such, researchers deciding which hypotheses to pursue towards publication will likely emphasize those with lower *psp*.

Recognize that the lower the *psp*, the less likely a "statistically significant" finding is to be true. As such, we are bound to a "seemingly inescapable trade-off" (Fiedler 2017) between the novel and the reliable. Journal editors face a difficult choice. Either publish studies that are

For a fixed resources, T, and a given (psp, pwr), the expected number of...	**Equation (note: throughout this paper, we take $B = 0$)**
True Positives published	$TP_{PUB}(psp, pwr) = psp \cdot n_S \cdot A \cdot Pr(TP)$
True Positives unpublished	$TP_{UN}(psp, pwr) = psp \cdot n_S \cdot (1 - A) \cdot Pr(TP)$
False Negatives published	$FN_{PUB}(psp, pwr) = psp \cdot n_S \cdot B \cdot Pr(FN)$
False Negatives unpublished	$FN_{UN}(psp, pwr) = psp \cdot n_S \cdot (1 - B) \cdot Pr(FN)$
False Positives published	$FP_{PUB}(psp, pwr) = (1 - psp) \cdot n_S \cdot A \cdot Pr(FP)$
False Positives unpublished	$FP_{UN}(psp, pwr) = (1 - psp) \cdot n_S \cdot (1 - A) \cdot Pr(FP)$
True Negatives published	$TN_{PUB}(psp, pwr) = (1 - psp) \cdot n_S \cdot B \cdot Pr(TN)$
True Negatives unpublished	$TN_{UN}(psp, pwr) = (1 - psp) \cdot n_S \cdot (1 - B) \cdot Pr(TN)$
Publications	$ENP(psp, pwr) = TP_{PUB} + FN_{PUB} + FP_{PUB} + TN_{PUB}$

Table 1: Equations for the expected number of studies (out of a total of n_S studies) for each of the eight categories; with A = prob. of publication for a positive result and B = prob. of publication for a negative result. The number of studies n_S, changes for different values of pwr. We have that: $n_S = k + n(pwr)^{-1}T$, where $n(pwr)$ is the required sample size to obtain a power of pwr.

surprising and exciting yet most likely false, or publish reliable studies which do little to advance our knowledge. Based on a belief that both ends of this spectrum are equally valuable, Higginson & Munafò (2016) conclude that, in order to increase reliability, current incentive structures should be redesigned, "giving less weight to strikingly novel findings." This is in agreement with the view of Hagen (2016) who writes: "If we are truly concerned about scientific reproducibility, then we need to reexamine the current emphasis on novelty and its role in the scientific process."

2.3 Model Framework

We describe our model framework in 5 simple steps.

(1) We assume, for simplicity, that all studies test a null hypothesis of equal population means against a two-sided alternative, with a standard two-sample Student t-test. Each study has an equal number of observations per sample ($n_1 = n_2$; $n = n_1 + n_2$). Furthermore, let us assume

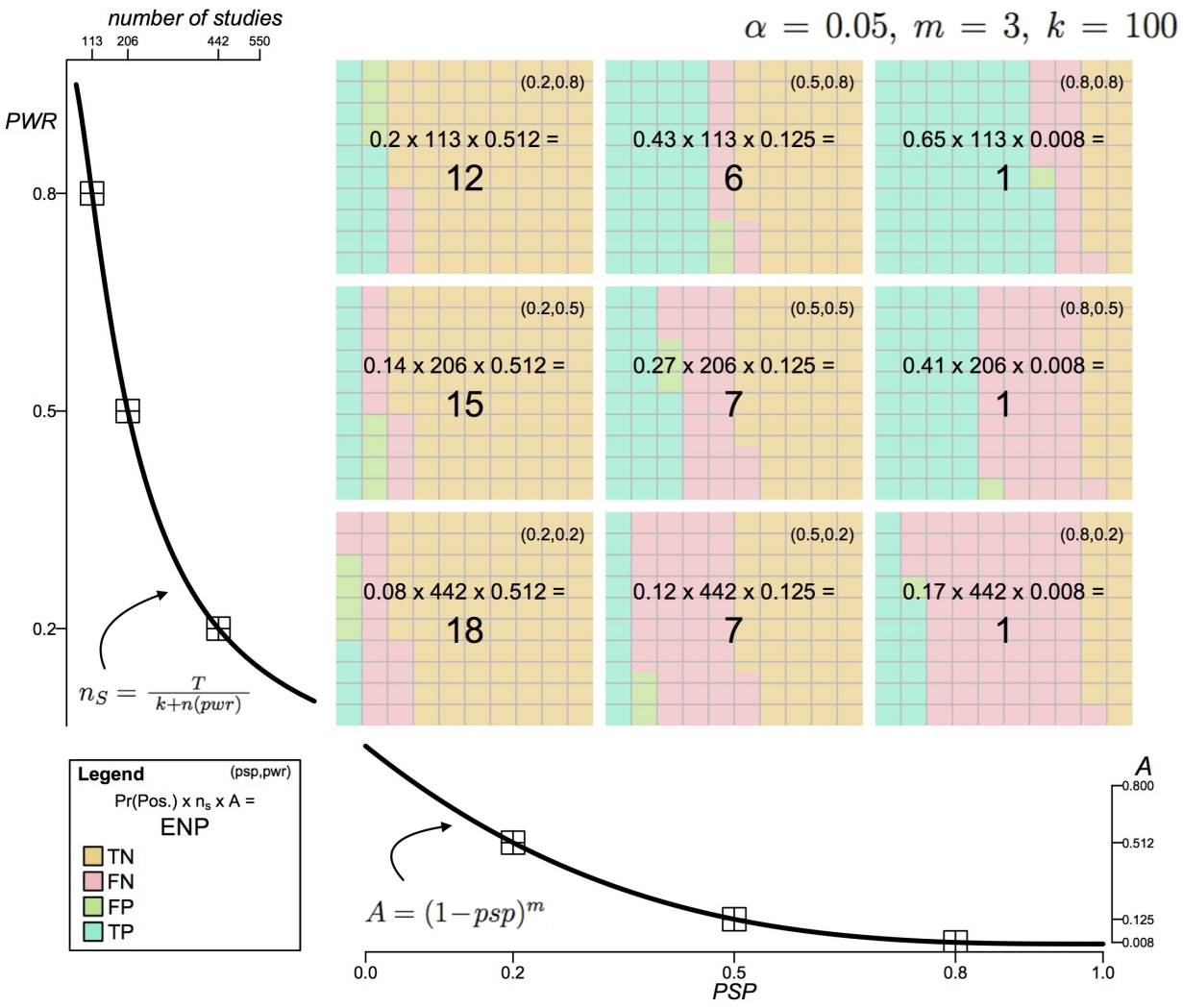

Figure 1: The nine gridded squares represent nine different research strategies; from left to right, we have $psp = 0.2$, 0.5, and 0.8; from bottom to top, we have $pwr = 0.2$, 0.5, and 0.8. For each pair of values for (psp, pwr), the different colored areas represent the probabilities of a true positive (TP), a false positive (FP), a true negative (TN), and a false negative (FN). The Expected Number of Publications (ENP) is calculated by multiplying (1) the probability of a positive finding ($Pr(Positive) = psp \cdot Pr(TP) + (1-psp) \cdot Pr(FP)$), (2) the number of studies ($n_S = \{k + n(pwr)\}^{-1}T$), and (3) the probability of publication ($A = (1 - psp)^m$). The figure corresponds to an ecosystem defined by $\alpha = 0.05$, $m = 3$, $k = 100$, $B = 0$, $\mu = 0.20$, $\sigma^2 = 1$, and $T = 100,000$.

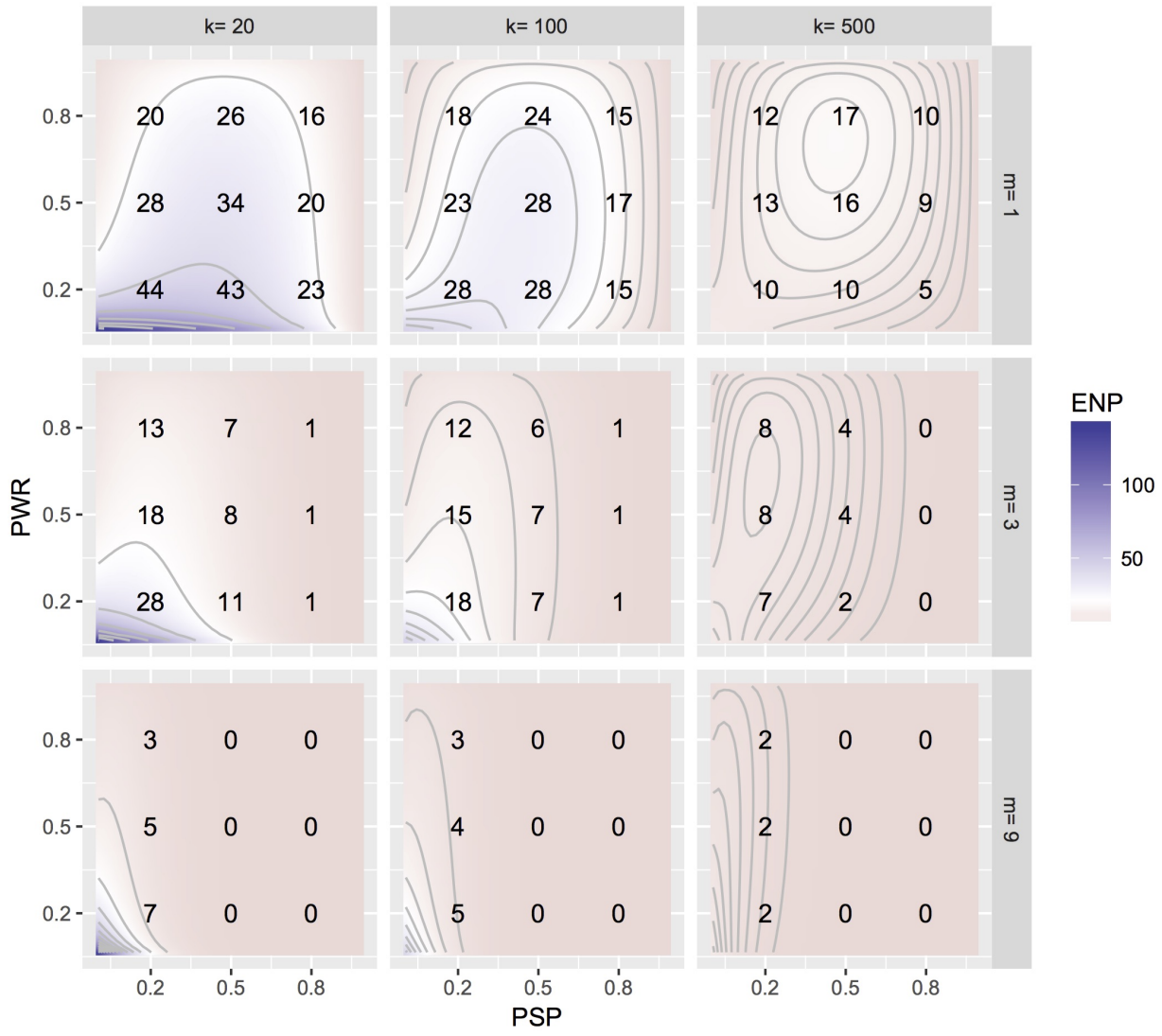

Figure 2: Plots show the number of expected publications, ENP, for different values of PSP and PWR. The printed numbers correspond to values of ENP at $psp = 0.2$, 0.5, and 0.8; and $pwr = 0.2$, 0.5, and 0.8. Each panel represents one ecosystem defined by $\alpha = 0.05$, $A = (1 - psp)^m$, and $B = 0$, with k and m as indicated by column and row labels respectively. Depending on k and m, the value of (PSP, PWR) that maximizes ENP can change substantially. With a higher k, higher-powered strategies will yield a greater ENP; with smaller m, optimal strategies are those with higher PSP. Assuming that researchers behave based on optimizing the use of their resources, it is interesting to observe how the"optimal strategy" changes under different scenarios.

that a study can have one of only two results: (1) *positive* (p-value $\leq \alpha$), or (2) *negative* (p-value $> \alpha$). Given the true effect size, $\mu_d = \mu_2 - \mu_1$ (the difference in population means), σ^2, the common population variance, and the sample size, n, we can easily calculate the probability of a significant result, using the standard formula for power.[1]

Then, for a given true effect size of δ or zero, we have the probability of a True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN), equal to: $Pr(TP) = Pr(Positive|\mu_d = \delta)$, $Pr(FN) = Pr(Negative|\mu_d = \delta)$, $Pr(FP) = Pr(Positive|\mu_d = 0)$, and $Pr(TN) = Pr(Negative|\mu_d = 0)$, respectively.

(2) Next we consider a large number of studies, n_S, each with a total sample size of n. Of these n_S studies, only a fraction, *psp* (where *psp* is the pre-study probability), have a true effect size of $\mu_d = \delta$. For the remaining $(1 - psp) \cdot n_S$ studies, we have $\mu_d = 0$. Note that for a given sample size, n, these n_S studies are each "powered" at level $pwr = Pr(TP)$. Throughout this paper, we focus on scenarios where $\sigma^2 = 1$ and $\delta = 0.20$ or $\delta = 0.43$. These choices reference the analyses of Lamberink et al. (2018) and Richard et al. (2003). Lamberink et al. (2018), based on an empirical analysis of over one hundred-thousand clinical trials conducted between 1975 and 2014, estimate that the median effect size of clinical trials is approximately a Cohen's $d = 0.20$. In a similar analysis based on data from over 25,000 social/personality studies, Richard et al. (2003) estimate that the mean effect size in social psychology research is of a Cohen's $d = 0.43$.

(3) We also label each study as either published (PUB) or unpublished (UN) for a total of 8 distinct categories (= 2 (positive, negative) x 2 (true and false) x 2 (published and unpublished)). One can determine the expected number of studies (out of a total of n_S studies) in each category by simple arithmetic. Table 1 lists the equations for each of the eight categories with A equal to the probability of publication for a positive result, and B equal to the probability of publication for a negative result. Throughout this paper, we will always assume that only positive studies are published, hence, we set $B = 0$. Initially, we will let the probability of publication of a positive study, A, depend on the *psp* according to the function $A = (1 - psp)^m$, where $m \in (0, \infty)$ is a

[1] The probability of a positive result is equal to:

$$Pr(Positive|\mu_d) = (1 - F_{n-2,\frac{\mu_d}{\sigma^*}}(t^*_{\alpha/2})) + F_{n-2,\frac{\mu_d}{\sigma^*}}(-t^*_{\alpha/2}), \quad (1)$$

where $t^*_{\alpha/2}$ is the upper $100 \cdot \frac{\alpha}{2}$-th percentile of the t-distribution with $n - 2$ degrees of freedom, $\sigma^* = \sigma\sqrt{(1/n_1 + 1/n_2)}$, and $F_{df,ncp}(x)$ is the cdf of the non-central t distribution with df degrees of freedom and non-centrality parameter ncp. We calculate the minimum required sample size, n, to obtain a desired power, pwr, as follows:

$$n = argmin_n(|pwr - (1 - F_{n-2,\frac{\mu_d}{\sigma^*}}(t^*_{\alpha/2})) + F_{n-2,\frac{\mu_d}{\sigma^*}}(-t^*_{\alpha/2})|) \quad (2)$$

tuning parameter. This simple, decreasing function of *psp* represents a system in which positive results with lower *psp* are more likely to be published on the basis of novelty. A higher value of m indicates that higher *psp* studies (i.e. low risk hypotheses) are less likely to be published. Consequently, a higher value of m implies a lower overall publication rate. In Section 4, we will define A differently, such that the probability of publication depends on the values of both *psp* and *pwr*.

(4) We determine the total number of studies, n_S, based on three parameters: T, the total resources available (in units of observations sampled); k, the fixed cost per study (also expressed in equivalent units of observations sampled); and n, the total sample size per study. Consequently, as in Higginson & Munafò (2016), $n_S = (k+n)^{-1}T$. Then for any given level of power, *pwr*, we can easily obtain the necessary sample size per study, n, (using established sample size formula for two-sample t-tests, see equations 1 and 2), and a resulting total number of studies, n_S. Keep in mind that n_S, the number of studies, is a function of *pwr*. Throughout this paper, when necessary, we take $T = 100,000$. However, note that when comparing the outcomes of different publication policies, this choice is entirely irrelevant. Consider that when k is small, the cost of data, relative to the total cost of a study, is large. Conversely, when k is large, the relative cost of increasing the study sample size is small. For example, suppose k=20 and $\delta = 0.20$, then increasing the power from 0.55 to 0.85, implies doubling the total cost of the study. If instead $k = 500$ (all else being equal), this increase in power corresponds to only a 50% increase in the total study cost.

(5) Finally, let us define a "research strategy" to be a given pair of values for (*psp*, *pwr*) within ([0,1] x [α,1]). Then, for a given research strategy we can easily calculate the total expected number of publications (ENP):

$$ENP(psp, pwr) = TP_{PUB} + FP_{PUB} + TN_{PUB} + FN_{PUB}. \tag{3}$$

Figure 1 illustrates how ENP is calculated for different values of *psp* and *pwr*, with fixed $\alpha = 0.05$, $k = 100$ and $m = 3$.

With this set-up in hand, suppose now a researcher pursues -consciously or unconsciously- strategies that maximize the expected number of publications (Charlton & Andras 2006). This may not be an entirely unreasonable assumption considering the influence of Goodhart's Law ("Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes") on academia, see Fire & Guestrin (2018), van Dijk et al. (2014). Figure 2 shows the analytical calculations of how ENP changes over a range of values of *psp* and *pwr* and

under different fixed values for k and m. (Note how the central panel of Figure 2 ($k = 100$, $m = 3$) corresponds to Figure 1.) Depending on k and m, the value of (psp, pwr) that maximizes ENP can change substantially. With larger k, higher-powered strategies will yield a greater ENP; with smaller m, optimal strategies are those with higher psp. It is interesting to observe how the optimal strategy changes under different scenarios. However, it may be more informative to consider a *distribution* of preferred strategies. This may also be a more realistic approach. While rational researchers may be drawn toward optimal strategies, surely scientists are not willing and/or able to precisely identify these.

Taking the incentive to publish as a starting point, we assume scientists are more likely to spend resources on studies whose psp and pwr produce a higher expected number of publications. So we consider a *distribution* of resource allocation across study characteristics whose density is proportional to the expected number of publications, i.e., $f_{RES}(psp, pwr) \propto ENP(psp, pwr)$. We emphasize that from a funder's viewpoint, $f_{RES}()$ describes where dollars are being spent on the (psp, pwr) plane. For example, consider the nine strategies illustrated in Figure 1. Given the relative ENP values, we expect the expenditure on $(psp, pwr) = (0.2, 0.2)$ studies to be 18 times greater than that on $(psp, pwr) = (0.8, 0.5)$ studies.

In the Appendix we give expressions for two distributions that follow on from $f_{RES}()$. The first is $f_{ATM}(psp, pwr)$ which describes where *attempted* studies (not resources) fall on the (psp, pwr) plane, presuming that resources are allocated according to $f_{RES}()$. Since higher-powered studies are most costly, $f_{RES}()$ and $f_{ATM}()$ are not the same. Hence $f_{ATM}()$ describes what kinds of studies are attempted more frequently.

Of course not all attempted studies are published. Thus in the Appendix we also give an expression for $f_{PUB}(pwp, pwr)$, the density of (psp, pwr) across *published* studies that is implied when $f_{ATM}()$ describes the characteristics of attempted studies. Note then that each of $f_{RES}()$, $f_{ATM}()$, and $f_{PUB}()$ describes relatively favored and disfavored (psp, pwr) combinations. However, the distinction between the three is important. While $f_{RES}()$ describes resource deployment, $f_{ATM}()$ describes the resulting constellation of attempted studies, and $f_{PUB}()$ describes the resulting constellation of published studies.

Armed with $f_{RES}()$, $f_{ATM}()$, and $f_{PUB}()$, we can investigate the properties of a given scientific ecosystem, and how these properties vary across ecosystems. Specifically, an ecosystem is specified by choices of α, k, m, A and B. For any specification, properties of the three distributions are readily computed via two-dimensional numerical integration using a fine (200 by 200) grid of

(psp, pwr) values. We do not require any simulation to obtain our results, (except for those results relating to the accuracy of published research, see details in the Appendix). To illustrate, consider the much coarser 3 by 3 grid plotted in Figure 1 and see how ENP could be easily calculated by summation across the two dimensions.

2.4 Ecosystem Metrics

We will evaluate the merits of different publication policies on the basis of the following six ecosystem metrics of interest. In the Appendix, we provide further details on these metrics and introduce some compact notation that will be useful for their calculation from $f_{RES}()$ and its by-products.

1. **The Publication Rate (PR).** The ratio of the number of published studies over the number of attempted studies is of evident interest.

2. **The Reliability (REL).** The proportion of published findings that are correct is a highly relevant metric for the scientific ecosystem. Ideally, we wish to see a literature with as few false positives as possible.

3. **The Breakthrough Discoveries (DSCV)**. The ability of the scientific ecosystem to produce breakthrough findings is an important attribute. Here, a true breakthrough result is defined as a true positive and published study that results from a *psp* value below a threshold of 0.05. Note that we will calculate the total number of true breakthrough discoveries, $DSCV$, for each ecosystem for T units of resources. We will also note the reliability of breakthrough discoveries (D_{REL}), the proportion of positive and published studies with $psp < 0.05$ that are true.

4. **The Balance Between Exploration and Confirmation ($IQpsp_{PUB}$).** The interquartile range of the distribution of *psp* values amongst published studies provides an assessment of the much discussed balance between exploratory and confirmatory research; see Sakaluk (2016) and Kimmelman et al. (2014).

5. **The Median Power of Published Studies ($Mpwr_{PUB}$).** As mentioned earlier, higher powered research leads to fewer yet more reliable publications.

6. **The Accuracy of Published Research.** We will report the mean absolute estimated effect size in published studies, $|\hat{\mu_d}|_{PUB}$, as well as the proportion of published studies

for which the estimated effect size is negative ($\hat{\mu}_{d_{PUB}}^{neg}$). It is well known that, due to a combination of publication bias and low powered studies, published effect sizes will be inflated, see Ioannidis (2008), Lane & Dunlap (1978). In addition, we will report the exaggeration ratio (the expected Type M error), and the Type S error rate (the proportion of published studies with an error in *sign*), see Gelman & Carlin (2014).

2.5 Model calibration

Let us begin by assessing whether or not our basic model framework is well calibrated. Are our modeling choices (specifically the choice of values to consider for parameters k and m) at all consistent with what is known about real conditions? In order to answer this important question, we considered a large number of different scenarios in which α is held fixed at 0.05, and both the fixed-cost parameter, k, and the novelty parameter, m, take various values; see Table 2.

There is a recognized lack of empirical research measuring the relationship between experimental costs and sample sizes (Roos 2017), and it is even more difficult to empirically access appropriate values for m. However, we were able to choose suitable values for m (= 1, 3, and 9) and k (= 20, 100, and 500) based on how our model outputs compared with empirical estimates in the meta-research literature. Specifically, values were chosen to exhibit a sufficiently wide range and so that the (1) reliability, (2) the power, and (3) the publication rate agreed, at least to a certain extent, with what has been observed empirically. Consider the following.

1. Due to the success of several recent large-scale reproducibility projects, we are able to obtain reasonable approximations for the *reliability* of published research in different scientific fields. For example, Begley & Ioannidis (2015) estimate that the reliability (i.e., true-positive rate) of social science experiments published in the journals *Nature* and *Science* is approximately 67%. Based on the large-scale reproducibility project of OpenScienceCollaboration (2015), Wilson & Wixted (2018) estimate that the reliability of published studies in well-respected social-psychology journals is approximately 49%, and in cognitive psychology journals, approximately 81%. For the chosen values of k and m, our model produces reliability measures ranging from 15% to 83%, for $\delta = 0.2$, and from 25% to 86%, for $\delta = 0.43$. Note that this wide range of values includes some reliability percentages that seem rather low (e.g., 15% and 25%).

2. There have also been a number of recent attempts to estimate the typical *power* of published

research using meta-analytic effect sizes. For example, Button et al. (2013*b*) estimate that the median statistical power in published neuroscience is approximately 21% and Dumas-Mallet et al. (2017), using a similar methodology, conclude that the median statistical power in the biomedical sciences ranges from about 9% to 30% depending on the disease under study. Most recently, Lamberink et al. (2018) analyzed the literature of RCTs and determined that the median power is approximately 9% overall (and about 20% in the subset of RCTs included in significant meta-analyses). For the chosen values of k and m, our model produces a simulated published literature with median power ranging from 9% to 60%, for $\delta = 0.2$, and from 21% to 71%, for $\delta = 0.43$.

3. While the *publication rate* numbers we obtain may appear rather low, ranging from 3% to 13%, consider that Siler et al. (2015), in a systematic review of manuscripts submitted to three leading medical journals, observed a publication rate of 6.2%. In a review of top psychology journals, Lee & Schunn (2011) found that acceptance rates ranged between 14% and 32%. It is important to recognize that these empirical estimates are calculated based on the subset of studies that are submitted for publication. If researchers are self-selecting their "most publishable" work for submission to journals, then these estimates will substantially overestimate the true publication rate of all studies. Song et al. (2014) estimate that about 85% of unpublished studies are unpublished for the simple reason that they are never submitted for publication. Our model also assumes that negative studies are never published ($B = 0$) and this (not entirely realistic) assumption also contributes to the low publication rate numbers we obtain. Senn (2012) provides substantial insights on both these related issues: the level of publication bias, and how researchers may choose to submit research based on the perceived probability of acceptance. Note that while the scenarios with higher publication rates may appear more plausible, these scenarios have very (perhaps unreasonably) high values for reliability and median power.

Regardless of the values chosen for k and m, the overall trends for our metrics of interest will be similar. See Table 2. As k increases, the reliability of published research (REL) will increase, while the number of true breakthrough discoveries ($DSCV$) decreases. As m increases, reliability (REL) decreases while the number of true breakthrough discoveries ($DSCV$) increases. This strikes us as both reasonable and realistic. When a greater emphasis is placed on novelty (i.e., when m is larger), there will be a greater number of smaller, high-risk studies published. While these publications are less reliable, they are more numerous and more likely to produce a

breakthrough finding. When data is more affordable relative to overall study cost (i.e., when k is larger), there will be fewer, larger studies and as a result the published literature will be more reliable but less likely to produce a breakthrough finding.

Before moving on, let us also briefly consider how the accuracy of published effect sizes changes with k and m, see Table 3. Overall, we see that the average of the absolute published effect size, $|\hat{\mu_d}|_{PUB}$, is always much larger than δ. Furthermore, we see that a large proportion of published effect sizes are negative. This is to be expected when *psp* values are small and many published studies are false positives. For the subset of studies that are positive, published, *and true*, we see that the Type M and the Type S errors are greatest when k is small, m is large, and δ is small. This is the same configuration that produces low powered publications. Indeed, Type M and Type S errors are known to occur more frequently when power is low (Gelman & Carlin 2014).

3 The effects of adopting lower significance thresholds

In this section, we investigate the impact of adopting lower significance thresholds. *How might the published literature be different if the* $\alpha = 0.05$ *threshold was lowered?* For positive studies, we will assume that the sample size of a study does not affect the likelihood of publication (at least not directly) and that studies with lower *psp* are more likely to be published according to the simple function introduced earlier, $A = (1 - psp)^m$. We compute the metrics of interest for 54 different ecosystems. Each ecosystem is uniquely defined with either $\delta = 0.2$ or $\delta = 0.43$, with one of three possible values for m (=1, 3, 9), with one of three possible values for k (=20, 100, 500), and most importantly, with one of three possible values for the α significance threshold (= 0.005, 0.020, 0.050).

3.1 Results

In Table 4, we note how the various metrics change with $\alpha = 0.005$ relative to $\alpha = 0.05$, by reporting ratios. While we focus our discussion specifically on scenarios for which $\delta = 0.20$, the results for $\delta = 0.43$ are all very similar. Figures in the Supplemental Material plot the complete results. Based on our results, we can make the following conclusions on the impact of adopting a lower, more stringent, significance threshold.

δ	k	m	PR	REL	$DSCV$	D_{REL}	$Mpwr_{PUB}$	$IQpsp_{PUB}$	$[q_{25th}, q_{75th}]$
0.20	20	1	0.06	0.62	0.17	0.07	0.20	0.35	[0.19 , 0.54]
0.20	20	3	0.03	0.36	0.38	0.07	0.12	0.19	[0.07 , 0.26]
0.20	20	9	0.03	0.15	0.86	0.06	0.09	0.07	[0.02 , 0.09]
0.20	100	1	0.08	0.75	0.09	0.12	0.42	0.33	[0.25 , 0.58]
0.20	100	3	0.04	0.52	0.22	0.12	0.32	0.20	[0.09 , 0.29]
0.20	100	9	0.03	0.25	0.51	0.10	0.21	0.08	[0.03 , 0.10]
0.20	500	1	0.11	0.83	0.04	0.18	0.60	0.31	[0.29 , 0.60]
0.20	500	3	0.05	0.65	0.10	0.18	0.55	0.21	[0.12 , 0.33]
0.20	500	9	0.03	0.36	0.26	0.16	0.44	0.09	[0.03 , 0.12]
0.43	20	1	0.08	0.76	0.42	0.12	0.42	0.33	[0.25 , 0.58]
0.43	20	3	0.04	0.53	1.00	0.12	0.32	0.20	[0.09 , 0.29]
0.43	20	9	0.03	0.25	2.37	0.11	0.21	0.08	[0.03 , 0.10]
0.43	100	1	0.11	0.83	0.21	0.18	0.59	0.31	[0.29 , 0.60]
0.43	100	3	0.05	0.65	0.50	0.18	0.54	0.21	[0.12 , 0.33]
0.43	100	9	0.03	0.36	1.24	0.16	0.43	0.08	[0.03 , 0.11]
0.43	500	1	0.13	0.86	0.07	0.23	0.71	0.31	[0.30 , 0.61]
0.43	500	3	0.06	0.71	0.17	0.22	0.68	0.21	[0.13 , 0.34]
0.43	500	9	0.03	0.43	0.44	0.20	0.60	0.09	[0.03 , 0.12]

Table 2: For ecosystems defined by fixed $\alpha = 0.05$, $A = (1-psp)^m$, and $B = 0$, and varying values of k, m, and δ, the table lists estimates for the metrics of interest. These include the publication rate (PR), the reliability (REL), the number of true breakthrough discoveries ($DSCV$), the median power of published studies ($Mpwr_{PUB}$), and the interquartile range of the distribution of psp values amongst published studies ($IQpsp_{PUB}$), with corresponding 25th and 75th quartiles. Additionally, the table lists the reliability amongst the subset of published studies for which $psp < 0.05$ (D_{REL}).

1. Reliability is substantially increased with a lower threshold. Based on our results, comparing $\alpha = 0.005$ to $\alpha = 0.05$, the impact on REL is greatest when δ is small, k is small, and m is large, see Table 4. This is due to the fact that with a lower significance threshold policy, attempted studies are typically of higher power (particularly so when k is small) and of higher pre-study probability. To understand why there would be more studies with higher power and higher pre-study probabilities, consider that with a lower α threshold, the probability of obtaining a significant result (i.e., p-value $< \alpha$) "by chance" is substantially

δ	m	k	$Mpwr_{PUB}$	$\hat{\mu_d}^{neg}_{PUB}$	$\lvert\hat{\mu_d}\rvert_{PUB}$	Type M err.	Type S err.
0.20	1	20	0.20	0.21	0.60	2.42	0.03
0.20	3	20	0.12	0.31	0.72	2.70	0.04
0.20	9	20	0.09	0.43	0.80	2.77	0.04
0.20	1	100	0.42	0.14	0.37	1.68	0.01
0.20	3	100	0.32	0.23	0.45	1.81	0.01
0.20	9	100	0.21	0.37	0.51	1.92	0.01
0.20	1	500	0.60	0.09	0.29	1.40	0.00
0.20	3	500	0.55	0.17	0.31	1.40	0.00
0.20	9	500	0.44	0.33	0.35	1.48	0.00
0.43	1	20	0.42	0.14	0.75	1.63	0.01
0.43	3	20	0.32	0.21	0.80	1.65	0.00
0.43	9	20	0.21	0.39	0.90	1.74	0.01
0.43	1	100	0.59	0.09	0.60	1.36	0.00
0.43	3	100	0.54	0.18	0.64	1.43	0.00
0.43	9	100	0.43	0.32	0.69	1.47	0.00
0.43	1	500	0.71	0.08	0.54	1.29	0.00
0.43	3	500	0.68	0.17	0.56	1.29	0.00
0.43	9	500	0.60	0.28	0.57	1.30	0.00

Table 3: For ecosystems defined by fixed $\alpha = 0.05$, $A = (1 - psp)^m$, and $B = 0$, and varying values of k, m, and δ, the table lists estimates for the metrics of interest related to the accuracy of the published effects sizes. These include the median power of published studies ($Mpwr_{PUB}$), the mean absolute published effect size ($\lvert\hat{\mu_d}\rvert_{PUB}$), the proportion of published effect sizes that are negative ($\hat{\mu_d}^{neg}_{PUB}$), the "Type S" error, and the "Type M" error.

reduced. To maximize one's expected number of publications (ENP) when $\alpha = 0.005$, it is a much better strategy to pursue higher *pwr* and higher *psp* strategies. This is particularly true when the effect size, δ, is small.

2. A disadvantage of the lower threshold is that the number of breakthrough discoveries is substantially lower, see Table 4 and Figure S3. This is due to the fact that with a lower α threshold, fewer "high risk" (i.e. low *psp*) studies are attempted. The chance that the p-value will fall below α, when α is lowered to 0.005, is already risky enough. The D_{REL} numbers we obtain suggest that, while fewer low *psp* studies will be published, these will

δ	k	m	Change in PR	Change in REL	Change in $DSCV$	Change in D_{REL}	Change in $Mpwr_{PUB}$	Change in $IQpsp_{PUB}$
0.20	20	1	1.14	1.58	0.12	7.57	2.73	0.83
0.20	20	3	0.71	2.58	0.15	7.68	4.57	1.05
0.20	20	9	0.32	5.38	0.21	8.01	5.11	1.46
0.20	100	1	0.99	1.30	0.19	5.47	1.37	0.86
0.20	100	3	0.73	1.81	0.23	5.57	1.77	0.95
0.20	100	9	0.39	3.47	0.31	5.89	2.52	1.27
0.20	500	1	0.91	1.18	0.28	4.06	1.07	0.89
0.20	500	3	0.75	1.48	0.33	4.14	1.17	0.95
0.20	500	9	0.46	2.49	0.43	4.40	1.42	1.12
0.43	20	1	0.99	1.30	0.19	5.49	1.36	0.86
0.43	20	3	0.74	1.81	0.22	5.59	1.78	0.95
0.43	20	9	0.39	3.46	0.31	5.91	2.55	1.27
0.43	100	1	0.92	1.18	0.28	4.10	1.08	0.89
0.43	100	3	0.75	1.49	0.32	4.18	1.19	0.95
0.43	100	9	0.46	2.51	0.42	4.44	1.44	1.19
0.43	500	1	0.92	1.14	0.37	3.49	1.02	0.90
0.43	500	3	0.78	1.37	0.42	3.55	1.07	0.95
0.43	500	9	0.51	2.15	0.55	3.78	1.19	1.12

Table 4: A comparison between ecosystems with $\alpha = 0.05$ and ecosystems with $\alpha = 0.005$. For all ecosystems, we have $A = (1 - psp)^m$, $B = 0$, and varying values of k, m, and δ. The table lists estimates for the ratios of the publication rate (PR), the ratio of the reliability (REL), the ratio of the number of true breakthrough discoveries ($DSCV$), the ratio of the median power of published studies($Mpwr_{PUB}$), and the ratio of the interquartile range of the distribution of psp values amongst published studies ($IQpsp_{PUB}$).

be much more reliable.

3. When increasing study power is less costly relative to the total cost of a study (i.e., when k is larger, or when δ is larger), the benefit of lowering the significance threshold (increased REL) is somewhat smaller. However, the downside (decreased $DSCV$) is *substantially* smaller, see left-panels of Figures S1 and S3. This suggests that a policy of lowering the significance threshold would perhaps be best suited in a field of research in which increasing one's sample size is less burdensome. This nuance recalls the suggestion of Ioannidis et al.

(2013): "Instead of trying to fit all studies to traditionally acceptable type I and type II errors, it may be preferable for investigators to select type I and type II error pairs that are optimal for the truly important outcomes and for clinically or biologically meaningful effect sizes."

4. When novelty is more of a requirement for publication (i.e., when m is larger), the benefit of lowering the significance threshold is larger and the downside smaller. This result is due to the fact that a smaller α will incentivize researchers to allocate resources in the direction towards *either* higher-powered *or* higher *psp* studies (i.e., away from the South-West corner of the plots in Figure 2). There is a choice between moving towards higher *pwr* (North) or towards higher *psp* (East), and different costs associated with each direction. This suggests that for a lower significance threshold policy to be most effective, editors should also adopt, in conjunction, stricter requirements for research novelty. To illustrate, consider three ecosystems of potential interest with their estimated REL, $DSCV$ and PR metrics (all with $\delta = 0.2$):
   (1) The *baseline* defined by $\alpha = 0.05$, $m = 3$, and $k = 500$ with:
   $REL = 0.650$, $DSCV = 0.105$, and $PR = 0.054$;
   (2) the *alternative* defined by $\alpha = 0.005$, $m = 3$, and $k = 500$ with:
   $REL = 0.963$, $DSCV = 0.034$, and PR =0.040; and
   (3) the *suggested* defined by $\alpha = 0.005$, $m = 9$, and $k = 500$ with:
   $REL = 0.898$, $DSCV = 0.112$, and PR =0.015.
   Note that the while the *suggested* has high REL and relatively high DSCV, the PR is substantially reduced.

5. As mentioned earlier, the balance between exploratory and confirmatory research is an important aspect of a scientific ecosystem. The results show that the interquartile range for *psp* does not change substantially with α. As such, we could conclude that even with a much lower significance threshold, there will still be a wide range of studies attempted in terms of their *psp*. However, *psp* values do tend to be substantially higher with smaller α. As such, we should expect that, with smaller α, research will move towards more confirmatory, and less exploratory studies.

6. With a lower α significance threshold, the published effect sizes are much more accurate. When $\alpha = 0.005$, the Type M error is reduced to at most 1.4 (i.e., effect size estimates are

inflated by at most 40%) and Type S error is negligible; see Figures 4, 5, S7 and S8. These reductions are greatest when data are relatively expensive and when the true effect size is small (i.e. when $k = 20$ and $\delta = 0.20$).

4 The effects of strict *a priori* power calculation requirements

In this section, we investigate the effects of requiring "sufficient" sample sizes. In practical terms, this means adopting publication policies that require studies to show *a priori* power calculations indicating that sample sizes are "sufficiently large" to achieve the desired level of statistical power, typically 80%. Whereas before, the chance of publishing a positive study in our framework depended only on *psp*, we now set the probability of publication, A, to also depend on power. In doing so however, we wish to acknowledge the fact that *a priori* sample size claims are often "wildly optimistic" (Bland 2009).

The "sample size samba" (Schulz & Grimes 2005) –the practice of "tweaking aspects of sample size calculation" (Hazra & Gogtay 2016) in order to obtain what is affordable– often results in a study with less than 80% power being advertised as having 80% power. Even in ideal circumstances, power can be exaggerated due to "optimism bias" (Djulbegovic et al. 2011), also known as the "illusion of power" (Vasishth & Gelman 2017), which occurs when the anticipated effect size is based on a literature filled with overestimates due to publication bias.

To take into account the problematic nature of *a priori* power calculations, our model is defined such that possessing only 50% power substantially reduces, but does not eliminate, the chance of publication. For studies which really do have 80% power and higher, there will be no notable reduction in the probability of publication. See the Appendix for details on how we define A as a continuous function of both *psp* and *pwr*.

We calculated the metrics of interest for the same 54 different ecosystems as in Section 3, with our new definition of A. To contrast these ecosystems with those discussed in the previous section, we refer to these ecosystems as "with SSR" (sample size requirements).

δ	k	m	Change in PR	Change in REL	Change in $DSCV$	Change in D_{REL}	Change in $Mpwr_{PUB}$	Change in $IQpsp_{PUB}$
0.20	20	1	1.46	1.44	0.32	3.97	3.57	0.87
0.20	20	3	1.04	2.08	0.35	4.01	6.13	1.08
0.20	20	9	0.65	3.33	0.41	4.12	7.61	1.38
0.20	100	1	1.21	1.18	0.53	2.40	1.74	0.92
0.20	100	3	1.03	1.44	0.56	2.42	2.25	1.00
0.20	100	9	0.77	2.03	0.63	2.48	3.33	1.20
0.20	500	1	1.08	1.07	0.75	1.63	1.27	0.95
0.20	500	3	1.00	1.18	0.79	1.64	1.39	1.00
0.20	500	9	0.85	1.44	0.85	1.67	1.68	1.06
0.43	20	1	1.21	1.18	0.53	2.39	1.72	0.92
0.43	20	3	1.03	1.44	0.56	2.41	2.25	1.00
0.43	20	9	0.77	2.02	0.63	2.47	3.33	1.20
0.43	100	1	1.08	1.07	0.74	1.65	1.29	0.95
0.43	100	3	1.00	1.19	0.78	1.66	1.41	1.00
0.43	100	9	0.85	1.45	0.85	1.69	1.72	1.12
0.43	500	1	1.04	1.04	0.89	1.38	1.14	0.97
0.43	500	3	1.00	1.10	0.92	1.38	1.19	1.00
0.43	500	9	0.90	1.25	0.98	1.40	1.32	1.12

Table 5: A comparison between ecosystems with a sample size requirement and ecosystems without. A is defined by equation 7 (with $c_{50} = 0.5$ and $c_{95} = 0.8$), $B = 0$, $\alpha = 0.05$, and varying values of k and m.

4.1 Results

See Table 5. Based on our results, we can make the following main conclusions on the measurable consequences of adopting a journal policy requiring an *a priori* sample size justification.

1. With SSR, we observed much higher powered studies, see Figure S5. Reliability is also increased, particularly when novelty is highly prized (m is large) and effect sizes are small. This result is as expected. As soon as having a small sample size jeopardizes the probability of publication, it is in the researcher's best interest to conduct higher-powered studies.

2. Requiring "sufficient sample sizes" for publication can be quite detrimental in terms of the number of breakthrough discoveries. The impact on $DSCV$ is greatest when δ, k, and m

are all small; see Table 5. The D_{REL} numbers show that reliability is particularly increased for low *psp* publications.

3. In conjunction with requiring larger sample sizes, it may be wise to place greater emphasis on research novelty. As in Section 3, such a combined approach could see an increase in reliability with only a limited decrease in discovery. This trade-off is most beneficial when k is small. Consider three ecosystems of potential interest (all with $\alpha = 0.05$, and $\delta = 0.2$) with their estimated REL, $DSCV$ and PR metrics:
   (1) The *baseline* defined by $m = 3$, $k = 100$, and without SSR; with $REL = 0.524$, $DSCV = 0.216$, $PR = 0.043$;
   (2) the *alternative* defined by $m = 3$, $k = 100$ and with SSR; with $REL = 0.757$, $DSCV = 0.122$, $PR = 0.044$; and
   (3) the *suggested* defined by $m = 6$, $k = 100$ and with SSR; with $REL = 0.502$, $DSCV = 0.325$, $PR = 0.022$.
   Note that the while the *suggested* has both higher REL and higher $DSCV$ than the *baseline*, the PR is reduced.

4. With SSR, the Type M error is at most 1.2 (i.e. effect size estimates are inflated by at most 20%) and Type S error is negligible; see Figures 4, 5, S7 and S8.

4.2 In tandem: The effects of adopting both a lower significance threshold and a power requirement

We are also curious as to whether lowering the significance threshold *in addition* to requiring larger sample sizes would carry any additional benefits relative to each policy innovation on its own. See Table 6 for results, and consider two main findings:

1. For ecosystems with SSR, the distribution of published studies does not change dramatically when α is lowered. Figure 3 shows the density describing the characteristics of published studies, $f_{PUB}(psp, pwr)$, for the four policies of interest, with $k = 500$ and $m = 3$: (1) $\alpha = 0.05$, without SSR; (2) $\alpha = 0.05$, with SSR; (3) $\alpha = 0.005$, without SSR; and (4) $\alpha = 0.005$, with SSR. The difference between the densities in (2) and (4) is primarily a matter of a shift in *psp*.

2. As expected, lowering the significance threshold further increases reliability and the number of breakthrough discoveries is further decreased; see Table 6, and Figures S1 and S3.

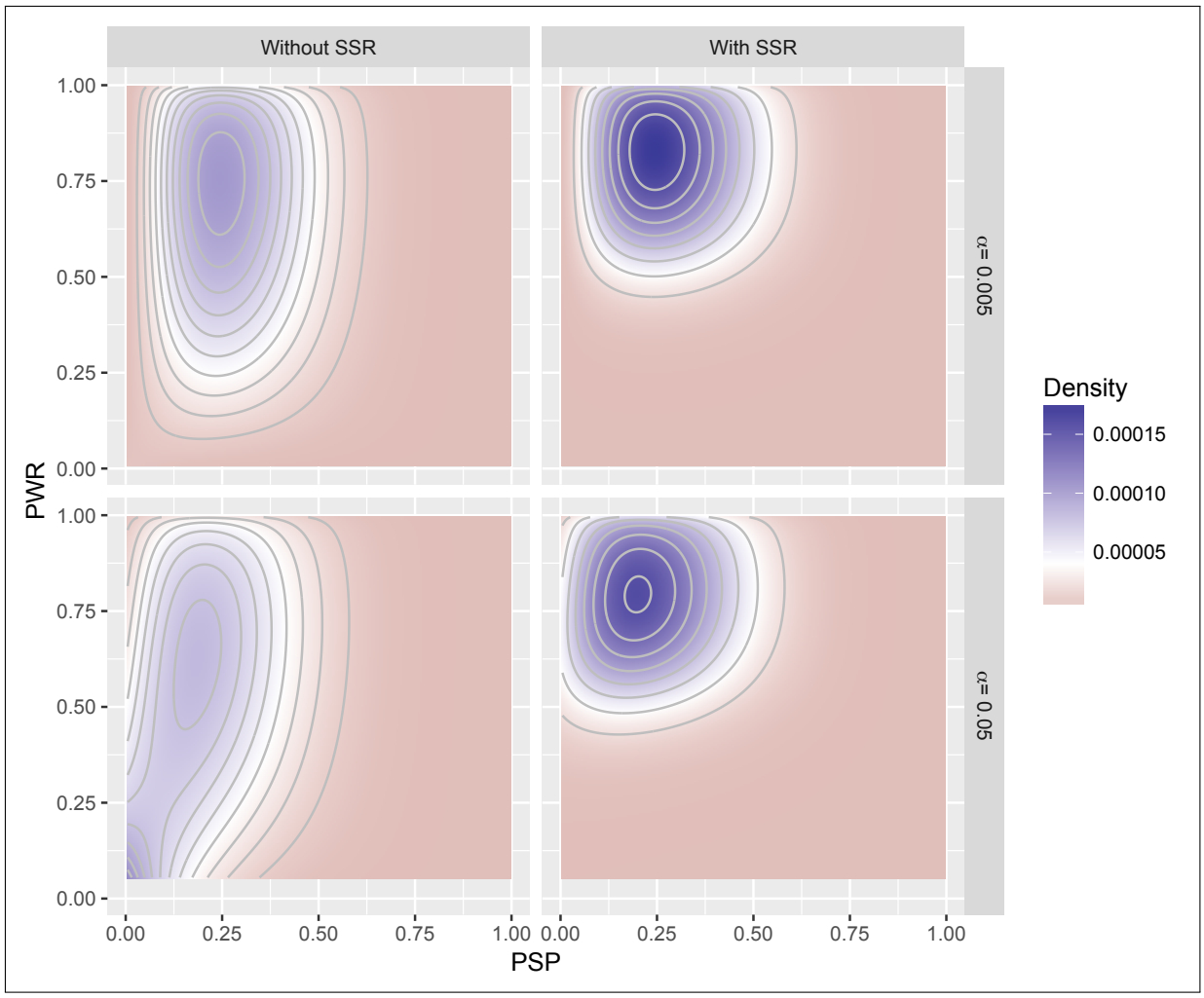

Figure 3: Heat-maps show the density of published papers, $f_{PUB}(psp, pwr)$, for the four policies of interest, with fixed $\delta = 0.2$, $k = 500$ and $m = 3$.

5 Conclusion

There remains substantial disagreement on the merits of requiring greater statistical stringency to address the reproducibility crisis. Yet all should agree that innovative publication policies can be part of the solution. Going forward, it is important to recognize that current norms for

Type 1 and Type 2 error levels have been driven (almost) entirely by tradition and inertia rather than careful coherent planning and result-driven decisions (Hubbard & Bayarri 2003). Hence, improvements should be possible.

In response to Amrhein et al. (2017) who suggest that a more stringent α threshold will lead to published science being *less* reliable, our results suggest otherwise. However, just as Amrhein et al. (2017) contend, our results indicate that the publication rate will end up being substantially lower with a smaller α. While going from $p < 0.05$ to $p < 0.005$ may be beneficial to published science in terms of reliability, we caution that there may be a large cost in terms of fewer true breakthrough discoveries. One must ask whether a large reduction in novel discoveries is an acceptable price to pay for a large increase in their reliability? Importantly, and somewhat unexpectedly, our results suggest that this can be mitigated (to some degree) by adopting a greater emphasis on research novelty. In practice, implementing stricter requirements for research novelty, could be accomplished by greater editorial emphasis on "surprising results" as a requisite for publication. This approach however, might be difficult to achieve unless one is willing to accept a much lower publication rate. In summary, publishing *less* may be the necessary price to pay for obtaining *more* reliable science.

Recently, some have suggested that researchers choose (and justify) an "optimal" value for α, for each unique study; see Mudge et al. (2012), Ioannidis et al. (2013) and Lakens et al. (2018). Each study within a journal would thereby have a different set of criteria. This is a most interesting idea and there are persuasive arguments in favor of such an approach. Still, it is difficult to anticipate how such a policy would play out in practice and how the research incentive structure would change in response. A model, like the one presented here, but with the α threshold set to vary with *psp*, could provide useful insight.

We are also cautious about greater sample size requirements. In summary, we found that the impacts of adopting a sample size requirement policy are similar to the impacts of lowering the α significance threshold. While improving reliability, requiring studies to show "sufficient power" will severely limit novel discoveries in fields where acquiring data is expensive. Is it beneficial for editors and reviewers to consider whether a study has "sufficient power"? How much should these criteria influence publication decisions? Answers to these questions are not at all obvious. Again, using the methodology introduced, we suggest that adopting a greater emphasis on research novelty may mitigate, to a certain extent, some of the downside of adopting greater sample size requirements at the cost of lowering the overall number of published studies. Given that, as a

result of publication bias, it can often be better to discard 90% of published results for meta-analytical purposes (Stanley et al. 2010), this may be an approach worth considering.

Our main recommendation is that, before adopting any (radical) policy changes, we should take a moment to carefully consider, and model how, these proposed changes might impact outcomes. The methodology we present here can be easily extended to do just this. Two scenarios of interest come immediately to mind.

First, it would be interesting to explore the impact of publication bias, i.e., the tendency of journals to reject non-significant results (Sterling et al. 1995). This could be done by allowing B to take different non-zero values. Based on simulation studies, de Winter & Happee (2013) suggest that publication bias can in fact be beneficial for the reliability of science. However, under slightly different assumptions, van Assen et al. (2014) arrive at a very different conclusion. Clearly, a better understanding of how publication bias changes a scientist's incentives is needed. If statistical significance becomes a much less important screening criterion for publication, how will published science change? Recently, a number of influential researchers have argued that to address low reliability, reviewers should "abandon statistical significance" (McShane et al. 2017). While our framework could be extended to explore this proposal (e.g., by taking $A = B$), these researchers suggest substituting statistical significance with a "more holistic view of the evidence" to inform publication decisions. Clearly this does not lend itself to our modeling framework, or similar meta-research paradigms.

Second, it would be worthwhile to investigate the potential impact of requiring study pre-registration. Coffman & Niederle (2015) use the accounting of Ioannidis (2005) to evaluate the effect of pre-registration on reliability and conclude that pre-registration will have only a modest impact. However, the impact on the publication rate and on the number of breakthrough findings is still not well understood. This is particularly relevant given the current trend to adopt "result-blind peer-review" (Greve et al. 2013) policies including, most recently, the policy of *Registered Reports* (Chambers et al. 2015). The conditional equivalence testing publication policy proposed in our earlier work, see Campbell & Gustafson (2018), could also be considered.

Our methodology assumes above all that researchers' decisions are driven exclusively by the desire to publish. But the situation is more complex. Publication is not necessarily the end goal for a scientific study and requirements with regards to significance and power are not only encountered at the publication stage. In the planning stages, before a study even begins, ethics committees and granting agencies will often have certain minimal requirements; see Ploutz-Snyder

et al. (2014) and Halpern et al. (2002). And after a study is published, regulatory bodies and policy makers will also often subject the results to a different set of norms.

We also assumed a framework of independent Bernoulli trials. Of course, in practice, most studies are not conducted independently. For example, in the context of clinical trials, phase 3 studies are conducted based on the success or failure of earlier phase 2 studies (see Burt et al. (2017) who model the advantages and disadvantages of different clinical development strategies). It is difficult to anticipate the consequences of failing to incorporate such dependencies in our model. A more elaborate model, in which studies are correlated and the *psp* of subsequent studies is updated appropriately, could prove very informative.

Finally, it is important to acknowledge that no publication policy will be perfect. Science is inherently challenging and we must always be willing to accept that a certain proportion of research is potentially false (Djulbegovic & Hozo 2007, Contopoulos-Ioannidis et al. 2003). Each policy will have its advantages and disadvantages. Our modeling exercise makes this all the more evident and forces us to carefully consider different potential trade-offs.

Code:

Please note that all code to produce the results, tables and figures in this article have been posted to a repository on the Open Science Framework, DOI 10.17605/OSF.IO/YQCVA.

Acknowledgements:

We wish to gratefully acknowledge Prof. Will Welch and Prof. John Petkau for their valuable suggestions and advice. Furthermore, we wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC grant number RGPIN 183772-13).

A Appendix

Here we introduce some compact notation that will be useful for expressing distributional quantities of interest. Particularly, the probabilities comprising the distribution of a study across the eight categories are expressed as $q_{abc}(psp, pwr)$, where $a \in \{0, 1\}$ indicates the truth ($a = 0$ for null, $a = 1$ for alternative), $b \in \{0, 1\}$ indicates the statistical finding ($b = 0$ for negative, $b = 1$ for positive), and $c \in \{U, P\}$ indicates publication status. As examples, we could write $FN_{UN} = q_{10U}$, or $TP_{PUB} = q_{11P}$. We also use a plus notation to add over subscripts, so, for instance, $q_{1+P} = TP_{PUB} + FN_{PUB}$.

As motivated above, we consider properties that result from a scientist or group of scientists stochastically allocating T *resources* (not studies per se) according to a distribution across (psp, pwr). We denote the function $n(pwr)$ as the required sample size to obtain a power of pwr. Presuming the incentive to publish, the density of this distribution is taken proportional to $ENP(psp, pwr)$, which we express as

$$\begin{aligned} f_{RES}(psp, pwr) &\propto ENP(psp, pwr) \\ &\propto \{k + n(pwr)\}^{-1} q_{++P}(psp, pwr). \end{aligned} \tag{4}$$

Consequently, the distribution of (psp, pwr) across *attempted* studies has density

$$\begin{aligned} f_{ATM}(psp, pwr) &\propto \{k + n(pwr)\}^{-1} f_{RES}(psp, pwr) \\ &\propto \{k + n(pwr)\}^{-2} q_{++P}(psp, pwr). \end{aligned} \tag{5}$$

In turn, the distribution of (psp, pwr) across *published studies* has density

$$\begin{aligned} f_{PUB}(psp, pwr) &\propto f_{ATM}(psp, pwr) q_{++P}(psp, pwr) \\ &\propto \{k + n(pwr)\}^{-2} \{q_{++P}(psp, pwr)\}^2. \end{aligned} \tag{6}$$

Note particularly that $f_{PUB}(psp, pwr) \propto \{f_{RES}(psp, pwr)\}^2$. Hence the distribution of (psp, pwr) across published studies is a concentrated version of the distribution describing how resources are deployed.

A.1 Ecosystem Metrics: further details

We evaluate each ecosystem of interest on the basis of the following five metrics.

A.1.1 Publication Rate and the Number of Studies Attempted/Published

If T units of resources are deployed according to $f_{RES}()$, then we expect that N_{ATM} studies will be attempted, where

$$T^{-1}N_{ATM} \quad = \quad E_{RES}\left[\{k+n(PWR)\}^{-1}\right].$$

Similarly, N_{PUB} studies will be published, with

$$T^{-1}N_{PUB} \quad = \quad E_{RES}\left[\{k+n(PWR)\}^{-1}q_{++P}(PSP,PWR)\right].$$

The ratio N_{PUB}/N_{ATM}, which does not depend on T, is of evident interest, as the *publication rate* (PR) for attempted studies.

A.1.2 The Reliability (REL).

A highly relevant metric for the scientific ecosystem is the proportion of published findings that are correct. In all the ecosystems we consider in this paper, we make the assumption that only positive results are published (i.e., $B = 0$). Therefore, we can express reliability (REL) simply as:

$$REL \quad = \quad \frac{E_{ATM}\left\{q_{11P}(PSP,PWR)\right\}}{E_{ATM}\left\{q_{+1P}(PSP,PWR)\right\}}.$$

More generally, in ecosystems where negative results might be published (i.e., $B(psp,pwr) \neq 0$), the reliability would equal the proportion of published papers that reach a correct conclusion, i.e.,

$$REL \quad = \quad \frac{E_{ATM}\left\{q_{00P}(PSP,PWR)+q_{11P}(PSP,PWR)\right\}}{E_{ATM}\left\{q_{++P}(PSP,PWR)\right\}}.$$

A.1.3 Breakthrough Discoveries (DSCV).

The ability of the scientific ecosystem to produce breakthrough findings is an important attribute. We quantify this in terms of spending T resource units yielding an expectation of $DSCV$ breakthrough results. Here a true breakthrough result is defined as a true positive and published study that results from a *psp* value below a threshold, i.e., a very surprising positive finding that gets published and also happens to be true. If we set the breakthrough threshold as $psp < 0.05$, then

$$T^{-1}DSCV \quad = \quad E_{RES}\left\{I_{(0,0.05)}(PSP)\frac{q_{11P}(PSP,PWR)}{k+n(PWR)}\right\}$$

Note that we can also write this out in terms of the density of attempted studies, f_{ATM}:

$$\begin{aligned} T^{-1}DSCV &= \frac{\int h(PSP,PWR) f_{ATM}(PSP,PWR)(k+n(PWR))}{\int f_{ATM}(PSP,PWR)(k+n(PWR))} \\ &= \frac{E_{ATM}[I_{(0,0.05)}(PSP) q_{11P}(PSP,PWR)]}{E_{ATM}[k+n(PWR)]}. \end{aligned}$$

where we define the function: $h(PSP,PWR) = I_{(0,0.05)}(PSP)\frac{q_{11P}(PSP,PWR)}{k+n(PWR)}$.

We also wish to quantify the reliability of those published studies with *psp* values below a threshold of 0.05. We have that:

$$D_{REL} = \frac{E_{ATM}\left\{I_{(0,0.05)}(PSP) q_{11P}(PSP,PWR)\right\}}{E_{ATM}\left\{I_{(0,0.05)}(PSP) q_{+1P}(PSP,PWR)\right\}}.$$

A.1.4 The Median Power of Published Studies ($Mpwr_{PUB}$).

We already mentioned the relevance of the *psp* marginals of (5) and (6). In a similar vein, the marginal distributions of *pwr* under each of these distributions are readily interpreted metrics of the ecosystem. We define $Mpwr_{PUB}$ as the median of *pwr* under the $f_{PUB}()$ distribution.

A.1.5 The Balance between Exploration and Confirmation ($IQpsp_{PUB}$).

There has been much discussion about the desired balance between researchers looking for *a priori* unlikely relationships versus confirming suspected relationships put forth by other researchers; see for example, Sakaluk (2016) and Kimmelman et al. (2014). The marginal distribution of *psp* arising from $f_{ATM}(psp, pwr)$ describes the balance between exploration versus confirmation for attempted studies, while the *psp* marginal from $f_{PUB}(psp, pwr)$ does the same for published studies. More specifically, we report the interquartile ranges of these marginal distributions for a given ecosystem.

A.1.6 The Accuracy of Published Research.

Consider extending $f_{ATM}(psp, pwr)$ to $f_{ATM}(psp, pwr, d, u)$, where:

$f_{ATM}(d, u|psp, pwr) = f_{ATM}(d|psp, pwr) f_{ATM}(u)$, and where

- $f_{ATM}(u)$ is the density of a standard uniform distribution; and

- $f_{ATM}(d|psp, pwr)$ is the density of a two-component mixture distribution such that:

$Pr(d = \delta) = psp$; and $Pr(d = 0) = 1 - psp$.

In order to obtain a sample from the distribution of published effect sizes, we proceed according to the following steps.

Set $w = 1$, $q = 1$ and $b = 1$. While $w < 1,000$:

1. Draw a Monte Carlo realization of $(psp_{[b]}, pwr_{[b]}, d_{[b]}, u_{[b]})$ from $f_{ATM}(psp, pwr, d, u)$.
2. Simulate a two-sample Normally distributed dataset, $data_{[b]}$, with a sample size of $n(pwr_{[b]})$, a true difference in population means of $\mu_d = d_{[b]}$, and a true variance of $\sigma^2 = 1$.
3. Run a standard t-test on $data_{[b]}$, and obtain the estimated effect size, $\hat{\delta}_{[b]}$, and the two-sided p-value, $pval_{[b]}$.
4. If $u_{[b]} < A(psp_{[b]}, pwr_{[b]})$ AND $pval_{[b]} < \alpha$, then do:
   $\hat{\delta}^{[q]}_{+1P} = \hat{\delta}_{[b]}$; and $q = q + 1$.
5. If $u_{[b]} < A(psp_{[b]}, pwr_{[b]})$ AND $pval_{[b]} < \alpha$, AND $d_{[b]} = \delta$, then do:
   $\hat{\delta}^{[w]}_{11P} = \hat{\delta}_{[b]}$; and $w = w + 1$.
6. Set $b = b + 1$.

Then we have that $\{\hat{\delta}_{+1P}\}$ is a Monte Carlo sample from the conditional distribution of effect sizes given that studies are both positive *and* published. We also have that $\{\hat{\delta}_{11P}\}$ is a Monte Carlo sample from the conditional distribution of effect sizes given that studies are positive, *and* published, *and* true.

Using these Monte Carlo samples, we can easily approximate the following metrics of interest:

- **Mean Absolute Published Effect Size ($|\hat{\mu_d}|_{PUB}$).** Approximated as the mean of $|\{\hat{\delta}_{+1P}\}|$.
- **Proportion of Published Effect Sizes that are Negative ($\hat{\mu_d}^{neg}_{PUB}$).** Approximated as the proportion of $\{\hat{\delta}_{+1P}\}$ that are less than zero.
- **The "Type S" error.** Approximated as the proportion of $\{\hat{\delta}_{11P}\}$ less than 0.
- **The "Type M" error.** Approximated as the mean of the absolute effect size divided by the true effect size: $|\{\hat{\delta}_{11P}\}|/\delta$.

A.2 The probability of publication as a function of both *psp* and *pwr*

In Section 3, the chance of publishing a positive study in our framework depended only on novelty via

$$A \;=\; (1-psp)^m.$$

For Section 4, a convenient choice of function to represent a journal policy of requiring an *a priori* sample size justification would be:

$$A \;=\; \frac{(1-psp)^m}{1+\exp\left\{-(\log 19)\frac{pwr-c_{50}}{c_{95}-c_{50}}\right\}}. \tag{7}$$

So A is reduced more when power is lower, with the extent of the reduction parameterized by (c_{50}, c_{95}). Specifically, c_{95} is the value for *pwr* at which the multiplicative reduction is near negligible (factor of 0.95), while c_{50} is the value for *pwr* at which the multiplicative reduction is a factor of 0.5. We conduct our experimentation using $c_{50}, c_{95} = (0.5, 0.8)$, with the following rationale. If a journal does require an a priori sample size justification, a claim of 80% power is the typical requirement. Hence a study which really attains 80% power is not likely to suffer in its quest for publication, motivating $c_{95} = 0.8$.

A.3 Additional results

δ	k	m	Change in PR	Change in REL	Change in $DSCV$	Change in D_{REL}	Change in $Mpwr_{PUB}$	Change in $IQpsp_{PUB}$
0.20	20	1	1.77	1.61	0.09	11.83	3.77	0.83
0.20	20	3	1.14	2.69	0.10	12.16	6.57	1.03
0.20	20	9	0.53	6.31	0.15	13.23	8.33	1.46
0.20	100	1	1.29	1.31	0.15	6.99	1.81	0.86
0.20	100	3	0.97	1.86	0.17	7.17	2.38	0.95
0.20	100	9	0.52	3.78	0.25	7.77	3.60	1.33
0.20	500	1	1.03	1.19	0.24	4.57	1.29	0.90
0.20	500	3	0.85	1.50	0.28	4.68	1.43	0.93
0.20	500	9	0.52	2.60	0.38	5.03	1.76	1.18
0.43	20	1	1.29	1.31	0.15	6.97	1.79	0.86
0.43	20	3	0.97	1.86	0.17	7.15	2.38	0.95
0.43	20	9	0.52	3.77	0.25	7.74	3.60	1.33
0.43	100	1	1.04	1.19	0.24	4.63	1.31	0.90
0.43	100	3	0.85	1.51	0.28	4.74	1.44	0.93
0.43	100	9	0.52	2.62	0.37	5.09	1.80	1.25
0.43	500	1	0.97	1.15	0.36	3.72	1.15	0.92
0.43	500	3	0.83	1.38	0.41	3.80	1.21	0.95
0.43	500	9	0.54	2.20	0.53	4.07	1.35	1.18

Table 6: A comparison between ecosystems with both a sample size requirement and $\alpha = 0.005$ and ecosystems without a sample size requirement and $\alpha = 0.05$. For those with a sample size requirement, A is defined by equation 7 (with $c_{50} = 0.5$ and $c_{95} = 0.8$). For those without a sample size requirement, $A = (1 - psp)^m$. For all we have $B = 0$, and varying values of k and m as indicated.

References

Amrhein, V., Korner-Nievergelt, F. & Roth, T. (2017), The earth is flat (p < 0.05): Significance thresholds and the crisis of unreplicable research, Technical report, PeerJ Preprints.

Aycaguer, L. C. S. & Galbán, P. A. (2013), 'Explicación del tamaño muestral empleado: una exigencia irracional de las revistas biomédicas', *Gaceta Sanitaria* **27**(1), 53–57.

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012), 'The rules of the game called psychological science', *Perspectives on Psychological Science* **7**(6), 543–554.

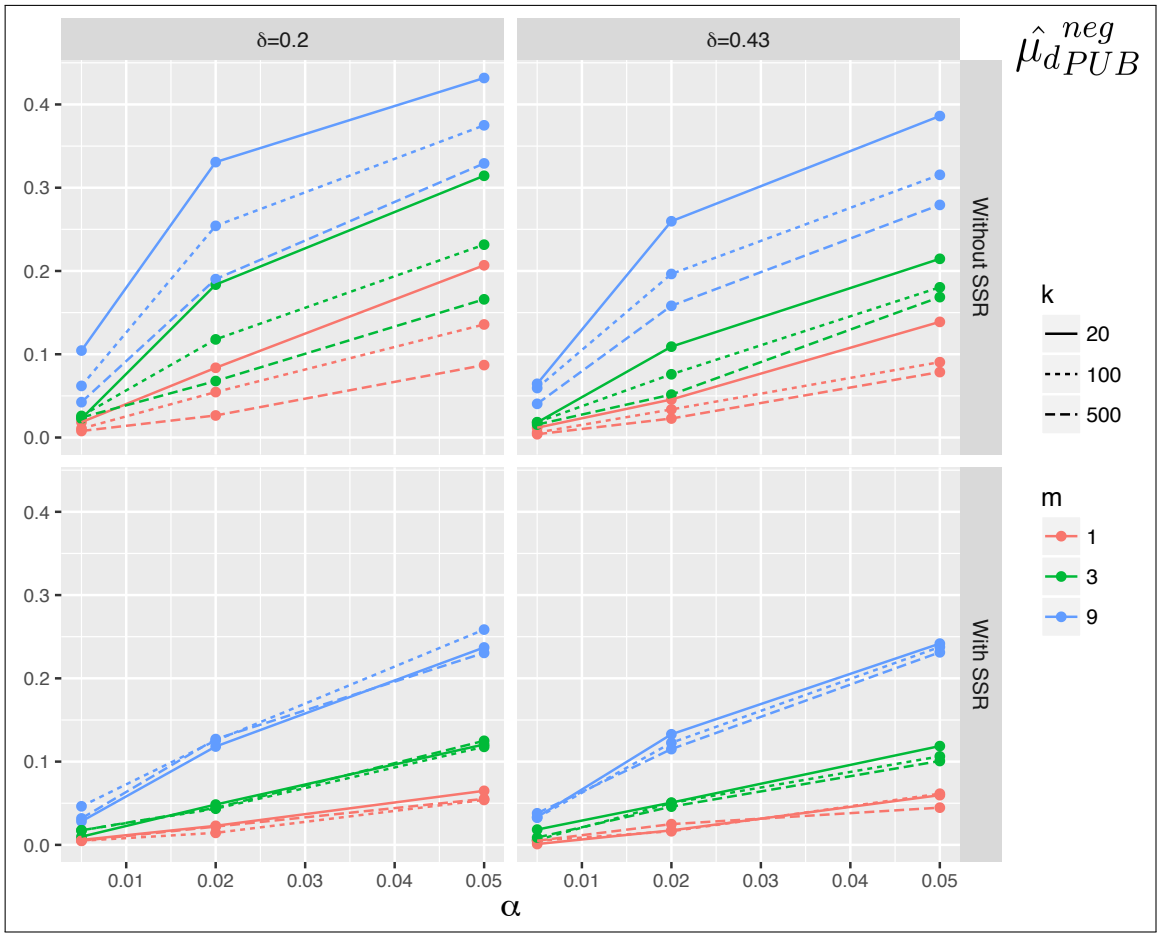

Figure 4: Values of $\hat{\mu}_{d_{PUB}}^{neg}$ for varying values of k, m and α; results for $\delta = 0.2$ and $\delta = 0.43$ on the top and bottom panels respectively. Left-panel shows results with no power requirement (i.e., $A = (1 - psp)^m$). Right-panel shows results with power requirement (i.e., A defined as per equation 7 (with $c_{50} = 0.5$ and $c_{95} = 0.8$)).

Begley, C. G. & Ellis, L. M. (2012), 'Drug development: Raise standards for preclinical cancer research', *Nature* **483**(7391), 531–533.

Begley, C. G. & Ioannidis, J. P. (2015), 'Reproducibility in science: improving the standard for basic and preclinical research', *Circulation research* **116**(1), 116–126.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al. (2018), 'Redefine statistical significance', *Nature Human Behaviour* **2**(1), 6.

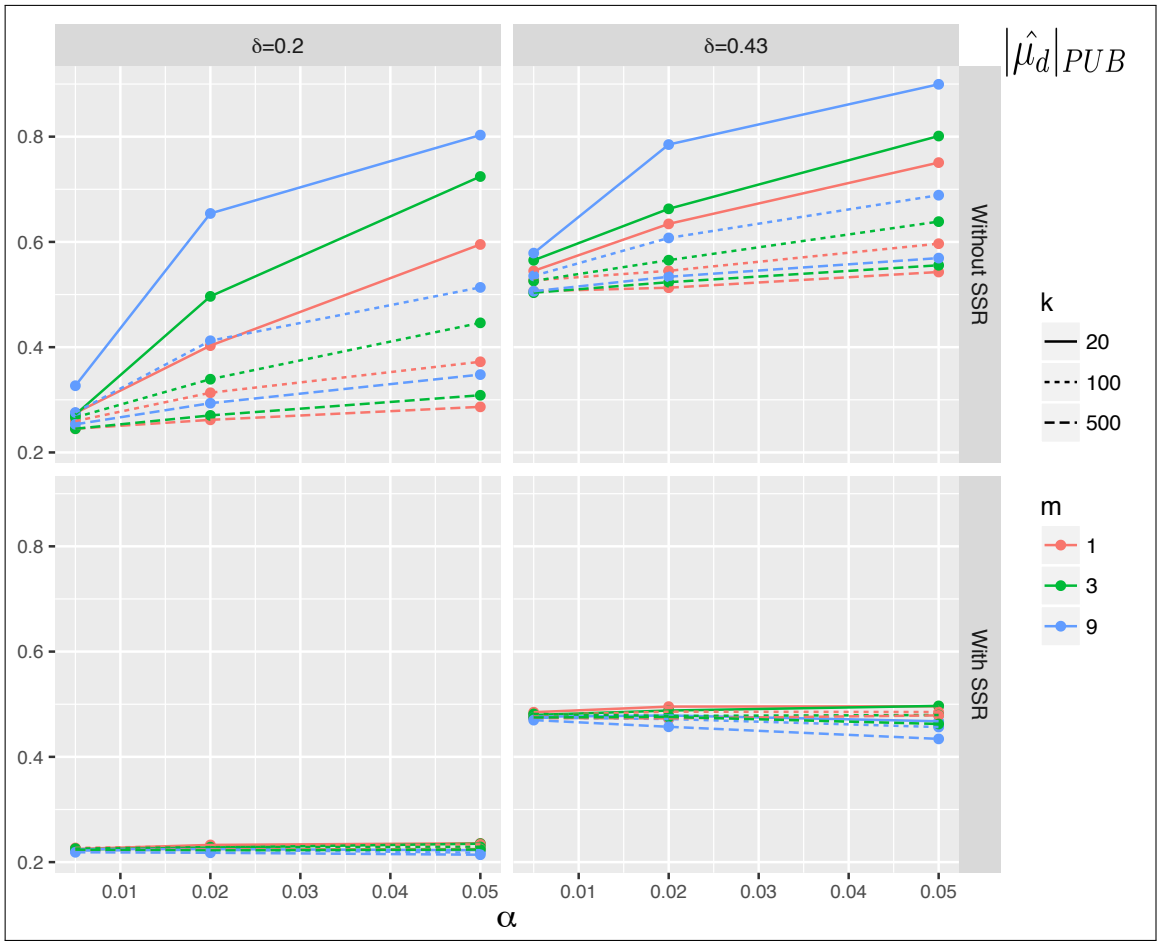

Figure 5: Values of $|\hat{\mu}_d|_{PUB}$ for varying values of k, m and α; results for $\delta = 0.2$ and $\delta = 0.43$ on the top and bottom panels respectively. Left-panel shows results with no power requirement (i.e., $A = (1 - psp)^m$). Right-panel shows results with power requirement (i.e., A defined as per equation 7 (with $c_{50} = 0.5$ and $c_{95} = 0.8$)).

Bland, J. M. (2009), 'The tyranny of power: is there a better way to calculate sample size?', *BMJ* **339**, b3985.

Borm, G. F., den Heijer, M. & Zielhuis, G. A. (2009), 'Publication bias was not a good reason to discourage trials with low power', *Journal of Clinical Epidemiology* **62**(1), 47–53.

Bosker, T., Mudge, J. F. & Munkittrick, K. R. (2013), 'Statistical reporting deficiencies in environmental toxicology', *Environmental toxicology and chemistry* **32**(8), 1737–1739.

Brembs, B., Button, K. & Munafò, M. (2013), 'Deep impact: unintended consequences of journal rank', *Frontiers in Human Neuroscience* **7**, 291.

Burt, T., Button, K., Thom, H., Noveck, R. & Munafò, M. R. (2017), 'The burden of the false-negatives in clinical development: Analyses of current and alternative scenarios and corrective measures', *Clinical and translational science* **10**(6), 470–479.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. & Munafò, M. R. (2013*a*), 'Empirical evidence for low reproducibility indicates low pre-study odds', *Nature Reviews Neuroscience* **14**(12), 877–877.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. & Munafò, M. R. (2013*b*), 'Power failure: why small sample size undermines the reliability of neuroscience', *Nature Reviews Neuroscience* **14**(5), 365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T. et al. (2016), 'Evaluating replicability of laboratory experiments in economics', *Science* **351**(6280), 1433–1436.

Campbell, H. & Gustafson, P. (2018), 'Conditional equivalence testing: An alternative remedy for publication bias', *PloS one* **13**(4), e0195145.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P. & Willmes, K. (2015), 'Registered reports: realigning incentives in scientific publishing', *Cortex* **66**, A1–A2.

Charles, P., Giraudeau, B., Dechartres, A., Baron, G. & Ravaud, P. (2009), 'Reporting of sample size calculation in randomised controlled trials', *BMJ* **338**, b1732.

Charlton, B. G. & Andras, P. (2006), 'How should we rate research?: Counting number of publications may be best research performance measure', *BMJ* **332**(7551), 1214–1215.

Coffman, L. C. & Niederle, M. (2015), 'Pre-analysis plans have limited upside, especially where replications are feasible', *Journal of Economic Perspectives* **29**(3), 81–98.

Cohen (1962), 'The statistical power of abnormal-social psychological research: a review.', *The Journal of Abnormal and Social Psychology* **65**(3), 145–153.

Cohen (2017), 'How should novelty be valued in science?', *eLife* **6**.

Contopoulos-Ioannidis, D. G., Ntzani, E. & Ioannidis, J. (2003), 'Translation of highly promising basic science research into clinical applications.', *The American journal of medicine* **114**(6), 477–484.

Dasgupta, P. & Maskin, E. (1987), 'The simple economics of research portfolios', *The Economic Journal* **97**(387), 581–595.

de Winter, J. & Happee, R. (2013), 'Why selective publication of statistically significant results can be effective', *PLoS One* **8**(6), e66463.

Djulbegovic, B. & Hozo, I. (2007), 'When should potentially false research findings be considered acceptable?', *PLoS Medcine* **4**(2), e26.

Djulbegovic, B., Kumar, A., Magazin, A., Schroen, A. T., Soares, H., Hozo, I., Clarke, M., Sargent, D. & Schell, M. J. (2011), 'Optimism bias leads to inconclusive results - an empirical study', *Journal of Clinical Epidemiology* **64**(6), 583–593.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F. & Munafò, M. R. (2017), 'Low statistical power in biomedical science: a review of three human research domains', *Royal Society Open Science* **4**(2), 160254.

Fanelli, D. (2011), 'Negative results are disappearing from most disciplines and countries', *Scientometrics* **90**(3), 891–904.

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R. & Thomason, N. (2006), 'Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology', *Conservation Biology* **20**(5), 1539–1544.

Fiedler, K. (2017), 'What constitutes strong psychological science? the (neglected) role of diagnosticity and a priori theorizing', *Perspectives on Psychological Science* **12**(1), 46–61.

Fire, M. & Guestrin, C. (2018), 'Over-optimization of academic publishing metrics: Observing goodhart's law in action', *arXiv preprint arXiv:1809.07841* .

Fraley, R. C. & Vazire, S. (2014), 'The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power', *PloS one* **9**(10), e109019.

Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. (2015), 'The economics of reproducibility in preclinical research', *PLoS Biology* **13**(6), e1002165.

Fritz, A., Scherndl, T. & Kühberger, A. (2013), 'A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough?', *Theory & Psychology* **23**(1), 98–122.

Gall, T., Ioannidis, J. & Maniadis, Z. (2017), 'The credibility crisis in research: Can economics tools help?', *PLoS Biology* **15**(4), e2001846.

Gaudart, J., Huiart, L., Milligan, P. J., Thiebaut, R. & Giorgi, R. (2014), 'Reproducibility issues in science, is p value really the only answer?', *Proc Natl Acad Sci USA* **111**, E1934.

Gelman, A. & Carlin, J. (2014), 'Beyond power calculations assessing type s (sign) and type m (magnitude) errors', *Perspectives on Psychological Science* **9**(6), 641–651.

Gervais, W. M., Jewell, J. A., Najle, M. B. & Ng, B. K. (2015), 'A powerful nudge? presenting calculable consequences of underpowered research shifts incentives toward adequately powered designs', *Social Psychological and Personality Science* **6**(7), 847–854.

Goodman, S. & Greenland, S. (2007), 'Assessing the unreliability of the medical literature: a response to 'why most published research findings are false?'', *Johns Hopkins University, Department of Biostatistics; Working Papers* .

Gornitzki, C., Larsson, A. & Fadeel, B. (2015), 'Freewheelin'scientists: citing Bob Dylan in the biomedical literature', *BMJ* **351**.

Greenwald, A. G. (1975), 'Consequences of prejudice against the null hypothesis.', *Psychological Bulletin* **82**(1), 1–20.

Greve, W., Bröder, A. & Erdfelder, E. (2013), 'Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture.', *European Psychologist* **18**(4), 286–294.

Hagen, K. (2016), 'Novel or reproducible: That is the question', *Glycobiology* **26**(5), 429–429.

Halpern, S. D., Karlawish, J. H. & Berlin, J. A. (2002), 'The continuing unethical conduct of underpowered clinical trials', *JAMA* **288**(3), 358–362.

Hazra, A. & Gogtay, N. (2016), 'Biostatistics series module 5: Determining sample size', *Indian journal of dermatology* **61**(5), 496.

Higginson, A. D. & Munafò, M. R. (2016), 'Current incentives for scientists lead to underpowered studies with erroneous conclusions', *PLoS Biology* **14**(11), e2000995.

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**(3), 171–178.

IntHout, J., Ioannidis, J. P. & Borm, G. F. (2016), 'Obtaining evidence by a single well-powered trial or several modestly powered trials', *Statistical Methods in Medical Research* **25**(2), 538–552.

Ioannidis, J. P. (2005), 'Why most published research findings are false', *PLoS Medicine* **2**(8), e124.

Ioannidis, J. P. (2008), 'Why most discovered true associations are inflated', *Epidemiology* pp. 640–648.

Ioannidis, J. P., Hozo, I. & Djulbegovic, B. (2013), 'Optimal type i and type ii error pairs when the available sample size is fixed', *Journal of Clinical Epidemiology* **66**(8), 903–910.

Johnson, V. E. (2013), 'Revised standards for statistical evidence', *Proceedings of the National Academy of Sciences* **110**(48), 19313–19317.

Johnson, V. E. (2014), 'Reply to Gelman, Gaudart, Pericchi: More reasons to revise standards for statistical evidence', *Proceedings of the National Academy of Sciences* **111**(19), E1936–E1937.

Kimmelman, J., Mogil, J. S. & Dirnagl, U. (2014), 'Distinguishing between exploratory and confirmatory preclinical research will improve translation', *PLoS biology* **12**(5), e1001863.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E. et al. (2018), 'Justify your alpha', *Nature Human Behaviour* **2**(3), 168.

Lamberink, H. J., Otte, W. M., Sinke, M. R., Lakens, D., Glasziou, P. P., Tijdink, J. K. & Vinkers, C. H. (2018), 'Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014', *Journal of clinical epidemiology* **102**, 123–128.

Lane, D. M. & Dunlap, W. P. (1978), 'Estimating effect size: Bias resulting from the significance criterion in editorial decisions', *British Journal of Mathematical and Statistical Psychology* **31**(2), 107–112.

Lee, C. J. & Schunn, C. D. (2011), 'Social biases and solutions for procedural objectivity', *Hypatia* **26**(2), 352–373.

Leek, J. T. & Jager, L. R. (2017), 'Is most published research really false?', *Annual Review of Statistics and Its Application* **4**, 109–122.

Lenth, R. V. (2001), 'Some practical guidelines for effective sample size determination', *The American Statistician* **55**(3), 187–193.

Makel, M. C., Plucker, J. A. & Hegarty, B. (2012), 'Replications in psychology research: How often do they really occur?', *Perspectives on Psychological Science* **7**(6), 537–542.

Martin, G. & Clarke, R. M. (2017), 'Are psychology journals anti-replication? a snapshot of editorial practices', *Frontiers in psychology* **8**, 523.

McLaughlin, A. (2011), 'In pursuit of resistance: pragmatic recommendations for doing science within one's means', *European Journal for Philosophy of Science* **1**(3), 353–371.

McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2017), 'Abandon statistical significance', *arXiv preprint arXiv:1709.07588* .

Miller, J. & Ulrich, R. (2016), 'Optimizing research payoff', *Perspectives on Psychological Science* **11**(5), 664–691.

Moonesinghe, R., Khoury, M. J. & Janssens, A. C. J. (2007), 'Most published research findings are falsebut a little replication goes a long way', *PLoS medicine* **4**(2), e28.

Mudge, J. F., Baker, L. F., Edge, C. B. & Houlahan, J. E. (2012), 'Setting an optimal α that minimizes errors in null hypothesis significance tests', *PloS one* **7**(2), e32734.

Nord, C. L., Valton, V., Wood, J. & Roiser, J. P. (2017), 'Power-up: a reanalysis of 'power failure' in neuroscience using mixture modelling', *Journal of Neuroscience* pp. 3592–16.

OpenScienceCollaboration (2015), 'Estimating the reproducibility of psychological science', *Science* **349**(6251), aac4716.

Ploutz-Snyder, R. J., Fiedler, J. & Feiveson, A. H. (2014), 'Justifying small-n research in scientifically amazing settings: challenging the notion that only ?big-n? studies are worthwhile', *Journal of Applied Physiology* **116**(9), 1251–1252.

Richard, F. D., Bond Jr, C. F. & Stokes-Zoota, J. J. (2003), 'One hundred years of social psychology quantitatively described.', *Review of General Psychology* **7**(4), 331.

Roos, J. M. (2017), 'Measuring the effects of experimental costs on sample sizes'.

Sakaluk, J. K. (2016), 'Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research', *Journal of Experimental Social Psychology* **66**, 47–54.

Schulz, K. F. & Grimes, D. A. (2005), 'Sample size calculations in randomised trials: mandatory and mystical', *The Lancet* **365**(9467), 1348–1353.

Senn, S. (2012), 'Misunderstanding publication bias: editors are not blameless after all', *F1000Research* **1**.

Siler, K., Lee, K. & Bero, L. (2015), 'Measuring the effectiveness of scientific gatekeeping', *Proceedings of the National Academy of Sciences* **112**(2), 360–365.

Smaldino, P. E. & McElreath, R. (2016), 'The natural selection of bad science', *Royal Society Open Science* **3**(9), 160384.

Song, F., Loke, Y. & Hooper, L. (2014), 'Why are medical and health-related studies not being published? a systematic review of reasons given by investigators', *PLoS One* **9**(10), e110418.

Spiegelhalter, D. (2017), 'Trust in numbers', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**(4), 948–965.

Stanley, T., Jarrell, S. B. & Doucouliagos, H. (2010), 'Could it be better to discard 90% of the data? a statistical paradox', *The American Statistician* **64**(1), 70–77.

Sterling, T. D., Rosenbaum, W. L. & Weinkam, J. J. (1995), 'Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa', *The American Statistician* **49**(1), 108–112.

Szucs, D. & Ioannidis, J. P. (2016), 'Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature', *bioRxiv* p. 071530.

van Assen, M. A., van Aert, R. C., Nuijten, M. B. & Wicherts, J. M. (2014), 'Why publishing everything is more effective than selective publishing of statistically significant results', *PLoS One* **9**(1), e84896.

van Dijk, D., Manor, O. & Carey, L. B. (2014), 'Publication metrics and success on the academic job market', *Current Biology* **24**(11), R516–R517.

Vasishth, S. & Gelman, A. (2017), 'The illusion of power: How the statistical significance filter leads to overconfident expectations of replicability', *arXiv preprint arXiv:1702.00556* .

Walker, A. M. (1995), 'Low power and striking results- a surprise but not a paradox', *Mass Medical Soc.* **332**(16), 1091–1092.

Wei, Y. & Chen, F. (2018), 'Lowering the p value thresholdreply', *JAMA* **320**(9), 937–938.

Wilson, B. M. & Wixted, J. T. (2018), 'The prior odds of testing a true effect in cognitive and social psychology', *Advances in Methods and Practices in Psychological Science* p. 2515245918767122.

Yeung, A. W. (2017), 'Do neuroscience journals accept replications? a survey of literature', *Frontiers in human neuroscience* **11**, 468.