INTERVENTION AND BACKFIRE IN THE REPLICATION CRISIS

AYDIN MOHSENI

Abstract. Scientific studies vary in their methodological soundness. Interventions in evidentiary standards and research practices can differentially affect studies as a function of their soundness. The conjunction of these facts has unrecognized implications for proposed interventions in the replication crisis. I argue that we should expect these facts to obtain, and demonstrate that, when accounting for differential effects of interventions as a function of soundness, several of the proposed interventions—lowering the significance threshold, promoting preregistration, and sample splitting—will produce less improvement than estimates would suggest and, in some cases, actually increase false discovery rates, sign error rates, and magnitude exaggeration ratios.

1. Introduction

Scientific studies vary in their methodological soundness. Less-sound studies are more likely to attain significance when the null hypothesis is true (Ioannidis 2005b), to produce a significant effect of the opposite sign as the true effect (Gelman and Carlin 2014), and to produce significant effect sizes many times larger than the true effect size (Ioannidis 2008).[1] These epistemic ills tend to coincide for the simple reason that they have a common cause: questionable research practices such as multiple comparisons, garden of forking paths, and underpowered studies (Gelman and Loken 2019).

Interventions in evidentiary standards and research practices can differentially affect studies as a function of their soundness. In particular, this can occur when less-sound methods make it so that less-sound studies are relatively more capable

[1] It is not, however, generally true that unsound studies are more likely to fail to attain significance when the null hypothesis is false. While under-powered studies tend to have higher false negative rates, questionable research practices tend to be strongly biased in favor of producing significant results, and so the false negative rates of less sound studies are not always less, on average, than those of more sound ones.

of meeting higher evidentiary standards or when more-sound studies are more likely to be voluntarily submitted to more stringent research practices.

In response to the findings of the recent replication crisis—that literatures in parts of the social and life sciences exhibit undesirable rates of failed replication (Arrowsmith 2011; Bogdan et al. 2017; Camerer et al. 2016; Ioannidis 2005a; Open Science Collaboration 2015)—various remedial interventions have been proposed (Benjamin et al. 2018; Crutzen and Peters 2017; Dwork et al. 2015; Gelman 2018; Ioannidis 2014; Munafò et al. 2017; Nosek et al. 2018b). Many of the proposed interventions are aimed at lowering the false discovery rate in a population of scientific studies either by raising a specific evidentiary standard—such as lowering the significance threshold (Benjamin et al. 2018) or raising the threshold for study power required for publication (Crutzen and Peters 2017)—or by promoting a research practice that constrains researcher degrees of freedom—such as preregistration (Nosek et al. 2018b) or sample-splitting (Dwork et al. 2015). These standards and practices need not affect all studies equally. Indeed, whenever the effects of interventions are not entirely uncorrelated with study soundness, one should be alert to the character of the interaction between the two.

This is the essential insight for understanding the following results. I consider the proposals to lower the significance threshold and to promote or require preregistration. In each case, I argue that there are facts that should make us expect that the interventions will filter more-sound studies more stringently than less-sound ones. I then proffer several demonstrations, all of which share a basic structure: under plausible assumptions, each of the interventions considered can lead to a decline in the mean replicability of the population of studies that attain statistical significance.

In all cases, my analysis reveals the interventions will lead to less improvement than estimates that don't account for soundness-dependent effects would suggest. And, in particular cases, I demonstrate that the interventions can, counterintuitively, actually increase rates of false discovery, sign error rates, and magnitude exaggeration ratios.

2. The model: Null Hypothesis Significance Testing for a Population of Studies with Varying Soundness

Consider a unit mass of independent studies. In each study, a researcher conducts a hypothesis test between a pair of null (H_0) and alternative (H_A) hypotheses. Let $\phi \in (0, 1)$ denote the proportion of null hypotheses that are true. The outcome of a study is significant if the hypothesis test yields a p-value less than the significance threshold α. I consider three types of errors: type I, type II, and type M errors.

A *false positive* (type I erorr) occurs when a study yields a significant outcome when the null is in fact true and the *false positive ratio* is equal to the number of false positive outcomes over the number of all significant outcomes. In expectation this is

$$Pr(H_0 \mid \text{significant}) = \frac{Pr(H_0, \text{significant})}{Pr(\text{significant})}.$$

Conversely, a *false negative* (type II error) occurs when a study yields a non-significant outcome when the null is in fact true and the *false negative ratio* is equal to the number of false negative outcomes over the number of all non-signficant outcomes. In expectation this is

$$Pr(H_A \mid \text{non-significant}) = \frac{Pr(H_A, \text{non-significant})}{Pr(\text{non-significant})}.$$

A *magnitude error* (type M error) occurs when the reported effect size of a study d is different than the true effect size D and the *magnitude exaggeration ratio* is equal to the average ratio of report effect sizes to true effect sizes. In expectation this is

$$\mathbb{E}\left(\frac{d}{D} \,\middle|\, \text{significant}\right) = \frac{\mathbb{E}(d, \text{significant})}{\mathbb{E}(D, \text{significant})}.$$

Studies vary in their methodological soundness. In particular, unsound studies are biased in the direction of obtaining significant results. For simplicity, imagine that studies can be divided into those employing maximally *sound* and *unsound* methods. *Sound* studies yield significant results when the null hypothesis is true with a frequency corresponding to the significance threshold α, while *unsound* studies always produce significant results regardless of the truth of the null hypothesis. Given that proportion θ of true null hypotheses, this yields rates for sound and unsound studies displayed in Table 1.

Some proportion λ of the population of studies are sound while the remaining $1 - \lambda$ are unsound. The aggregate statistical properties of the population of studies then are weighted averages of those produced by the sub-populations of sound and unsound studies. For example, the false positive ratio of all studies P can be expressed as

$$P = \lambda w P_u + (1 - w_u) P_s \tag{1}$$

	Sound		Unsound	
	H_0	H_A	H_0	H_A
Significant	$\theta\alpha$	$(1-\theta)\beta$	θ	$1-\theta$
Non-significant	$\theta(1-\alpha)$	$(1-\theta)(1-\beta)$	θ	$1-\theta$

Table 1. Given maximally sound and unsound studies, the above table details expected rates of attaining significant and non-significant results given that either the null hypothesis H_0 or alternative hypothesis H_A is true for a population of studies where proportion θ of the null hypotheses are true.

where P_u and P_s are the false positive ratios of sound and unsound studies and $w = \frac{\lambda}{\lambda+(1-\lambda)(\phi\alpha+(1-\phi)(1-\beta))}$ is the fraction of significant outcomes contributed by unsound studies. The same holds for each false negative ratios N, and—given the additivity of expectation—magnitude exaggeration ratios M. Thus, we have the following aggregates statistics for the populations of studies

$$N = wN_u + (1-w_u)N_s, \tag{2}$$

$$M = wM_u + (1-w_u)M_s. \tag{3}$$

More generally, there will be some continuous distribution f_s of the methodological soundness of studies. Soundness s varies from maximal soundness at $s = \alpha$ to maximla unsoundness at $s = 1$ which is understood as the minimum p-value of the study outcome which, via researcher degrees of freedom, can be manipulated to attain significance at the default significance threshold α. The the previously described statistical properties (1-3) of a population of studies is then given by

$$T = \int_{\alpha}^{1} w_s T_s \mathrm{d}f_s, \quad \text{for each } T \in \{P, N, S, M\}.$$

To see how this works, consider two studies of differing soundness where the first is maximally sound $s = \alpha$ and the second is less sound $s = 3\alpha$. Let each study consist in a single-tailed hypothesis test. When the test yields an outcome with effect size $z_{\frac{\alpha}{2}}$, both the sound and unsound study will report a statistically significant outcome with p-value $\frac{\alpha}{2}$. But when the test yields an effect size of $z_{2\alpha}$ the sound study will report a non-significant outcome with p-value 2α while the less-sound study will still report a significant outcome with p-value α. For an effect size larger than the range of manipulation $z_{3\alpha}$ of the less sound study, say $z_{4\alpha}$, both the sound and unsound studies will report a non-significant p-value of 4α. Note that this sort of "thresholding" of effect sizes by less-sound studies is drawn from and supported by empirical distributions of reported p-values (Simmons et al. 2011).

An intervention in such cases can be seen as producing two simultaneous effects. First, the intervention will change the statistical properties of the sub-populations of sound and unsound studies. Taken in isolation, within any subset of studies sorted by soundness, this will typically appear to be a desirable effect. For example, the false discovery rate of each subset will decrease. Second, however, the intervention will change the *relative proportion* of all significant outcomes which are contributed by studies of varying soundness. If the intervention filters sound studies more stringently than unsound studies, this can yield characteristically undesirable effects. For example, the subset of less-sound studies may produce a greater fraction of the significant results and so—under a regime of publication bias—be published much more often and so compose a relatively larger fraction of the literature.

3. Promoting or requiring preregistration

Preregistration is one of the most promising and broadly-endorsed of the proposed interventions in the replication crisis (COS 2015; Munafò et al. 2017; Nosek et al. 2018b). Yet, it is not without its costs or its detractors (Scott 2013; Ledgerwood 2018; Nosek et al. 2018a).

Preregistration of a study consist in a researcher specifying her study design and analysis plan prior to conducting the study in order to avoid the confounding effects of researcher degrees of freedom and to enhance scientific transparency and credibility (Lindsay et al. 2016).

But studies are costly, and preregistering a study may mean that the researcher cannot report a statistically significant effect that she did not anticipate and include in her study design. Given this, preregistration may deter researchers from conducting important but risky exploratory studies when the associations or causes at play are not yet well understood.

In response to this concern, proponents of preregistration have proposed that preregistration be introduced on a voluntary basis, allowing researchers to eschew preregistration when they think it appropriate (Nosek et al. 2018b). The thought behind this compromise is that any portion of a scientific community actively engaged in preregistration will contribute positively to the overall replicability of the population of studies.

However, I will show that this may well not be the case. The voluntary aspect of preregistration can correlate its adoption with studies that are already less likely to produce false positive outcomes thus produce soundness-dependent filtering of studies that can diminish or reverse desired improvement.

This correlation can be realized in at least two ways. The first possible source of correlation is that researchers who are already more likely to be employing sound methods may be more likely to preregister their studies. Here, it is the researcher herself that correlates more sound studies with more stringent filtration.

The argument is as follows. It is highly plausible that the same attitudes, incentives, educational backgrounds, institutional resources, and commitments that make a researcher already more likely to employ sound research methods would make her more likely to adopt yet more sound methods in the form of preregistration.

Importantly, this relationship need not hold in every case, but simply on average. Not every scientist who is already using more-sound methods need be more likely to preregister her studies, only that use of more-sound methods positively correlates with adoption of preregistration at the population level. We state this assumption explicitly as follows.

Assumption 1. *Researchers who are more likely to already use sound methods are, on average, more likely to preregister their studies.*

This is already sufficient for us draw two conclusions. The first conclusion regards the lessening of improvements from preregistration, and the second regards the possibility of reversal of improvement.

Claim 1. *If more-sound studies are more likely to be preregistered, then less-sound ones, then voluntary preregistration will produce less improvement in the false discovery rate of a literature than predicted by an estimates that fails to account for this fact. (All proofs provided in the Mathematical Appendix.)*

To see this, consider the simple where there are two sub-populations of *more-sound* and *less-sound* sound with respective soundness s, u where $s < u$. Neither is perfectly sound, so that preregistration will lower the false positive rate of both, making it so that any study that preregisters will exhibit a false positive rate s_0 closer to that of the significance threshold α. This can be summarized by the ordering $\alpha \leq s_0 < s < u \leq 1$. Further, assume that more-sound studies are more likely than less-sound ones by a factor of k.

Recall that preregistration decreases both the rate of false positive results but also the absolute number of significant results. Given this, voluntary preregistration of studies makes it so that more-sound studies will contribute relatively fewer significant outcomes. Given that less-sound studies tend to produce both more false positive and greater magnitude errors, this effects the statistics of the literature on both counts.

To see how this works, consider the extreme case where less-sound studies always obtain significance ($u = 1$), more-sound studies have modestly low false positive rates ($s = 2\alpha = 0.2$), only more-sound studies are likely to be preregistered ($k = \infty$), and where the population is evenly split between more-sound and less-sound studies ($\lambda = \frac{1}{2}$). Prior to the introduction of preregistration, the false discovery rate of the population of studies will be given by

$$P = wP_u + (1-w)P_s = w\theta + (1-w)\frac{0.2\theta}{0.2\theta + (1-\theta)(1-\beta(0.2))}$$

where $w = \frac{1}{1+(0.2\theta+(1-\theta)(1-\beta(0.2))}$ is the fraction of significant results contributed by less-sound studies.

If we further assume that three quarters of null hypotheses are true ($\theta = \frac{3}{4}$), and that studies are reasonably well-powered with average effect sizes ($d = 0.5$) across the literature, then this yields a false discovery rate of roughly 0.72 for the literature.

An estimate that fails to account for soundness-dependence will predict that, if half of studies are preregistered, then the false discovery rate will be lowered to 0.63. (That is, halfway to the rate of 0.54 obtained if *all* studies were preregistered.) Whereas, accounting for the correlation between more sound studies and the likely to preregister, we see that the false positive rate will actually increase to 0.74.

This example just discussed is illustrated in 1. What is shown is the two countervailing effects of preregistration. The false discovery ratio of sound studies P_s decreases as desired. But the proportion w of significant results contributed by unsound studies increases as more sound studies are submitted to the more stringent standards produced by preregistration. Ultimately, this entails that the false positive rate P of the literature will not decrease as much as would be predicted without account for the correlation between study soundness and

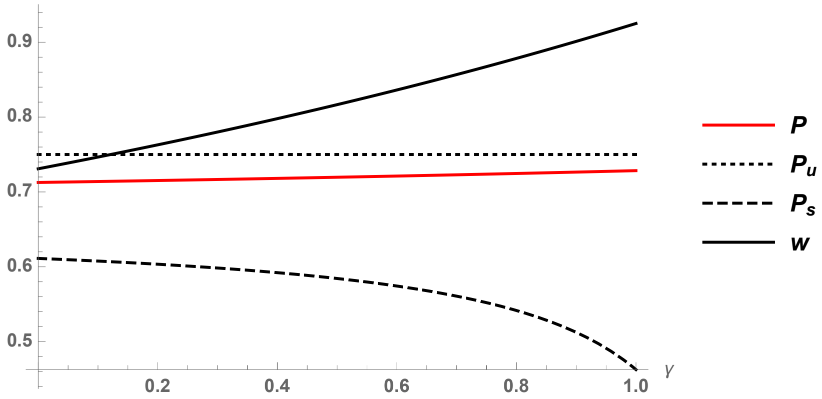

Figure 1. The false discovery rate of a literature can increase if more-sound studies are more likely to be preregistered than less-sound ones. γ denotes the proportion of more-sound studies preregistered; P is the false discovery rate of the literature; P_u is the false discovery rate of unsound studies; P_s is the false discovery rate of sound studies; and w is the proportion of significant results contributed by unsound studies.

voluntary preregistration. Indeed, here we see that it increases as more studies are preregistered.The source code for all models is provide in **??**.

Claim 2. *If sound methods and voluntary preregistration are sufficiently strongly correlated, then voluntary preregistration can increase the false discovery rate of a literature.*

This is important given that, thus far, models of the expected effects of preregistration have not accounted for correlation between sounds methods and voluntary preregistration (Nosek et al. 2018a,b,c). Rather, by failing to correct for such a covariate, they have implicitly assumed that voluntary preregistration will occur essentially at random.

We have seen that, if the difference in more and less-sounds in methods is sufficiently great, and the correlation between soundness and preregistration sufficiently strong, then voluntary preregistration can make it so that the false discovery rate and across a literature can actually increases.

Whether such a reversal can obtain is an empirical matter; it will depend on the distribution of questionable research practices and on the actual correlation between soundness of methods and voluntary preregistration. Neither may be sufficiently strong to produce a full reversal. But, as we will see, this is not the only potential source of correlation.

A second plausible source of correlation resides in the fact that researchers who are more confident in the truth of their hypotheses may be more likely to voluntarily preregister their studies.

Assumption 2. *Researchers who are more confident in the truth of their studies are, on average, more likely to preregister their studies.*

It is widely understood that researchers make (implicit or explicit) cost-benefit assessments in choosing which studies to undertake and which methods to employ. Indeed, it is reasonable to assume that this is one source of the prevalence of questionable research practices. One factor in such assessments will likely be a researcher's confident in obtaining significant, publication-worthy results.

For correlation to exist by this route, it need not be the case that researchers prioritize attaining significant results above all other considerations—such as the potential impacts of a line of study or sheer curiosity—but only that, all else being equal, researchers are on average more likely to undertake studies if they think they will be successful in attaining significant results.

If this is right,then then one should expect the benefit for refraining from preregistration will be larger for researchers precisely when they are less confident in the truth of their hypotheses. Conversely, researchers with higher priors on their alternative hypotheses should be more likely to preregister. Indeed, this is advertised as a feature of voluntary preregistration: if one is less sure as to which hypotheses should be formulated prior to data-collection, or one is engaged in more "exploratory" research, then one can opt out of preregistration (Lindsay et al. 2016).

This need only be coupled with a further assumption for positive correlation to obtain between the likelihood of a study to be voluntarily preregistered and the prior probability of its hypotheses to be true.

Assumption 3. *Researchers' confidence in the truth of their study hypotheses is, on average, positively correlated with the truth of their study hypotheses.*

It follows immediately from the conjunction of Assumptions 2 and 3 that the prior probability of a hypothesis will be positively correlated with the likelihood of its preregistration. Notice that Assumption 3 does not require that researchers be particularly good predictors of the truth of their hypotheses. Though recent work by Camerer et al (2016) which enrolled researchers in prediction markets to bet on which studies would fail to replicate strongly suggests that researchers may, in fact, be surprisingly good judges of the truth of, at least other researchers', study outcomes. Indeed, it compatible with this assumption that researcher do significantly worse than chance in predicting whether a hypothesis will yield a significant result. All that is required is that, all else being equal, researcher confidence is neither entirely random nor anti-correlated with the truth of their hypotheses. It seems quite plausible that this might be so.

Claim 3. *If studies testing hypotheses will greater prior odds are more likely to be preregistered, then voluntary preregistration will produce less improvement in the false discovery rate of a literature than predicted by an estimates that fails to account for this fact.*

Once again, we see that facts about researcher can act as latent variables, correlating studies less likely to produce false discoveries with more stringent filtering. With this in hand, very similar dynamics as we have seen before emerge. The only difference is that higher false discovery rates are being produced not by differentials in questionable research practices, but by differentials in prior of hypotheses and the greater likelihood of more plausible hypotheses to be preregistered.

To see how this works, consider the simple case where there are two types of hypotheses in the population: *plausible* ones and *implausible* ones which exhibit lower and higher frequencies of their null hypotheses being true ($\theta_p = \frac{1}{3}, \theta_i = \frac{2}{3}$), and conventionally medium and small average effect sizes ($d_p = 0.5, d_i = 0.2$), respectively. Here, all researchers employ similarly sound methods ($s_p = s_i = 6\alpha = 0.3$), but only more-plausible studies are likely to be preregistered ($k = \infty$).

If we further assume the population is evenly split between studies with more-plausible and less-plausible hypotheses ($\lambda = \frac{1}{2}$), then this yields a false discovery rate of roughly 0.40 across the literature. An estimate that fails to account for soundness-dependence will predict that, if half of all studies are preregistered, then the false discovery rate will be lowered to 0.36. (That is, halfway to the rate of 0.30 obtained if *all* studies were preregistered.) Whereas, accounting for the correlation between more plausible studies and probability of preregistration, we see that the false positive rate will actually increase to 0.51.

This example just discussed is illustrated in 2. Once again, we see that two countervailing effects are produced by preregistration. The false discovery ratio

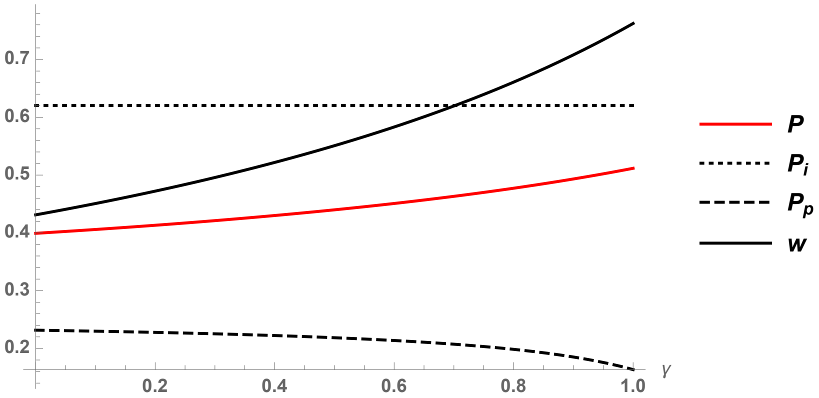

FIGURE 2. The false discovery rate of a literature can increase if studies with plausible hypotheses are more likely to be preregistered than those with less plausible ones. γ denotes the proportion of plausible studies preregistered; P is the false discovery rate of the literature; P_i is the false discovery rate of implausible hypotheses; P_p is the false discovery rate of plausible hypotheses; and w is the proportion of significant results contributed by implausible hypotheses.

of studies testing plausible hypotheses P_s decreases as desired. But the proportion w of significant results contributed by studies testing implausible hypotheses increases as more plausible studies are submitted to the more stringent standards of preregistration. Ultimately, this entails that the false discovery rate P of the literature will not decrease as much as would be predicted without account for the correlation between study soundness and voluntary preregistration. And here we see it increase as more studies are preregistered.

Claim 4. *If prior odds of study hypotheses and voluntary preregistration are sufficiently strongly correlated, then voluntary preregistration can increase the false discovery rate of the literature.*

Notice that the effects described in Claims 1 and 3 can simultaneously be at play. If study soundness and prior odds of hypotheses are independent, then the two effects are additive. However, the state of affairs may be worse than this. Plausibly, hypotheses with higher prior odds obtain significant results more often without the need for p-hacking or other questionable research practices. In this way, each sound methods, high priors, and likelihood of preregistration become positively correlated, exacerbating the problems detailed thus far.

4. Discussion

There is a well-known moral from social epistemology: epistemic methods that appear to be optimal at the individual level may be suboptimal when considered within the context of collective inquiry.[2] Analogously, here we see that interventions in research practices that appear to have favorable effects when applied to individual studies may have unfavorable effects for the broader population of studies.

These results have consequences for the proposed interventions. Knowledge of the possibility of soundness-dependent backfire effects may ultimately inform our choice of which interventions to prioritize and ultimately pursue. One may use the understanding of the possibility of such effects to guide implementation of the interventions.

In the case of promoting voluntary preregistration, if researchers who already employ more-sound methods are more likely to voluntarily submit their studies to more stringent standards, then these researchers will produce fewer significant results in expectation. Simultaneously, if researchers who are more confident in the truth of their hypotheses are more likely to preregister their studies, then studies which would be less likely to produce false discoveries will again be more stringently filtered.

In each case, the statistical properties of (on average) less-sound or less-plausible studies who opt out of preregistration will more strongly determine those of the literature. This entails that improvement in the false discovery rate in the literature will be substantially less than would be anticipated if one fails to account for soundness-dependent effects. Moreover, as we have seen, if the correlation in each case are sufficiently strong, partial participation in preregistration can actually be worse than no participation at all.

Appendix A. Mathematical Appendix

All the R code for this project is available at GitHub at: (ANONYMIZED FOR REVIEW) https://anonymous.4open.science/r/3f5bc2af-ded9-4314-b7b9-bb46ebc4cf97/.

[2] This observation sometimes goes by the name of the Independence Thesis: that social rationality is independent of individual rationality (Mayo-Wilson et al. 2013).

Appendix B. Mathematical Appendix

We show that the magnitude exaggeration ratio of studies increases as the significance threshold decreases.

Proof. The outcome (sample mean) of a given study is given by a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. The magnitude exaggeration ratio of a study R_M is defined as the ratio of the reported effect size conditional on having received a significant result over the true effect size. This can be written as a function of the significance threshold α.

$$R_M(\alpha) \equiv \frac{\mathbb{E}[|X| \mid |X| > k_\alpha]}{\mu}$$

where k_α is the critical value at a significance threshold of α and is thus positive and decreasing in α.

Without loss of generality, let $\mu > 0$. The magnitude exaggeration ratio then simplifies to the inverse Mills ratio

$$2\mu^{-1}\left(\mu - \sigma\frac{\phi(k_\alpha - \mu)}{\Phi(k_\alpha - \mu)}\right) \tag{4}$$

by a well-known property of the truncated normal distribution.

It suffices to show that (4) is increasing in k_α. Since $\sigma > 0$, this is accomplished by showing that $\frac{\phi(k-\mu)}{\Phi(k-\mu)}$ is decreasing in k_α. Using the property of the normal density that $\phi'(z) = -z\phi(z)$ and the quotient rule for derivatives we observe that

$$\frac{\mathrm{d}}{\mathrm{d}k_\alpha}\left[\frac{\phi(k_\alpha - \mu)}{\Phi(k_\alpha - \mu)}\right] = -k_\alpha\left[\frac{\phi(k_\alpha - \mu)[\Phi(k_\alpha - \mu) + \phi(k_\alpha - \mu)]}{\Phi(k_\alpha - \mu)^2}\right] \tag{5}$$

which much be negative since $-k_\alpha$ is negative and each ϕ and Φ are positive functions. Hence R_m increases as α decreases. □

We show that the sign error rate of studies decreases as the significance threshold decreases.

Proof. The sign error rate of a study is defined as the probability that the sign of the measured effect opposes that of the true effect conditional on having received a significant result.

$$R_S \equiv Pr(sgn(X) \neq sgn(\mu) \mid |X| > k)$$

where k_α is the critical value at a significance threshold of α.

Without loss of generality, let $\mu > 0$. Then the sign error rate can be stated as

$$\frac{Pr(X < -k_\alpha)}{Pr(X < -k_\alpha) + Pr(X > k_\alpha)}. \tag{6}$$

Given that each probability term in (6) is decreasing in k, it suffices to show that $\frac{Pr(X=-k_\alpha)}{Pr(X=-k)+Pr(X=k_\alpha)}$ is decreasing in k for all $k > 0$. Further, it suffices to show simply that $\frac{Pr(X=-k_\alpha)}{Pr(X=k_\alpha)}$ is decreasing in k. We unpack the expression

$$\frac{Pr(X = -k_\alpha)}{Pr(X = k_\alpha)} = \frac{\phi(-k_\alpha)}{\phi(k_\alpha)} = e^{-\frac{1}{\sigma^2}(\mu^2+2k\mu)}.$$

and integrate with respect to k_α to get

$$\frac{\mathrm{d}}{\mathrm{d}k_\alpha}[e^{-\frac{1}{\sigma^2}(\mu^2+2k\mu)}] = -\left(\frac{2\mu}{\sigma^2}\right) e^{-\frac{1}{\sigma^2}(\mu^2+2k_\alpha\mu)}$$

which is negative as each the exponential function, μ, and σ are positive. Hence R_S decreases as α decreases. □

References

Arrowsmith, J. (2011). Phase II failures: 2008–2010. *Nature Reviews Drug Discovery*.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2018). Redefine statistical significance.

Bogdan, R., B. J. Salmeron, C. E. Carey, A. Agrawal, V. D. Calhoun, H. Garavan, A. R. Hariri, A. Heinz, M. N. Hill, A. Holmes, N. H. Kalin, and D. Goldman (2017). Imaging Genetics and Genomics in Psychiatry: A Critical Review of Progress and Potential.

Camerer, C. F., A. Dreber, E. Forsell, T. H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016). Evaluating replicability of laboratory experiments in economics. *Science*.

COS (2015). Guidelines for transparency and openness promotion (TOP) in journal policies and practices “The TOP Guidelines”. Technical report.

Crutzen, R. and G.-J. Y. Peters (2017). Targeting Next Generations to Change the Common Practice of Underpowered Research. *Frontiers in Psychology*.

Dwork, C., V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*.

Gelman, A. (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*.

Gelman, A. and J. Carlin (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*.

Gelman, A. and E. Loken (2019). The Statistical Crisis in Science. In *The Best Writing on Mathematics 2015*.

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*.

Ioannidis, J. P. A. (2005b). Why most published research findings are false.

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLoS Medicine*.

Ledgerwood, A. (2018). The preregistration revolution needs to distinguish between predictions and analyses. *Proceedings of the National Academy of Sciences*.

Lindsay, D. S., D. J. Simons, and S. O. Lilienfeld (2016). Research Preregistration 101. *Association for Psychological Science*.

Mayo-Wilson, C., K. Zollman, and D. Danks (2013). Wisdom of crowds versus groupthink: Learning in groups and in isolation. *International Journal of Game Theory 42*(3), 695–723.

Munafò, M. R., B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie Du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis (2017). A manifesto for reproducible science.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor (2018a). Reply to Ledgerwood: Predictions without analysis plans are inert. *Proceedings of the National Academy of Sciences*.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor (2018b). The preregistration revolution. *Proceedings of the National Academy of Sciences*.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor (2018c). The preregistration revolution. *Proceedings of the National Academy of Sciences*.

Open Science Collaboration (2015). Estimating the reproducibility of psychological. *Science*.

Scott, S. (2013). Pre-registration would put science in chains. Time Higher Education.e. *Time Higher Education*.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). False-Positive Psychology. *Psychological Science*.