Running head: *p*-HACKING, PUBLICATION BIAS, AND META-ANALYTIC EFFECTS

p-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates

Malte Friese & Julius Frankenbach

Saarland University

2020, *Psychological Methods, 25*, 456-471.

Author Note

Malte Friese and Julius Frankenbach, Department of Psychology, Saarland University.

Both authors contributed equally to this work and share the first authorship. We thank Michael Inzlicht, David D. Loschelder, Dorota Reis, Simine Vazire, and an anonymous reviewer for valuable comments on an earlier version of this article. All code is available at https://osf.io/phwne/?view_only=14077670341a4ab08b4d75b8a13aaf01.

Correspondence concerning this article should be addressed to Malte Friese or Julius Frankenbach, Department of Psychology, Saarland University, Campus A2 4, 66123 Saarbrucken, Germany. Email: malte.friese@uni-saarland.de or julius.frankenbach@gmail.com

Abstract

Science depends on trustworthy evidence. Thus, a biased scientific record is of questionable value because it impedes scientific progress, and the public receives advice on the basis of unreliable evidence that has the potential to have far-reaching detrimental consequences. Meta-analysis is a valid and reliable technique that can be used to summarize research evidence. However, meta-analytic effect size estimates may themselves be biased, threatening the validity and usefulness of meta-analyses to promote scientific progress. Here, we offer a large-scale simulation study to elucidate how *p*-hacking and publication bias distort meta-analytic effect size estimates under a broad array of circumstances that reflect the reality that exists across a variety of research areas. The results revealed that, first, very high levels of publication bias can severely distort the cumulative evidence. Second, *p*-hacking and publication bias interact: At relatively high and low levels of publication bias, *p*-hacking does comparatively little harm, but at medium levels of publication bias, *p*-hacking can considerably contribute to bias, especially when the true effects are very small or are approaching zero. Third, *p*-hacking can severely increase the rate of false positives. A key implication is that, in addition to preventing *p*-hacking, policies in research institutions, funding agencies, and scientific journals need to make the prevention of publication bias a top priority to ensure a trustworthy base of evidence.

Word count: 220

Keywords: meta-analysis, *p*-hacking, publication bias, meta-science

p-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates

Science depends on trustworthy evidence. If the published scientific record is biased, its value is seriously compromised: Researchers are led to believe in phenomena that are frail or might not even exist at all. Theory development is led astray. The ability to explain the world to the public is undermined, and public trust in science is compromised. In short: If science fails to deliver trustworthy, reliable evidence, a society may wonder why it should invest in scientific endeavors at all.

In recent years, the trustworthiness of psychological science has been seriously questioned (Lilienfeld & Waldman, 2017). One important reason for the doubt and criticism has been the observation that many published psychological studies cannot be replicated in a straightforward fashion (e.g., Nosek & Lakens, 2014; Open Science Collaboration, 2015). Several problems that may contribute to this lamentable status have been identified, including low statistical power (Bertamini & Munafò, 2012; Maxwell, 2004), the use of questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), publication bias (Bakker, van Dijk, & Wicherts, 2012; Fanelli, 2010), and hypothesizing after the results are known (HARKing; Kerr, 1998). Together, these problems may lead researchers to seriously overestimate the robustness of the cumulative evidence in a field of investigation. True effect sizes can be critically smaller and less stable than the available evidence suggests. As a consequence, Psychology has started to experience all of the detrimental consequences alluded to above.

The most important methodological tool that can be used to quantitatively summarize the available evidence in a given research literature is a meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009; Gurevitch, Koricheva, Nakagawa, & Stewart, 2018; Johnson & Eagly, 2014). Meta-analyses summarize the results of multiple studies addressing the same research question to reach an overall understanding of the state of the evidence. Thus, the unit

of analysis changes from the individual level to the aggregated level—ideally, the complete body of evidence that has been collected with respect to a particular research question (Murad & Montori, 2013).

Meta-analyses have several strengths. One salient strength is that due to the greater statistical power, meta-analyses can be conducted to reliably detect even small effects that are not as easy to detect with single primary studies. Meta-analyses can also be used to estimate (summary) effect sizes with greater precision (i.e., narrower confidence intervals) than single primary studies. Importantly, meta-analyses can also estimate variation in underlying true effects (e.g., when different populations are investigated across studies, different manipulations are employed, or different dependent variables are used) and shed light on moderating factors that may have been missed or were impossible to investigate in the primary studies. These and other properties make the meta-analysis a powerful tool that researchers can use to obtain a comprehensive overview of what is known and not yet known in a given field of research.

In times of doubt about the replicability and robustness of individual primary studies, researchers are even more likely to rely on meta-analyses to obtain a trustworthy picture of the state of the evidence. Importantly, the validity of meta-analyses may also be threatened by the problems that lead to a lack of replicability and robustness in primary studies. For example, the quality of a meta-analysis crucially depends on the quality of the primary studies it is composed of. In a field featuring many poorly conducted studies, a meta-analysis may be unable to level out the biases of primary studies if these biases are systematic rather than unsystematic (Borenstein et al., 2009). Thus, it is imperative to examine the impact that various sources of bias can have on meta-analytic effect size estimates.

In recent years, two problems in particular have received considerable attention as presumably the leading causes of deficient robustness in psychological science: Questionable

research practices—often referred to as *p*-hacking—and publication bias (Bakker et al., 2012; Munafò et al., 2017; Nelson, Simmons, & Simonsohn, 2018). It is widely assumed that both *p*-hacking and publication bias can seriously distort the cumulative evidence and consequently the meta-analyses that are conducted to summarize this evidence.

There has been an active meta-scientific debate about the *prevalence* of *p*-hacking and publication bias (e.g., Dubben & Beck-Bornholdt, 2005; Hartgerink, 2017; Head, Holman, Lanfear, Kahn, & Jennions, 2015; Kuhberger, Fritz, & Scherndl, 2014). What has been surprisingly neglected are the quantifiable *consequences* of *p*-hacking and publication bias with respect to cumulative knowledge formation. Some of the most important questions are: To what extent do different degrees of *p*-hacking and publication bias distort meta-analytic effect size estimates? What are the relative impacts of *p*-hacking and publication bias in bringing about these distortions? And how might the consequences of *p*-hacking and publication bias depend on the extent to which the other exists; that is, how might they interact to jointly distort cumulative scientific evidence? This knowledge is crucial: In order to implement the structural and procedural changes in research institutions, publishing, funding, and policy that promise the greatest progress for obtaining a realistic reflection of reality from the published literature, the field needs to know which problems cause the greatest harm under which circumstances.

Here, we addressed these important questions about the quantifiable consequences of *p*-hacking and publication bias for cumulative knowledge formation by conducting a large-scale simulation study. In this study, we made no assumptions about the prevalence rates of *p*-hacking and publication bias. Rather, we simulated their consequences using a broad range of potential severities, thus accounting for (a) potential realities across a diverse array of research and (b) diverging assumptions about these prevalence rates by different researchers.

What are *p*-Hacking and Publication Bias?

Definition of *p*-hacking. The concept of *p*-hacking refers to nonprincipled decisions during data analysis that are aimed at reducing the *p*-value of a significance test and thus make the data look more robust than they actually are. Examples are selectively excluding outliers, collecting additional data without controlling for inflated error rates, or selectively controlling for covariates (John et al., 2012; Simmons et al., 2011). Thus, although there are several different *p*-hacks, they all serve as functionally equivalent means to the same end: To reduce an originally nonsignificant *p*-value to significance. Such *p*-hacking can be caused by bad intentions but may often be driven by good intentions to help the data reveal the insights that are presumably hidden in them and are otherwise not as clearly observable (Nelson et al., 2018). Also, it is likely that many researchers are not aware of the extent to which their data-analytic practices increase false-positive rates (Simmons et al., 2011).

The prevalence of *p*-hacking in (psychological) science is a subject of debate (Fiedler & Schwarz, 2016; John et al., 2012). Some researchers have argued that *p*-hacking is omnipresent and is so pervasive that it helps researchers get around a file drawer because they will *p*-hack (almost) any study into publishable significance (Nelson et al., 2018). Large-scale analyses have sought (and found) indirect evidence for *p*-hacking by examining empirical *p*-value distributions in the published literature that suggested a cluster of *p*-values just below .05 (e.g., Head et al., 2015; Masicampo & Lalande, 2012). These findings are consistent with the assumption that most *p*-hacking researchers stop once they reach an outcome that barely crosses the crucial .05 border. These analyses and their underlying assumptions have been criticized on methodological and logical grounds (e.g., Hartgerink, 2017; Lakens, 2015). They might also not be specific enough because they lump together all *p*-values reported across a large array of publications, including the many for which there was little publication pressure (e.g., manipulation checks, sanity checks, follow-up analyses, nonfocal hypothesis tests) with the few focal tests for which there was publication pressure

and therefore the incentive to *p*-hack. In sum, the true prevalence of *p*-hacking is unknown (Bruns & Ioannidis, 2016) and most likely varies across the different literatures.

Definition of publication bias. Publication bias occurs when many studies that did not produce the desired outcomes are not published (Fanelli, 2012; Franco, Malhotra, & Simonovits, 2014). Authors are less likely to submit "failed" studies for publication, and if they do, reviewers and editors are less likely to support the publication of such studies compared with "successful" studies that produced statistically significant outcomes. As a result, most studies in Psychology that get published report hypotheses that "worked" (Fanelli, 2010; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995).

Publication bias is a major threat to the validity of meta-analytic results. To reflect the true state of the evidence, meta-analyses require access to the full evidence base, or at least a representative sample of this evidence. If studies with certain characteristics are more likely to be included in a meta-analysis than others, this introduces systematic bias that distorts the conclusions that will be drawn. Meta-analyses enjoy a good reputation and are particularly trusted by many researchers due to their often seemingly authoritative data base. If they paint a misleading picture of the evidence because the evidence base is biased, scientific progress may be hampered because incorrect theories and beliefs will remain popular (Ferguson & Heene, 2012).

There is little disagreement in the literature that publication bias exists, but the actual prevalence of bias has been debated and tends to vary across subdisciplines and different areas of research (Fanelli, Costas, & Ioannidis, 2017). Some analysts have suggested that publication bias is pervasive and particularly so in the Social Sciences such as Psychology (Bakker et al., 2012; Fanelli, 2010, 2012; Ferguson & Brannick, 2012).

The Detrimental Impact of *p*-Hacking and Publication Bias

There is a general consensus that both *p*-hacking and publication bias exist in the psychological literature. What is under debate and unknown is their factual prevalence rates in Psychology as a whole and its subdisciplines. For the present study, the actual prevalence rates of *p*-hacking and publication bias were not our primary interest (and we did not seek to determine the actual prevalence rates). Instead, we sought to model the *consequences* of *p*-hacking and publication bias in terms of meta-analytic effect size distortions *as a function of* wide ranges of potential severities of *p*-hacking and publication bias.

How do *p*-hacking, publication bias, and their interplay distort meta-analytic effect size estimates? This can be conveniently illustrated with a funnel plot. Consider Figure 1. Let us assume researchers suspect a difference between two conditions but do not know whether or not this difference actually exists. Panel A depicts 1,000 simulated studies with a true population effect of zero. (Hence, in this example, the suspected difference does not exist.) Larger studies are located toward the top of the funnel and are more closely distributed around the true effect size. By contrast, smaller studies are located toward the bottom of the funnel and are more widely distributed around the true effect size. By definition, only 2.5% of all studies produce significant effects in the expected direction (genuine false positives) when a two-tailed significance test with $\alpha = .05$ is applied (i.e., studies that fall outside the funnel and to the right). These significant findings that fall in the expected direction have a high probability of getting published. All other studies have a lower probability of getting published (Bakker et al., 2012; Fanelli, 2012; Sterling, 1959; Sterling et al., 1995). This also includes 2.5% of all studies that produce significant effects in the unexpected direction (i.e., studies that fall outside the funnel and to the left).

The yellow-to-red colored dots on the right within the funnel represent studies that are "in danger of being *p*-hacked." These studies revealed nonsignificant effects in the expected direction that researchers might be able to push below the significance level through *p*-

hacking. Panel B depicts the same funnel plot as Panel A with the exception that about 50% of the studies that were "in danger of being *p*-hacked" were hacked to significance with resulting *p*-values that fell between .05 and .001. For these studies, the effect sizes ended up being inflated. Such *p*-hacking may have occurred in a variety of ways (e.g., unplanned inclusion of covariates, flexible outlier treatment). For the present purposes, it is inconsequential which specific *p*-hacks were used. They all serve the same purpose: to reduce an originally nonsignificant *p*-value to significance. (Again, this does not necessarily imply intentionally inappropriate behavior but may occur when a researcher runs multiple analyses while searching for a coherent story that the data may tell and without a clear awareness of the extent to which this approach can increase the false positive rate.)

The two funnel plots in Figure 1 reveal three important insights: First, imagine that researchers only had access to studies that fell outside the funnel to the right. All other studies would be lost to the file drawer. Summarizing this subset of studies in a meta-analysis (i.e., the only evidence available: only significant studies in the expected direction) would lead to a vastly exaggerated meta-analytic effect size estimate. This is the consequence of publication bias. Second, in this case, when only significant studies in the expected direction are available to summarize, it seems that it would not matter much whether a large number of these significant studies were *p*-hacked to significance (Panel B, black dots plus colored dots) or not (Panel A, only black dots): Both subsets of studies that fell outside the funnel to the right would cluster relatively closely together and therefore yield similar summary effect sizes. In other words, *p*-hacking would seriously increase the rate of false positive studies in Panel B (i.e., all colored dots outside the funnel in Panel B are false positives). However, despite the larger number of false positives, the estimated summary effect would not increase to a notable extent. Third, imagine there was no publication bias, and researchers had access to *all* 1,000 studies in Figure 1. It would be obvious that this literature would reveal no effect.

Again, it would not matter much whether *p*-hacking was absent (Panel A) or present (Panel B). In this case, there would simply be too many black-dot studies both inside and outside the funnel for the hacked colored-dot studies to make an appreciable impact on the meta-analytic effect size estimate.

In summary, the visual inspection of Figure 1 suggests that, perhaps surprisingly, *p*-hacking might not matter much for the estimation of meta-analytic effect sizes when the publication bias is close to 0% or close to 100%. Yet, what happens between these extremes is less clear. As the ratio of significant to nonsignificant studies changes, *p*-hacking may contribute additional bias.

The Present Research

We set out to formally examine these provisional observations in a large-scale simulation study. In this study, we generated sets of simulated studies and systematically varied the degrees of both *p*-hacking and publication bias. More specifically, we varied the probability with which a study in danger of being *p*-hacked would actually be *p*-hacked to significance. This reflects the pervasiveness with which researchers in a field (intentionally or unintentionally) *p*-hack an originally nonsignificant *p*-value to significance if this is in principle possible. We also varied the degree of publication bias by moving different proportions of nonsignificant studies to the file drawer so that they would be unavailable for researchers interested in meta-analyzing the respective (simulated) literature. Finally, we conducted random-effects meta-analyses across the remaining (*p*-hacked and non-*p*-hacked) studies and calculated the meta-analytic effect sizes. In a real research literature, these meta-analytic effect sizes would be used to approximate the true sizes of the effects of interest in the population.

Factors of influence. Of course, meta-analytic effect size estimates in the actual literature are influenced by many more factors than *p*-hacking and publication bias. To

generalize the findings from varying levels of *p*-hacking and publication bias, we also systematically varied several such factors of influence that may be present in the actual literature:

(1) Danger zone: How many studies are "in danger of being *p*-hacked"? Some researchers may believe that it is only possible to *p*-hack relatively small original *p*-values to significance (e.g., $p = .200$). Everything else may be unfeasible and reminiscent of the intentional fabrication of data. However, other researchers may believe that it is possible to *p*-hack even very large original *p*-values to significance (e.g., $p = .800$), for example, by employing complex combinations of various *p*-hacks (e.g., treatment of outliers, peeking at the data, inclusion of covariates).

The larger the danger zone for original nonhacked *p*-values is, the greater the influence of *p*-hacking on meta-analytic summaries. This is because a larger danger zone encompasses a larger number of studies that can be hacked and that make a more extensive horizontal movement toward significance in the funnel (i.e., they particularly distort the meta-analytic summary effect). The actual size of a danger zone in a given literature is impossible to know. It is therefore important to examine the influence of the size of the danger zone across a broad range of possible values.

(2) True effect size: When there is a true effect, nonhacked studies falling outside the funnel to the right will not represent false positives but will instead provide evidence of a real effect. The larger the true effect, the larger this proportion of studies (Simonsohn, Nelson, & Simmons, 2014a). Consequently, larger true effects should decrease the influence of *p*-hacking because the proportion of *p*-hacked studies in the set of all significant studies will be smaller than in a field

with smaller true effects. Similarly, larger true effects will mean that publication bias will have less of an effect because, out of all the studies that were conducted, a larger proportion will be significant and will have a high probability of getting published.

We examined a broad range of true effect sizes to allow for a comprehensive understanding of how the true effect size impacts the biases that *p*-hacking and publication bias exert.

(3) Heterogeneity: In the psychological research literature, there is not one true fixed effect size. Instead, true effects vary: One manipulation of a construct may be more effective than another manipulation; the same manipulation may be more effective in one population of participants than in another population; one dependent variable used to measure a construct of interest may be more sensitive to an experimental manipulation than another dependent variable, and so forth (Borenstein et al., 2009). The funnel plot (and a meta-analysis for that matter) specifies one mean effect across all studies. If heterogeneity is acknowledged (random-effects model), the effect size estimate reflects the mean of the underlying true effects. Thus, heterogeneity increases the variability of studies on the x-axis in the funnel depicted in Figure 1. This may lead to a larger number of genuinely significant studies. A recent analysis of between-study heterogeneity based on more than 700 meta-analyses provided evidence for substantial heterogeneity in Psychology and variability in the levels of heterogeneity across the various research literatures (van Erp, Verhagen, Grasman, & Wagenmakers, 2017). It is thus important to consider a broad range of values of heterogeneity when examining the impact of *p*-hacking and publication bias.

(4) Typical sample sizes: The more precise a study, the more accurately it can estimate the true underlying effect. Effect sizes based on smaller samples vary more strongly. Research literatures differ in how the sample sizes of individual studies are distributed: Some literatures typically feature larger, more precise studies than others (Marszalek, Barber, Kohlhart, & Cooper, 2011). This may influence the impact of *p*-hacking and publication bias, for example, because smaller studies require larger effect sizes to achieve statistical significance. Thus, it is important to consider the influence of various typical sample sizes when trying to understand the impact of *p*-hacking and publication bias on meta-analytic effect size estimates.

(5) The probability that significant studies will be published: Studies with "positive" results (i.e., significant results in the expected direction) are more likely to be published than studies with "negative" (i.e., nonsignificant) results (Bakker et al., 2012; Fanelli, 2012; Sterling, 1959). However, not all studies that "worked" will be published. For example, authors may be reluctant to submit a study for publication if they feel the study did not provide strong enough evidence in support of the favored hypotheses (Giner-Sorolla, 2012). Also, reviewers and editors may be reluctant to advocate the publication of studies that might not extend previous knowledge far enough to warrant publication (Nosek, Spies, & Motyl, 2012). Thus, there may be variability in a significant study's probability of getting published.

The lower the probability that significant studies will be published, the smaller the impact of *p*-hacking because, with a lower probability of publication, fewer of the *p*-hacked studies that could bias the meta-analytic estimate will be published. Also, the lower the probability of publication, the smaller the impact of

> publication bias because the distortion introduced by the nonpublication of nonsignificant results is offset to the extent that significant findings are also not published. (Everything else being equal, there would be no bias in the mean effect size estimate if the same proportions of significant and nonsignificant findings were not published.)

The interplay of *p*-hacking, publication bias, and all factors of influence can conveniently be graphically examined by means of two freely accessible interactive online applications that will be discussed in the Method section. Although we focused on the meta-analysis of two-group comparison designs using Cohen's *d* as the measure of effect size (Cohen, 1988), we believe that the insights gained by our study can be readily applied to various kinds of meta-analyses using different effect sizes.

Method

We simulated the effects of varying degrees of *p*-hacking and publication bias on the distortion of meta-analytic effect size estimates as a function of the five factors of influence identified in the Introduction: danger zone, true effect size, heterogeneity, typical sample sizes, and the probability that significant studies will be published. Essentially, this process involved simulating many different versions of the sets of studies depicted in the funnel plots in Figure 1, henceforth referred to as configurations. Figure 1 displays two of the many possible configurations. Simulating each configuration was a multistep process. In the first major step, studies were generated with varying levels of true effect sizes and heterogeneity and were based on different sample size distributions. In the second major step, varying levels of *p*-hacking and publication bias were introduced to the studies generated in the first step. In the third step, the meta-analytic summary effects of the configurations (a precision-weighted average of all studies in a configuration) were graphically depicted in outcome figures that illustrate how the five factors of influence changed the interplay between *p*-

hacking and publication bias in distorting the cumulative evidence base in our simulations. All simulations were conducted using R (R Core Team, 2017). The meta-analytic models were fit using the rmeta package (Lumley, 2012).

In total, we simulated 282,240 different configurations. For reasons of clarity, we cannot report the results of all levels of the factors of influence (and their various combinations). However, we offer two interactive online applications that provide visual representations of the effects of the simulations: Interactive online application 1 (https://bit.ly/2LIvRX7) visually represents how the factors of influence impact the funnel plot depicted in Figure 1. Interactive online application 2 (https://bit.ly/2Vno8gH) visually represents effects of the factors of influence on the graphical displays of the results akin to Figure 2. Both applications offer the opportunity to examine the results as a function of additional values of the factors of influence not reported in the manuscript (e.g., additional values of true effect sizes, danger zone, severity of *p*-hacking, heterogeneity).

Study Generation

The first step was to simulate individual studies in which two independent groups were compared. The true between-group mean difference per study was set to the sum of a fixed effect δ and a random effect τ_i, where values for τ_i were randomly drawn from a normal distribution with mean 0 and standard deviation τ. Fixed-effects models are based on the assumption that there is one true effect size underlying all studies included in a meta-analysis. By contrast, random-effects models are based on the assumption that true effect sizes may differ across studies due to, for example, different effects in different populations or different experimental manipulations. In our simulation we simulated random-effects (by introducing heterogeneity) and accordingly used random-effects meta-analysis for modeling. We assume that for most of the psychological literature, a random-effects model is more plausible than a fixed-effects model (Borenstein et al., 2009).

The τ and δ values were varied across configurations. We entered 0, 0.2, or 0.5 for the δ values (true effect size, Factor 2 listed above) and 0.10, 0.2, or 0.32 for the τ values (heterogeneity, Factor 3). Selected values for δ were based on Cohen's conventions for small and medium effects (Cohen, 1988). We assume that these cover the majority of effects in Psychology (e.g., Bosco, Aguinis, Singh, Field, & Pierce, 2015; Gignac & Szodorai, 2016; Richard, Bond, & Stokes-Zoota, 2003). According to a recent meta-analysis, our chosen values for τ represent the 25%, 50%, and 75% quantiles in an empirical distribution of τ estimates in Psychology (van Erp et al., 2017). For additional values, see interactive online application 2.

Samples sizes of individual studies were set to $n_i = m_j + \chi_i^2 * k$, where χ_i^2 was randomly drawn from a χ^2 distribution with three degrees of freedom, k was set to 8, and m_j was varied across configurations. Values more than 80 points above m_j were truncated (about 1.9% of the distribution). The default value for m_j was 20. The resulting distribution was right-skewed, with skewness = 1.03, *Mdn* = 39, *M* = 42.61, *SD* = 16.80, Min = 20, Max = 100. Hence, the distribution included both small and large sample sizes, but small sample sizes were more prevalent. In addition to the default of $m_j = 20$, we also realized configurations with $m_j = 10$ and $m_j = 50$ (typical sample sizes, factor of influence 4). Thus, we utilized sample size distributions with *Mdn* = 29, 39, and 79 per condition. This approach enabled us to shift the central tendency of the distribution without changing its shape.[1] See Figure S1 in the supplemental materials for a graphical depiction of the three sample-size distributions. We preferred a synthetic sample-size distribution over an empirically derived distribution for two reasons. First, investigations of historical sample sizes in Psychology (e.g., Fraley & Vazire, 2014; Marszalek et al., 2011) typically do not report the study design,

[1] The specific shape of the sample size distribution was arbitrary and based on plausibility assumptions. In additional simulations, we systematically varied *df* and *k* to ensure that our results were not artifacts of the shape of the selected sample-size distribution. This was the case. Variations in the shape of the distribution affected the results to a negligible extent.

rendering it impossible to infer the typical sample size per condition. Second, sample sizes in Psychology are changing rapidly (Nelson et al., 2018; Sassenberg & Ditrich, 2019), and we aimed to make projections for the present (and future) rather than the past.

For each configuration, we started with a set of 1,000 simulated studies and computed the standardized mean difference (*d*, Cohen, 1988), the standard error of the standardized mean difference (SE_d), and the *p*-value for each study.[2]

Introduction of Biases

Next, *p*-hacking and publication bias were applied to the set of studies.

***p*-hacking.** Studies were defined as "in danger of being *p*-hacked" if *d* was positive and the *p*-value fell above .05 and below a predefined cut-off value. The upper border of this danger zone was linearly increased across the full range of standard errors so that the danger zone was smallest for studies with minimum SE_d (top of the funnel, Figure 1) and largest for studies with maximum SE_d (bottom of funnel, Figure 1). This approach resulted in the curved danger zone border visible in Figure 1. The danger zone is smaller for precise studies and larger for imprecise studies. For example, one danger zone we considered was .4/.6, such that studies with the lowest standard errors (i.e., the largest sample size in the set) were in danger if their *p*-values fell between .050 and .400, and studies with the highest standard errors (i.e., the smallest sample size in the set) were in danger if their *p*-values fell between .050 and .600. This reflects the fact that studies with small sample sizes are easier to *p*-hack compared with studies with larger sample sizes (Bakker et al., 2012). Note that the specific largest and smallest *n* per cell in a given configuration depends on which 1,000 values (for 1,000 studies) are randomly selected from the sample size distribution and, of course, on the selected

[2] Note that the results for summary effect sizes, the primary outcome of our simulation, are independent of the number of simulated studies. This is the case because simulated studies in the meta-analysis are independent, such that studies already existing in the set have no impact on new studies being added. Every simulated configuration can be viewed as a data-generating mechanism with a specific underlying distribution of effect sizes. The presented results are estimates for the expected values (means) of the distributions. This expected value is the same, whether we draw $k = 100$, $k = 1{,}000$, or $k = 10{,}000$ from the distribution.

sample size distribution (small, standard, large). We report results for three levels of the danger zone factor: .2/.4, .4/.6, or .6/.8 (factor of influence 1). The effects of additional smaller and larger danger zones can be examined with interactive online application 2.

Any study identified as in danger of being *p*-hacked was then hacked with a certain probability. It did not matter which specific *p*-hacking technique was used to reduce the *p*-value of a study because different *p*-hacks serve as means to the same goal, that is, to lower the *p*-value of an originally nonsignificant study to significance. If a study was *p*-hacked, its *p*-value was replaced with a value randomly drawn from a triangular distribution with Max = 0.049, Mode = 0.049, and Min = 0.001, thus satisfying the assumptions that (a) *p*-hacking leads to a left-skewed distribution of significant *p*-values (more *p*-values close to .05 than expected given a true effect; Simonsohn et al., 2014a; Simonsohn, Nelson, & Simmons, 2014b), and (b) some studies nevertheless achieve quite low *p*-values after hacking (Simonsohn, Simmons, & Nelson, 2015). (For robustness checks of this procedure, see the Discussion section.) For *p*-hacked studies, *d* was then recomputed from the new *p*-value and the sample size. Average *p*-hacking probabilities were set to 0, 0.4 (only depicted in the illustrative example in Figure 2A) and 0.8 (*p*-hacking probability). Again, the choice of these values does not indicate that we deem these *p*-hacking probabilities particularly likely to reflect reality. Instead, we chose rather extreme values (no *p*-hacking at all, 80% of all studies in danger of being *p*-hacked were in fact hacked, respectively) and a moderate value to illustrate the potential range of influences that the probability of *p*-hacking can exert. (Although possible, we deemed *p*-hacking probabilities even higher than 0.8 unlikely to be representative of the psychological research literature.)

To account for the fact that *p*-values closer to $p = .05$ are easier to push over the significance threshold than larger *p*-values, we introduced a linear gradient to the probability of *p*-hacking: Values that fell exactly on the (horizontal) middle of the danger zone received

the nominal average *p*-hacking probability (i.e., 80% at the 0.8 level). Values that fell at either (horizontal) end of the danger zone, that is, almost exactly at the significant threshold of $p = .05$ or almost exactly at the left-hand border of the zone, received *p*-hacking probabilities of 0.2 above and 0.2 below the nominal probability, respectively. The probability was linearly decreased from the left to the right border of the danger zone. In Figure 1, this is visually represented via the yellow-to-red color gradient. For simplicity, we refer to the *p*-hacking probabilities by their nominal (center) level. The gradient did not apply to *p*-hacking probabilities of 0 and 1 (only included in the online applications).

Publication bias. We distinguished between two different kinds of publication bias: First, the nonpublication of studies that "did not work." Second, the nonpublication of studies that "worked." Following common conventions, we reserve the term publication bias for the former and call the latter "the probability of publishing significant studies."

Publication bias was simulated by removing a random subset of studies that were either nonsignificant with positive *d*-values or had negative *d*-values. The severity of publication bias was operationalized by varying the percentage of studies removed in steps of 5% from 0% (no publication bias) to 100% (perfect publication bias).

The probability of the publication of significant studies was modeled analogously. Specifically, significant studies with positive *d*-values were included in the final set with probabilities of 100%, 90%, or 80% (factor of influence 4). For the effects of a smaller probability of the publication of significant studies see interactive online application 2.

Meta-analysis and graphical displays of the simulation results. In total, we simulated 282,240 unique configurations. For each configuration, we fit random-effects meta-analysis models. Because the simulations are probabilistic, running the same configuration twice never yields perfectly identical results. To stabilize the results, we ran all configurations reported in this article 1,000 times and averaged the resulting estimates. All

other configurations that are accessible in the interactive online application 2 were run 100 times. This procedure reduced the influence of chance to a negligible amount. Finally, the summary effects were retrieved, some of which are presented in graphical displays of the simulation results.

Figures 2B-2F depict summary effects of 126 configurations each (630 in total): Twenty-one levels of publication bias severity, two levels of *p*-hacking probability, and three levels of one selected factor of influence. All other factors of influence were held constant within each figure. In Figure 2A, all factors of influence were held constant at a set of default values.

Results

In a first step, we will expand on one representative example that illustrates how varying severities of *p*-hacking and publication bias distort the meta-analytic effect size estimate of one hypothetical simulated literature (Figure 2A). In this example, the true effect was set to $d = 0.2$, heterogeneity was set to the median value in psychological science, $\tau = 0.2$ (van Erp et al., 2017), the danger zone was set to .4/.6, and the probability that significant studies would be published was set to 90%. Although the true values from any given area of the literature are unknown, this configuration was intended to reflect one plausible approximation of reality for some research areas. Many additional configurations are reported in the following section, and other ones can conveniently be examined with interactive online application 2.

In a second step, we examined the unique effects of each of the five factors of influence—danger zone, true effect size, heterogeneity, typical sample sizes, and the probability of publishing significant studies—by separately varying the values of one factor while holding the other factors constant (Figures 2B-2F).

Step 1: Illustrative Example

The three lines in Figure 2A reflect the (biased) estimated meta-analytic effect size when *p*-hacking was absent (*p*-hacking probability = 0, solid line), moderate (*p*-hacking probability = 0.4, dashed/dotted line), and when *p*-hacking was severe (*p*-hacking probability = 0.8, i.e., 80% of all studies in the danger zone were *p*-hacked to significance, dashed line).

The simulations revealed several interesting insights: First, in the absence of *p*-hacking (solid line), the effect size bias increased exponentially as the severity of publication bias increased. Even when 50% of all "failed" studies were lost to the file drawer, there was only a relatively small effect size bias (true effect $d = 0.2$, estimated effect $d_{est} = 0.26$, effect size bias $d_{bias} = 0.06$). This pattern changed dramatically as publication bias became very severe (e.g., 95%). When this happened, the estimated effect was $d_{est} = 0.50$ ($d_{bias} = 0.30$), indicating that researchers would conclude that this literature represents a robust phenomenon with an average effect size that is more than twice the size of the true effect. An implication of this observation is that even modest reductions in publication bias can greatly reduce the effect size bias in the literature when researchers suspect very severe publication bias.

Second, the effect of moderate *p*-hacking (dashed/dotted line, *p*-hacking probability = 0.4) was only modest compared with no *p*-hacking (*p*-hacking probability = 0). Even assuming that 80% of all studies in the danger zone would be *p*-hacked to significance (dashed line, *p*-hacking probability = 0.8) did not dramatically increase the effect size bias. In the latter case of *p*-hacking probability of 0.8, the additional bias due to *p*-hacking (and its interaction with publication bias) remained below $d_{hack} = 0.1$ across the various levels of publication bias (d_{hack} equals the difference between the solid and dashed lines; see Figure S2). At very high levels of publication bias (i.e., 90-100%), the additional bias due to *p*-hacking was negligible and even became negative (i.e., *p*-hacking led to a slight *reduction* in the overall degree of bias in these cases). Between 0% and approximately 80% of publication bias *p*-hacking had the greatest relative biasing effect (see Figure S2).

A different way to interpret Figure 2A is to examine what it takes to produce a certain effect. For example, let us assume that a meta-analysis revealed an average effect of d_{est} = 0.4. A researcher who is very experienced in the respective field may believe the true effect is d = 0.2 at best. This researcher is interested in what degrees of *p*-hacking and publication bias may have produced the supposedly biased effect size estimate of d_{est} = 0.4. In the absence of publication bias (very left part of Figure 2A), *p*-hacking alone would be unable to even come close to a bias of d_{hack} = 0.2. The actual effect size bias in the absence of publication bias would be only $d_{hack} = d_{bias}$ = 0.01 (i.e., d_{est} = 0.21) if 40% of all studies in danger of being *p*-hacked were *p*-hacked, and only $d_{hack} = d_{bias}$ = 0.04 (i.e., d_{est} = 0.24) even when 80% of all studies in danger of being *p*-hacked were hacked. By contrast, in the absence of *p*-hacking, publication bias alone would be able to seriously bias the estimated effect: It would take a severity of 85% publication bias to double the true effect size to an estimated d_{est} = 0.40. Illustrating the interaction of *p*-hacking and publication bias, assuming a *p*-hacking probability of 80%, a publication bias of "only" 75% would also lead to a biased effect size of d_{est} = 0.40. In sum, to introduce serious meta-analytic effect size bias into the literature, publication bias is necessary, but *p*-hacking is not. However, in combination with publication bias *p*-hacking may exert considerable effect size bias.

Step 2: Factors of Influence

Next, we examined the specific effects of each of the five factors of influence and how variations in these factors could affect the general conclusions drawn from the illustrative example. For ease of interpretation, we only considered the two most extreme values of *p*-hacking probability for the following analyses (0% and 80%).

Danger zone. The illustrative example was based on a danger zone of p = .050 to .400 (larger, more precise studies) to p = .050 to .600 (smaller, less precise studies). It is possible that *p*-hacking is less or more feasible, and, consequently, fewer or more studies are

in danger of being *p*-hacked. Figure 2B depicts the results when the danger zone was set to .2/.4 (blue line), .4/.6 (black line), and .6/.8 (red line). Reducing the danger zone decreased the impact of *p*-hacking slightly; expanding the danger zone increased the impact of *p*-hacking slightly. However, even under the most severe circumstances (danger zone .6/.8, *p*-hacking probability = 0.8, red dashed line), the additional bias due to *p*-hacking (and its interaction with publication bias) remained smaller than $\Delta d_{hack} = 0.10$ (difference between solid black and dashed red line, see also Figure S3). Overall, strongly varying the size of the danger zone from .2/.4 to .6/.8 had a modest impact on the effect size bias. Obviously, in the absence of *p*-hacking (solid line), the size of the danger zones did not impact the estimated effect size.

True effect size. Figure 2C illustrates that in literatures with a true effect size of $d = 0$, the exponential relationship between publication bias and the estimated effect size was greatly amplified (Figure 2C, blue lines). At a publication bias of 90% and in the absence of *p*-hacking, the estimated biased effect size was $d_{est} = d_{bias} = 0.21$; at a publication bias of 95%, it was already $d_{est} = 0.31$. In addition, publication bias and *p*-hacking interacted more strongly than when the true effect was $d = 0.2$ as in the default example: The additional biasing effect of *p*-hacking was much stronger at high degrees of publication bias (e.g., 80% or 90%) than at low degrees of publication bias (e.g., 0% or 10%; difference between the solid and dashed blue lines; see also Figure S4).

By contrast, in literatures with a true effect size of $d = 0.5$, the exponential relationship between publication bias and the estimated effect size was greatly dampened (red lines). In the absence of *p*-hacking, extreme levels of publication bias (e.g., 90 or 95%) still biased the effect size by approximately $d_{bias} \approx 0.15$, but the distortion was much smaller than at lower true effect sizes. At high levels of publication bias (e.g., 90 or 95%), severe *p*-

hacking even had a slightly dampening influence on the estimated effect size, effectively leading to *less* biased effect size estimates.

A true effect size of $d = 0$ may be regarded as deserving special attention because in this case, publication bias and *p*-hacking may "produce" an effect out of thin air. Figure 2C illustrates that the overall level of bias may be particularly large when $d = 0$. Therefore, we additionally explored the effects of all factors of influence specifically for the case of $d = 0$ (Figure S5). Results revealed that the particularly pronounced interaction between *p*-hacking and publication bias at $d = 0$ remained intact and the factors of influence exerted similar influences as in our illustrative example with $d = 0.2$. As before, the bias introduced by *p*-hacking alone was small, but the relative amount of bias attributable to *p*-hacking and its interaction with publication bias was noticeably increased for $d = 0$ compared to $d = 0.2$.

Heterogeneity. Lower heterogeneity (i.e., $\tau = 0.10$, 25% quantile) slightly reduced the effect size bias at high levels of publication bias (Figure 2D, blue lines). Higher heterogeneity (i.e., $\tau = 0.32$, 75% quantile) slightly increased the effect size bias at high levels of publication bias (red lines). Both effects were rather modest.

Typical sample size. Entering smaller (*Mdn* = 29 participants per condition, Figure 2E, blue lines) rather than standard sample sizes (*Mdn* = 39 participants per condition, black lines) into the simulation led to slightly stronger biasing effects of *p*-hacking and publication bias, especially at high levels of publication bias (> 80%). This was plausible because smaller samples require larger effect sizes to achieve statistical significance. Conversely, larger sample sizes (*Mdn* = 69 participants per condition, red lines) were associated with smaller effect size biases, especially at high levels of publication bias.

Probability of publishing significant studies. The assumption that 80% or 100% rather than 90% of all studies that "worked" would be published had only negligible effects on the estimated effect size (Figure 2F).

Supplemental analyses. We additionally explored the effects of publication bias and *p*-hacking on the number of studies in the meta-analysis (*k*, see Figure S6) and the precision of the summary effect (standard error, see Figure S7). Besides the switch of outcomes, Figures S6 and S7 are identical to Figure 2. The results for *k* are straightforward. The number of studies decreased linearly with increasing levels of publication bias, from 1,000 studies at 0% publication bias to about 200 studies at 100% publication bias in the default configuration. This effect was attenuated by *p*-hacking because studies were "saved" from ending up in the file drawer. At 80% *p*-hacking and 100% publication bias, about 500 studies remained, depending on the configuration.

When standard errors were set as outcomes, an interesting pattern emerged. With increasing publication bias and no *p*-hacking, standard errors also increased. This is intuitive, because publication bias removes studies from the meta-analysis and smaller meta-analyses are less precise. However, when *p*-hacking was added, this effect was effectively canceled out. When *p*-hacking was at 80%, increasing publication bias ded not decrease precision. Rather, precision remained approximately stable across the range of publication bias (0% - 100%). The cancellation effect occured because *p*-hacking shifts effect sizes to a relatively narrow corridor outside the border of the funnel, thus creating a tightly-packed cluster of effects that results in a high-precision estimate.

When the true effect was zero (Figure S7C, solid, blue line) the impact of publication bias on precision was especially pronounced, because with no true effect, only 2.5 percent of studies reach significance by chance. Even here, adding *p*-hacking to the mix increased precision notably. By this way, publication bias and *p*-hacking were working in concert to create the illusion of a precise non-zero effect that was in truth zero.

Discussion

Single studies are rarely conclusive. Therefore, scholars rely on meta-analyses that shift the focus from single studies to aggregated evidence: Different researchers contribute to the same research question through replication and extension. This collaborative approach to science holds great promise for the advancement of knowledge, but it is sensitive to distortions. If the cumulative evidence base is severely biased, researchers run the risk of drawing false conclusions. With the present simulations, we examined how a broad range of levels of severity with respect to *p*-hacking and publication bias could distort cumulative science as indicated by meta-analytic effect size estimates, considering broad variation in various factors of influence that are likely to exist in realistic research literatures. This study provides several key insights: First, *p*-hacking and publication bias interact: A high level of publication bias can greatly distort the available evidence base. At relatively low and high levels of publication bias, even severe *p*-hacking contributes little additional bias. At medium levels of publication bias, however, *p*-hacking can contribute considerable additional bias, especially when true effects are negligible or even approach zero.

The reason underlying this interaction is simple: When publication bias is low, there are so many studies that contribute to a valid estimation of effect sizes that *p*-hacking is not able to seriously distort such a robust body of evidence. When publication bias is high, effect sizes are already considerably distorted. Hacking some of the few remaining nonsignificant studies to significance only adds to the body of significant studies that already dominate the meta-analytic effect size estimate, but such *p*-hacking does not necessarily change the already biased mean estimate. When publication bias is not particularly low or high, however, considerable *p*-hacking will inflate the effect sizes of the originally nonsignificant studies so much that the overall effect size estimate will be more affected than under other circumstances. This is particularly the case when there is only a negligible or no true effect.

Second, the factors of influence we considered impact the interplay of *p*-hacking and publication bias to varying degrees. No factor of influence fundamentally changes the general interaction pattern, but larger sample sizes, lower heterogeneity, and particularly larger true effects protect against the biasing influence of *p*-hacking and publication bias. We briefly elaborate on why this is the case.

Smaller sample sizes are associated with stronger effect size biases because, in small samples, large effect sizes are needed to achieve statistical significance, leading to greater deviations from the true effect. Everything else being equal, *p*-hacking therefore contributes particularly much effect size bias in studies with small samples. By contrast, in larger samples, smaller effect sizes can achieve statistical significance, which is usually the critical threshold for publication.

When heterogeneity is high, effect sizes are more widely dispersed around the mean effect. This implies that more results will achieve statistical significance (see interactive online application 1 for a graphical illustration). Of these, the ones that are significant in the expected directions will not be affected by publication bias. Instead, they will likely be published as evidence in support of the expected phenomenon. This leads to a greater effect size bias. Put differently, low heterogeneity offers partial protection against extreme effect size biases.

By far, the strongest protective role against effect size bias emerged for larger true effects. Both the maximum level of bias introduced by publication bias alone and the maximum additional bias from *p*-hacking and its interaction with publication bias were strongly reduced as the true effects increased. The reason for this is simple: The stronger the true effect, the greater the number of studies that will achieve (high levels of) statistical significance even without any *p*-hacking (Simonsohn et al., 2014b). In this case, *p*-hacked studies contribute to this pool of significant studies without seriously changing the effect size

(see interactive online applications). The large numbers of significant studies with substantial true effects offer another beneficial consequence: Even when nonsignificant studies are sent to the file drawer, this will not have as severe an impact on the overall effect size estimate because the large number of significant studies attests to the real effect size.

These observations about large true effects have a flip side: Maximum bias is greatest when there is no true effect at all ($d = 0$). In this case, publication bias has the most leverage to distort meta-analytic effect size estimates, and *p*-hacking can add considerable additional bias when publication bias is high (e.g., higher than 50%). In other words, *p*-hacking and publication bias may produce seemingly robust meta-analytic effects out of thin air when in reality the effect is zero.

Taken together, the observations relating to variations in true effect sizes lead to a central and reassuring insight offered by the present simulations: Researchers working in areas involving phenomena with healthy, robust true effects need to worry much less about *p*-hacking and publication bias compared with those investigating small effects. In the case of strong effects, meta-analytic effect size bias will remain modest even under unfortunate conditions. Unfortunately, in many cases, researchers will not know with certainty whether or not an effect of interest has a robust or a negligible true effect before conducting a meta-analysis.

In supplemental analyses, we investigated how *p*-hacking and publication bias affect the precision of a summary effect. This analysis revealed that *p*-hacking can make meta-analytic estimates appear more precise than they actually are, because effect sizes are shifted into a narrow corridor just outside the significance border.

The Role of *p*-Hacking

Some researchers have intensively argued that *p*-hacking—not publication bias—is *the* major threat to psychological science (e.g., Nelson et al., 2018; Simmons et al., 2011).

Why do our simulations paint a more nuanced picture? The major difference between previous work on *p*-hacking and the present approach is that previous work focused on how dramatically *p*-hacking increased the rate of *false positives*, whereas, with the latter, we focused on the *distortion of meta-analytic effect sizes*. We argue that considering distortions of meta-analytic effect size estimates is important because not only false positives, but also substantial distortions in such estimates can impede scientific progress, lead researchers and practitioners astray, and result in a waste of resources if research is based on invalid inferences.

Is this meant to imply that researchers should not worry about *p*-hacking and false positives? Not at all. False positives are a major concern and can exert detrimental influences: For example, in small, emerging literatures, a few prominently published false positives may prematurely lead to the impression of a robust phenomenon. Likely, the more abundant a research literature, the less influence will a handful of prominent studies exert on the researchers in a field (provided there is no extreme publication bias). However, if—for whatever reason—a literature never develops to a substantial size, the detrimental influence of a few prominent false positive findings may remain strong for a long time.

One reason why a literature might never develop to a substantial size is when the true effect is zero, and there is no *p*-hacking in that particular literature. In this case, researchers would not be likely to produce a large number of studies because it would quickly become clear that there is nothing to find. This would consequently affect the number of studies included in any meta-analysis of this literature or would even determine whether a meta-analysis gets conducted in the first place. If, however, there is *p*-hacking in a literature with a true null or a negligible effect then *p*-hacking may make this literature appear as if there is something there. This may encourage other researchers to conduct further studies on this effect – a lamentable waste of resources. Preventing *p*-hacking would be an important

prophylactic against the risk of contributing further confidence in a potentially non-existing effect and ultimately against conducting meta-analyses of many false-positive studies, because the literature would never develop to a substantial size if researchers gave up on effects that would rather consistently spread around null. Precluding *p*-hacking is therefore also important to avoid a literature to accumulate when there is no interesting effect.

We emphasize that with the present work, we did not aim to question previous assertions about *p*-hacking and its deleterious effect on the rate of false positives. In fact, our work corroborates them: In our illustrative example (Figure 2A), *p*-hacking added a large number of false positives: About 55% (!) of all significant studies were false positives due to *p*-hacking. Nevertheless, in the absence of publication bias, *p*-hacking biased the effect size estimate only by $d_{bias} = 0.04$. This demonstrates that the "rate of false positives" and "effect size bias" are two different ways to examine the consequences of *p*-hacking and publication bias. Both are important, but a high rate of false positives does not necessarily imply a strong bias in cumulative effect sizes.

Implications

The importance of preventing *p*-hacking has rightfully received considerable attention in recent years (Nelson et al., 2018), but the prevention of publication bias has not received as much attention. This needs to change. Both need to become top priorities in psychological research. Both an increase in false positives (the most tangible consequence of *p*-hacking) and distortions of meta-analytic effect sizes (the most tangible consequence of publication bias) have potentially deleterious consequences for psychological science.

Estimates of the prevalence of *p*-hacking vary widely (Fiedler & Schwarz, 2016; Hartgerink, 2017; Head et al., 2015; John et al., 2012; Nelson et al., 2018). The severity of publication bias likely varies considerably by research area, but evidence suggests it may be high in general in Psychology (and higher than in many other sciences; Bakker et al., 2012;

Fanelli, 2010; Fanelli, 2012; Fanelli et al., 2017; Ferguson & Brannick, 2012). Therefore, the field may want to take a conservative position, assume drastic severities, and consider what should be done to curtail the consequences. The good news is: Even moderate reductions of a potentially strong publication bias will greatly reduce its biasing effects (see Figure 2). How can this be accomplished? We suggest four easy and cost-efficient solutions.

First, funding institutions have a strong interest in their money being used in the most efficient way to foster scientific progress. Therefore, they may enforce the transparent reporting of data and the results of all studies that have been paid for by a grant. In the case that not all studies end up being published, the remaining studies could mandatorily be placed in repositories such as the Open Science Framework and made public after a certain period of time. Funding institutions could even go so far as to make the allocation of future grants partly contingent on such transparent reporting for studies paid for by past grants. Second, more journals should require authors to clarify whether they have conducted any additional studies that addressed the same research question. These should be reported, and the results and data should be made available (e.g., in a supplement, via a link to a data repository). Journals could introduce standardized tables in which authors report all studies that have been conducted and identify them as pilot studies, actual tests of the hypotheses, and so forth. Even if not all authors responded truthfully to these requirements, this measure alone would unearth a considerable share of otherwise hidden studies and thereby considerably improve the robustness and validity of meta-analytic findings. Third, registered reports lessen both *p*-hacking and publication bias (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015; Jonas & Cesario, 2016). To the extent that more journals give this format ample journal space, publication bias will be reduced. Fourth, not only should preregistrations of studies (Nosek, Ebersole, DeHaven, & Mellor, 2018; van 't Veer & Giner-Sorolla, 2016) incentivize and regularly entail information about the study design, outcome variables, and analysis

plans, but they should also encourage or even mandate to make the data public, even if the study remains unpublished. In this way, not only will preregistrations reduce *p*-hacking, but they will also make research discoverable for meta-analysts, even if the data are never published in a journal. More generally, preregistration and adherence to open science standards that emphasize the unbiased access to materials, data and code will improve the replicability not only of individual studies but also of meta-analyses in the long run (Gurevitch et al., 2018; Nosek et al., 2018). In sum, the field needs a shift in culture. Researchers need to be aware not only of the detrimental consequences of *p*-hacking, but also publication bias, and they also need to be incentivized as well as required to minimize it. Minimizing publication bias comes with an extra premium: If all well-conducted studies are published independent of the outcome, incentives to *p*-hack will be drastically reduced. Preventing publication bias will indirectly reduce *p*-hacking as well.

The present simulations reveal how strikingly a strong severity of publication bias may distort meta-analytic effect size estimates even in the (albeit unrealistic) absence of *p*-hacking. One implication of these findings is that the field needs techniques that validly and reliably correct for the effects of publication bias under realistic circumstances (e.g., varying true effects, heterogeneity, scarcity of nonsignificant studies).

Various techniques have been proposed to correct for publication bias (e.g., trim and fill, Duval & Tweedie, 2000; *p*-curve, Simonsohn et al., 2014; *p*-uniform, van Assen, van Aert, & Wicherts, 2015; PET-PEESE, Stanley & Doucouliagos, 2014; selection models, Iyengar & Greenhouse, 1988). Two recent large-scale simulation studies compared the performance of several methods in correcting for publication bias under conditions that were realistic in various psychological research literatures (Carter, Schönbrodt, Gervais, & Hilgard, 2019; Renkewitz & Keiner, 2018). Using slightly different approaches, both studies concluded that no single method consistently performed well under diverse circumstances

and no method consistently outperformed the others (but see van Aert & van Assen, 2018, for an advancement of the *p*-uniform method). Ideally, future versions of these techniques will also be suited for modern meta-analytic methods, such as robust variance estimation (Hedges, Tipton, & Johnson, 2010) or multilevel meta-analysis (Van den Noortgate & Onghena, 2003), which can account for effect size dependencies. These meta-analysis techniques are increasingly used with data sets that include more than one effect size from the same study, but as yet researchers using these methods have to resort to bias correction methods developed for traditional meta-analysis that only allows the inclusion of one effect size per study (e.g., Coles, Larsen, & Lench, 2019; Friese, Frankenbach, Job, & Loschelder, 2017).

For researchers who rely on meta-analyses, our findings provide a starting point from which to estimate the combinations of *p*-hacking, publication bias severity, and factors of influence that could produce a given meta-analytic effect size estimate. An expert in a given literature may be able to make informed guesses about realistic values of the factors of influence. Using interactive online application 2, a researcher may run sensitivity analyses of various values of the factors of influence to check which combinations of *p*-hacking and publication bias may realistically produce a given meta-analytic estimate.

Potential Objections to the Simulations

In this section, we address some arguments against some assumptions underlying the present study and the conclusions than can be drawn from it.

The severity of *p*-hacking was overestimated. Critics may argue that it is unrealistic to *p*-hack initial *p*-values of .8 to significance without deliberately manipulating the data and that assuming a prevalence rate of .8 for *p*-hacking—80% of all studies in danger are in fact hacked—is far too pessimistic. In this case, the impact of *p*-hacking in our simulations would be overstated. In response, we refer to Figure 2, which illustrates that we simulated the impact of *p*-hacking across a broad range of severities in terms of both the probability of *p*-

hacking and the size of the danger zone. Our simulations are thus informative with respect to a broad range of potential realities in different research areas.

The severity of *p*-hacking was underestimated. One potential objection is that *p*-hacking was not modeled severely enough because researchers will basically do everything to avoid losing a study (and the resources they invested in it) to the file drawer (Nelson et al., 2018). Had *p*-hacking been modeled severely enough, the effects on meta-analytic effect size estimates would have been much stronger. The "severity of *p*-hacking" may refer to different aspects, either in isolation or together: (a) How many studies of those in danger of being *p*-hacked will in fact be *p*-hacked? (b) How large is the "danger zone" of studies that can be *p*-hacked? (c) If researchers *p*-hack, how far below the .05 threshold do they *p*-hack their studies? The smaller the final *p*-value, the more strongly the effect size of this particular study will be biased (given a constant sample size).

In response to the concern that *p*-hacking might not been modeled severely enough, we refer to Figure 2B, which (also) illustrates the effects of very severe *p*-hacking: 80% of studies with original *p*-values of up to .800 will be *p*-hacked to significance. The results (red line) reveal considerable distortions caused by this very severe *p*-hacking, but much less than that caused by severe degrees of publication bias. Interactive online application 2 allows interested researchers to simulate the effects of even more severe *p*-hacking.

A critic may go further and object that researchers will even *p*-hack studies to significance when they originally showed mean differences in the unexpected direction. We argue that it is unlikely that this happens on a large scale. There will be both statistical and moral boundaries that prevent researchers from going this far. Even if it is technically possible in some cases, such behavior would more adequately be labeled fraud and data fabrication rather than *p*-hacking, and even pessimistic estimates have identified such misdeeds as very rare (John et al., 2012).

Alternatively, might researchers *p*-hack original findings that went in the "wrong" direction so that they become significant in this unexpected direction and later claim this result was expected (Nelson et al., 2018)? On the basis of individual studies, this may be possible. However, in taking the broader perspective of a cumulative literature, we deem this unlikely. Take studies on the contested stereotype threat effect as an example (Flore & Wicherts, 2015; Spencer, Logel, & Davies, 2016; Stoet & Geary, 2012). Any description of a study that was originally planned to examine stereotype threat will be identified as such with a decent probability. Any dependent variable included to indicate a stereotype threat effect will be difficult to defend as consistent with a priori assumptions if the result ends up going in the other direction. Dependent variables are selected for a purpose. We believe that in the long run, it is unlikely that *just any* effect on a dependent variable could be argued to be in line with a priori assumptions about a given phenomenon under investigation. After all, researchers do not only need to tell a coherent story across (potentially) several studies in a single manuscript, but their findings also have to be coherent over time with the authors' other publications and the respective literature more generally.[3]

Perhaps researchers regularly *p*-hack original findings that went in the "wrong" direction to force them to achieve significance and then come up with a new post hoc explanation that is void of any reference to the literature that was the starting point for the study (e.g., stereotype threat). Again, on the level of individual studies, this could conceivably happen (Kerr, 1998). However, on the level of a cumulative literature, this is more difficult to envision. If this process were to happen on a large scale, this would lead to highly diverse publication lists of researchers in terms of topics. According to this argument, (almost) every study that "did not work" would be used to tell a completely different story to

[3] Note that we have no stakes in the debate about stereotype threat and do not take any position here. This is only one of many examples for which this argument could be made.

avoid making reference to the initial research question. In reality, most researchers conduct research on a few focused areas rather than a wide variety of different research areas. This is an indirect indication that studies that "did not work" often end up in the file drawer rather than being *p*-hacked to just *any* publishable finding for addressing a randomly and a posteriori fitted research question.

Finally, a critic may be concerned that if researchers *p*-hack, they will bring the *p*-value down to even smaller levels than modeled here. Remember that in the present studies, *p*-hacked *p*-values lay between .049 and a minimum of .001 (a triangular distribution with the mode at .049). According to this distribution, the *p*-values are three times more likely to fall between .05 and .025 than between .025 and .001. The general notion that hacked *p*-values are more likely to fall just below $p = .05$ is generally accepted and engrained in common bias detection methods such as a *p*-curve analysis (Simonsohn et al., 2014a, 2014b). To examine the consequences of this particular distribution of *p*-hacked *p*-values, we ran sensitivity analyses. First, we re-ran the simulations with the same triangular distribution, but we set the lower end of the *p*-values to .000001 instead of .001. The results were nearly identical. Second, we changed the distribution of *p*-hacked *p*-values from a triangular distribution to two uniform distributions ranging from .05 to .025 and from .025 to .000001, and we made the first uniform distribution twice as prevalent as the second. Even under these very conservative assumptions, the changes in the results were practically negligible and did not affect the conclusions in any way. Taken together, the presented results are remarkably robust even when assuming that *p*-hacked studies are very often *p*-hacked to excessively low *p*-values.

Limitations

In our simulations, we considered the effects of several different factors of influence that plausibly could have seriously affected the results of the study. This approach resulted in

a comprehensive set of findings. However, it is impossible to map every facet of reality. Modeling inevitably requires a reduction in complexity compared with reality. In the following, we discuss some plausible deviations from reality.

One limitation of the present work was that it was based on the assumption that all studies are valid observations of real (null) effects. In reality, low-quality study designs and executions may lead to systematically or unsystematically biased estimates of true effects. The possibility of unsystematic bias was partially captured by heterogeneity as one factor of influence that introduced the possibility of more than one true effect in a given literature, but we did not explicitly model systematic bias (e.g., high-quality studies are more likely to reveal true effects). In a similar vein, some studies that we modeled as independent might in fact be partly dependent in reality, for example, because they were conducted in the same lab or drew participants from the same participant pool. As we explained in the discussion section on *p*-hacking, social factors more generally may contribute to the emergence of literatures with partly dependent studies. For example, if *p*-hacked studies appeared to elucidate an effect that in fact may not exist this may promote further studies on the effect that would likely not have been conducted had there been no *p*-hacking in the literature. These observations might not question the insights about the interplay of *p*-hacking and publication bias revealed by the present study, but they highlight that some literatures plagued by *p*-hacked false positives might never accumulate and be meta-analyzed if *p*-hacking could be effectively prevented.

A second limitation is that our results may be contingent on the assumption that the probability that "failed" studies will be published is independent of effect size. This assumption could be challenged. For example, it is reasonable to expect that studies that are significant in the opposite direction from what was predicted may be easier to publish than studies that found no difference whatsoever between conditions (Ioannidis & Trikalinos,

2005). The solutions to this problem mentioned earlier, such as registered reports or editors asking for additional studies, are focused on decoupling the probability of publication from the results, thereby ameliorating this problem.

Third, we assumed that the probability that nonsignificant studies will be published is independent of sample size. This assumption may be challenged on the basis of the argument that large nonsignificant studies may have a higher probability of getting published than small nonsignificant studies because researchers have greater incentives to publish studies in which they invested many (as compared with little) resources.

Finally, we operationalized *p*-hacking essentially as lowering an originally non-significant *p*-value to significance. However, how exactly researchers may end up with polished results that initially did not look as promising is not fully understood and could also entail processes not captured by our operationalization of *p*-hacking. For example, if multiple outcomes are assessed in a study, but only a subset of these is reported this does not fit our operationalization of *p*-hacking.

Mild forms of outcome selection, say, some researchers selecting from two or three outcomes, may be covered by the more extreme settings of the "danger zone" influence factor. With these settings, *p*-values shift from up to 0.8 to below 0.05. We may conceptualize the initial *p*-value as some average of multiple outcomes, while the final *p*-value stems only from a subset of outcomes. When outcome switching is more extreme, say, a substantial percentage of researchers consistently selected a small number of outcomes from an originally large number of outcomes, this go beyond the conceptual boundaries of our simulation. In this case, suppression of evidence in a literature would be so widespread that we would be hesitant to place the phenomenon under the conceptual umbrella of *p*-hacking. In fact, it would rather be a form of publication bias on the level of outcomes: Some outcomes are lost to the file drawer, others are reported. In a possible next step, should the

selected original *p*-values not be significant, they may undergo one or several treatments that let it drop under $p = .05$ as is assumed in the current study (e.g, by transformation of variables, recoding or exclusion of outliers etc.). Future studies may want to add selective reporting of outcomes as an additional separate step in the simulations that determines which outcomes (and their *p*-values) then undergo *p*-hacking in the way operationalized here and examine the consequences on meta-analytic effect size biases as a function of the number of assessed and selected outcomes.

Conclusion

Meta-analysis is an essential tool for scientific progress that is considered more trustworthy and robust against various biasing influences than individual studies (Gurevitch et al., 2018). However, the validity of meta-analyses is threatened by systematic sources of error. The present study highlighted how *p*-hacking and publication bias can interact to bias meta-analytic effect size estimates under a large number of circumstances. In recent years, the increase in false-positive findings due to *p*-hacking and, in turn, ways to prevent *p*-hacking have rightfully received considerable attention. The present results highlight that in order to increase the trustworthiness of psychological science, the reduction of publication bias also needs to become a primary objective. Otherwise researchers would run the risk of drawing vastly incorrect conclusions from bodies of evidence. Such a trend would have significant implications for theories, the robustness of practical interventions, the allocation of resources in both research and practice, and, last but not least, trust in our discipline.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554. https://doi.org/10.1177/1745691612459060

Bertamini, M., & Munafò, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science, 7*, 67-71. https://doi.org/10.1177/1745691611429353

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431-449. https://doi.org/10.1037/a0038047

Bruns, S. B., & Ioannidis, J. P. A. (2016). p-curve and p-hacking in observational research. *Plos One, 11*, e0149144. https://doi.org/10.1371/journal.pone.0149144

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*, 115-144. https://doi.org/10.1177/2515245919847196

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex, 66*, A1-A2. https://doi.org/10.1016/j.cortex.2015.03.022

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. 2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin, 145*, 610-651. https://doi.org/10.1037/bul0000194

Dubben, H. H., & Beck-Bornholdt, H. P. (2005). Systematic review of publication bias in studies on publication bias. *British Medical Journal, 331*, 433-434. https://doi.org/10.1136/bmj.38478.497164.F7

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455-463. https://doi.org/DOI 10.1111/j.0006-341X.2000.00455.x

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *Plos One, 5*, e10068. https://doi.org/10.1371/journal.pone.0010068

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

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America, 114*, 3714-3719. https://doi.org/10.1073/pnas.1618569114

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*, 120-128. https://doi.org/10.1037/a0024445

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological sxcience's aversion to the null. *Perspectives on Psychological Science, 7*, 555-561. https://doi.org/10.1177/1745691612459059

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*, 45-52. https://doi.org/10.1177/1948550615612150

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology, 53*, 25-44. https://doi.org/10.1016/j.jsp.2014.10.002

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*, e109019. https://doi.org/10.1371/journal.pone.0109019

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502-1505. https://doi.org/10.1126/science.1255484

Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science, 12*, 1077-1099. https://doi.org/10.1177/1745691617697076

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*, 562-571. https://doi.org/10.1177/1745691612457576

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature, 555*, 175-182. https://doi.org/10.1038/nature25753

Hartgerink, C. H. J. (2017). Reanalyzing Head et al. (2015): investigating the robustness of widespread p-hacking. *Peerj, 5*, e3068. https://doi.org/10.7717/peerj.3068

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*, e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65. https://doi.org/10.1002/jrsm.5

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58*, 543-549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science, 3*, 109-117

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. https://doi.org/10.1177/0956797611430953

Johnson, B. T., & Eagly, A. H. (2014). Meta-analysis of social-personality psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 675-707). London: Cambridge

Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology, 1*, 1-7. https://doi.org/10.1080/23743603.2015.1070611

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kuhberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *Plos One, 9*. https://doi.org/ARTN e105825

10.1371/journal.pone.0105825

Lakens, D. (2015). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology, 68*, 829-832. https://doi.org/10.1080/17470218.2014.982664

Lilienfeld, S. O., & Waldman, I. D. (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. Chichester, UK: Wiley.

Lumley, T. (2012). rmeta: Meta-analysis. R package version 2.16 [Computer Software]. Retrieved from https://CRAN.R-project.org/package=rmeta

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*, 331-348. https://doi.org/10.2466/03.11.pms.112.2.331-348

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology, 65*, 2271-2279. https://doi.org/10.1080/17470218.2012.711335

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163. https://doi.org/10.1037/1082-989x.9.2.147

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. https://doi.org/10.1038/s41562-016-0021

Murad, M. H., & Montori, V. M. (2013). Synthesizing evidence shifting the focus from individual studies to the body of evidence. *Jama-Journal of the American Medical Association, 309*, 2217-2218. https://doi.org/10.1001/jama.2013.5616

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 2600-2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137-141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. https://doi.org/10.1126/science.aac4716

R Core Team. (2017). R: A language and environment for statistical computing [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Renkewitz, F., & Keiner, M. (2018, December 20). *How to detect publication bias in psychological research? A comparative evaluation of six statistical methods*. https://doi.org/10.31234/osf.io/w94ep

Richard, F. D., Bond, C. F., & Stokes-Zoota. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363. https://doi.org/10.1037/1089-2680.7.4.331

Sassenberg, K., & Ditrich, L. (2019). Research in Social Psychology Changed Between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies. *Advances in Methods and Practices in Psychological Science, 2*, 107-114. https://doi.org/10.1177/2515245919838781

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-681. https://doi.org/10.1177/1745691614553988

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: A key to the file-drawer. *Journal of Experimental Psychology-General, 143*, 534-547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, A reply to Ulrich and Miller (2015). *Journal of Experimental Psychology-General, 144*, 1146-1152. https://doi.org/10.1037/xge0000104

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67*, 415-437. https://doi.org/10.1146/annurev-psych-073115-103235

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*, 60-78. https://doi.org/10.1002/jrsm.1095

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association, 54*, 30-34

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication cecisions revisited - the effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician, 49*, 108-112. https://doi.org/10.2307/2684823

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology, 16*, 93-102. https://doi.org/10.1037/a0026617

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12

van Aert, R. C. M., & van Assen, M. A. L. M. (2018, October 2). *Correcting for publication bias in a meta-analysis with the p-uniform* method*. https://doi.org/10.31222/osf.io/zqjr9

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*, 293-309. https://doi.org/10.1037/met0000025

Van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*, 765-790. https://doi.org/10.1177/0013164402251027

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E. J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in psychological bulletin from 1990–2013. *Journal of Open Psychology Data, 5*, 4

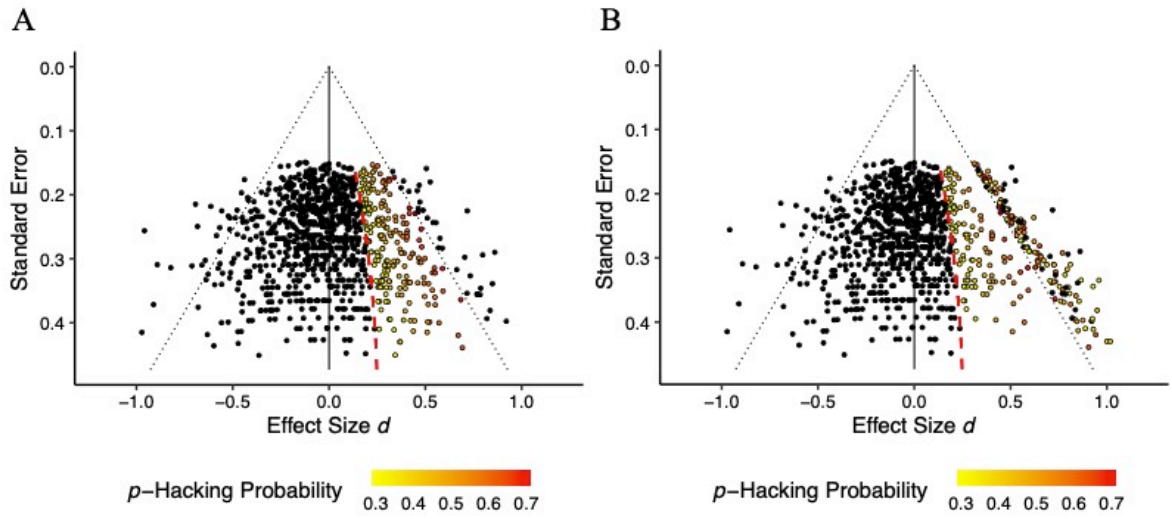
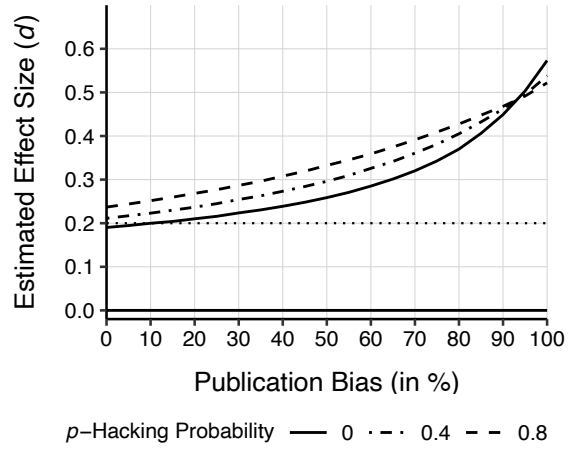
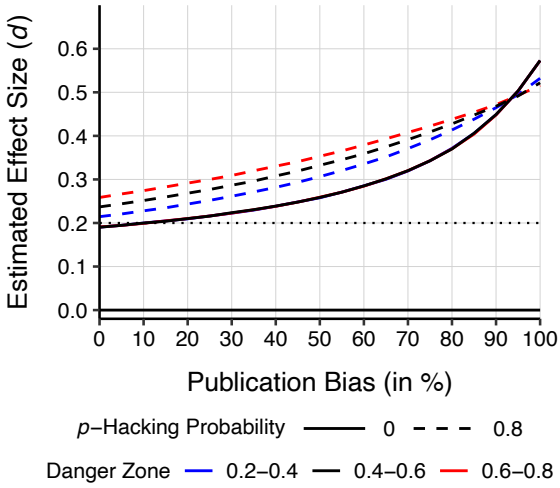


Figure 1. Funnel plots of 1,000 hypothetical studies with a true underlying effect size of zero. The outer dotted lines indicate the triangular region within which 95% of studies are expected to fall in the absence of *p*-hacking, publication bias, and heterogeneity. Effect sizes (Cohen's *d*) are represented on the *x*-axis. Precision (SE_d) is represented on the *y*-axis. The dashed red line indicates the left border of the *p*-hacking danger zone. Studies that fall between the dashed red line and the right border of the funnel are "in danger of being *p*-hacked." The probability that a study that is in danger of being *p*-hacked is indeed *p*-hacked is indicated by color, such that the yellow colored studies are hacked with a probability of 0.3 and the red colored studies are hacked with a probability of 0.7. In Panel A, the probability of *p*-hacking is depicted for illustrative purposes only. In Panel B, the studies that are in danger have actually been hacked according to their assigned probability. Thus, most of the red studies but only a few of the yellow studies have been hacked.

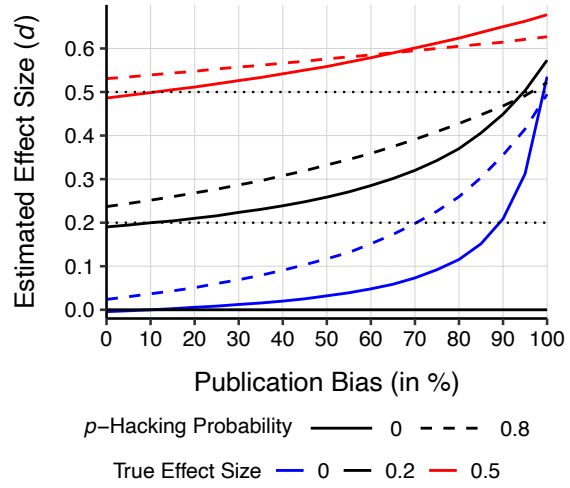
A – Default Settings



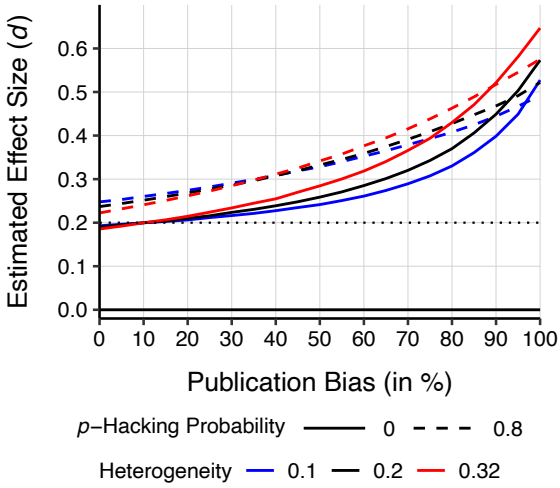
B - Danger Zone



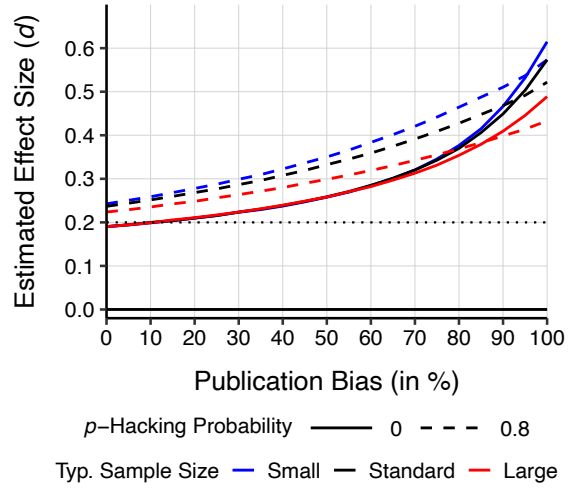
C - True Effect Size



D - Heterogeneity



E – Typical Sample Size



F – Publication Probability

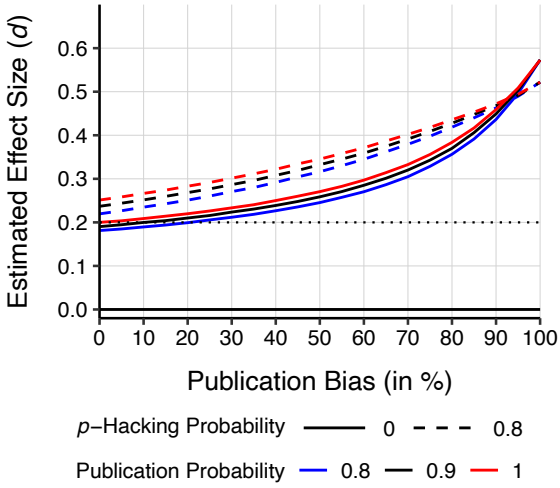


Figure 2. Results of the full simulation. Figure 2A displays the results of the default configuration. The settings for this configuration are $d = 0.2$, $\tau = 0.2$, and danger zone = .4/.6. Typical sample sizes per condition were drawn from the standard distribution. The default probability of the publication of significant studies was .9. The thin horizontal black line represents the null effect. The dotted black line represents the true effect. The curved solid line represents the estimated biased effect d_{est} at a *p*-hacking probability of 0. The curved dashed/dotted line represents the estimated biased effect d_{est} at a *p*-hacking probability of 0.4. The curved dashed line represents the estimated biased effect d_{est} at a *p*-hacking probability of 0.8. Figures 2B-2F display results when the five factors of influence were systematically varied. For these figures, all default settings were used and only the respective factor was varied. Different colors indicate different levels of the factors as described in the legend below each figure. Solid lines always indicate a *p*-hacking probability of zero, whereas dashed lines always indicate a *p*-hacking probability of 0.8. Hacking probabilities of 0.4 are not depicted.