Fothcoming: JEP:G
This version: June 2019.

99% Impossible: A Valid, or Falsifiable, Internal Meta-Analysis

Joachim Vosgerau
Bocconi
joachim.vosgerau@unibocconi.it

Uri Simonsohn
ESADE
urisohn@gmail.com

Leif D. Nelson
UC, Berkeley - Haas
Leif_nelson@berkeley.edu

Joseph P. Simmons
Penn - Wharton
jsimmo@upenn.edu

Abstract

Several researchers have relied on, or advocated for, internal meta-analysis, which involves statistically aggregating multiple studies in a paper to assess their overall evidential value. Advocates of internal meta-analysis argue that it provides an efficient approach to increasing statistical power and solving the file-drawer problem. Here we show that the validity of internal-meta-analysis rests on the assumption that no studies or analyses were selectively reported. That is, the technique is only valid if (1) all conducted studies were included (i.e., an empty file-drawer), and (2) for each included study, exactly one analysis was attempted (i.e., there was no *p*-hacking). We show that even very small doses of selective reporting invalidate internal-meta-analysis. For example, the kind of minimal *p*-hacking that increases the false-positive rate of one study to just 8% increases the false-positive rate of a 10-study internal meta-analysis to 83%. If selective reporting is *approximately* zero, but not *exactly* zero, then internal meta-analysis is invalid. To be valid, (1) an internal meta-analysis would need to exclusively contain studies that were properly pre-registered, (2) those pre-registrations would have to be followed in all essential aspects, and (3) the decision of whether to include a given study in an internal meta-analysis would have to be made before *any* of those studies are run.

Keywords: meta-analysis, *p*-hacking, file-drawer, false-positives, falsification, replicability

In recent years, experimental psychologists have learned that many of their published findings do not replicate, and that they need to adopt better research practices (for a review, see Nelson, Simmons, & Simonsohn, 2018). Inevitably, some of the methodological changes that have been proposed are more likely than others to effectively increase the replicability of psychological science. Some methodological changes, however, are not merely ineffective, but are, despite the good intentions of their proponents, harmful. Of all the changes that have been proposed, we believe that reliance on *internal meta-analysis* is the most harmful of all, posing a significant threat to the integrity of our discipline.

Internal meta-analysis involves statistically aggregating multiple studies reported in a paper (see e.g., Braver, Thoemmes, & Rosenthal, 2014; Cumming, 2012, 2014; Inzlicht 2015; Maner 2014; McShane & Böckenholt, 2017; Rosenthal, 1990; Stanley & Spence, 2014; Tuk, Zhang, and Sweldens, 2015), usually to examine whether the overall effect is statistically significant.[1] By aggregating across many studies, internal meta-analysis increases statistical power, potentially encouraging researchers to report more of their studies, particularly those that did not yield conventional levels of significance. These purported advantages – more statistical power and less "file-drawering" – have helped to make internal meta-analysis increasingly popular.

For example, Tuk et al. (2015) published a paper containing 18 studies, of which only two were significant (study 5 with p=.046 and study 9 with p=.038). However, an internal meta-analysis of all 18 studies showed a highly significant average effect: Cohen's d=0.22, Z=3.81, p<0.001. The editor who accepted Tuk et al.'s (2017) article for publication wrote, in a blog post,

[1] Internal meta-analysis can also be used to assess "whether" an effect is heterogeneous, but of course studies using different designs almost always involve different true effect sizes (if any of the effects are not zero). Internal meta-analysis can also be used to assess which factors moderate an effect. Although most papers would have an insufficient number of studies to identify moderators, this goal seems potentially useful to us. This exercise, however, is necessarily exploratory and correlational, requiring confirmatory follow-ups for any discovered association to be taken at face value.

that “This paper […] is now my favorite as editor […], because it is […] a template for the kinds of things we should be seeing more of in our top journals” (Inzlicht, 2015). With similar enthusiasm, Braver et al. (2014) wrote, “If a nonsignificant Study 2 can be published as part of a package of studies that together produce a significant [...] Meta-Analysis, researchers will be more likely to include nonsignificant individual findings in their papers” (p. 340). These perceived benefits have led advocates of internal meta-analysis to make general recommendations such as, “in multi-study papers, focus on meta-analytical findings rather than on individual significance tests” (Maner, 2014, p. 345), and “The key is meta-analytical thinking: Appreciate any study as part of a future meta-analysis” (Cumming, 2014, p. 27). Stanley and Spencer (2014) have even gone so far as to “suggest that the replication crisis perhaps exists only for those who do not view research through the lens of meta-analysis” (p. 316). Researchers have been receptive of these recommendations. In an 18-month span between January 2017 and June 2018, 16 papers published online or in print in the *Journal of Experimental Psychology: General* used internal meta-analysis to calculate an average effect across studies (for a list of these papers see the Appendix).

In this article, we propose that adopting the practice of internal meta-analysis is likely to have unintended and dire consequences for the credibility of published findings. The validity of internal meta-analysis hinges on the assumption that *none* of the analyzed findings were selectively reported; the method is valid only if one confirmatory analysis was conducted in each original study, and only if every study was included in the internal meta-analysis.

In what follows, we will (1) explain why we believe this assumption is frequently violated, as some amount of selective reporting is in most cases inevitable, and (2) show that even minimal levels of selective reporting *dramatically* increase the probability of obtaining a false-positive result in an internal meta-analysis. For example, the kind of minimal selective reporting that

inflates an *individual study's* false-positive rate to just 8% will inflate the false-positive rate of a 10-study internal meta-analysis to an astonishing 83%.

Internal meta-analysis should only be used when researchers can convincingly demonstrate that absolutely none of the results were selectively reported. This standard could be met if every study was pre-registered, no study deviated from its pre-registration, and the set of studies to run was also pre-registered. This standard appears to have been met by "many-lab" replication efforts, for which the investigators predetermined both the set of studies to be run as well as the critical analysis for each study (e.g., Alogna et al., 2014; Ebersole et al., 2016; Hagger et al., 2016; Klein et al., 2014; O'Donnell et al., 2018; Verschuere et al., 2018; Wagenmakers et al., 2016).[2]

Selective Reporting Is (Almost) Inevitable

There are two ways for researchers to selectively report findings. They can either selectively report whole studies, a practice called *file-drawering*, and/or they can selectively report analyses within a study, a practice called *p-hacking*. We will start by discussing *p*-hacking.

P-hacking Is (Almost) Inevitable

We worry that some researchers may believe that the selective reporting of favorable analyses is immoral, a malevolent form of dishonesty. For someone who holds this belief, the claim that selective reporting is almost inevitable is tantamount to the claim that almost all researchers are bad people. We believe that the selective reporting of favorable analyses and studies only rarely springs from an intentional decision to be dishonest, but is in most cases the inevitable consequence of (moral) human beings' tendency to interpret ambiguous information in ways that are consistent

[2] These are the 1%.

with their desires and beliefs (Kunda, 1990; Vazire, 2015).

For example, when a team of researchers collects and analyzes two dependent measures, and only one of those analyses yields support for their hypothesis, they are going to try to understand why. They are likely to conclude that the one that yielded a more favorable result was probably the better, more sensitive of the two measures, and to choose to report that superior measure in their meta-analysis. In this mundane example two things are true. First, these perfectly moral researchers are making a completely reasonable inference, as it may actually be the case that one measure is better than the other. Second, this is a (consequential) form of *p*-hacking, because the researchers' decision about which measure to report was influenced by the results.

To appreciate the near inevitability of *p*-hacking, consider what it would take to *not* do it. To not *p*-hack, researchers would either have to be indifferent to the outcome of the study, or they would have to (1) plan out in advance exactly how they were going to run their key analysis, and (2) remember and execute that plan when it comes time to write up the results. This means deciding in advance exactly which measures to analyze, how to score those measures, how to deal with outliers or inattentive participants, which covariates to include in the analysis, etc. Motivated researchers who do not perfectly plan out their analyses in advance will have to make *ex post* decisions about which analyses are the best ones. And, because they are human, we can expect them to make those decisions in ways that benefit them rather than in ways that harm them. Indeed, in the presence of desire and in the absence of perfect planning (i.e., pre-registration), some amount of *p*-hacking is virtually inevitable.

As we will show, although minimal levels of *p*-hacking have minimal effects on the validity of individual studies, they have large effects on the validity of internal meta-analysis.

File-drawering Is (Almost) Inevitable

But what about file-drawering? Wouldn't it be immoral for researchers to claim that they are reporting all of their studies, only to then withhold some of them? Wouldn't moral researchers avoid doing this?

The problem with this line of thinking is that it assumes that what counts as a valid study for a particular project is unambiguous. In many cases, it is not.[3]

The research process frequently leads researchers down unexpected paths of inquiry, so that by the time a project is completed the investigation focuses on a research question that is at least a little different from the research question that motivated the project in the first place. It is not easy to decide which studies belong to the project we ended up studying rather than the project we began studying. Similarly, when investigating new phenomena, we often learn (or perhaps merely convince ourselves) that there are bad ways to study it. When deciding which studies to include in an internal meta-analysis, we must determine whether a failed study did not work because of bad design or execution (in which case it does not belong in the meta-analysis) or whether it did not work despite being competently designed and executed (in which case it belongs in the meta-analysis). These are necessarily ambiguous decisions. In the real world, deciding which studies belong to a project is often a messy business, and those decisions are likely to be resolved in ways that help the researchers rather than in ways that harm them.

In the absence of perfectly planned research projects, it is almost impossible for researchers not to engage in at least a modicum of selective reporting (Vazire, 2017). And, as we shall see, if a meta-analysis is infused with even a modicum of selective reporting, it becomes an invalid and

[3] In this section, we are assuming that decisions as to which studies to file-drawer in an internal meta-analysis are made primarily by authors (who can observe all of the studies they have run), rather than by journal editors (who observe the studies that were reported in the original submission).

dangerously misleading tool.

Internal meta-analysis amplifies the effects of selective reporting

The Effects of P-hacking

P-hacking is the selective reporting of data and analyses that obtained more favorable results. Combining common forms of *p*-hacking (e.g., adding observations after obtaining a nonsignificant result or cherry-picking dependent variables) can greatly increase the probability that an investigation will produce a false-positive result. For example, Simmons et al. (2011) showed that a conservative combination of *p*-hacking attempts could increase the false-positive rate from the nominal 5% to over 60%.

The effect of *p*-hacking on the false-positive rate of an individual study pales in comparison to its effects on the false-positive rate of an internal meta-analysis. Let's start by noting that when a set of studies finds consistent results, an internal meta-analysis of those studies will usually obtain a much lower *p*-value than any of the individual studies. For example, meta-analyzing a paper with two studies of similar sample size, each significant at p=.049, would produce a *p*-value that is about ten times smaller than those in the original studies, p=.005. If you add a third study with p=.049, the meta-analytic *p*-value drops almost 10 times again, down to p=.0006.[4] Meta-analysis can turn barely significant results into extremely significant results.

In the absence of selective reporting, the fact that meta-analysis turns barely significant results into extremely significant results is a good feature; two legitimate p=.049 results do indeed constitute more evidence for the existence of an effect than either one does on its own. But this

[4] If some studies have much larger sample sizes than the others, a random-effects meta-analysis may reverse this general pattern. For example, if one study has n=20 and the other n=1000, and both are significant at p=.049, then the meta-analytic p-value would be p=.24 (see R Code; https://osf.io/dpwyg).

good feature becomes a bad feature in the presence of selective reporting. If the studies contain even trivial amounts of bias, the internal meta-analysis will exhibit substantial levels of bias.

Figure 1 may help build an intuition for how relatively inconsequential levels of *p*-hacking within individual studies combine to generate a large bias in a meta-analysis. This figure depicts the results of 20 simulated studies investigating a true effect of zero. In each study the researcher p-hacks by conducting two analyses instead of one, and reports the better of the two (even if it is not significant and even if it is in the wrong direction). The figure shows the results for the first analysis that was conducted (smaller circles) and for the better of the two (larger squares).

For example, in the first simulated study, the figure shows that the first analysis obtained a weaker result than the second, and thus the researchers would report (only) the second; as a result, the square is above the circle. In the second simulated study, the first analysis produced the stronger result, and thus the circle and the square occupy the same position. Looking at the individual studies, we see that the effect of this minor form of *p*-hacking is small (and, of course, half the time it makes no difference). In fact, in this particular simulation, *p*-hacking *did not alter the statistical significance of a single study*, as all 20 studies remained non-significant. This is indeed very mild *p*-hacking. At the same time, one can see how the squares are (necessarily) systematically above the circles. Thus, when one meta-analyzes the squares, one (over)estimates the effect to be $\hat{d}$=.20, Z=2.81, *p*=.0049, rather than accurately estimating it to be zero.[5]

[5] One could ask whether this level of p-hacking is truly minimal or truly realistic. A reviewer suggested that researchers might follow a different strategy: for any given study, try to p-hack to reach statistical significance, but if they fail to achieve significance, un-hack their analyses back to whichever analysis was conducted first. If researchers p-hack/un-hack in that manner, then the consequences of p-hacking could be more modest for any internal meta-analysis. Although this model of researcher behavior is worthwhile to consider, we do not think it is realistic. We believe that the p-hacking that we simulate is both realistic and minimal because (1) it arises from attempting at most two analyses per study (see Figures 1 & 4), and (2) it is quite inconsequential at the individual study level (with an individual study false-positive rate below 8%, for a directional hypothesis). Such a level of p-hacking will feel minimal to the researcher regardless of what motivates their behavior. Regardless of whether p-hacking arises from motivated reasoning, or from an intentionally misleading act, keep in mind that the researcher who p-hacks is the same researcher who meta-analyzes, and thus is unlikely to un-hack prior to meta-analyzing. If p-hacking arises because of motivated reasoning,

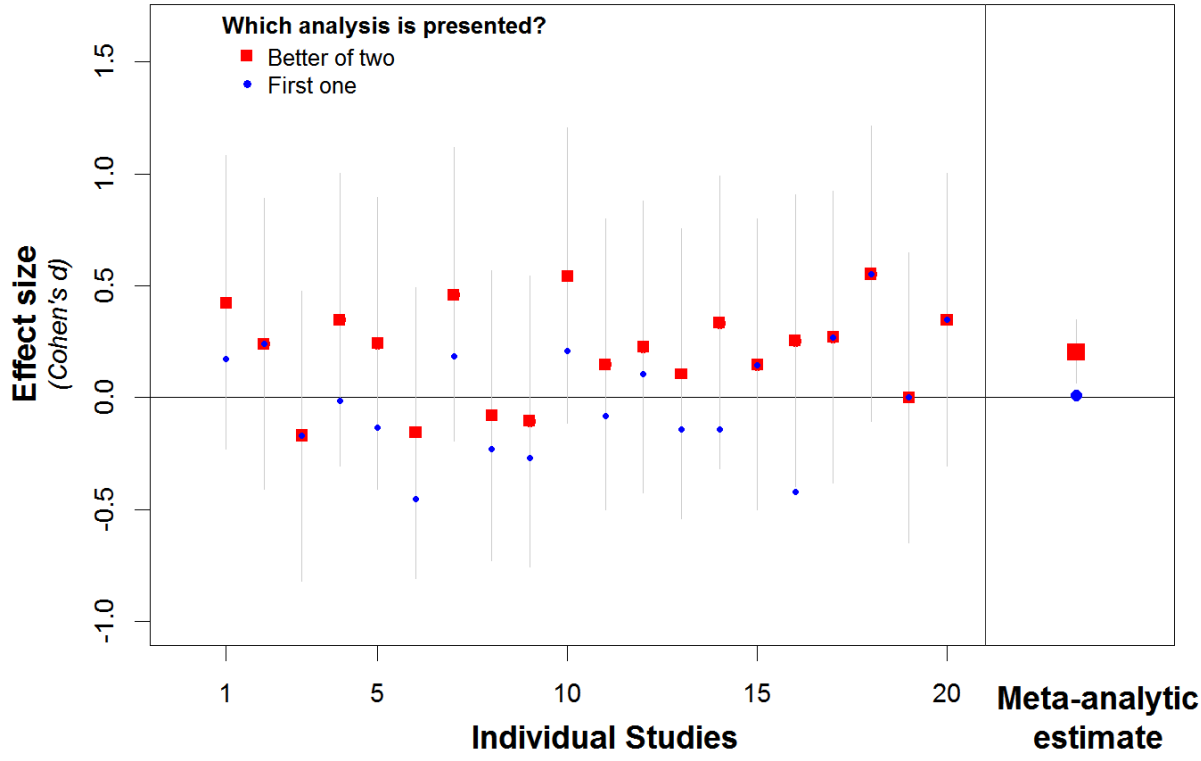

Figure 1. An illustration of the effect of *p*-hacking on individual studies vs. internal meta-analysis. We simulated twenty studies investigating a true effect of zero. For each study, a researcher performs two statistically independent analyses and either always reports the first one (blue circles) or *p*-hacks by reporting the best (higher value) of the two (red squares). Vertical lines correspond to 2 standard errors from the better (more positive) of the two results. This simulation assumes each pair of analyses results in uncorrelated *p*-values. It does not assume that the two attempted analyses were of the same type across studies. R Code to reproduce figure: https://osf.io/ejp5r/

Going beyond this example, we can simulate how increasing an individual study's false-positive rate more generally affects the meta-analytic false-positive rate. For example, Figure 2 shows that when one *p*-hacks a directional hypothesis in a way that slightly increases the false-positive rate from 2.5% to 6%, the 10-study meta-analytic false-positive rate increases to 52%.

the researcher will still be motivated to believe that an analysis that yields a more desirable result is the one that should be meta-analyzed. If p-hacking arises because of an intention to mislead, there will still be an intention to mislead when conducting the meta-analysis.

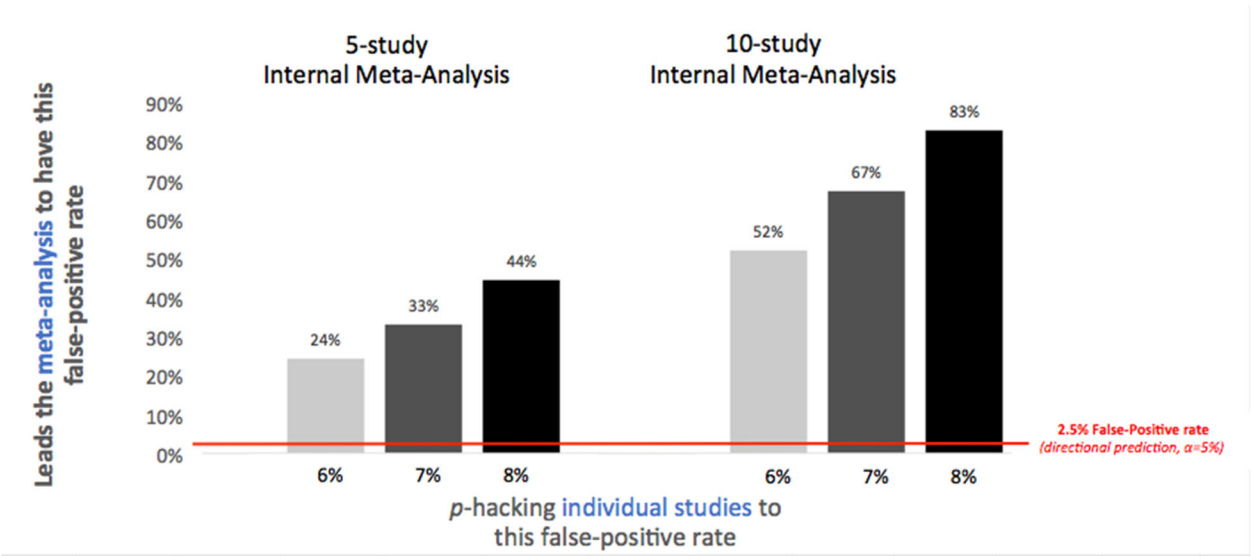

Fig 2. *How minimal levels of p-hacking individual studies affects the false-positive rates of 5-study and 10-study meta-analyses.* We simulated researchers conducting a maximum of four analyses on any given data. We sequentially drew *p*-values from a uniform distribution, U(0,1), correlated r=.69, r=.57, or r=.40 with the previously drawn one.[6] These correlations were calibrated to obtain false-positive rates of 6%, 7%, and 8%, respectively, for individual studies. The meta-analysis includes, for each study, the analysis with the first significant *p*-value obtained, or the lowest non-significant one out of the four. The obtained t-values are converted to Cohen d's and then analyzed with a random-effects meta-analysis that assumes equal sample sizes across studies (n=20, without loss of generality). R Code to reproduce figure: https://osf.io/k3zs2.

File-drawering

Imagine that researchers only conduct internal meta-analyses on sets of studies that reported the one pre-registered analysis. In other words, assume there is no *p*-hacking at all. Even under these exceptional circumstances, internal meta-analysis would be invalid if the decision about which *studies* to include in the meta-analysis was at all influenced by the studies' results. As we described above, some selective reporting of studies is virtually inevitable for most researchers most of the time. But readers may have the hope that, by encouraging researchers to relegate *fewer* of their studies to the file drawer, internal meta-analysis will at least make things better. After all,

[6] For more information on how *p*-hacking can be operationalized as the sequential drawing of correlated *p*-values, see Supplement 3 in Simonsohn et al (2014).

isn't it better to retrieve some of our failed studies rather than none of our failed studies? The answer is, "Well, no, actually; it is almost certainly worse." The reason is that motivated researchers are unlikely to retrieve failed studies at random, but instead are more likely to retrieve (and justify including) studies that produce more favorable results. Ironically, this means that internal meta-analysis is likely to *exacerbate* the consequences of the file-drawer problem.

As a preliminary example, imagine a one-study paper with a p=.049 result. Imagine that the file-drawer contains two similar studies that did not "work," a p=.20 in the right direction and a p=.20 in the wrong direction. If they are both added to the meta-analysis, then the overall effect would be non-significant. But if instead the researcher only partially emptied the file drawer, including only the *right-direction* p=.20 in the paper (perhaps because the effect in the wrong direction was identified as testing a different effect, or as having a flawed design, etc.), the evidence will now seem *stronger*, generating an internal meta-analytic result of p=.021, rather than the single study result of p=.049.

Once we consider more realistic numbers of studies in projects and file drawers, the negative consequences of the partially emptied file drawers for internal meta-analysis become more obvious. Figure 3 shows how easy or difficult it would be to publish a five-study paper supporting a false hypothesis, depending on (1) whether the researcher needs all five studies to be significant or simply needs the five-study meta-analysis to be significant, and (2) whether the researcher drops 0, 1, 2, 3, 4, or 5 additional studies. In this scenario, we are assuming that the researcher does not p-hack at all, and thus that in each study only one pre-planned analysis was conducted.

The solid black line in Figure 3 shows that the probability of publishing a false-positive five-study manuscript is extremely small when a researcher needs all five studies to be significant, even when she file-drawers some of her attempted studies. In contrast, the solid blue line shows that

file-drawering studies has a profound effect on the probability of publishing a false-positive five-study paper if the researcher simply needs the overall internal meta-analysis to be significant, rather than all of the individual studies. For example, Figure 3 shows that even if three studies are unreported (i.e., eight studies were run, and the best five were reported), the meta-analytic false-positive rate increases from 5% to 19%.

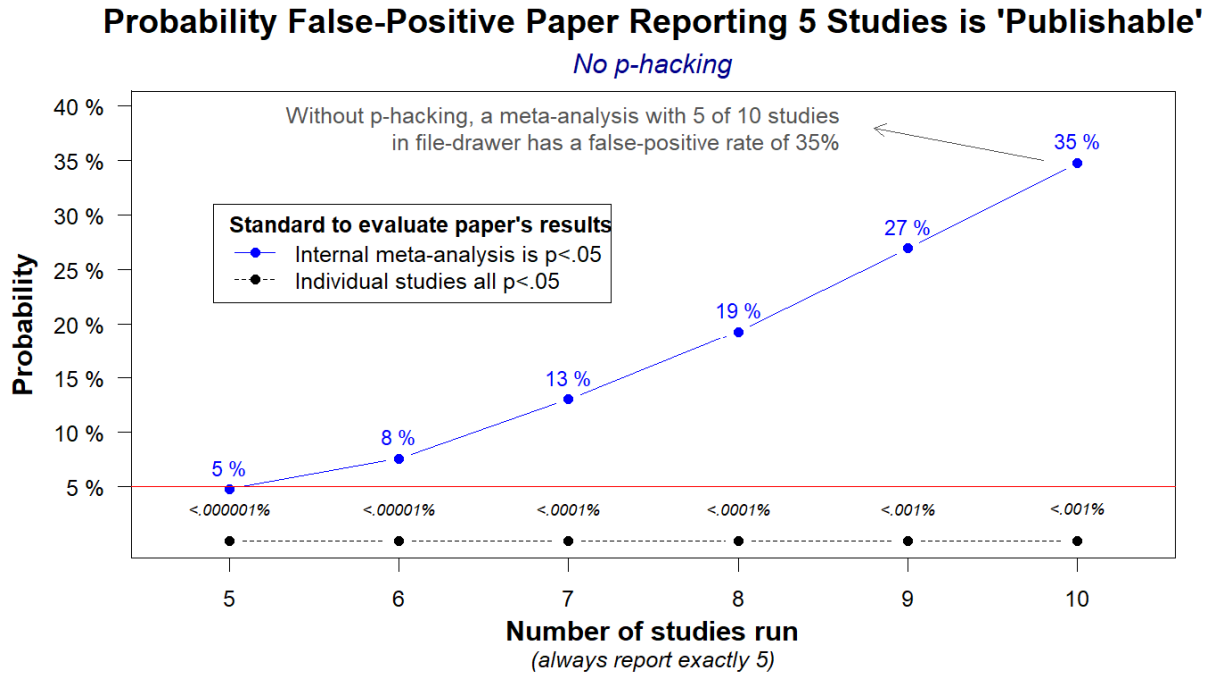

Fig 3. *File-drawering and the probability of obtaining a false-positive five-study internal meta-analysis vs. five individual false-positive studies.* The results were obtained by drawing the number of studies indicated on the x-axis under the null, generating individual t-values for those individual studies, and keeping the five largest t-values that were generated. Those five t-values were then converted into Cohen's ds, and aggregated using random effects meta-analysis. This was repeated 10,000 times for each of the 5, 6, 7, 8, 9, and 10 total t-values. The y-axis depicts the percentage of simulations for which the overall meta-analysis obtained $p < .05$, and for which all five of the studies were individually significant. For example, the two right-most dots show that when 10 studies are attempted, and the best 5 are reported, there is a 35% chance that the internal meta-analysis will be $p < .05$, but there is less than a 1/10,000 chance that the five studies will be individually significant. R Code to reproduce figure: https://osf.io/pdyhv/.

The problem would be even worse than this figure suggests if researchers behave as Ueno, Fastricht, and Murayama (2017) assume they behave, conducting studies until the meta-analysis is significant. Our analyses show that even if researchers were to pre-register every single study

and adhere to their pre-registrations, partial file-drawering would still make it much too easy to generate a false-positive internal meta-analysis.

It seems that the simple solution to this problem is to ask researchers to include *all* the studies that belong in the meta-analysis. Unfortunately, as discussed earlier, this will be very difficult because, in our experience, it is often ambiguous which studies do and do not belong in the internal meta-analysis, particularly since studies with bad designs should not be included. Thus, the way to surmount this problem is for researchers to decide, before any studies are run, that they are going to conduct an internal meta-analysis and that it is going to include studies that satisfy a predetermined rule that does not hinge on the obtained results (e.g., "we will meta-analyze all studies we run between now and December 2019 with effort as a dependent variable"). Ideally, the predetermined rule would itself be pre-registered.

As bad as the results of Figure 3 are, it is important to emphasize that these results hinge on an assumption that will be only rarely met: that there is absolutely no *p*-hacking, and thus that exactly one pre-planned analysis was conducted for every study. Internal meta-analysis looks much worse when you relax this assumption. To illustrate, Figure 4 shows how the results depicted in Figure 3 change when you introduce mild *p*-hacking in the form of conducting two analyses instead of just one.

Reinforcing the fact that such mild forms of p-hacking can invalidate meta-analyses even when all studies are reported, the left-most dots show that internal meta-analysis generates an unacceptably high false-positive rate even in the absence of file-drawering (20% for the level of p-hacking depicted). More importantly, it gets very bad very quickly if this mild form of p-hacking is combined with file-drawering just one or two studies. For example, researchers who report the best 5 out of 8 studies, while engaging in this mild form of *p*-hacking, will be able to produce a

false-positive five-study paper 65% of the time.

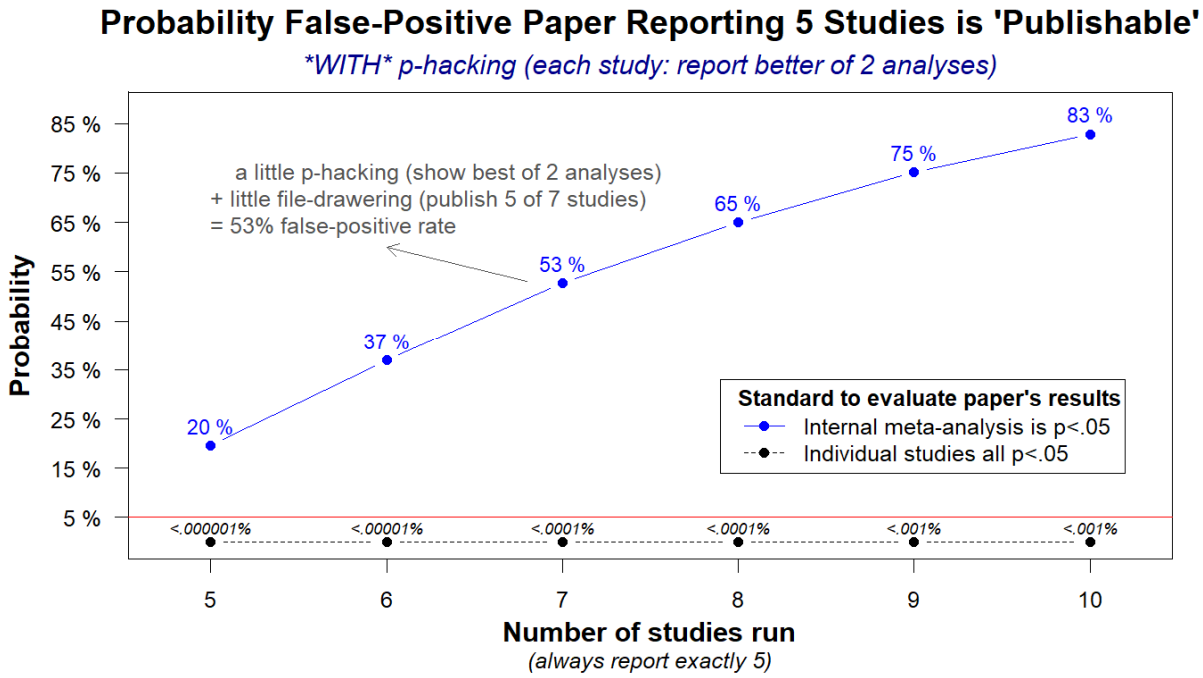

Fig 4. *File-drawering + p-hacking and the probability of obtaining a false-positive five-study internal meta-analysis vs. five individual false-positive studies.* The results were obtained by drawing the number of studies indicated in the x-axis (under the null). For each study, two separate independent t-values are drawn, and the higher one is kept (minimal p-hacking: two instead of one analysis attempted). The five studies with the larger t-values among these were converted onto Cohen's d, and aggregated via a random-effects meta-analysis. This was repeated 10,000 times. The top line depicts the percentage of simulations where the overall meta-analysis obtained $p<.05$, and the bottom one where all five of the selectively reported studies would be expected to be significant, based on the binomial distribution. For example, the two left-most dots show that when the 5 studies reported are the only ones attempted, and for each, the best of two independent analysis run is reported, there is a 20% chance that the internal meta-analysis will be $p<.05$, and only 1 in a 100 million chance that the five studies will be individually significant. R Code to reproduce figure: https://osf.io/pdyhv/.

We believe that our simulations underestimate rather than overestimate the intensity of selective reporting that one should expect from a well-intentioned researcher who does not pre-register her/his studies. The fact that we cannot prove that we are right about this is part of the point. For a given set of studies, the intensity of *p*-hacking and/or file-drawering is unobservable, and thus the (single-study and meta-analytic) false-positive rates are unknown. This fact, in combination with the extremely high false-positive rates that result from meta-analyzing studies

that were to some extent selectively reported, mean that, upon observing an internal meta-analysis, we cannot know whether or not to believe its conclusions. Unless one can truly rule out the possibility that the individual studies were *p*-hacked or file-drawered at all – by, for example, showing that every study in the meta-analysis was properly pre-registered, that the pre-registered plans were followed, and that all studies were included according to a preregistered rule – the results of an internal meta-analysis cannot be trusted. If selective reporting is *approximately* zero, but not *exactly* zero, then internal meta-analysis is invalid.[7]

In sum, an internal meta-analysis is plausibly valid only if (1) it exclusively contains studies that were properly pre-registered, (2) those pre-registrations were followed in all essential aspects, and (3) the decisions of whether to include a given study in an internal meta-analysis is made before any of the studies are run.

False-Positive Meta-Analyses Are Difficult To Falsify

If false-positive internal meta-analyses were as easy to falsify/correct as they are to produce, the problems we have identified may be manageable in the long run. Unfortunately, the opposite is true. It is considerably harder to correct a false-positive internal meta-analysis than it is to correct a false-positive single study.

When we consider individual studies as stand-alone units of evidence, a false-positive finding can be corrected (or verified) by conducting a convincing and well-executed replication. But when we consider bundles of studies – internal meta-analyses – as holistic units of evidence, the path to correcting a false-positive result is not obvious. What should it take to correct a false-positive finding obtained via internal meta-analysis? We consider two possible strategies: (1) a (highly

[7] In these cases, internal meta-analysis also (and obviously) leads to biased (inflated) effect size estimates.

powered) replication of one meta-analyzed study, and (2) replicating all of the studies in the meta-analysis.

Replicating One Study

An interesting challenge when selecting one of the meta-analyzed studies for replication is that many (or possibly all) of the original individual studies will be non-significant to begin with. How does one "replicate" or "fail to replicate" a study that never worked in the first place?

Leaving that aside, consider a researcher who sets out to replicate *one* study included in an internal meta-analysis, perhaps the one that produced the smallest *p*-value, the largest effect size, or the one with the cleanest design. How should the result from that replication be analyzed?

The possibility that seems most consistent with the underlying philosophy of internal meta-analysis is to *combine* the replication results with the existing evidence, computing a new meta-analytic estimate. However, the addition of an otherwise convincing failure-to-replicate will seldom change the results of a multi-study meta-analysis. Building on the simulations reported in the right columns of Figure 2, in which bundles of 10 studies were *p*-hacked, we consider what would happen if we added an 11th study to the meta-analysis that had an 83% false-positive rate. If that replication had the same sample size as the original study, the meta-analytic result would remain significant 89% of the time (i.e., 89% of the 83% significant meta-analyses would be significant). If the replication had 2.5 times the original sample size (Simonsohn, 2015), the result would remain significant 81% of the time. Even if the replication had the same sample size as all ten original studies combined(!), the majority of false-positive internal meta-analyses (about 54%) would survive the replication effort (R Code https://osf.io/k3zs2/).[8]

Instead of *combining* the replication and all of the original studies into a single meta-analysis,

[8] These calibrations are only trivially impacted by switching from random to fixed effects.

one could instead interpret the original study that failed to replicate as "annulled," and to rerun the original internal meta-analysis without it (and without its replication). Under the same assumptions from the previous paragraph, 78% of false-positive meta-analysis results would remain significant.[9] In sum, it is discouragingly unlikely that replicating a single study from a false-positive internal meta-analysis would allow us to correct the erroneous meta-analysis.

Re-running All Studies

Perhaps the most natural – albeit exceedingly costly – approach to correcting a false-positive internal meta-analysis is to re-run every study that it includes, and to then combine the resulting replications with the original studies. The problem is that if the original studies are distorted by selective reporting, the combination of original and replication studies will also be distorted by it, leading to elevated false-positive rates even after re-running every single study. We conducted simulations to assess the magnitude of this problem. We started from the false-positive 10-study internal meta-analysis simulated for Figure 2 and proceeded to add 10 studies drawn under the null, re-running the meta-analyses now with 20 studies each (10 original and 10 replication studies). When all ten new replications had the same sample size as the original, 47% of false-positive internal meta-analyses remained significant. When all ten replications had 2.5 times the original sample size, still 30% of internal meta-analysis remained significant.

Keep in mind that all of this assumes something extremely (and perhaps unrealistically) optimistic – that replicators could afford (or would bother) to replicate every single study in a meta-analysis and that others would judge all of those replication attempts to be of sufficient

[9] To arrive at that 78% we proceed as follows. First, 20% of meta-analyses survive because, if the null is true, replications have 80% power to accept the null when we run 2.5 times the original sample size (Simonsohn, 2015). Thus, in 20% of cases we would not even drop the original study. To estimate the share of the remaining 80% that survives, we re-ran the simulated meta-analyses that obtained the 83% false-positive rate in Figure 2, excluding the single study with the largest effect size in each paper, finding that 72% of meta-analysis would survive such exclusion. We multiply this by 80% (the probability we drop the study in the first place) and we add 20%, arriving at 78%.

quality. Absent this wild assumption, we are left with the possibility that one cannot ever realistically attempt to falsify a false-positive internal meta-analysis.

Implications for External Meta-analysis

In this paper, we have focused on internal meta-analysis because it is being championed as a way to reduce false-positive results caused by the selective reporting of statistically significant studies (and to reduce false-negative results caused by running individually underpowered studies). To what extent do our concerns extend to traditional, "external" meta-analysis, which includes studies that were originally reported in separate papers, standing on their own?

The concerns expressed in this paper do also apply to external meta-analysis, for the logic is the same. If you average several studies that are even slightly biased in the same direction, the meta-analytic summary will be very biased. Thus, selective reporting also poses a big threat to the validity of external meta-analysis.

So why are we focused on internal meta-analysis in this paper? There are three main reasons.

Reason #1: Internal Meta-Analysis Is A Bigger Threat

Internal meta-analysis has become very popular, and we think there are three reasons why: it is easy to perform, it is often used as a way to turn a hard-to-publish set of results into an easier-to-publish set of results, and researchers have been told that it is the right thing to do. We are worried that if researchers continue to believe in the merits of this problematic procedure, its popularity is likely to increase even further, threatening the integrity of our discipline.

Although external meta-analysis may also be rendered invalid by the inclusion of selectively reported results, it is not easy to perform and it is usually not used to turn a hard-to-publish set of results into an easier-to-publish set of results. This means that the potential harm caused by external meta-analyses will be contained within the small subset of quantitative review articles. In contrast,

internal meta-analyses may reside within, and thus threaten the integrity of, all multi-study empirical papers.

Reason #2: The Problems We Have Identified Are Less Consequential In External Meta-Analysis

Although the problems we have outlined in this paper also apply to external meta-analysis, there are at least four reasons to believe that they will sometimes be less consequential in this context.

First, in most external meta-analyses, we can observe meta-analysts' reasons for including versus excluding a set of studies, and in many cases the excluded studies will be observable. In contrast, in internal meta-analysis, we cannot see which studies the authors may have decided to exclude, nor the reasons for those exclusions.

Second, whereas an internal meta-analyst can always make the decision to continue running new studies so as to achieve meta-analytic significance (Ueno et al., 2017), the external meta-analyst typically does not run new studies.

Third, although all studies in an internal meta-analysis will be designed to show the same effect, some external meta-analyses will contain studies designed to show opposite or null effects, as well as studies that were not designed to test to the effect of interest to the meta-analyst. Thus, some studies in an external meta-analysis may be biased in opposite directions (thus partially cancelling out those biases), and others may be unbiased (or biased in ways that are not consequential for the meta-analysis).

Fourth, often the goal of external meta-analysis is not to achieve statistical significance, but rather to estimate effect sizes, and try to explain why different studies have found different effects. In contrast, internal meta-analysis is often used as a way to achieve statistical significance. Thus,

we would expect external meta-analysts to be more likely to conclude that an overall effect is nonsignificant.

In sum, although our concerns apply to both internal and external meta-analysis, the problems arising from selective reporting are probably less consequential for external meta-analysis.

Reason #3: External Meta-Analysis Has A Whole Host Of Different Problems

Although external meta-analysis is less affected by the problems that we have described in this paper, it is plagued by other difficult-to-solve problems that are less likely to manifest for internal meta-analysis, including the problem of ensuring that every included study is properly designed (e.g., free of confounds), as well as free of errors and fraud. For example, we are unaware of any meta-analyses that have excluded studies because they were poorly designed, and yet that is a common reason to reject individual articles in the first place. Indeed, the same researcher may reject a poorly designed study as a reviewer, but accept it as a meta-analyst.

The problems with external meta-analysis are unique enough and severe enough to require their own paper.

What Is Meta-Analysis Good For?

Despite our misgivings about using meta-analysis to conduct statistical inference, we believe that it does represent a valuable way to conduct *exploratory* research.

Imagine, for example, that a researcher were interested in studying the conditions under which people are optimistic vs. pessimistic. Since so much research has been done on this topic, the researcher could systematically extract the results from all competently conducted and adequately reported studies of optimism/pessimism, order them from “most evidence for optimism” to “most evidence for pessimism,” and then look to see how the studies finding evidence for optimism differ from the studies finding evidence for pessimism. From this endeavor she may learn, for example,

that the studies finding the best evidence for optimism tended to ask participants about their own health and relationships, whereas those finding the best evidence for pessimism tended to ask participants about politics. She could then hypothesize that optimism is more likely to emerge for questions about the self, whereas pessimism is more likely to emerge for questions about the world at large.

Of course, it is critical to remember that this is a hypothesis that needs the support of confirmatory research to become a valid scientific conclusion. The hypothesis is necessarily speculative, as the correlation that she is observing could have many causes – for example, perhaps the studies that found evidence for optimism used a different measure than the studies that found evidence for pessimism – or it could be an artifact of the stimuli chosen by the original experimenters (see Wells & Windschitl, 1999), or a by-product of selective reporting. But this is still a valuable enterprise, as the meta-analysis can help researchers use existing data to generate new hypotheses that can then be examined in new, confirmatory, pre-registered studies.

Recommendations

We recommend to never draw inferences about the existence of an effect from internal meta-analyses, unless we know that none of the studies and analyses were selectively reported. We should decide whether to believe a particular hypothesis not by conducting an internal meta-analysis, but by judging the quality of the individual studies that seem to support that hypothesis. We don't believe in the robustness of anchoring effects or motivated reasoning or preference projection because someone meta-analyzed these literatures; we believe in them because the studies supporting them are well designed and because exact replications of these effects have been overwhelmingly successful. Scientific knowledge advances one replicable study at a time.

At the same time, we cannot expect all published papers to substantially contribute to scientific

knowledge. Sometimes a tentative/suggestive result, or even the puzzling absence of a result, is all a researcher can deliver, and such contributions can eventually prove valuable. A researcher may, at the necessary termination of a project, end up with inconclusive evidence. What should she do? If she believes the results are informative, she should by all means publish them.

Whether a paper contains enough of a contribution to merit publication is something that researchers and reviewers should decide on a case-by-case basis, as is currently done. The core message of this paper is not that only perfectly convincing individual studies – studies that are highly powered, significant, and already replicated – should be published. Rather, the core message of this paper is *that evaluations about the contribution and conclusiveness of the results presented in a paper should not be based on an internal meta-analysis of such studies, unless all studies were pre-registered and an ex-ante study rule was set that governs the inclusion of studies in the internal meta-analysis.*

Internal meta-analysis could be used for exploratory purposes, to compare results across different studies in the search for possible moderators. Such endeavors may lead to interesting research *questions* that future confirmatory research can attempt to address, but they do not, under the vast majority of circumstances, provide scientifically valid *answers*.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., ... & Buswell, K. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556-578.

Braver, Sanford L., Felix J. Thoemmes, and Robert Rosenthal (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, 9(3), 333-342.

Cumming, Geoff (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming, Geoff (2014). The new statistics why and how. *Psychological Science*, 25(1), 7-29.

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*(2), 455-463.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68-82.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573.

Inzlicht, Michael (2015). A tale of two papers, blogpost, https://web.archive.org/web/*/http://sometimesimwrong.typepad.com/wrong/2015/11/guest-post-a-tale-of-two-papers.html, accessed January 29th, 2017.

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

Clinical Trials, 4(3), 245.

John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23(5), 524-532.

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Cemalcilar, Z. (2014). Data from investigating variation in replicability: A "many labs" replication project. *Journal of Open Psychology Data*, *2*(1).

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498.

Maner, Jon K. (2014). Let's put our money where our mouth is if authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, 9(3), 343-351.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research*, *43*(6), 1048-1063.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511-534.

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., ... & Balatekin, N. (2018). Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*(2), 268-294.

Rosenthal, R. (1990). Replication in behavioral research. In J. W. Neulip (Ed.), *Handbook of replication research in the behavioral and social sciences* (pp. 1–30). Corta Madera, CA: Select Press.

Schimmack, U. (2014). The test of insufficient variance (TIVA): A new tool for the detection of questionable research practices. Retrieved from

https://web.archive.org/web/20170504174949/https://replicationindex.wordpress.com/2014/12/30/the-test-of-insufficient-variance-tiva-a-new-tool-for-the-detection-of-questionable-research-practices/

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*(11), 1359-1366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569.

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

Stanley, David J., and Jeffrey R. Spence. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science* 9, no. 3, 305-318.

Tuk, Mirjam A., Kuangjie Zhang, and Steven Sweldens (2015). The propagation of self-control: Self-control in one domain simultaneously improves self-control in other domains. *Journal of Experimental Psychology: General*, 144(3), 639-655.

Ueno, Taiji, Greta Fastrich, and Kou Murayama (2017). Meta-analysis to integrate effect sizes within a paper: Possible misuse and Type-1 error inflation. *Journal of Experimental Psychology: General*, forthcoming.

Vazire, Simine (2015). This is what p-hacking looks like, blogpost, https://web.archive.org/web/20181018065906/http://sometimesimwrong.typepad.com/wrong/2015/02/this-is-what-p-hacking-looks-like.html, accessed February 19th 2018.

Vazire, Simine (2017). Be your own a**hole, blogpost, https://web.archive.org/web/20181008152627/http://sometimesimwrong.typepad.com/wr

ong/2017/05/be-your-own-ahole.html, accessed January 25th 2018.

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., ... & Barbosa, F. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299-317.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered replication report: strack, martin, & stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917-928.

Wells, G. L. & P. D. Windschitl (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*, 1115-1125.

Appendix

Table: Papers published since 2017 in the *Journal of Experimental Psychology: General* referencing internal meta-analysis to calculate an average effect across studies.

Year/ Publication	Title	Authors	Quote
2017 *JEP:G* *146*(2) p. 269-285	Social Affiliation in Same-Class and Cross-Class Interactions	Côté et al.	Here, we test the prediction that social affiliation among same-class partners is stronger at the extremes of the class spectrum, given that these groups are highly distinctive and most separated from others by institutional and economic forces. An internal meta-analysis of 4 studies ($N = 723$) provided support for this hypothesis. (Abstract)
2017 *JEP:G* *146*(1) p. 1-19	The Dynamic Effect of Incentives on Postreward Task Engagement	Goswami & Urminsky	We test the effect of ending an incentive in Study 1, as well as in an internal meta-analysis of all data we have collected. (p. 2)
2017 *JEP:G* *146*(1) p. 134-153	Imagining wrong: Fictitious contexts mitigate condemnation of harm more than impurity	Sabo & Giner-Sorolla	After the last experiment, we also report the outcome of a meta-analysis across a set of comparable experiments that included the same set of fictional contexts, looking at differences in fictive pass effects between the contexts. (p. 135) . . . To condense and systematically analyze these results, we conducted a first meta-analysis of Experiments 3, 4, and 5 to get a more holistic understanding of the fictive pass effects by context across our experiments. (p. 148).
2017 *JEP:G* *146*(2) p. 250-268.	The accuracy of less: Natural bounds explain why quantity decreases are estimated more accurately than quantity increases	Chandon & Ordabayeva	To obtain a more reliable estimate of the power exponents for increasing and decreasing quantities, we conducted a meta-analysis of the results obtained in the control conditions (estimation of the final size) of the five studies reported in the paper and in the three studies reported in the supplementary material. (p. 262-263)

2017 *JEP:G* *146*(1) p. 123-133	Two paths to blame: Intentionality directs moral information processing along two distinct tracks	Monroe & Malle	A meta-analysis of the present three studies would allow us to estimate a more reliable effect size and to examine whether switch costs are symmetric across tracks. To test the reliability of our findings we derived effect sizes (Cohen's d) for comparing switch RTs with match RTs within each of the three studies and computed average effect sizes, using inverse variance weights. In both fixed and random effects models, the average effect size of overall switch costs was d = 0.32 (corresponding to 435 ms), 95%CI [0.17; 0.47], z= 4.2, p< .0001. (p. 129)
2017 *JEP:G* *146*(4) p. 512-528	The "common good" phenomenon: Why similarities are positive and differences are negative	Alves et al.	In order to estimate the average effect sizes for our main predictions, we conducted a mini meta-analysis using fixed effects, in which the mean effect sizes were weighted by sample size (Goh, Hall, & Rosenthal, 2016). (p. 524)
2017 *JEP:G* *146*(11) p. 1574-1585	Children understand that agents maximize expected utilities.	Jara-Ettinger, Floyd, Tenenbaum, & Schulz	Power analyses with parameters estimated from a meta analysis on our data confirm that our experiments' power is over .95, with a .04 chance of producing a false positive (see supplemental text). (p. 1576)
2017 *JEP:G* *146*(10) p. 1379-1401	Power as an Emotional Liability: Implications for Perceived Authenticity and Trust After a Transgression	Kim et al.	After completing these studies, we also conducted meta-analyses of our data. (p. 1393)
2017 *JEP:G* *146*(8) p. 1086-1105	When Good Is Stickier Than Bad: Understanding Gain/Loss Asymmetries in Sequential Framing Effects	Sparks & Ledgerwood	Given recent calls to move from evaluating single studies in isolation to considering the information provided by a cumulative body of research evidence (e.g., Braver, Thoemmes, & Rosenthal, 2014; Ledgerwood, 2014; Maner, 2014), we conducted a meta-analysis to quantitatively synthesize the results from the studies that tested familiarity as a moderator of reframing effects in the gain domain (Studies 3s, 3, 4, and 5). (p. 1099)

2017 *JEP:G* *146*(8) p. 1164-1188	The role of empathy in experiencing vicarious anxiety	Shu, Hassell, Weber, Ochsner, & Mobbs	Internal meta-analyses of data across studies 1, 2, and 3. As the designs for Studies 1–3 were similar, we conducted internal meta-analyses to estimate the average effect sizes of the main correlations reported in these studies (Braver, Thoemmes, & Rosenthal, 2014; Cumming, 2014). (p. 1178)
2018 *JEP:G* *147*(2) p. 190-208	The importance of awareness for understanding language	Rabagliati et al.	Our null findings for sentence processing are corroborated by a meta-analysis that aggregates our studies with the prior literature. (Abstract)
2018 *JEP:G* *147*(1) p. 93-112	Mindfulness increases prosocial responses toward ostracized strangers through empathic concern	Berry et al.	Similarities across study procedures, including outcome measures, dispositional measures, and the use of no instruction control conditions in Studies 3 and 4 allowed for meta-analysis of effect sizes pertaining to the observed relations of mindfulness (dispositional and briefly trained) to prosocial responsiveness. Meta-analytically derived summary mean effects across studies have greater precision than do single study results; thus, we first asked whether effect sizes for the relations of mindfulness to empathic concern and to both helping behavior outcomes remained stable (and secondarily, statistically significant) across studies using various active and inactive control conditions. Second, we sought more precise estimates of the effect sizes of experimentally manipulated brief mindfulness training on empathic concern and on both helping behavior outcomes across all three experiments (Studies 2–4). (p. 105-106)
2018 *JEP:G* *147*(5) p. 747-781	What do short-term and long-term relationships look like? Building the relationship coordination and strategic timing (ReCAST) model	Eastwick, Keneski, Morgan, McDonald, & Huang	In the section Aggregated Results Across Studies, we meta-analyze all 10 short-term vs. long-term desired behavior differences across studies and test whether differences across studies reflect between-study heterogeneity or simply sampling variability. (p.761)

2018 *JEP:G* *146*(2) p. 194-213.	Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes?	Kurdi & Banaji	Because Studies 1–5 shared the same basic design, we aggregated the results across these studies meta-analytically (see Figure 1). (p. 206)
2018 *JEP:G* *147*(3) 377-397	Implications of individual differences in on-average null effects	Miller & Schwarz	It would also be possible to adapt traditional methods of meta-analysis for this purpose, since testing for effect-size heterogeneity across studies (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Hartung, Argac, & Makambi, 2003; Higgins, Thompson, Deeks, & Altman, 2003) is quite analogous to testing for it across individuals. (p. 387)
2018 *JEP:G* *147*(4) p. 514-544	The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically	Pedersen, McAuliffe & McCullough	Finally, we conclude this article with a meta-analytic summary of the results of these five experiments. (abstract)