Annual Review of Psychology

Psychology's Renaissance

Leif D. Nelson,[1] Joseph Simmons,[2] and Uri Simonsohn[2]

[1]Haas School of Business, University of California, Berkeley, California 94720; email: Leif_Nelson@haas.berkeley.edu

[2]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; email: jsimmo@upenn.edu, urisohn@gmail.com

Annu. Rev. Psychol. 2018. 69:511–34

First published as a Review in Advance on October 25, 2017

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-122216-011836

Keywords

p-hacking, publication bias, renaissance, methodology, false positives, preregistration

Abstract

In 2010–2012, a few largely coincidental events led experimental psychologists to realize that their approach to collecting, analyzing, and reporting data made it too easy to publish false-positive findings. This sparked a period of methodological reflection that we review here and call Psychology's Renaissance. We begin by describing how psychologists' concerns with publication bias shifted from worrying about file-drawered studies to worrying about *p*-hacked analyses. We then review the methodological changes that psychologists have proposed and, in some cases, embraced. In describing how the renaissance has unfolded, we attempt to describe different points of view fairly but not neutrally, so as to identify the most promising paths forward. In so doing, we champion disclosure and preregistration, express skepticism about most statistical solutions to publication bias, take positions on the analysis and interpretation of replication failures, and contend that meta-analytical thinking *increases* the prevalence of false positives. Our general thesis is that the scientific practices of experimental psychologists have improved dramatically.

Contents

INTRODUCTION

If a team of research psychologists were to emerge today from a 7-year hibernation, they would not recognize their field. Authors voluntarily posting their data. Top journals routinely publishing replication attempts, both failures and successes. Hundreds of researchers preregistering their studies. Crowded methods symposia at many conferences. Enormous increases in sample sizes. Some top journals requiring the full disclosure of measures, conditions, exclusions, and the rules for determining sample sizes. Several multilab replication efforts accepted for publication before any data were collected. Overall, an unprecedented focus on replicability. What on earth just happened?

Many have been referring to this period as psychology's "replication crisis." This makes no sense. We do not call the rain that follows a long drought a water crisis. We do not call sustained growth following a recession an economic crisis. Experimental psychologists spent several decades relying on methods of data collection and analysis that make it too easy to publish false-positive, nonreplicable results. During that time, it was impossible to distinguish between findings that are true and replicable and those that are false and not replicable. *That* period, when we were unaware of the problem and thus did nothing about it, constituted the real replication crisis.

That crisis appears to be ending. Researchers now understand that the old ways of collecting and analyzing data produce results that are not diagnostic of truth and that a new, more enlightened approach is needed. Thousands of psychologists have embraced this notion. The improvements to our field have been dramatic. This is psychology's *renaissance*.

THE RENAISSANCE BEGINS: 2010–2012

Psychology's renaissance began when, in 2010–2012, a series of events drove psychological scientists into a spiral of methodological introspection. In this section, we review the events that, in our view, were the most consequential.

1. Social psychology's most prestigious journal, the *Journal of Personality and Social Psychology*, published an article by Daryl Bem (2011) in which he presented nine experiments supporting a transparently outlandish claim—that people can be influenced by an unforeseeable future event. For example, in his study 8, Bem found that participants were better able to recall words that they were later randomly assigned to rehearse. The reaction to this article was one of widespread disbelief, both inside (e.g., Wagenmakers et al. 2011) and outside (e.g., Carey 2011b) of academia; not surprisingly, these results did not replicate (Galak et al. 2012).[1] Bem's article prompted psychologists to start wondering how such a well-respected and well-intentioned scientist could have amassed a large body of evidence for an obviously false hypothesis.
2. Diederik Stapel, one of social psychology's most prominent and prolific contributors, confessed to decades of data fabrication. He eventually retracted dozens of articles. Predictably, this attracted attention from psychologists, non–psychological scientists, and the media (see, e.g., Achenbach 2011, Carey 2011a). Almost simultaneously, two other psychologists—Lawrence Sanna at the University of Michigan and Dirk Smeesters at Erasmus University Rotterdam—were discovered to have been authors on multiple articles containing fabricated results (see Simonsohn 2013). Stapel, Smeesters, and Sanna were separately investigated for academic misconduct, and all three resigned from their tenured positions. Importantly, the investigation into Stapel's work uncovered problematic methodological practices even in studies that were not fabricated.[2] As did the publication of Bem's (2011) article, these events prompted many psychologists to re-evaluate how the field conducts research.
3. In 2011, we wrote "False-Positive Psychology" (Simmons et al. 2011), an article reporting the surprisingly severe consequences of selectively reporting data and analyses, a practice that we later called *p-hacking*. In that article, we showed that conducting multiple analyses on the same data set and then reporting only the one(s) that obtained statistical significance (e.g., analyzing multiple measures but reporting only one) can dramatically increase the likelihood of publishing a false-positive finding. Independently and nearly simultaneously, John et al. (2012) documented that a large fraction of psychological researchers admitted engaging in precisely the forms of *p*-hacking that we had considered; for example, about 65% of respondents indicated that they had dropped a dependent variable when reporting a study. Identifying these realities—that researchers engage in *p*-hacking and that *p*-hacking makes it trivially easy to accumulate significant evidence for a false hypothesis—opened psychologists' eyes to the fact that many published findings, and even whole literatures, could be false positive.
4. Doyen et al. (2012) reported a failure to replicate one of the most famous findings in social psychology, that priming people with elderly stereotypes made them walk more slowly (Bargh et al. 1996). This prompted a lively and widely publicized debate, which, in turn, prompted Nobel Prize winner Daniel Kahneman to write a widely circulated email calling for researchers to resolve the debate by conducting systematic replications. (We have archived

[1] Bem did not identify flaws in the replication study; instead he conducted a meta-analysis with 90 studies that he interpreted as corroborating his findings (Bem et al. 2016). The meta-analysis is available on the online platform F1000Research.

[2] For example, the report from the three Dutch university committees investigating Stapel said, "The following situation also occurred. A known measuring instrument consists of six items. The article referred to this instrument but the dataset showed that only four items had been included; two items were omitted without mention. In yet another experiment, again with the same measuring instrument, the same happened, but now with two different items omitted, again without mention. The only explanation for this behavior is that it is meant to obtain confirmation of the research hypotheses" (Levelt et al. 2012, p. 50). We have archived a copy of the report at **https://osf.io/eup6d/**.

some of these exchanges at **https://osf.io/eygvz/**.) Perhaps not coincidentally, replication attempts soon became much more common.

5. In 2011, psychologist Brian Nosek began organizing several collaborative replication efforts, in which multiple independent labs attempted to replicate previously published research (Open Sci. Collab. 2012). He and Jeffrey Spies also worked to develop an online platform, the Open Science Framework (OSF), which was originally released in 2012; it allowed researchers to more transparently record, share, and report their work. This effort culminated, in 2013, in their launch of the Center for Open Science, a nonprofit organization that has heavily influenced the movement toward better research practices in psychology.

P-HACKING EXPLAINS AN OLD PARADOX

Psychologists have long been aware of two seemingly contradictory problems with the published literature. On the one hand, the overwhelming majority of published findings are statistically significant (Fanelli 2012, Greenwald 1975, Sterling 1959). On the other hand, the overwhelming majority of published studies are underpowered and, thus, theoretically unlikely to obtain results that are statistically significant (Chase & Chase 1976, Cohen 1962, Sedlmeier & Gigerenzer 1989). The sample sizes of experiments meant that most studies should have been failing, but the published record suggested almost uniform success.[3]

There is an old, popular, and simple explanation for this paradox. Experiments that work are sent to a journal, whereas experiments that fail are sent to the file drawer (Rosenthal 1979). We believe that this "file-drawer explanation" is incorrect. Most failed studies are not *missing*. They are published in our journals, masquerading as successes.

The file-drawer explanation becomes transparently implausible once its assumptions are made explicit. It assumes that researchers conduct a study and perform one (predetermined) statistical analysis. If the analysis is significant, then they publish it. If it is not significant, then the researchers give up and start over. This is not a realistic depiction of researcher behavior. Researchers would not so quickly give up on their chances for publication, nor would they abandon the beliefs that led them to run the study, just because the first analysis they ran was not statistically significant. They would instead explore the data further, examining, for example, whether outliers were interfering with the effect, whether the effect was significant within a subset of participants or trials, or whether it emerged when the dependent variable was coded differently. Pre-2011 researchers did occasionally file-drawer a study, although they did not do so when the study failed, but rather when *p*-hacking did. Thus, whereas our file drawers are sprinkled with failed *studies* that we did not publish, they are overflowing with failed *analyses* of the studies that we did publish.

Prior to 2011, even psychologists writing about publication bias and statistical power had ignored the fact that *p*-hacking occurs, let alone the fact that it is a first-order problem for the validity of psychological research. Psychology suffered from *p*-hacking neglect.[4] For example, in the article that introduced the term "file-drawer problem," Rosenthal (1979, p. 638, emphasis

[3] In a satirical piece, Brian A. Nosek, under the pseudonym Ariana K. Bones, proposed a precognition explanation: Researchers are able to predict which studies will produce an unlucky result and, thus, they know not to run these studies (Bones 2012).

[4] Methodologists in other fields had brought up the problem we now know as *p*-hacking (Cole 1957, Ioannidis 2005, Leamer 1983, Phillips 2004). However, perhaps because they did not demonstrate that this was a problem worth worrying about or because they did not propose concrete and practical solutions to prevent it (i.e., they did not demonstrate that this was a solvable problem), their concerns did not have perceivable consequences on how research was conducted and reported in their fields.

added) wrote, "The extreme view . . . is that the journals are filled with the 5% of the *studies* that show Type I errors, while the file drawers back at the lab are filled with the [other] 95% of the *studies*." Similarly, Greenwald (1975, table 1), in his seminal (and excellent) "Consequences of prejudice against the null hypothesis," considered four things researchers may do upon obtaining a nonsignificant result: (*a*) submitting the null finding for publication, (*b*) conducting an exact replication, (*c*) conducting a modified replication, or (*d*) giving up. His list excludes conducting additional analyses of the same data.

Vul et al. (2009), in their article about "voodoo correlations," represent an interesting partial exception to *p*-hacking neglect.[5] Focusing on fMRI research, they discussed the influence of choosing to report only the analyses of voxels that were statistically significant. However, the exception is only partial because Vul et al. suggested that the problem did not apply to psychological research more generally. For example, they wrote, "We suspect that the problems brought to light here are ones that most editors and reviewers of studies using purely behavioral measures would usually be quite sensitive to" (Vul et al. 2009, p. 285).[6]

Over the years, researchers repeatedly documented the fact that psychologists were running dramatically underpowered studies (e.g., Chase & Chase 1976, Cohen 1962), and they repeatedly called for sample sizes to be increased to levels that simple math revealed to be necessary. Eventually, Sedlmeier & Gigerenzer (1989) decided to see whether any of these articles had been successful in getting researchers to increase their sample sizes. They concluded that they were not; sample sizes had remained inadequately low. They speculated that this was in part because researchers overestimate the statistical power of small samples and, thus, the statistical power of their own small-sampled studies (citing, interestingly, Tversky & Kahneman 1971).[7] But this explanation is necessarily incomplete. How could researchers possibly maintain wrong beliefs about required sample sizes in the face of years of feedback from experience? It might make sense for new graduate students to erroneously think 12 participants per cell will be a sufficiently large sample size to test a counterintuitive attenuated interaction hypothesis, but it would not make sense for a full professor to maintain this belief after running hundreds of experiments that should have failed. It is one thing for a very young child to believe that 12 peas are enough for dinner and quite another for a chronically starving adult to do so.

P-hacking provides the real solution to the paradox. *P*-hacking is the only honest and practical way to *consistently* get underpowered studies to be statistically significant. Researchers did not learn from experience to increase their sample sizes precisely because their underpowered studies *were not failing*.[8] *P*-hacking allowed researchers to think, "I know that Jacob Cohen keeps saying that we need to increase our sample sizes, but most of my studies work; he must be talking about other people. They should really get their act together."

[5]Their article became well known with the title referencing "voodoo correlations," but shortly before publication, they replaced "voodoo" with "puzzlingly high."

[6]Fiedler (2011), in his article "Voodoo Correlations Are Everywhere," appears at first glance to extend Vul et al.'s (2009) conclusions to all psychological research. However, he actually makes a completely different point, expressing the concern that researchers may systematically design studies that are expected to generate larger effects (e.g., building in moderators that inflate effects). Fiedler's article was, therefore, about increasing generalizability rather than reducing false positives. Even his subsection on "Biases from the analyses" (Fiedler 2011, p. 166) exclusively pertained to decisions made *before* data collection.

[7]Sedlmeier & Gigerenzer (1989) also speculated that underpowered studies persisted because psychologists first learned about Fisher's approach to inference, which did not include discussions of statistical power, and only later learned about Neyman-Pearson's approach, which did include discussions of statistical power (see Sedlmeier & Gigerenzer 1989, p. 314, last paragraph). This explanation seems sufficiently implausible to us to be relegated to a footnote.

[8]For readers enamored with the peas analogy, *p*-hacking is like eating a seven-course meal after you eat the peas. Believing that $n = 12$ is enough is like attributing your satiety to the peas.

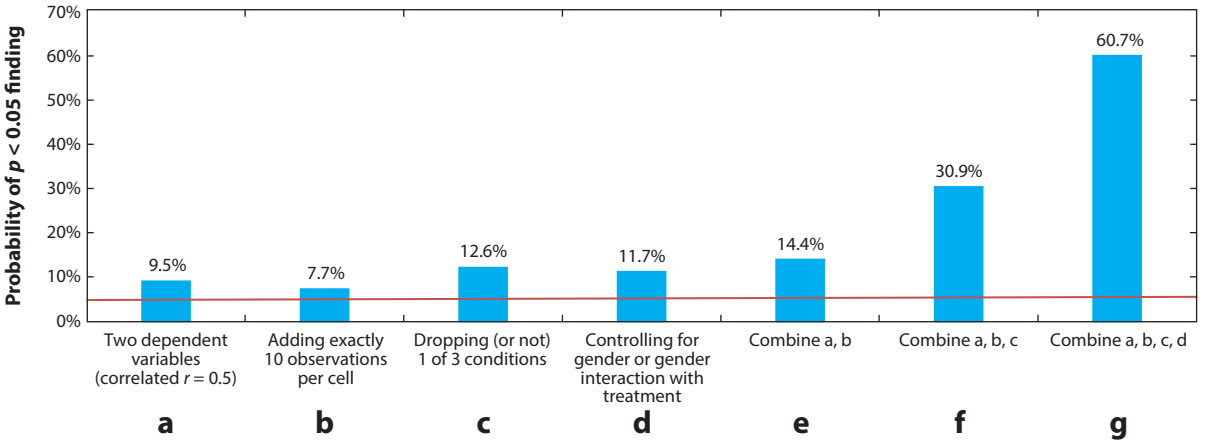

Figure 1

Probability that a study obtains statistical significance ($p < 0.05$) as a function of the types of *p*-hacking a researcher is willing to engage in (originally reported in Simmons et al. 2011, table 1). The figure is based on 15,000 simulated studies. The baseline study was a two-condition between-subjects design with 20 observations per cell drawn from the same normal distribution (thus, under the null hypothesis of no difference between conditions). The *y* axis depicts the share of studies for which at least one attempted analysis was significant. Results were obtained by (*a*) conducting three t-tests, one on each of two dependent variables and a third on the average of these two variables; (*b*) conducting one t-test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell; (*c*) conducting t-tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low $= -1$, medium $= 0$, high $= 1$); and (*d*) conducting a t-test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). For bar *d*, we report a significant effect if the effect of the condition was significant in any of these analyses or if the gender $\times$ condition interaction was significant. The R code to reproduce these results can be found at **https://osf.io/a67ft/**.

In "False-Positive Psychology" (Simmons et al. 2011), we demonstrated that *p*-hacking represents a major threat to the validity of *all* empirical research that relies on hypothesis testing. Specifically, we showed how acceptable levels of *p*-hacking could raise false-positive rates to unacceptable levels (e.g., from 5% to 61%; see **Figure 1**). Furthermore, by moderately *p*-hacking two real experiments, we demonstrated how easy it is to obtain statistically significant evidence for a transparently false hypothesis: that listening to a song can change a person's age. It is now clear that, with enough analytic flexibility, *p*-hacking can turn any false hypothesis into one that has statistically significant support.[9]

It is tempting to take comfort in the fact that psychology publications usually contain more than one experiment (see, e.g., Stroebe 2016, section 1.6.1). Even if a single *study's* false-positive rate is as high as 61%, the odds of getting four false positives for a single *article* (in the predicted direction) is nevertheless quite low: $(0.61/2)^4 = 0.8\%$. However, this framing of the problem assumes two things—one that is never true and one that is sometimes not true. The first assumption is that researchers publish every study. They do not. If a researcher is willing to file-drawer even a small number of studies, a false-positive four-study article becomes easy enough to produce [Pashler & Harris (2012, argument 2, p. 533) provide an excellent discussion of this issue]. Framed differently, to find four experiments worth of significant evidence for a false (directional)

[9]The effect of *p*-hacking on Bayesian hypothesis testing and on both frequentist and Bayesian confidence intervals is just as severe (see Simonsohn 2014).

hypothesis, researchers need to run 160 studies if they do not *p*-hack but only 13 if they do.[10] The other assumption is that 61% represents some sort of upper bound, when, in fact, that estimate is likely to be conservative, particularly if researchers are able to flexibly contort the hypotheses to fit the data that they observe (Kerr 1998). In truth, it is not that hard to get a study's false-positive rate to be very close to 100%, in which case even a multistudy article's false-positive rate will be close to 100%.

Irrespective of these details, it is a mathematical fact that *p*-hacking makes it dramatically easier to generate false-positive findings, so much so that, for decades, *p*-hacking enabled researchers to achieve the otherwise mathematically impossible feat of getting most of their underpowered studies to be significant. *P*-hacking has long been the biggest threat to the integrity of our discipline.

FALSE POSITIVES ARE BAD, WE DO NOT KNOW HOW MANY THERE ARE, AND THAT DOES NOT MATTER

False positives are bad. Publishing them can cause scientists to spend precious resources chasing down false leads, policy makers to enact potentially harmful or ineffective policies, and funding agencies to allocate their resources away from hypotheses that are actually true. When false positives populate our literatures, we can no longer distinguish between what is true and what is false, undermining the very goal of science.

It is sometimes argued that any policy that reduces false positives will necessarily increase false negatives (e.g., Fiedler et al. 2012). However, because studying false positives diverts resources and attention away from the study of true hypotheses, the reduction of false positives can actually decrease false negatives. For instance, psychologists who realize that they should not study precognition because it is a false positive may instead discover true effects that would not have been discovered otherwise.

The need to prevent the publication of false positives is not controversial. What are controversial are the questions of whether the existing literature contains a large number of false-positive findings and, thus, how much of that literature should be believed. Some researchers believe that *p*-hacking and false positives are common, whereas others believe that they are relatively rare.

The evidence in favor of the commonness of false positives comes in two forms. First, as outlined in the previous section, it is extremely unlikely for researchers to have succeeded for so long in getting so many underpowered studies to be statistically significant without engaging in *p*-hacking; because *p*-hacking makes it nearly as easy to publish a false-positive as a true-positive finding, it seems reasonable to assume that many published findings are false positives. Second, systematic attempts to replicate existing findings have not been overwhelmingly successful (see, e.g., Alogna et al. 2014, Cheung et al. 2016, Open Sci. Collab. 2015, Wagenmakers et al. 2016).

The counterargument tends to come in two forms. One consists of dismissing failures to replicate as either poorly executed or misguided, an argument that we discuss in more detail in the section titled Replications. The other is grounded in the belief that researchers are well intentioned and that *p*-hacking is not. According to this view, asserting that researchers engage in *p*-hacking is tantamount to asserting that researchers are unethical. For example, in an editorial about these issues, Luce et al. (2012, p. iii) wrote, "we are concerned that the present tenor of the discussions and the structure of the solutions may produce an environment that presumes abuse."

[10] If the false-positive rate is 5% (1 in 20), one needs an average of 80 attempts to get four that are $p < .05$ in either direction and 160 attempts to get four in the predicted direction. If the false-positive rate is 61%, these numbers are 6.6 and 13.1, respectively.

Along the same lines, Fiedler & Schwarz (2016) criticized John et al.'s (2012) survey assessing the prevalence of questionable research practices on the grounds that the survey did not sufficiently distinguish between selective reporting that was well intentioned and selective reporting that was ill intentioned.

As we see it, both sides of this debate agree that the vast majority of researchers are honest and well intentioned. The disagreement is, instead, about whether honest and well-intentioned researchers will engage in *p*-hacking. As (honest and well-intentioned) researchers who *p*-hacked for many years, we strongly believe that they do.

P-hacking is a pervasive problem precisely because researchers usually do not realize that they are doing it or appreciate that what they are doing is consequential (Vazire 2015). It is not something that malevolent researchers engage in while laughing maniacally; it is something that benevolent researchers engage in while trying to understand their otherwise imperfect results. *P*-hacking is the byproduct of the very human tendency to justify actions that produce desirable outcomes (Kunda 1990, Mahoney 1979). To suspect that researchers *p*-hack is merely to suspect that they are human.

Fortunately, at the end of the day, this debate surrounding the prevalence of false positives is irrelevant. Whether one believes that false positives are common or rare, we can all agree that we should embrace research methods that prevent the publication of false positives. Thus, we do not need to agree on the prevalence of false positives, but only on the fact that the methods we used for many decades did not adequately defend against them.

To illustrate this point, imagine a surgeon who chooses at random which of a patient's diseased fingers to amputate. After the procedure, observers could consider whether the surgeon had amputated the correct or incorrect finger, but that consideration is obviously irrelevant for determining whether the surgeon should adopt a different, less error-prone method going forward. Like the surgeon, regardless of what happened previously, we should embrace the method that is demonstrably more likely to increase our chances of getting things right in the future.

Accordingly, in the following sections, we focus not on the question of whether false positives are common or rare, but on what we can do to prevent and correct them.

PREVENTING *P*-HACKING IN FUTURE RESEARCH

In this section, we discuss two policies to reduce *p*-hacking: disclosure and preregistration.

Disclosure

The most straightforward way to prevent researchers from selectively reporting their methods and analyses is to require them to report less selectively. At the bare minimum, this means requiring authors to disclose all of their measures, manipulations, and exclusions, as well as how they determined their sample sizes (Simmons et al. 2011). This allows reviewers to better diagnose whether authors' results are likely to have been obtained by *p*-hacking.

Although requiring scientists to report what they actually did in their studies should be uncontroversial, most journals have refrained from adopting this requirement. Still, important progress has been made. As editor, Eric Eich (2014) labored for months to implement these disclosure requirements at *Psychological Science*, and his successor, Stephen Lindsay (2015), extended them. Simine Vazire (2016) implemented similar requirements at *Social Psychological and Personality Science*.

Even journals that do not require authors to disclose all of their important methodological details have seen an uptick in authors' propensity to fully disclose their methods, both because of an

increase in voluntary disclosure[11] and because many reviewers now demand it, often including the following text in their reviews (see **http://osf.io/hadz3**): "I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science. I include it in every review." Disclosure is slowly becoming the norm, although it is not yet a ubiquitous requirement.

Preregistration

Whereas disclosure represents the bare minimum defense against *p*-hacking, proper preregistration represents the cure. Preregistrations are time-stamped plans for data analysis written before any data are analyzed. Preregistrations identify, in advance, which analyses are confirmatory and which are exploratory (Wagenmakers et al. 2012), greatly reducing the prevalence and influence of *p*-hacking (Bakker et al. 2012, van't Veer & Giner-Sorolla 2016).

One concern often raised about preregistrations is that they may hinder exploration. For example, Association for Psychological Science president Goldin-Meadow (2016) wrote, "[I] fear that preregistration will stifle discovery. Science isn't just about testing hypotheses—it's also about discovering hypotheses . . . . Aren't we supposed to let the data guide us in our exploration? How can we make new discoveries if our studies need to be catalogued before they are run?"

Overcoming this concern requires realizing that preregistrations do not tie researchers' hands, but merely uncover readers' eyes. Preregistering does not preclude exploration, but it does communicate to readers that it occurred. Preregistering allows readers to discriminate between confirmatory analyses, which provide valid *p*-values and trustworthy results, and exploratory analyses, which provide invalid *p*-values and tentative results (Moore 2016). Preregistration allows confirmatory results to be given the full credit that they deserve.

As we recently wrote (Simmons et al. 2017), preregistration has two key advantages over disclosure. First, it gives researchers the freedom to conduct analyses that could, if disclosed afterwards, seem suspicious, such as excluding participants who failed an attention check or running an unusual statistical test. Similarly, preregistration allows researchers to add observations to a study. For example, a researcher could specify a plan to add 100 observations if the key analysis is not significant after collecting the first 100 (see also Lakens 2014). Second, preregistration is the only way for authors to convincingly demonstrate that their key analyses were not *p*-hacked.

Psychologists currently have two main options for preregistration: AsPredicted (**http://AsPredicted.org**) and the OSF (**http://osf.io**). On AsPredicted, researchers preregister by completing a standardized form containing eight short questions. The platform then generates an easy-to-read one-page PDF (for an example, see **https://Aspredicted.org/nfj4s.pdf**), which can be shared anonymously during the review process and made public if and when the authors are ready to do so. On the OSF, researchers with accounts can collaborate, share, and archive files. These files are stored as elements within projects. Individual elements within a project can be locked and time stamped, registered, and used as study preregistrations if researchers provide the appropriate information or if they select a preregistration template (for step-by-step instructions, see **https://osf.io/sgrk6/**; for the set of templates, see **https://osf.io/zab38/**).

[11] We have suggested that authors include a standardized disclosure statement in their articles: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study" (Simmons et al. 2012).

Although very few published studies in psychology have been preregistered, this is likely attributable to the considerable lag between study design and study publication. Indeed, there is ample evidence that preregistration is starting to catch on. AsPredicted was launched in December 2015. In its first 14 months, 1,631 different people completed an AsPredicted preregistration. We expect that in 3–5 years, published preregistered experimental psychology studies will be common.

ADDRESSING *P*-HACKING IN PAST RESEARCH

In this section, we discuss the two main approaches that researchers take to identify which published findings are true positives and which are false positives: attempts to replicate individual studies and statistical analyses of collections of studies.

Replications

To a scientist, a true effect is one that replicates under specifiable conditions (K12 Reader 2012, Popper 1963). Not surprisingly, widespread concern about the truth of existing effects has generated a surge of interest in replicating studies under the conditions originally specified. Accordingly, there have been discussions of how to increase the frequency of replications (Asendorpf et al. 2013, Koole & Lakens 2012), discussions of how to properly design replications (Brandt et al. 2014, Schmidt 2009), special issues dedicated to replications (Nosek & Lakens 2013, Pashler & Wagenmakers 2012), multilab replication efforts (Ebersole et al. 2016, Klein et al. 2014), and the creation of registered replication reports (Simons et al. 2014). Indeed, of CurateScience.org's database of over 1,000 attempts to replicate psychological studies (which we have archived at **http://web.archive.org/web/20170227214829/http://curatescience.org/**), approximately 96% of them have been conducted since 2011 (E.P. LeBel, personal communication).

Instead of reviewing this impressive body of largely unrelated studies, in this section, we discuss issues pertaining to the interpretation and analysis of replication results.

How to interpret failures to replicate. Just as it is impossible to bathe in the same river twice, it is impossible to run the same study twice. This unfortunate fact generates the same debate almost every time a failure to replicate becomes public, with some arguing that the replication casts doubt on the veracity of the original finding and others attributing the failure to substantive differences between the replication and the original (or to poor execution of the replication attempt). How should such debates be productively resolved?

Some of the debates are about differences in design. For example, the Open Science Collaboration (Open Sci. Collab. 2015) made painstaking efforts to precisely replicate the procedures from 100 published psychology studies. Nevertheless, some differences inevitably occurred and have been the subject of intense debate regarding how consequential they may be (Anderson et al. 2016, Gilbert et al. 2016, Inbar 2016, Van Bavel et al. 2016).

Even when the original design and the replication design are strictly identical, people may debate whether they are psychologically identical. For example, Doyen et al. (2012) failed to replicate the seminal study by Bargh et al. (1996), in which US participants primed with words associated with the elderly were reported to have walked more slowly when exiting the experiment. Stroebe & Strack (2014, p. 62), citing differences in the times and locations of the studies, wrote, "it is also possible that the concept of 'walking slowly' is not a central part of the stereotype of [the] elderly in Belgium some 20 years later." Similarly, when a registered replication report (Wagenmakers et al. 2016) failed to replicate Strack et al.'s (1988) finding that participants evaluate cartoons as funnier when biting a pencil in a way that makes them smile, Strack (2016, p. 929) responded that "despite

the [similar] obtained ratings of funniness [M = 4.73 in original, M = 4.59 in replication], it must be asked if Gary Larson's The Far Side cartoons [used in these studies] that were iconic for the zeitgeist of the 1980s instantiated similar psychological conditions 30 years later."

Whose responsibility is it to empirically test whether differences between the original and the replication are responsible for a replication failure? Answers to this question fall on a continuum between two extremes. One extreme treats every hypothesized moderator as consequential until proven otherwise (e.g., assuming that only US participants stereotype the elderly as slow walkers unless data falsify this hypothesis). The other extreme treats every hypothesized moderator as inconsequential until proven otherwise (e.g., assuming that Belgians also stereotype the elderly as slow walkers unless data falsify this hypothesis). For instance, Simons (2014, p. 77) writes, "When researchers posit a moderator explanation for discrepant results, they make a testable claim . . . . They then can conduct a confirmatory study to manipulate that moderator and demonstrate that they can reproduce the effect and make it vanish. In fact, they have a responsibility to do so."

We propose a middle ground, whereby the burden of proof is on the researcher espousing the least plausible claim. If an original study manipulated hunger by having participants fast for 8 hours, whereas a failed replication manipulated hunger by having participants fast for 8 minutes, then the burden of proof is on the replicator to show that the failure persists when hunger is more effectively manipulated. On the other hand, if original researchers attribute a failure to replicate their research to the fact that the replication was conducted on a different day of the week, then the burden of proof is on them to show that the day-of-the-week moderator is relevant. Although original and replication researchers often disagree about the plausibility of hypothesized moderators, neutral observers often agree. Thus, neutral observers often agree on who has the burden of proof.

Sometimes, the plausibility of an original author's explanation for a replication failure can be assessed by conducting additional analyses on existing data (Simonsohn 2016b). For example, Carney et al. (2015, table 2, row 2) suggested that Ranehill et al. (2015) may have failed to replicate their power poses effect because Ranehill et al.'s Swiss participants, unlike Carney et al.'s (2010) US participants, may not have associated the same poses with power. This explanation predicts that the power manipulation should not have worked in the replication. However, in the Swiss replication, "results showed a significant effect of power posing on self-reported feelings of power," (Ranehill et al. 2015, p. 653) indicating that the power poses manipulation was, in fact, successful in Switzerland. To be clear, not all of Carney et al.'s (2015) hypothesized moderators are ruled out by this particular analysis. Our purpose is simply to illustrate how new analyses of existing data can be used to assess the plausibility of hypothesized moderators.

Regardless of who has the burden of proof, it is absolutely essential for those who are not convinced by a failure to replicate to make testable claims about the circumstance(s) under which the effect is expected to replicate and, therefore, the circumstances under which a failure to replicate would be informative. For example, to move the debate forward, if Strack (2016) believes that The Far Side cartoons are no longer sensitive to smile-inducing manipulations, he would need to specify a precise set of procedures for choosing stimuli that will produce the effect. If critics of a replication cannot specify conditions under which the effect is expected to replicate, then they are not making a scientific claim (Lakatos 1970).

What is a failure to replicate? In the previous section, we discussed how failures to replicate should be interpreted. But how do we even know when a replication attempt has failed? This question is harder to answer than it appears.

Replications have traditionally been deemed failures when the effect described by the original study is not statistically significant in the replication. This approach has two obvious flaws. First, a

replication attempt could be nonsignificant simply because its sample size is too small (Asendorpf et al. 2013, Patil et al. 2016, Valentine et al. 2011, Verhagen & Wagenmakers 2014). Second, a replication attempt could be significant even if the effect size is categorically smaller than in the original. For example, imagine that an original study produces a significant effect size of $d = 0.80$ with 20 observations per cell, and a replication produces a significant effect size of $d = 0.01$ with 20 *million* observations per cell. Despite the statistically significant effect in the replication, it is clear that the original study could not have detected such a small effect. It thus seems odd to treat the replication as successful (Simonsohn 2015c).

An alternative approach is to examine whether the replication effect size is significantly different from the original effect size. This approach similarly suffers from two major problems. First, when original effects are barely significant (e.g., $p = 0.049$), replications with nearly infinite sample sizes may not be significantly different from the original, even if the true effect size is exactly zero (Asendorpf et al. 2013, Simonsohn 2015c). This means that replications of *p*-hacked false-positive findings are particularly unlikely to fail, as such findings tend to have *p*-values that are barely significant (Simonsohn et al. 2014a). Second, if the original and replication studies are both highly powered, a statistically significant difference may not be meaningful. For instance, if an original study found an effect size of $\hat{d} = 0.53$ and a replication found a significantly different effect size of $\hat{d} = 0.51$, it would be bizarre to conclude that the replication had failed.[12]

Other approaches to comparing original and replication results are also unsatisfactory. The Reproducibility Project (Open Sci. Collab. 2015) relied on a few different tests to assess how many of 100 replication attempts were failures. One such test examined whether the original study's effect size estimate was within the confidence interval of the replication. However, this approach ignores the uncertainty around the original study's effect size estimate and, consequently, inflates failure rates, making psychology seem *less* replicable than it is. Critical of the Reproducibility Project's conclusion, Gilbert et al. (2016) reanalyzed their data and tested whether the replication's effect size estimate was within the confidence interval of the original. However, this approach ignores the uncertainty around the replication's effect size estimate and, consequently, deflates failure rates, making psychology seem *more* replicable than it is (for a detailed discussion, see Simonsohn 2016a).

All of these approaches lack two critical features. First, they fail to account for the fact that replication attempts may be *inconclusive*, rather than definitive successes or failures (Etz & Vandekerckhove 2016, Simonsohn 2015c, Verhagen & Wagenmakers 2014). Second, they fail to test whether the effect size observed in the replication is small enough for us to call it a failure. To accomplish this, one must define what "small enough" means. One approach is to define "small enough" in absolute terms. For example, one could test whether the replication study allows one to reject an effect size of $|d| \geq 0.10$ (see, e.g., Greenwald 1975; Serlin & Lapsley 1985, 1992). This approach requires replication sample sizes to be very large, as one needs approximately 1,500 observations per cell to have an 80% chance to reject $d \geq 0.10$ if the true effect size is zero. A more practical definition of "small enough" incorporates a relative assessment of size. Two approaches have been proposed along these lines. One is a Bayesian approach that tests whether the replication is more consistent with there being no effect or with the conclusions from the original study, given an assumed prior (Verhagen & Wagenmakers 2014).[13] The other assesses whether the effect size

[12] Patil et al. (2016) propose concluding that a replication has failed when its point estimate lies outside of the "prediction interval" of the original study. This is mathematically equivalent to testing whether the standardized point estimates of the original and replication studies are significantly different from each other.

[13] The default Bayesian t-test (Rouder et al. 2009) could also be used to assess whether a replication has failed. It tests whether the results are more consistent with the null or with a default hypothesis. Unfortunately, when one uses the default hypothesis

observed in the replication would have been detectable with the sample size used in the original.[14] This second approach leads to easy-to-interpret results and requires replicators to run 2.5 times the sample size of the original to have an 80% chance of concluding that the replication has failed if the true effect is zero (Simonsohn 2015c).[15]

Returning to the Reproducibility Project, only 36 of its 100 replication attempts obtained a significant result (Open Sci. Collab. 2015). Many observers, including those in the popular press, interpreted this as meaning that the rest had failed. For example, an article in *The New York Times* indicated that "More than 60 of the studies did not hold up" (Carey 2015). However, once we apply a better test of failure to replicate that allows for replications to be categorized as inconclusive, the interpretation differs. Again, 36% were conclusive successes; however, among the remaining 64%, only 25% were conclusive failures, and 49% were inconclusive, i.e., neither failures nor successes (Simonsohn 2016a; for similar results using the Bayesian approach to evaluating replications, see Etz & Vandekerckhove 2016). This high share of inconclusive results reflects the fact that many of the 100 replications were underpowered, with samples smaller than 2.5 times the sample sizes of the original studies.

Analyzing a Collection of Studies

Whereas replications aim to elucidate whether *individual* studies contain evidence, they cannot easily elucidate whether *collections* of studies contain evidence. To accomplish this, researchers have resorted to statistical reanalyses of existing data. Some of these approaches are designed to tell us whether a collection of studies contains any evidence of selective reporting; others are designed to tell us whether a collection of studies contains any evidence of the effect of interest after controlling for selective reporting. We discuss each of these approaches in turn.

Have the results we observed been selectively reported? Researchers have used tools such as the funnel plot (Egger et al. 1997) and the excessive significance test (Francis 2012, Ioannidis & Trikalinos 2007, Schimmack 2012) to assess the likelihood that a literature (or individual article) is missing at least one nonsignificant result. This approach suffers from at least two shortcomings. First, it does not seek to answer a question that needs to be answered. We already know that researchers do not report 100% of their nonsignificant studies and analyses. Some researchers have reported some null findings, but none have reported all of them. Second, and even more important, knowing that a literature contains selective reporting tells us nothing about whether the effects in that literature are actually true (Morey 2013, Simonsohn 2012). Thus, this approach is unable to answer the question we are actually interested in: Does the literature contain evidence once you correct for selective reporting?

After controlling for selective reporting, do the data suggest the effect exists? Trim-and-fill is the most common approach for correcting for selective reporting in a collection of studies (Duval & Tweedie 2000). This technique aims to create an unbiased set of studies by removing some seemingly biased studies from the set (trimming) and replacing them with fictitious studies

that proponents of the test have advocated for, the test classifies even highly significant small-to-moderate effects as accepting the null. It is prejudiced against small effects (Simonsohn 2015b).

[14]Simonsohn (2015c) has specifically proposed testing whether the original study would have had 33% power to detect the effect size observed in the replication.

[15]There are Bayesian approaches to these types of tests as well. Kruschke (2013) provides an approach that adopts an absolute definition of "small enough."

that are seemingly missing from the set (filling). Unfortunately, trim-and-fill vastly undercorrects for selective reporting. For example, when the true effect size is zero ($d = 0$) and selective reporting inflates the estimate to $\hat{d} = 0.72$, trim-and-fill corrects the estimate to $\hat{d} = 0.66$, thus producing an effect size estimate that is nearly as biased as it would have been without the correction (Simonsohn et al. 2014b, figure 2).[16]

A different approach for correcting for selective reporting is *p*-curve analysis (Simonsohn et al. 2014a). *P*-curve is the distribution of statistically significant *p*-values from a set of studies.[17] We can think of *p*-curve as the histogram of significant *p*-values (what share are 0.01s, what share are 0.02s, etc.), although *p*-curve analysis treats *p*-values as continuous. *P*-curve's shape can be used to diagnose whether a literature contains replicable effects. True effects of any size ($d \neq 0$), studied with samples of any size, generate right-skewed *p*-curves (more very significant than barely significant *p*-values). Truly nonexistent effects ($d = 0$), studied with samples of any size, generate flat *p*-curves (just as many very significant as barely significant *p*-values). Thus, significantly right-skewed *p*-curves tell us that a literature contains at least some effects that are expected to replicate, and significantly flat *p*-curves tell us that a literature's studies are not expected to replicate. (Nonsignificant *p*-curves are inconclusive.) Although *p*-curve analysis performs much better than other methods that aim to correct for publication bias, it is not infallible.[18] Like all other methods, *p*-curve analysis may lead to inaccurate results when the analyzed collection of studies includes some that are fraudulent, erroneously reported, or ambitiously *p*-hacked (e.g., seeking $p < 0.01$ instead of $p < 0.05$).[19] *P*-curve analysis may also be inaccurate if *p*-curvers incorrectly choose *p*-values from individual studies or if they select which studies to analyze after looking at the results (see Footnote 17).[20]

Simmons & Simonsohn (2017) provide an example of *p*-curve analysis that illustrates its ability to diagnose the replicability of effects. As mentioned above, Ranehill et al. (2015) failed to replicate Carney et al.'s (2010) effects of power poses on behavioral and hormonal outcomes. However, they did replicate Carney et al.'s effects of power poses on feelings of power (the manipulation check). Carney et al. (2015) responded to Ranehill et al. by identifying 33 articles purporting to show that power poses do affect downstream outcomes or feelings of power. Simmons and Simonsohn (2017) *p*-curved those articles, and the *p*-curve results matched those of the replication. Although a *p*-curve of studies analyzing self-reported feelings of power was directionally right skewed (suggesting the presence of replicable effects), the *p*-curve of studies analyzing downstream outcomes was significantly flat (suggesting the absence of replicable effects).

[16] PET-PEESE (Stanley & Doucouliagos 2014) is an alternative but less popular tool. It performs poorly both in the presence of selective reporting (Gervais 2015) and in its absence (Simonsohn 2017). Psychologists should not use PET-PEESE.

[17] These *p*-values must come from an analysis that tested the original authors' hypothesis of interest, be statistically independent of each other, and be expected to have a uniform distribution under the null; we provide detailed guidelines on how to select *p*-values in ways that ensure that these statistical requirements are met (Simonsohn et al. 2014a, figure 4).

[18] *P*-curve analysis can also be used to estimate effect size correcting for publication bias (Simonsohn et al. 2014b).

[19] Ulrich & Miller (2015) first pointed out the problem with ambitious *p*-hacking. In "Better *P*-curves" (Simonsohn et al. 2015), we modified *p*-curve analysis to make it more robust.

[20] Some critics have claimed that *p*-curve analysis is distorted in the presence of effect size heterogeneity, that is, when different studies included in *p*-curve investigate effects of different sizes (McShane et al. 2016, van Aert et al. 2016). However, *p*-curve is actually robust to heterogeneity (see, e.g., Simonsohn et al. 2014a, figure 2; Simonsohn et al. 2014b, supplement 2). The different perspectives arise because of how the critics define the question of interest. *P*-curve estimates the average true effect of the studies that are included in the analyses, which is the result that we believe to be of interest. The critics would like *p*-curve to estimate the average effect of all studies that could ever be attempted, a result that we do not believe to be of interest. In any case, for any method to achieve the critics' objective, researchers would need to design and attempt their studies at random, which they do not.

INNOCENT ERRORS AND (NOT-SO-INNOCENT) FRAUD

Although *p*-hacking is arguably the biggest threat to the validity of published research, it is not the only threat. Results can also be false positive because of unintentional errors or fraud. In this section, we review efforts to address these important problems.

Unintentional Errors

Many parts of the publication process are susceptible to human error, including data entry and analysis, the copying of results from statistical software into manuscripts, and the copyediting process. Most errors are impossible for readers and reviewers to detect without having access to the raw data, making it impossible to assess their prevalence. However, some errors are visible to those who are looking for them. For example, one article with more than 600 Google Scholar citations contains the line, "...showing an increase in expected willingness to help when payment level increased from low to medium, $F(1, 607) = 3.48$, $p < 0.001$." That phrase includes all the information needed to detect the reporting error, as the *p*-value associated with $F(1, 607) = 3.48$ is not $p < 0.001$, but rather $p = 0.063$. It is clear that there is an error with the *F*-value, *p*-value, or both.[21]

To see how common it is for research articles to contain these types of errors, Bakker & Wicherts (2011) checked the internal consistency of 4,077 test results reported in various psychology journals. They found that 54% of articles contained at least one error, and 12% had at least one error that altered the statistical significance of a reported result.[22] In a separate effort, Brown & Heathers (2016) developed the granularity-related inconsistency of means (GRIM) test, which checks whether reported means of integer data (e.g., Likert scales) are consistent with reported sample sizes. For instance, if an article's sample size is $n = 10$, the mean of a Likert scale variable could include up to one decimal; thus, for example, 2.32 is not a possible mean when $n = 10$. Applying the GRIM test to a small sample of 71 articles, Brown & Heathers (2016) found that more than 50% had at least one error.

Although some researcher errors are slight and inconsequential, others are more serious and may invalidate the researchers' conclusions or expose deeper problems with the researchers' methods and analyses. For example, seemingly inconsequential errors detected by applying the GRIM test to a set of four articles by the same author led to the discovery of potentially more serious errors in unrelated articles by that author (see van der Zee et al. 2017).

Perhaps the simplest solution to this problem is to require authors to post their data and materials. Because even the most well-intentioned researcher is bound to make an occasional mistake, it makes sense to allow those mistakes to be identified and corrected. Public data posting not only allows others to verify the accuracy of the analyses, but also incentivizes authors to more carefully avoid errors (Miguel et al. 2014, Wicherts & Bakker 2012). We started posting data, code, and original materials for our own research a few years ago; it is sobering how often these actions lead us to catch errors shortly before we submit our manuscripts. Our own "False-Positive Psychology" article (Simmons et al. 2011) had a (fortunately inconsequential) error that was discovered, 5 years after publication, by a careful reader relying on our posted materials.[23]

[21] Every test result in this article is incorrect; our best guess is that the reported *F*-values are actually *t*-values. For instance, the quoted $F(1, 607) = 3.48$ should have been reported as $t(607) = 3.48$.

[22] Similar analyses and results are reported by Berle & Starcevic (2007), García-Berthou & Alcaraz (2004), and Nuijten et al. (2016).

[23] The error was discovered by Aurélien Allard; the updated R code to reproduce the simulations in the article explains the nature of the error (**https://osf.io/a67ft/**). **Figure 1** reports the corrected results.

The OSF has dramatically facilitated the posting of raw data and materials. Moreover, the journal *Psychological Science*, under the editorship of Eric Eich (2014) and in collaboration with the Center for Open Science, has introduced a badge system that places a small icon on the first page of articles that have made their data and materials publicly accessible. Kidwell et al. (2016) report that, before the badge system began (in 2014), authors publishing in *Psychological Science* were just as unlikely to make their data publicly available as were authors publishing in other leading journals (about 3% of authors did so). By the following year, authors publishing in *Psychological Science* were posting their data almost 40% of the time, whereas the other journals saw no increase at all.

Fraud

Many see fraud as the asteroid collision of social science, locally catastrophic but so rare that prevention is hardly relevant. Suspicions of fraud in, and the accompanying retractions of, articles published by Stapel, Smeesters, Sanna, Chiou, Förster, LaCour, and others suggest that this view of fraud may be naive. Discussions of fraud typically focus on two questions: How common is it and how can we stop it?[24]

Estimating the frequency of fraud is very difficult. Some blatantly detectable fraud is prevented by vigilant coauthors, reviewers, or editors and, thus, not typically observed by the rest of the field. The fraud that gets through those filters might be noticed by a very small share of readers. Of those readers, a very small number might ever make their concerns known. And of those concerns, fewer still will eventually be made public, as bringing forward concerns of fraud imposes reputational risks, enormous time commitments, and no tangible reward. This is merely to say that, for every case of fraud that is identified, there are almost certainly many more that are not. Even restricting the prevalence estimate to publicly identified cases, it is increasingly clear that fraud is frequent enough to require more than a passing consideration. To our knowledge, there is no perfect or excellent solution. We discuss three imperfect ones, from least to most promising.

Do not worry about it. Section 8.10 of the American Psychological Association code of ethics simply states, "psychologists do not fabricate data" (**http://web.archive.org/web/20160803035613/http://www.apa.org/ethics/code/manual-updates.aspx**). That statement clarifies the aspirations of the APA but not how it aims to detect or adjudicate violations of the ethical code. The status quo is to simply not worry about these violations. Although this has the benefits of being socially agreeable (e.g., suggesting that we assume that everyone is honest) and easy to implement, it is a dangerous and untenable approach.

Make data public. The APA code of ethics also states (in Section 8.14) that "after research results are published, psychologists do not withhold the data on which their conclusions are based." This is a good start to preventing fraud. As some have argued (Miguel et al. 2014, Simonsohn 2013, Wicherts 2011), the sharing of data serves the dual function of encouraging researchers to be accurate and honest and providing readers with the tools to detect errors and fraud. Requesting data from another researcher—particularly for the stated justification of suspecting fraud—is socially taxing. Furthermore, although the APA prescribes the sharing of data, there is no enforcement mechanism. We have heard many stories from other researchers who were told that the requested

[24]For information on Chiou's retracted article, see **http://datacolada.org/1**; for information on one of Förster's retracted articles, see **http://datacolada.org/21**. We refer to the other retracted articles in other sections of this review.

data were coming soon (but then never arrive), were impossible to share, had been lost, or were legally impounded. We have personally been denied data explicitly because the authors wished to avoid criticism; the authors wrote bluntly, "no data for you." The APA already says that data should be available upon request, but in practice, they are not (see e.g., Wicherts et al. 2006, 2011). Indeed, psychologists' reluctance to share data has been systematically documented (Wicherts et al. 2006). It seems insincere to equate being a psychologist with being a data sharer (as the APA does in the above quote) and to then fail to require psychologists to post their data.

Make studies auditable. Data-sharing policies would not have been enough to stop Mike LaCour from publishing a (since retracted) article in *Science* based on apparently fraudulent data (LaCour & Green 2014). The sequence that led to the development of the project and the publication of the article illustrates just how hard fraud is to detect and how destructive it can be. LaCour, a graduate student, approached Green, a renowned professor, with an early set of data. Green was skeptical and requested a replication, and LaCour delivered it; at that point, Green, a famously skeptical methodologist, was persuaded to coauthor the article. Although methodological expertise, a replication, the peer review process, and data sharing did not detect the apparent fraud, an audit of the study did. When a pair of graduate students were attempting a replication and needed more details than were included in the published article, they contacted the survey firm that had supposedly collected the data. The firm had never heard of LaCour and did not have the capacity to collect the type of data that LaCour had claimed to have obtained from them. It was not some sophisticated procedure that led to the retraction (none of those were sufficient), but rather something akin to the type of basic fact checking that journalists routinely do (Singal 2015; D. Broockman, J. Kalla, and P. Aronow, unpublished manuscript, **https://web.archive.org/web/https://stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf**).

The US Internal Revenue Service does not audit all taxpayers, but all taxpayers are expected to file taxes as if they will be audited. We think that researchers should similarly behave as if they will be audited. For example, journals could require authors to provide information on exactly when (i.e., specific dates and times), exactly where, and by whom the data were collected. Journals could then do the routine fact checking that newspapers do. As it currently stands, if *The New York Times* writes an article about a publication in *Science*, only the former does any fact checking.

OTHER ATTEMPTS AT REFORM

In this review, we have focused on methodological reforms designed to diagnose or decrease the publication of false positives. However, not all of the methodological reforms proposed in recent years will, or even attempt to, accomplish this. In this section, we briefly discuss two such categories of reform: an emphasis on meta-analysis and a de-emphasis on *p*-values.

Meta-Analyses

The acknowledged fallibility of individual studies has inspired some researchers to advocate for considering aggregations of many studies instead (Braver et al. 2014, Stanley & Spence 2014, Tuk et al. 2015). We can characterize this perspective as advocating for meta-analytic thinking (Cumming 2014). Meta-analytic thinking has its benefits. It allows inferences to be based on larger and potentially more diverse samples, promotes collaboration among scientists, and incentivizes more systematic research programs. Nevertheless, meta-analytic thinking not only fails to solve the problems of *p*-hacking, reporting errors, and fraud, it dramatically exacerbates them. In our

view, meta-analytic thinking would make the false-positives problem worse, not better (for other concerns about meta-analysis, see Sharpe 1997, Ueno et al. 2016).

What is so problematic about meta-analytical thinking? To answer this question, it is useful to distinguish between meta-analyses that aggregate studies across versus within articles. When aggregating studies across articles, the main goal is to compute an overall average effect. In fields like medicine, different research teams will sometimes conduct functionally identical studies of the same effect; in these cases, the average result is a more precise answer to the question everyone is asking (e.g., what is the average effect of Prozac on depression?). In psychology, however, different research teams study different effects; in these cases, the average result is a more precise answer to a question nobody is asking (e.g., what is the average effect of all possible psychological interventions that could ever be attempted on depression?). Indeed, meta-analyses in psychology typically estimate a parameter that does not exist (Simonsohn 2015a).

Leaving the meaningfulness of the question aside, there is the issue of the credibility of the answer. Meta-analysts are further from data collection than are the original researchers. That distance, in combination with the sheer number of studies included in meta-analyses, makes it infeasible for the meta-analyst (and for reviewers and readers of the meta-analysis) to assess the quality of the original data. For example, the meta-analyst cannot assess if an original result was caused by errors of data collection or analysis or if it was a product of a methodological detail that was not divulged by the original authors. Even assessing the quality of the original design is often infeasible given the number of studies involved (e.g., the presence of confounds or demand effects). Furthermore, the harder a meta-analyst works to avoid publication bias by including unpublished work, the more the quality control problem is amplified by the addition of content that has either never been reviewed or been reviewed and rejected. All of these factors are especially consequential given that errors in original studies are not random—they are usually biased in the direction of finding statistical significance—and thus do not cancel out across studies. The end result of a meta-analysis is as strong as the weakest link; if there is *some* garbage in, then there is *only* garbage out.

On the surface, aggregating studies *within* articles would seem to solve some of these problems, but, in fact, it makes some of those problems much more severe. Currently, the expectation is that, even in multistudy articles, an individual study must be judged on its own merits. Thus, if one study is deemed to be problematic, it need not affect what one infers about the others. When we evaluate evidence by relying on the statistical aggregation of all studies in an article, a single problematic study will contaminate the inferences of the remainder. Thus, a multistudy article may hinge on the impact of a single flawed study that carries all other studies through the meta-analytic filter.

If someone is presented with four cups of juice and one cup of poison, they can drink safely by avoiding the poisoned cup. However, if all five cups are poured into a single pitcher, they are going to get sick. Similarly, if one out of five studies is poorly done or obviously *p*-hacked, that study can be easily and safely ignored. However, if it is combined with the other studies into one analysis, it will taint them all (and that taint is likely to go unnoticed). In addition, combining many studies into one analysis can dramatically exacerbate the consequences of *p*-hacking. Consider a researcher who minimally *p*-hacks by simply choosing the best of three uncorrelated dependent variables. If they conduct one study, the false-positive rate increases to just 7.3%. However, if the same researcher were to apply the same behavior across 10 studies and then meta-analyze them, then the false-positive rate increases to a staggering 83% (Vosgerau et al. 2017).[25]

Thus, meta-analysis does not solve the problems the field faces; it exacerbates them.

[25] These calculations assume a directional prediction submitted to a two-sided test and, thus, a 2.5% false-positive rate without *p*-hacking.

P-Value Bashing

When, during the Lyndon Johnson administration,[26] Bakan (1966) wrote an article in opposition to the use of *p*-values in psychology, he raised complaints that will sound familiar to readers today: "the null hypothesis is generally false anyway" (p. 425), "papers in which significance has not been obtained are not submitted" (p. 427), "this wrongness is based on the commonly held belief that the p value is a 'measure' of degree of confidence" (p. 430), "we would be much better off if we were to attempt to estimate the magnitude of the parameters in the population" (p. 436), and, "In terms of a statistical approach which is an alternative, the various methods associated with the theorem of Bayes. . .may be appropriate" (p. 436). Strikingly, Bakan (1966, p. 423) opens this 50-year-old article by saying, "what will be said in this paper is hardly original. It is, in a certain sense, what 'everybody knows.'" We may accurately call these arguments the "old statistics critiques."

These old statistics critiques and the counterarguments put forward by those espousing the continued use of *p*-values have been rehashed by every generation of psychological researchers. Psychology's renaissance did not interfere with this tradition. Cumming (2014) has recently argued for point estimation and confidence intervals to be used over *p*-values, Kruschke (2013) for Bayesian estimation over *p*-values, and Wagenmakers et al. (2011) for Bayes factors over *p*-values.

Despite their repeated republication, the influence of the old statistics critiques on actual practice has not been significant. Like researchers in Bakan's era, researchers in our era rely heavily on *p*-values. Why is this the case?

We think it is because there is actually no compelling reason to abandon the use of *p*-values. It is true that *p*-values are imperfect, but, for the types of questions that most psychologists are interested in answering, they are no more imperfect than confidence intervals, effect sizes, or Bayesian approaches. The biggest problem with *p*-values is that they can be mindlessly relied upon; however, when effect size estimates, confidence intervals, or Bayesian results are mindlessly relied upon, the results are at least as problematic. It is not the statistic that causes the problem, it is the mindlessness. We suspect some readers will strongly disagree with our position (both today and 50 years from today), but they probably will not disagree with the fact that these alternative approaches do not address the fundamental problem that psychology's renaissance has concerned itself with: conducting research in ways that make false positives less likely.

CONCLUSION

Roughly seven years ago, our field realized that the credibility of our discipline hinged on changing the way we collect and analyze data. And the field has responded. Practices that promise to increase the integrity of our discipline—e.g., replications, disclosure, preregistration—are orders of magnitude more common than they were just a short time ago. Although they are not yet common enough, it is clear that the Middle Ages are behind us, and the Enlightenment is just around the corner.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

[26] For our European friends, think Charles de Gaulle.

LITERATURE CITED

Achenbach J. 2011. Diederik Stapel: the lying Dutchman. *The Washington Post Blog*, Nov. 1. **http://web.archive.org/web/20170418235730/https://www.washingtonpost.com/blogs/achenblog/post/diederik-stapel-the-lying-dutchman/2011/11/01/gIQA86XOdM_blog.html**

Alogna VK, Attaya MK, Aucoin P, Bahník Š, Birch S, et al. 2014. Registered replication report: Schooler and Engstler-Schooler (1990). *Perspect. Psychol. Sci.* 9:556–78

Anderson CJ, Bahník Š, Barnett-Cowan M, Bosco FA, Chandler J, et al. 2016. Response to comment on "Estimating the reproducibility of psychological science." *Science* 351:1037

Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJ, et al. 2013. Recommendations for increasing replicability in psychology. *Eur. J. Personal.* 27:108–19

Bakan D. 1966. The test of significance in psychological research. *Psychol. Bull.* 66:423–37

Bakker M, van Dijk A, Wicherts JM. 2012. The rules of the game called psychological science. *Perspect. Psychol. Sci.* 7:543–54

Bakker M, Wicherts JM. 2011. The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods* 43:666–78

Bargh JA, Chen M, Burrows L. 1996. Automaticity of social behavior: direct effects of trait construct and stereotype activation on action. *J. Personal. Soc. Psychol.* 71:230–44

Bem D, Tressoldi PE, Rabeyron T, Duggan M. 2016. Feeling the future: a meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Res.* 4:1188

Bem DJ. 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J. Personal. Soc. Psychol.* 100:407–25

Berle D, Starcevic V. 2007. Inconsistencies between reported test statistics and *p*-values in two psychiatry journals. *Int. J. Methods Psychiatric Res.* 16:202–7

Bones AK. 2012. We knew the future all along. *Perspect. Psychol. Sci.* 7:307–9

Brandt MJ, IJzerman H, Dijksterhuis A, Farach FJ, Geller J, et al. 2014. The replication recipe: What makes for a convincing replication? *J. Exp. Soc. Psychol.* 50:217–24

Braver SL, Thoemmes FJ, Rosenthal R. 2014. Continuously cumulating meta-analysis and replicability. *Perspect. Psychol. Sci.* 9:333–42

Brown NJ, Heathers JA. 2016. The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Soc. Psychol. Personal. Sci.* 8(4):363–69

Carey B. 2011a. Fraud case seen as a red flag for psychology research. *The New York Times*, Nov. 2

Carey B. 2011b. Journal's paper on ESP expected to prompt outrage. *The New York Times*, Jan. 5

Carey B. 2015. Many psychology findings not as strong as claimed, study says. *The New York Times*, Aug. 27. **https://web.archive.org/web/20170714054021/https://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html**

Carney DR, Cuddy AJ, Yap AJ. 2010. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychol. Sci.* 21:1363–68

Carney DR, Cuddy AJ, Yap AJ. 2015. Review and summary of research on the embodied effects of expansive (versus contractive) nonverbal displays. *Psychol. Sci.* 26:657–63

Chase LJ, Chase RB. 1976. A statistical power analysis of applied psychological research. *J. Appl. Psychol.* 61:234–37

Cheung I, Campbell L, LeBel EP. 2016. Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon 2002. *Perspect. Psychol. Sci.* 11:750–64

Cohen J. 1962. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65:145–53

Cole LC. 1957. Biological clock in the unicorn. *Science* 125:874–76

Cumming G. 2014. The new statistics: why and how. *Psychol. Sci.* 25:7–29

Doyen S, Klein O, Pichon CL, Cleeremans A. 2012. Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE* 7:e29081

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* 56:455–63

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, et al. 2016. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* 67:68–82

Egger M, Smith GD, Schneider M, Minder C. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315:629–34

Eich E. 2014. Business not as usual. *Psychol. Sci.* 25:3–6

Etz A, Vandekerckhove J. 2016. A Bayesian perspective on the reproducibility project: psychology. *PLOS ONE* 11:e0149794

Fanelli D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90:891–904

Fiedler K. 2011. Voodoo correlations are everywhere—not only in neuroscience. *Perspect. Psychol. Sci.* 6:163–71

Fiedler K, Kutzner F, Krueger JI. 2012. The long way from α-error control to validity proper: problems with a short-sighted false-positive debate. *Perspect. Psychol. Sci.* 7:661–69

Fiedler K, Schwarz N. 2016. Questionable research practices revisited. *Soc. Psychol. Personal. Sci.* 7:45–52

Francis G. 2012. Too good to be true: publication bias in two prominent studies from experimental psychology. *Psychon. Bull. Rev.* 19(2):151–56

Galak J, LeBoeuf RA, Nelson LD, Simmons JP. 2012. Correcting the past: failures to replicate ψ. *J. Personal. Soc. Psychol.* 103:933–48

García-Berthou E, Alcaraz C. 2004. Incongruence between test statistics and P values in medical papers. *BMC Med. Res. Methodol.* 4:13

Gervais W. 2015. Putting PET-PEESE to the test. *Will Gervais Blog*, June 25. **https://web.archive.org/web/20170326200626/http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1**

Gilbert DT, King G, Pettigrew S, Wilson TD. 2016. Comment on "Estimating the reproducibility of psychological science." *Science* 351:1037

Goldin-Meadow S. 2016. Why preregistration makes me nervous. *Association for Psychological Science Observer*, Sept. **https://web.archive.org/web/20170227180244/http://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous**

Greenwald AG. 1975. Consequences of prejudice against the null hypothesis. *Psychol. Bull.* 82:1–20

Inbar Y. 2016. Association between contextual dependence and replicability in psychology may be spurious. *PNAS* 113(43):E4933–34

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med.* 2:696–701

Ioannidis JPA, Trikalinos TA. 2007. An exploratory test for an excess of significant findings. *Clin. Trials* 4:245–53

John L, Loewenstein GF, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychol. Sci.* 23:524–32

K12 Reader. 2012. The scientific method. In *2nd Grade Reading Comprehension Worksheets*. **http://www.k12reader.com/worksheet/the-scientific-method/view**

Kerr NL. 1998. HARKing: hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* 2:196–217

Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, et al. 2016. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLOS Biol.* 14:e1002456

Klein RA, Ratliff K, Vianello M, Reginald B, Adams J, Bahnik S, et al. 2014. *Investigating variation in replicability: a "Many Labs" replication project*. Open Sci. Found. Proj., Cent. Open Sci./Va. Commonwealth Univ., Charlottesville, VA/Richmond, VA

Koole SL, Lakens D. 2012. Rewarding replications: a sure and simple way to improve psychological science. *Perspect. Psychol. Sci.* 7:608–14

Kruschke JK. 2013. Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.* 142:573–603

Kunda Z. 1990. The case for motivated reasoning. *Psychol. Bull.* 108:480–98

LaCour MJ, Green DP. 2014. When contact changes minds: an experiment on transmission of support for gay equality. *Science* 346:1366–69

Lakatos I. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the Growth of Knowledge*, ed. I Lakatos, A Musgrave, pp. 170–96. Cambridge, UK: Cambridge Univ. Press

Lakens D. 2014. Performing high-powered studies efficiently with sequential analyses. *Eur. J. Soc. Psychol.* 44:701–10

Leamer EE. 1983. Let's take the con out of econometrics. *Am. Econ. Rev.* 73(1):31–43
Levelt WJ, Drenth P, Noort E. 2012. *Flawed science: the fraudulent research practices of social psychologist Diederik Stapel*. Rep., Tilburg Univ./Univ. Amsterdam/Univ. Groningen, Tilburg, Neth./Amsterdam/Groningen, Neth.
Lindsay DS. 2015. Replication in psychological science. *Psychol. Sci.* 26:1827–32
Luce MF, McGill A, Peracchio L. 2012. *Promoting an Environment of Scientific Integrity: Individual and Community Responsibilities*. Oxford, UK: Oxford Univ. Press
Mahoney MJ. 1979. Review paper: psychology of the scientist: an evaluative review. *Soc. Stud. Sci.* 9:349–75
McShane BB, Böckenholt U, Hansen KT. 2016. Adjusting for publication bias in meta-analysis: an evaluation of selection methods and some cautionary notes. *Perspect. Psychol. Sci.* 11:730–49
Miguel E, Camerer CF, Casey K, Cohen J, Esterling K, et al. 2014. Promoting transparency in social science research. *Science* 343:30–31
Moore DA. 2016. Preregister if you want to. *Am. Psychol.* 71:238–39
Morey RD. 2013. The consistency test does not—and cannot—deliver what is advertised: a comment on Francis 2013. *J. Math. Psychol.* 57:180–83
Nosek BA, Lakens DE. 2013. Call for proposals: special issue of *Social Psychology* on "replications of important results in social psychology." *Soc. Psychol.* 44:59–60
Nuijten MB, Hartgerink CH, van Assen MA, Epskamp S, Wicherts JM. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* 48:1205–26
Open Sci. Collab. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* 7:657–60
Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716
Pashler H, Harris CR. 2012. Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* 7:531–36
Pashler H, Wagenmakers EJ. 2012. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7:528–30
Patil P, Peng RD, Leek JT. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* 11:539–44
Phillips CV. 2004. Publication bias in situ. *BMC Med. Res. Methodol.* 4:20
Popper KR. 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books
Ranehill E, Dreber A, Johannesson M, Leiberg S, Sul S, Weber RA. 2015. Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women. *Psychol. Sci.* 26(5):653–56
Rosenthal R. 1979. The "file drawer problem" and tolerance for null results. *Psychol. Bull.* 86:638–41
Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16:225–37
Schimmack U. 2012. The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* 17:551–66
Schmidt S. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* 13:90–100
Sedlmeier P, Gigerenzer G. 1989. Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105:309–16
Serlin RC, Lapsley DK. 1985. Rationality in psychological research: the good-enough principle. *Am. Psychol.* 40:73–83
Serlin RC, Lapsley DK. 1992. Rational appraisal of psychological research and the good-enough principle. In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, ed. G Keren, C Lewis, 199–228. Mahwah, NJ: Lawrence Erlbaum Assoc.
Sharpe D. 1997. Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. *Clin. Psychol. Rev.* 17:881–901
Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22:1359–66
Simmons JP, Nelson LD, Simonsohn U. 2012. A 21 word solution. *Dialogue* 26(2):4–7

Simmons JP, Nelson LD, Simonsohn U. 2017. False-positive citations. *Perspect. Psychol. Sci.* In press
Simmons JP, Simonsohn U. 2017. Power posing: p-curving the evidence. *Psychol. Sci.* 28(5):687–93
Simons DJ. 2014. The value of direct replication. *Perspect. Psychol. Sci.* 9:76–80
Simons DJ, Holcombe AO, Spellman BA. 2014. An introduction to registered replication reports at *Perspectives on Psychological Science*. *Perspect. Psychol. Sci*. 9:552–55
Simonsohn U. 2012. It does not follow: evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspect. Psychol. Sci.* 7:597–99
Simonsohn U. 2013. Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychol. Sci.* 24:1875–88
Simonsohn U. 2014. [13] Posterior-hacking. *Data Colada*, Jan. 13. **https://web.archive.org/web/http://datacolada.org/13**
Simonsohn U. 2015a. [33] The effect size does not exist. *Data Colada*, Feb. 9. **https://web.archive.org/web/http://datacolada.org/33**
Simonsohn U. 2015b. [35] The default Bayesian test is prejudiced against small effects. *Data Colada*, Apr. 9. **https://web.archive.org/web/http://datacolada.org/35**
Simonsohn U. 2015c. Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.* 26:559–69
Simonsohn U. 2016a. [47] Evaluating replications: 40% full ≠ 60% empty. *Data Colada*, March 3. **https://web.archive.org/web/http://datacolada.org/47**
Simonsohn U. 2016b. Each reader decides if a replication counts: reply to Schwarz and Clore 2016. *Psychol. Sci.* 27:1410–12
Simonsohn U. 2017. [59] PET-PEESE is not like homeopathy. *Data Colada*, Apr. 12. **https://web.archive.org/web/http://datacolada.org/59**
Simonsohn U, Nelson LD, Simmons JP. 2014a. *P*-curve: a key to the file drawer. *J. Exp. Psychol. Gen.* 143:534–47
Simonsohn U, Nelson LD, Simmons JP. 2014b. *P*-curve and effect size: correcting for publication bias using only significant results. *Perspect. Psychol. Sci.* 9:666–81
Simonsohn U, Simmons JP, Nelson LD. 2015. Better *p*-curves: making *p*-curve analysis more robust to errors, fraud, and ambitions *p*-hacking, a reply to Ulrich and Miller (2015). *J. Exp. Psychol. Gen.* 144:1146–52
Singal J. 2015. The case of the amazing gay-marriage data: how a graduate student reluctantly uncovered a huge scientific fraud. *New York Magazine*, May 29
Stanley DJ, Spence JR. 2014. Expectations for replications: Are yours realistic? *Perspect. Psychol. Sci.* 9:305–18
Stanley T, Doucouliagos H. 2014. Meta-regression approximations to reduce publication selection bias. *Res. Synth. Methods* 5:60–78
Sterling TD. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* 54(285):30–34
Strack F. 2016. Reflection on the smiling registered replication report. *Perspect. Psychol. Sci.* 11(6):929–30
Strack F, Martin L, Stepper S. 1988. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *J. Personal. Soc. Psychol.* 54:768–77
Stroebe W. 2016. Are most published social psychological findings false? *J. Exp. Soc. Psychol.* 66:134–44
Stroebe W, Strack F. 2014. The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9:59–71
Tuk MA, Zhang K, Sweldens S. 2015. The propagation of self-control: Self-control in one domain simultaneously improves self-control in other domains. *J. Exp. Psychol. Gen.* 144(3):639–54
Tversky A, Kahneman D. 1971. Belief in the law of small numbers. *Psychol. Bull.* 76:105–10
Ueno T, Fastrich GM, Murayama K. 2016. Meta-analysis to integrate effect sizes within an article: possible misuse and Type I error inflation. *Am. Psychol. Assoc.* 145(5):643–54
Ulrich R, Miller J. 2015. P-hacking by post hoc selection with multiple opportunities: detectability by skewness test? Comment on Simonsohn, Nelson, and Simmons (2014). *J. Exp. Psychol. Gen.* 144:1137–45
Valentine JC, Biglan A, Boruch RF, Castro FG, Collins LM, et al. 2011. Replication in prevention science. *Prev. Sci.* 12:103–17
van Aert RC, Wicherts JM, van Assen MA. 2016. Conducting meta-analyses based on *p* values: reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspect. Psychol. Sci.* 11:713–29

Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA. 2016. Contextual sensitivity in scientific reproducibility. *PNAS* 113(23):6454–59

van der Zee T, Anaya J, Brown NJL. 2017. Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab. *PeerJ Preprints* 5:e2748v1

van't Veer AE, Giner-Sorolla R. 2016. Pre-registration in social psychology—a discussion and suggested template. *J. Exp. Soc. Psychol.* 67:2–12

Vazire S. 2015. This is what *p*-hacking looks like. *Sometimes I'm Wrong*, Feb. **https://web.archive.org/web/http://sometimesimwrong.typepad.com/wrong/2015/02/this-is-what-p-hacking-looks-like.html**

Vazire S. 2016. Editorial. *Soc. Psychol. Personal. Sci.* 7:3–7

Verhagen J, Wagenmakers EJ. 2014. A Bayesian test to quantify the success or failure of a replication attempt. *J. Exp. Psychol. Gen.* 143:1457–75

Vosgerau J, Simonsohn U, Nelson LD, Simmons JP. 2017. *Don't do internal meta-analysis: it makes false-positives easier to produce and harder to correct*. Open Sci. Found. Proj., Cent. Open Sci., Charlottesville, VA

Vul E, Harris CR, Winkielman P, Pashler H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4:274–90

Wagenmakers EJ, Beek T, Dijkhoff L, Gronau QF, Acosta A, et al. 2016. Registered replication report: Strack, Martin, & Stepper 1988. *Perspect. Psychol. Sci.* 11(6):917–28

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL. 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem 2011. *J. Personal. Soc. Psychol.* 100(3):426–32

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. 2012. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* 7:632–38

Wicherts JM. 2011. Psychology must learn a lesson from fraud case. *Nature* 480:7

Wicherts JM, Bakker M. 2012. Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence* 40:73–76

Wicherts JM, Bakker M, Molenaar D. 2011. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE* 6:e26828

Wicherts JM, Borsboom D, Kats J, Molenaar D. 2006. The poor availability of psychological research data for reanalysis. *Am. Psychol.* 61:726–28