American Psychological Association

Psychological Bulletin

0033-2909/18/$12.00

2018, Vol. 144, No. 12, 1325–1346
http://dx.doi.org/10.1037/bul0000169

What Meta-Analyses Reveal About the Replicability of Psychological Research

T. D. Stanley
Deakin University

Evan C. Carter
U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland

Hristos Doucouliagos
Deakin University

Can recent failures to replicate psychological research be explained by typical magnitudes of statistical power, bias or heterogeneity? A large survey of 12,065 estimated effect sizes from 200 meta-analyses and nearly 8,000 papers is used to assess these key dimensions of replicability. First, our survey finds that psychological research is, on average, afflicted with low statistical power. The median of median power across these 200 areas of research is about 36%, and only about 8% of studies have adequate power (using Cohen's 80% convention). Second, the median proportion of the observed variation among reported effect sizes attributed to heterogeneity is 74% (I^2). Heterogeneity of this magnitude makes it unlikely that the typical psychological study can be closely replicated when replication is defined as study-level null hypothesis significance testing. Third, the good news is that we find only a small amount of average residual reporting bias, allaying some of the often-expressed concerns about the reach of publication bias and questionable research practices. Nonetheless, the low power and high heterogeneity that our survey finds fully explain recent difficulties to replicate highly regarded psychological studies and reveal challenges for scientific progress in psychology.

Public Significance Statement
A survey of 12,065 estimated effects from nearly 8,000 research papers finds that the average statistical power in psychology is 36% and only 8% of studies have adequate power. Typical heterogeneity is nearly three times larger than reported sampling error variation. Heterogeneity this large easily explains recent highly publicized failures to replicate in psychology. In most cases, we find little evidence that publication bias is a major factor.

Keywords: power, bias, heterogeneity, meta-analysis, replicability

Supplemental materials: http://dx.doi.org/10.1037/bul0000169.supp

Is psychology in crisis? Recently, there have been highly publicized failures to replicate seemingly well-established psychological phenomena—that is, studies designed to be identical do not produce statistically significant results in the same direction as the original work (e.g., Open Science Collaboration, 2015).[1] These failed replications are especially problematic because many regard replication as the hallmark of science (Popper, 1959). The most pessimistic interpretation of these findings is that such high rates of failed replication invalidate psychological science. Understandably then, these findings have received a large amount of attention and many authors have offered explanations for this difficulty in replicating research in psychology (e.g., Fabrigar & Wegener, 2016; Open Science Collaboration, 2015; Pashler & Harris, 2012; Patil, Peng, & Leek, 2016; Schmidt & Oh, 2016). Despite the various opinions on the topic, the frequent practice of defining replication success in terms of null hypothesis significance testing means that three key dimensions—statistical power, selective reporting bias, and between-study heterogeneity—are likely to play key roles. Here, we survey these three aspects of psychological research across nearly 12,000 studies from 200 areas (or subjects) of empirical research to help understand what is reasonable to expect from replication in psychology and what might be done to improve psychological science.

This article was published Online First October 15, 2018.

T. D. Stanley, Deakin Lab for the Meta-Analysis of Research (DeLMAR), School of Business, Deakin University; Evan C. Carter, Human Research and Engineering Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland; Hristos Doucouliagos, Department of Economics, DeLMAR, and Alfred Deakin Institute for Citizenship and Globalisation, Deakin University.

Correspondence concerning this article should be addressed to T. D. Stanley, School of Business, Deakin University, Burwood, VIC 3125, Australia. E-mail: stanley@hendrix.edu

[1] See the websites http://curatescience.org/ and http://psychfiledrawer.org/ for growing lists of replications in psychology.

We calculate statistical power retrospectively using meta-analytic estimates of the *true effect*—a term we use as shorthand to refer to the mean of the distribution of true effects. Specifically, we examine 200 previously published meta-analytic data sets and calculate two simple weighted averages of reported effect sizes plus one bias-corrected estimate to serve as proxies for the relevant mean of the distribution of true effects. As we report below, we find that: (a) only about one in 12 studies is adequately powered (not surprising, given previous work on power in psychology: Cohen, 1962, 1977; Fraley & Vazire, 2014; Maxwell, 2004; Rossi, 1990; Sedlmeier & Gigerenzer, 1989); (b) there is typically only a small amount of selective reporting bias; and (c) the variance among reported effects due to heterogeneity is nearly three times larger than the reported sampling variance. Such substantial heterogeneity implies that attempted replication studies will frequently and correctly produce a markedly different finding from the original study. Combine this issue with chronically low statistical power and some degree of selective reporting bias, and failures to replicate in psychology are inevitable.

Our findings add further weight to the call for researchers in psychology to take statistical power seriously (Fraley & Vazire, 2014; Maxwell, 2004; Pashler & Wagenmakers, 2012; Rossi, 1990; Tressoldi & Giofré, 2015) and to think carefully about the implications of heterogeneity for the planning and interpretations of replications (Klein et al., 2015; McShane & Böckenholt, 2016). Our results highlight that meaningful replication in psychological research will likely only be achieved through carefully planned, multisite, preregistered efforts.

Reproducibility, Replicability, and Generalizability

Obviously, the findings of a specific study can be verified with different levels of rigor and generalizability. We follow others in distinguishing among three related concepts concerning the generalizability and trustworthiness of research findings (Asendorpf et al., 2013; LeBel, Vanpaemel, McCarthy, Earp, & Elson, 2017). A study may be considered reproducible if other researchers are able to produce the exact same results using the same data and statistical analyses. Reproducibility is the reason that many researchers make their data and codes freely available to others. Reproducibility is narrower than replicability, but it helps to identify and remove some errors from our scientific knowledge. Reproducibility is critical if results from experimental science are to be believed.

Replicability means that a previous finding will be obtained in a new random sample "drawn from a multidimensional space that captures the most important facets of the research design" (Asendorpf et al., 2013, p. 5). A successful replication occurs when the differences in results are insubstantial. Critically, replicability requires that the replication study does in fact capture the important facets of the original study's design. A replication is typically defined as an exact replication if it is thought to capture all of these critical facets and as a conceptual replication if these components are only similar but not quite exact. For example, if the latent construct being measured by the dependent variable in both the original and the replication study is the same, but its operationalization is notably different, the subsequent study would be considered a conceptual replication. Historically, researchers in psychology have tended to publish more conceptual replications than exact replications. For example, using a random sample of 342 replications in psychological research published between 1946 and 2012, Makel, Plucker, and Hegarty (2012) found that 81.9% of all replications are conceptual replications.

Generalizability requires further that subsequent findings are independent of unmeasured factors (e.g., age, gender, culture) in the original study. For example, we would not label a finding as generalizable if it is only replicable in studies conducted in English with samples of US college students. If our goal is to gain an understanding of human psychology in general, then any result that only exists under a very narrow set of conditions is likely to be of little practical importance.

Generalizability is critical to the discussion of replicability because contextual sensitivity (i.e., results are influenced by "hidden moderators") can make a replication falsely appear unsuccessful (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). If a replication produces a different finding from the original study because the effect is contextually sensitive, it need not be a "failure," but instead, the effect may not be generalizable to the replication study's context. For example, if a replication study is conducted online rather than in the laboratory as the original study was, this operational choice might produce a substantially different result. However, effects that can only be reproduced in the laboratory or under only very specific and contextual sensitive conditions may ultimately be of little genuine scientific interest.

Statistically, one expects that the variation between an original finding and an exact replication will only be due to unsystematic sampling error. In contrast, if the context of an original finding is not fully captured by the replication attempt, or if the replication attempt is a conceptual replication, then variation between the original finding and the replication might be due to both between-study heterogeneity and random sampling error. Below, we argue that large between-study heterogeneity is one of the main sources of the observed difficulty in replicating psychological studies.

The foregoing discussion of replicability does not provide a specific definition of replication success, although many have been proposed. For example, the Open Science Collaboration (2015) compared the results of their 100 replication studies directly to the results of the original studies using three quantitative definitions of success. A success was claimed when (a) the replication results matched the original results in both effect direction and statistical significance (using the conventional $\alpha = .05$); (b) the effect size estimate provided by the original study was within the 95% confidence interval of the estimate from the replication study; or (c) a meta-analytic estimate based on both the original and replication results was distinguishable from zero. Other researchers have suggested further ways of assessing replications (e.g., Braver, Thoemmes, & Rosenthal, 2014; Patil et al., 2016).

Although our analysis does not depend on any specific definition of successful replication, following the Open Science Collaboration (2015), we believe that replication success must somehow involve the sign and significance of the reported effect size. We prefer to view replication as related to the sign and practical significance of the reported effect size, rather than its statistical significance. Below, we will discuss successful replication from both perspectives. However, from the reactions to the Open Science Collaboration (2015), replication success is most often viewed as the duplication the original effect's direction and statistical significance. This view of replication success is found in the popular press (Patil & Leek, 2015), *Science* (Bohannon, 2015),

Nature (Baker, 2015), and in many subsequent scientific articles (e.g., Dreber et al., 2015; Lindsay, 2015; van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016).

For any definition of replication success that involves both direction and significance, there are three governing factors for a successful replication: statistical power, bias, and between-study heterogeneity.[2] In the following sections, we describe how each of these relates to the replicability of a finding. We then analyze a set of 200 previously published meta-analyses to provide estimates of statistical power, bias, and heterogeneity, and discuss what these estimates imply for what one should expect when conducting replication in psychology.

Statistical Power

Statistical power is the probability of finding a statistically significant result if the effect in question is truly nonzero (i.e., a correct rejection of the null hypothesis). A study is adequately powered if it has a high probability of finding an effect when one exists, and since Cohen (1965), adequate power has been widely accepted to be 80%.[3] Psychological professional organizations and journals have formally recognized the importance of statistical power. For example, the APA Publication Manual states, "When applying inferential statistics, take seriously the statistical power considerations associated with your tests of hypotheses. . . . [Y]ou should routinely provide evidence that your study has sufficient power to detect effects of substantive interest (e.g., see Cohen, 1988)" (American Psychological Association, 2010, p. 30). According to the Psychonomic Society: "Studies with low statistical power produce inherently ambiguous results because they *often fail to replicate*. Thus it is highly desirable to have ample statistical power" (Psychonomic Society, 2012, p. 1, emphasis added). Moreover, the past 50 years have seen many surveys and calls for greater use of prospective power calculations in psychology—that is, planning research so as to ensure adequate power to detect the effect of interest (e.g., Cohen, 1962, 1977; Fraley & Vazire, 2014; Maxwell, 2004; Pashler & Wagenmakers, 2012; Rossi, 1990; Sedlmeier & Gigerenzer, 1989; Tressoldi & Giofré, 2015). In spite of such statements and frequent admonitions to increase power, prospective power calculations remain quite rare (Tressoldi & Giofré, 2015).

When successful replication is seen as replication of results that are statistically significant and in the same direction as the original, low power will frequently cause replication failures. First, if a replication attempt itself has low power, then by definition it will not be likely to succeed because it has a low probability of reaching statistical significance. Second, original studies with insufficient power will tend to be overestimated to obtain statistical significance (Open Science Collaboration, 2015). As a result, planned replications that use prospective power calculations (based on inflated effect size estimates) are likely to underestimate the required sample size and thereby be insufficiently powered. That is, low power begets low power. Third, if the original study has low power, the poststudy odds of a statistically significant finding reflecting a true effect can be quite low (Ioannidis, 2005). Bayes formula demonstrates that if the original study had low power, then a statistically significant finding will not produce a high probability that there is actually a genuine nonzero effect (Ioannidis, 2005). In this case, a replication should "fail."

Statistical power is determined by sample size, desired significance level, α, and the magnitude of the true effect investigated. The first two quantities are widely known, whereas the magnitude of the true effect must be estimated. This raises the obvious question: How can researchers know the effect size when research is conducted for the very purpose of estimating this effect size? One option is to calculate post hoc power using the reported effect size(s)—that is, using the result of a study's test of an effect to calculate the power of that test. Critically, post hoc calculations are circular and tell us little beyond these studies' reported *p*-values (Fraley & Vazire, 2014; Hoenig & Heisey, 2001; Yuan & Maxwell, 2005). This post hoc circularity is especially pernicious if statistically significant estimates are preferentially reported (i.e., selective reporting bias, discussed below). "Small-*N* studies that actually produce significant results tend to report larger effect sizes than comparable large-*N* studies, thereby biasing their observed (post hoc) power estimates upwards" (Fraley & Vazire, 2014, p. 6, parentheses added).

A better option is to calculate hypothetical power on the basis of arbitrarily defined, but widely used, small, medium or large effect sizes. Such hypothetical power has been the preferred approach used in several previous surveys of psychology, which have shown that the typical power to detect a medium effect in psychological research is inadequate; see Maxwell (2004, p. 148) and his citations to past power surveys. For example, two classic power surveys found that the average power to detect a correlation of 0.2 to be quite low: 14% (Cohen, 1962) or 17% (Sedlmeier & Gigerenzer, 1989), but a more recent and encouraging survey of social psychology and personality journals finds that the power to detect a correlation of 0.2 has at least doubled, though it remains inadequate and typically less than 50% (Fraley & Vazire, 2014).

A third and, we think, highly useful option is to calculate power retrospectively using an estimate of the effect calculated from a meta-analysis. This has been done previously for at least two different fields. Button et al. (2013) reviewed 730 studies from 49 meta-analyses in neuroscience and found that the average retrospective power was 21%, and Ioannidis, Stanley, and Doucouliagos (2017) found that, typically, only about 6% of economic studies have adequate power. Our survey calculates exactly this kind of retrospective power, because doing so has the advantage of using a meta-analysis and thus the entire relevant research record to assess power. The potential vicious circle of calculating power referred to above is further broken when the chosen meta-analysis

[2] The magnitude of the mean of the true effect distribution may, of course, be considered a fourth factor. Very large effects, like the phase of the moon on a given day, are easy to replicate. Because the size of the underlying psychological phenomenon is entirely beyond the control of psychological research, we do not focus on this dimension in our survey. However, our survey and others show that the typical effects of interest to psychology are practically small, $0.2 \leq \text{SMD} \leq 0.5$, from the point of view of Cohen's guidelines.

[3] At this conventional 80% level, the likelihood of a Type II error (or a "false negative") is four times the conventional .05 probability of a Type I error (or a "false positive"). To some, a 20% Type II error is still too high (Schmidt & Hunter, 2015). For a given application, statisticians have long realized that researchers should adjust their tests and thereby the Type I and Type II errors to account for the relative costs of these two errors—see, for example, Ioannidis, Hozo, and Djulbegovic (2013). However, the information needed to do so properly is often beyond the researchers' knowledge.

methods are resilient to selective reporting bias, which is the topic to which we now turn.

Selective Reporting Bias

The second research dimension that we survey is bias. Here, we use the term *selective reporting bias* to refer collectively to situations in which the significance and magnitude of a study's results have been exaggerated by choices in data collection, analysis, or reporting. Thus, our use of *selective reporting bias* encompasses more specific terms such as the file drawer problem, publication bias, reporting bias, *p*-hacking, and questionable research practices (Gelman & Carlin, 2014; John, Loewenstein, & Prelec, 2012; Scargle, 2000; Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014; Stanley, 2008; Stanley & Doucouliagos, 2012, 2014; Wicherts et al., 2016). Here, we are only concerned about the aggregate effects that these various research practices might have on the research record, rather than the details of their specific pathways. Regardless of whether statistically nonsignificant findings are suppressed (traditionally called "publication bias" or the "file drawer problem"), only some selected outcome measures or comparisons are reported or whether any of a number of other questionable research practices are employed, the net effects on the research record are largely the same—an exaggeration of the size and significance of reported effects.

These biases are distinct from outright scientific fraud in that it is almost certainly motivated by researchers' desires to "go where the data lead," or by reviewers' and editors' motivations to use limited journal space for findings that move the field forward. However, even well-meaning motivations can undermine the validity of research findings, potentially producing convincing evidence of a psychological effect that does not exist. Selective reporting biases can cause problems for replicability because replications may be doomed from the start if the original reported finding is spurious or grossly inflated. For this reason, understanding the degree of bias in psychology is important when assessing the credibility of research or when planning replications.

Selective reporting bias or publication bias, as it is usually called, is said to occur if the dissemination of findings depends on the specifics of those findings. For example, findings with statistically significant *p* values or theory-consistent findings are more likely to be published, reported and promoted than other findings. As a result, any review of a literature (including meta-analysis) will tend to overestimate the evidence for an effect because such positive findings will be overrepresented in the observed sample of reported findings.

It seems quite clear that selective reporting or publication bias exists in psychology, though it is difficult to estimate its true prevalence. For example, several reviews have found that the number of statistically significant results reported in psychology is larger than what should be expected given the level of statistical power (e.g., Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995; Fanelli, 2010). Moreover, Bakker, van Dijk, and Wicherts (2012) reviewed 13 meta-analyses in psychology and identify evidence consistent with publication bias in seven. Kühberger, Fritz, and Scherndl (2014) investigate a random sample of 500 studies published in 2007 and found several statistical indicators of publication bias, including a persistent correlation between sample size and effect size. Quite recently, Fanelli, Costas, and Ioannidis (2017) corroborated the prevalence of a negative correlation between sample size and effect size among 430 meta-analyses from psychology and psychiatry.

An inverse correlation between the magnitude of the effect size and sample size would be expected when there is selective reporting for statistical significance. If there is a tendency for some researchers to selectively report statistically significant findings, then greater efforts will be required by those researchers who have only small samples to work with. Because small samples produce larger standard errors, correspondingly larger effects must be found to overcome these large standard errors and obtain statistical significance. With the benefit of larger samples, it is likely that the first small effect found will automatically be statistically significant. When they are not, researchers who wish to report statistically significant findings will require much less manipulation because even small effects (or biases) will be statistically significant when there are large samples. Thus, this often-observed, inverse correlation between sample size and reported effect size is an implication of and is, therefore, consistent with the tendency by some researchers to report statistically significant findings selectively.

In contrast, some have suggested that small-sample studies are somehow better, more able to find big effects. But is it plausible to believe that small-sample psychological studies are typically conducted with more care or at a higher level of quality than large-sample studies? We think not. First, we know that small-sample studies are less able to distinguish effects from background by the very definition of statistical power. Second, it is unlikely that researchers with small samples have first conducted prospective power calculations. Past surveys have found that very few studies (3–5%) choose their sample sizes on the basis of power calculations (Tressoldi & Giofré, 2015). Yet such prospective power calculations with an associated 80% power level are required by the APA manual and have been accepted as critical in the field for decades. Third, even if these small-sample researchers believe that there are large effects to be found in their area of research, they know that whatever they find (large or small) will be unreliable and "buzzing with confusion" (Maxwell, 2004, p.161). Thus, by widely known standards of psychological research, those who use small samples know that they are conducting lower-quality, less-rigorous research. Fourth, in contrast, large labs and research programs tend to conduct larger studies, *ceteris paribus*, and they also have more resources to better design their instruments and more carefully execute their protocols. Consider, for example, large replications efforts such as the Open Science Collaboration (2015) and Hagger et al. (2016) where the care and execution of experiments are demonstrably of higher quality and their sample sizes are larger. However, nothing in this article assumes that larger studies are in any way better than smaller studies, other than their demonstrably higher power. Nor, does our concern that small-sample studies tend to exhibit larger bias depend on small-sample studies being somehow of a lower quality (aside from statistical power). Nonetheless, if some researchers have a preference for statistically significant results, this alone will cause the inverse correlation between sample size and reported effect size that has often been observed. Direct evidence of publication bias has also been found in psychology. When data from an intended research protocol are available, Franco, Malhotra, and Simonovits (2016) find that published effects have a median *p* value of .02,

compared to the median unreported p value of 0.35, suggesting that statistically significant findings are selectively reported.

Recently, much attention has been given to selective reporting through the use of undisclosed, flexible approaches to data collection and analysis, often called p-hacking or questionable research practices (John et al., 2012; Simmons et al., 2011; Wicherts et al., 2016). Examples of these behaviors include analyzing data as they are collected and stopping data collection once a desired result has been achieved, deciding whether to exclude outliers, investigating experimental conditions or moderators on the basis of the results, and reporting one's exploratory findings as if they were the result of confirmatory research (see Wicherts et al., 2016 for a comprehensive list). Like publication bias, it is extremely difficult to determine the prevalence of such behaviors, but several findings are of note. For example, John et al. (2012) surveyed 2,000 researchers in psychology using a questionnaire-based approach designed to correct for social desirability and found that the use of questionable research practices may represent the norm (but see Fiedler & Schwarz, 2016). In addition, LeBel et al. (2013) created an online database, PsychDisclosure.org, which allows authors to disclose whether their published articles include all the methodological details that went into the work. They found that of the authors who participated, 11.2% had not fully reported all excluded observations, 11.8% had not reported all experimental conditions, 45.3% had not reported all measures that had been collected, and 88.8% had not reported their data collection strategy. Franco et al. (2016) compared recorded research intentions to the associated published results and found that about 70% of studies did not disclose every outcome measured and 40% did not disclose every experimental condition tested. Moreover, there is evidence indicating that researchers rely on problematic intuitions about data collection that can further lead to practices that inflate their findings through bias (Erica, Sprenger, Thomas, & Dougherty, 2014; Bakker, Hartgerink, Wicherts, & van der Maas, 2016). For the purposes of our survey, the source or exact methods of selective reporting for statistical significance is immaterial. We merely wish to document any evidence of an overall tendency to exaggerate psychological effects should it exist.

Heterogeneity

The third governing factor for a successful replication is low heterogeneity. As mentioned above, *heterogeneity* refers to variance in the reported findings that results from there being no single true effect size. Rather, there is a distribution of true effects. To compare heterogeneity from one area (or subject) of research to another and from one measure of empirical effect size to another, systematic reviewers often compute I^2 (Higgins & Thompson, 2002, pp.1546–7). I^2 is the proportion (or percentage) of observed variation among reported effect sizes that cannot be explained by the calculated standard errors associated with these reported effect sizes. It is a relative measure of the variance among reported effects that is due to differences between, for example, studies' experimental methods, measures, population, cohorts and statistical methods, relative to the total variance found among the reported effects.

For most researchers, I^2 provides an easy to understand, descriptive summary, much like R^2 in regression analysis. However, because I^2 is a relative measure of heterogeneity, its magnitude can be misleading. If I^2 is high (e.g., 0.9 or 90%) but all studies have large samples and high power, heterogeneity in terms of effect sizes might still be low with little practical consequence. However, even a small I^2 can have considerable practical consequence for topics of psychological research that are dominated by small samples and low power, which has often been found to be typical in psychology (e.g., Button et al., 2013; Cohen, 1962, 1977; Fraley & Vazire, 2014; Maxwell, 2004; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). Because we collect information on the random sampling variance on all of these 12,065 estimated effect sizes, we can also calculate heterogeneity in terms of standardized mean differences or correlations to assess the practical consequences of the heterogeneity that our survey finds.

Importantly, considerable heterogeneity has been found even in carefully conducted exact replications in psychology—those designed to minimize differences between the original study design and the replication. For example, two massive efforts using preregistered protocols, in which different teams of researchers ran the same study as closely as possible, uncovered statistically significant amounts of heterogeneity—I^2 = 36% (Hagger et al., 2016) and 45% (Eerland et al., 2016). Furthermore, Klein et al. (2015) reported a large-scale effort to replicate 15 findings across 36 different sites that intentionally differed in a variety of characteristics (e.g., studies completed online or in the laboratory, samples from the United States or elsewhere). Klein et al. (2015) found significant amounts of heterogeneity in eight of the 16 effects that were replicated (I^2 = 23% to 91%); however, a comparison of the intraclass correlation among effects to the intraclass correlation among sites found that very little of this heterogeneity in effect sizes was accounted for by differences in the sites, suggesting that heterogeneity was genuinely a characteristic of the phenomena being studied. Of course, heterogeneity would be expected to be higher when replications are conceptual (i.e., those making little attempt to duplicate all of the relevant design characteristics of the original study) or when "hidden moderators" influence research outcomes.

In the face of large heterogeneity, replicability will be severely compromised. For example, suppose that the true mean correlation is 0.2, which is roughly consistent with what past surveys in psychology have found (Richard, Bond, & Stokes-Zoota, 2003), and that the standard deviation in the distribution of true effects due to heterogeneity is approximately the same size. Then, the probability that a replication will correctly return a medium-to-large true effect ($r > 0.3$) or a negative or negligible true effect ($r < .1$) is 62%. In other words, if an original study measures an effect that is influenced by notable heterogeneity, a replication attempt of that study can easily appear to have failed when, in fact, both accurately measure different true effects. With high heterogeneity, the psychological effect in question is itself too variable or context sensitive, regardless of bias or sample size, to be successively replicated frequently.

The Present Meta-Analytic Survey

In this study, we survey the statistical power, heterogeneity and residual selection bias of psychological research through meta-analyses published in recent issues of the *Psychological Bulletin.* These meta-analyses define the sampling frame from which we collect over 12,000 effect sizes and their standard errors in the 200

most recently published *Psychological Bulletin* meta-analyses where these effect sizes and their standard errors were reported. To calculate power, retrospectively, from a systematic review, we must first have an estimate of the mean of the true effect distribution. We use three different estimators for each of these 200 areas of research. Two are simple weighted averages, the unrestricted weighted least squares (WLS) and (WAAP) the weighted average of the adequately powered (Ioannidis et al., 2017; Stanley & Doucouliagos, 2015; Stanley et al., 2017). WLS gives the same point estimate as conventional fixed-effect meta-analysis; however, WLS assumes a random-effects model and provides standard errors and confidence intervals that reflect the heterogeneity that is found in the area of research under examination. Both WLS and WAAP give conservative estimates in the sense that they give the benefit of doubt to the credibility of psychological research by overestimating the magnitude of the mean of the true effect distribution (and hence overestimating statistical power) if there is some selective reporting bias. These weighted averages are unbiased otherwise. Our third proxy, PET-PEESE, for the mean of the true effect distribution is more "activist," because it makes an effort to correct for selective reporting bias and might, therefore, underestimate true effect in some cases (Stanley, 2017; Stanley & Doucouliagos, 2014; Stanley et al., 2017). See the Meta-Analytic Estimates of True Effects section for a detailed discussion of these estimators, their statistical properties, and their role in this survey.

Next, we calculate median power and the proportion of studies with adequate power (≥80%) for each of these 200 meta-analyses using these three proxies for the true effect. To gauge residual selective reporting bias, we then compare the median absolute value of the mean (i.e., the unweighted simple average) found in each of these 200 meta-analyses to the median absolute value of these same three proxies for the true effect.[4] From past surveys of hypothetical power, one would expect to find low statistical power to be typical in psychological research, regardless of the estimator employed, and our survey does not disappoint. Last, we survey heterogeneity across these 200 meta-analyses in two ways: relatively, using I^2 and also in units of SMDs. In sum, thousands of statistics were calculated in 600 meta-regression analyses and over a dozen meta-meta-analyses. Next, we turn to a detailed account of the calculation of all of these statistics and the collecting of these 200 meta-analyses and their associated 12,065 effect sizes.

Method

Data Collection

To assess power, bias, and heterogeneity in psychological research, we require both effect sizes and their standard errors over a wide range of psychological studies. Because only past meta-analyses are likely to have collected the requisite information to calculate power, and because *Psychological Bulletin* is the premier journal for meta-analysis in psychology, we use it to define our sampling frame. We took a convenience sample of the 200 most recently published meta-analyses (as of June 1, 2016) for which we could acquire the necessary statistics. Thus, our unit of analysis is a meta-analysis and, with it, all of the dozens of individual estimates that lie therein. To the extent that research reported in *Psychological Bulletin* is representative of empirical psychology, our findings will also be more broadly representative of psychological research. We intentionally selected *Psychological Bulletin* as our sampling frame, in part, to reflect what many would regard as the best, most influential, research in the field. However, we can only be sure that our survey is representative of the research that the editors and reviewers of the *Psychological Bulletin* consider to be of top quality and relevant to psychology. Because our survey is descriptive, we make *no* inferences to any population—that is, our survey assesses several important dimensions of psychological research as reported in *Psychological Bulletin*, but our results should not be taken as representing all psychological research. We focused on the most recent issues of the *Psychological Bulletin,* ending June 1, 2016, because the topics covered there are more likely to reflect contemporary psychological research interests. As we discuss in greater detail below, the findings from our survey are also consistent with past surveys of psychological research; thus, we have some reason to believe that they might be more generally representative.

Before our survey of 200 meta-analyses commenced, we posted a pre-analysis plan at Open Science Framework on June 1, 2016 (https://osf.io/2vfyj/). In December of 2016, we filed an amended pre-analysis plan to increase the number of meta-analyses surveyed from the originally planned 100 to 200, while keeping everything else the same. We made this adjustment to maintain a broad coverage across areas of research even though the typical article published in *Psychological Bulletin* contains more than three separate meta-analyses. The 200 meta-analyses are reported in 61 *Psychological Bulletin* articles; see details below. In all cases, we use the highest level of aggregation of meta-analyses reported in these articles. That is, we use the first level of aggregation that authors report. Some authors refer to this as the "overall" meta-analysis at the study level. Others refer to this as "primary meta-analysis." This is the level of aggregation before the sample of studies is broken up to conduct moderator or subgroup analysis. Some authors report a single overall meta-analysis; others report several such overall meta-analyses. When no overall, single meta-analysis is reported, we did not combine separate meta-analyses together. In all cases, we follow the judgment of the authors of these 200 meta-analyses reported in 61 *Psychological Bulletin* articles about which estimated effect sizes are appropriate to meta-analyze—that is, we made no decisions about which studies should be included in the 200 overall meta-analyses. Furthermore, we did not subdivide any of these 200 meta-analyses. In summary, we neither combined nor subdivided any meta-analysis reported in these 61 articles.[5] Details of the level of aggregation and descriptions of the 200 meta-analyses included in our survey can be found in Supplemental Table A1 in the online supplemental

[4] We use the absolute value here to account for psychological effects that are expected to be negative (i.e., inverse or negatively correlated). When calculating I^2 and estimates of effect sizes for each of these 200 meta-analyses, no absolute values are computed; effect sizes are used as reported, positive or negative.

[5] There may be some differences in what these *Psychological Bulletin* authors choose to be the overall or primary meta-analyses across studies. We follow the first, or most aggregated, level reported. Rather than imposing our own notion of the appropriate level of aggregation, we rely on the authors who are experts in the specific research domain. Below we report how there are no important differences in power or heterogeneity between those *Psychological Bulletin* papers that combine all effect sizes into one overall meta-analysis and those that do not.

material. Our survey satisfies the Meta-Analysis Reporting Standards.

Search strategies. As noted above, only meta-analyses published in *Psychological Bulletin* were considered. We manually searched all issues of *Psychological Bulletin* from 2011 to 2016, as detailed in Table B1. We began with the June 2016 issue of *Psychological Bulletin* and worked backward until we obtained the required statistics from 200 meta-analyses. When necessary, we contacted authors for this information (10 of 28 provided it). The data collection ended when 200 meta-analysis data sets with the needed information were gathered from 61 articles published between 2011 to 2016 in the *Psychological Bulletin.* These 61 papers are marked in the references by an asterisk.

Inclusion and exclusion criteria. All the studies are in English. Studies were eligible for inclusion in our study if they reported the estimated effect sizes and their standard errors, either in the article or in a supplement. We excluded four categories of meta-analyses. First, we exclude any meta-analysis that did not report both effect sizes and their standard errors. Consequently, we also exclude systematic reviews, as opposed to meta-analyses, because they typically do not fully report all effect sizes and their standard errors. Second, to avoid double-counting, we exclude any meta-analysis from a reply or comment to a published meta-analysis that was already part of our database. Third, we exclude a couple of meta-analyses that used partial eta squared as the effect size. Partial eta squared cannot be used in conventional power calculations, a central outcome of our survey, nor can they be converted and compared to other types of effect sizes (correlations and standardized mean differences) used in all of the other 200 meta-analyses. Fourth, we excluded the two meta-analyses with fewer than five observations because all statistical calculations are unreliable when based on less than a handful of measures. The Appendix presents the distribution of studies across the 6-year period included in our sample. During this period, there were 115 meta-analysis articles published in *Psychological Bulletin*. Hence, we include meta-analyses from 53% of these articles. It is also worth noting that our survey therefore contains over 80% of the articles published in 2015 and 2016, as a larger proportion of more recent publications report the necessary data.

Coding procedures. The data used in the survey, effect sizes, and their standard errors were reported by authors of published meta-analyses in *Psychological Bulletin*. These were extracted by all three authors, all of whom are experienced in meta-analysis. There were no disputes in coding as we used the data supplied by authors themselves. No issues of study quality arose, as all the meta-analyses are published in *Psychological Bulletin* and have already undergone a rigorous review process.

Fifty-four percent of the effect sizes in our sample are reported as correlations. All of our calculations of power, heterogeneity and bias are made for each of the 200 meta-analyses in the originally reported measure of effect size. Hence, all summary calculations are independent of the type of effect size or of the transformation of one to the other. However, for descriptive clarity and simplicity, meta-averages in terms of correlations were converted to standardized mean differences (Cohen's *d*) to be comparable to the others. A minority of meta-analyses report Hedges' *g* correction of standardized mean differences (SMD) for degrees of freedom. We make no conversion between Cohen's *d* and Hedges' *g*, because it does not make a practical difference to any of our results. The signs of effect sizes were left as the original meta-analysts reported them.

To highlight the types of meta-analytic studies included in our survey and to discuss some of the issues that naturally arise, we discuss more fully two of these 61 *Psychological Bulletin* articles: North and Fiske (2015), "Modern attitudes toward older adults in the aging world: A cross-cultural meta-analysis" and Williams and Tiedens (2016), "The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior." Williams and Tiedens (2016) reported six separate meta-analyses of "backlash" to women's dominance behavior, and all six are included in our survey of 200 meta-analyses. Three of these meta-analyses concern the simple effect of gender from dominance behavior on likability, competence, and "downstream" outcomes such as hireability. The other three meta-analyses involve the interaction of gender and dominance on these same three types of outcome measures. Williams and Tiedens (2016) include a mix of experimental and correlational studies but most (85%) are experimental. Because all six separate meta-analyses are reported and discussed by Williams and Tiedens (2016), all six are included in our survey.

In contrast, North and Fiske (2015) report only one meta-analysis of observational studies about East–West attitudes toward older adults. North and Fiske (2015) combined quantitative outcomes involving attitude measures on: ageism, the aging process, perceived wisdom, warmth and so forth, as well as behavior-based measures from contact with older adults (p. 999). Clearly, North and Fiske (2015) thought these different outcomes measures to be sufficiently similar or homogeneous to be compared and treated as the same phenomenon. However, like most meta-analytic studies in psychology, North and Fiske (2015) also conduct a moderator analysis that investigated the effect of different outcomes measures among others.[6]

As mentioned, it is important to note that we do not choose the level of aggregation of the meta-analyses that are included on our survey. Rather, in all cases, we follow the professional judgment of experts in these specific areas of psychological research. In the example of gender and dominance research, Williams and Tiedens do not judge that the effects on likability, competence, and downstream outcomes to be sufficiently homogeneous to be analyzed together. It seems sensible that any measure of likability, for example, reflects the same phenomenon as any other measure of likability when evaluating the effects of gender and dominance. Regardless, the judgment about where to draw the line is better made by those experts who have read, coded, summarized and analyzed the entire relevant research literature(s)—Williams and Tiedens in this case. Similarly, Williams and Tiedens (2016) consider interaction effects to be too different from the simple effects and from each other to be aggregated. Because these are different psychological phenomenon containing nonoverlapping

[6] Separate meta-analyses cannot be generally constructed for all reported moderator variables in these meta-analysis studies; thus, we do not include further subdivided moderator meta-analyses. Often, meta-regression is used for moderator analysis, and only summary results are reported. Even when separate moderator subsets are investigated, it is often the case that insufficient information is reported for us to make these separations. Furthermore, moderator analysis is conducted with the expressed intent to significantly reduce or entirely remove heterogeneity.

effect sizes, we include these interaction meta-analyses as well. This issue is potentially important because the level of aggregation of the meta-analyses may have implications about how much heterogeneity is found. The broader, more encompassing the level of aggregation, the higher one would expect heterogeneity to be.

Our sample of 200 meta-analyses comes from 61 *Psychological Bulletin* articles. This raises the issue of dependence among multiple observations from the same study and/or from multiple authors across studies. However, this dependence does not affect our survey. Sample overlap affects only the standard errors and confidence intervals and causes no biases. Our survey makes no claim about the probability of our reported statistics; therefore, the variance of the sampling distribution is irrelevant to our reported descriptive statistics. Any dependence of meta-analyses within articles or estimate overlaps (which we do not have) or research reports overlaps across meta-analyses within a meta-analysis paper do not affect the validity of our summary *descriptive* statistics.

Meta-Analytic Estimates of True Effect

To be both conservative and robust, we use three meta-estimators of true effect because the meta-estimate that one chooses might, in some individual cases, makes a practically notable difference in how a given area of research is characterized. We investigate three reasonable alternative proxies for true effect drawn from all the reported results in a given meta-analysis to be sure that our survey results are, in fact, robust to the chosen meta-method of estimating effect and also to potential selective reporting bias.

To be clear, there is no perfect estimate of the true effect size (or the mean of the distribution of true effect sizes) when some authors, reviewers, or editors preferentially select statistically significant effects (Ioannidis et al., 2017; Stanley, 2017; Stanley, Doucouliagos, & Ioannidis, 2017). With selective reporting bias (viz., publication bias, the file drawer problem, small-sample bias and *p*-hacking), all meta-estimates are biased because the data from which they are calculated are themselves biased to an unknowable degree. However, a series of statistical simulation studies have documented how some estimators are more biased than others when there is selective reporting bias (Moreno et al., 2009; Stanley, 2008, 2017; Stanley & Doucouliagos, 2014, 2015; Stanley et al., 2017).

Conventional meta-analysis typically estimates true effects using either fixed-effect (FE) or a random-effects (RE) weighted average, or both (Cooper & Hedges, 1994; Hedges & Olkin, 1985; Stanley & Doucouliagos, 2015). The FE weighted average employs optimal weights that are the same as those used by a recently proposed unrestricted weighted least squares (WLS) weighted average (Stanley & Doucouliagos, 2015). We prefer the unrestricted WLS version of the conventional fixed-effect meta-analysis for inferential purposes because the unrestricted WLS weighted average automatically accounts for heterogeneity when calculating confidence intervals or significance tests (Stanley & Doucouliagos, 2015). This WLS estimator is also consistently less biased than RE when there is selective reporting bias. The point estimates of WLS and FE must always be exactly the same; thus, using WLS is exactly equivalent to using FE in our survey.[7]

We have chosen not to use the RE weighted average to assess power because RE is widely known to be more biased than FE and thereby WLS when there is selective reporting bias (Henmi & Copas, 2010; Poole & Greenland, 1999; Stanley, 2017; Stanley & Doucouliagos, 2014; Stanley & Doucouliagos, 2015; Stanley et al., 2017; Sutton, Song, Gilbody, & Abrams, 2000). WLS and FE give less weight than RE to small-sample studies, where selective reporting is likely to be the most severe. In the aggregate, giving more weight to the largest studies and less weight to small studies will reduce selective reporting bias if it is present and is statistically sensible even when it is not. Besides, WLS estimates remain practically as good as conventional RE meta-analysis when there is no selective reporting for statistical significance (Stanley & Doucouliagos, 2015; Stanley et al., 2017). Like RE, WLS assumes a random-effects model and is interpreted in the same way as RE.

Our second estimator exploits the importance of statistical power by overweighting the most precise, largest studies. The WAAP uses the same formulas as does WLS but applies them only on those estimates found to be adequately powered relative to WLS as the proxy for true effect. WAAP is more resilient to selective reporting biases because adequately powered studies are more reliable and require fewer questionable research practices to achieve statistical significance. Simulations show that WAAP is as good as random-effects when there are no selective reporting biases and is superior to RE when there is selective reporting for statistical significance (Stanley et al., 2017). When half of the reported experimental results have been selected for their statistical significance, WAAP consistently reduces bias, on average, by 50% (Stanley et al., 2017). The weakness of WAAP is that it cannot be computed if there are no studies with adequate power, a condition found in 35% of the 200 areas of psychological research that comprise our survey. Thus, Stanley et al. (2017) proposed using WLS when WAAP cannot be computed, giving a WAAP-WLS weighted average. In the below assessments of power and bias, WAAP-WLS is the second approach that we employ. WAAP-WLS has the added value of forcing meta-analysts to seriously consider and report the statistical power found in their area of research.

WLS and WAAP-WLS passively moderate selective reporting bias. In contrast, simple meta-regression models have been shown to reduce selective reporting bias more aggressively when it is present (Moreno et al., 2009; Stanley, 2005, 2008, 2017; Stanley & Doucouliagos, 2014, 2015, 2017; Stanley et al., 2017; Stanley, 2017). The precision-effect test-precision effect estimate with standard error (PET-PEESE) is a conditional estimate of average effect from simple WLS meta-regressions of each estimated effect size on its standard error (PET) or, alternatively, on its variance (PEESE)—Stanley and Doucouliagos (2014).[8] When only statistically significant positive results are reported, selective reporting bias is known to be equal to the reported estimate's standard error times the inverse Mills' ratio (Stanley & Doucouliagos, 2014, p. 61). The inverse Mills' ratio is a complex function of the true effect and the standard error, which

[7] To see how to calculate WLS's standard error or confidence interval, consult Stanley and Doucouliagos (2015) and Stanley et al. (2017). However, any basic regression routine will automatically calculate our unrestricted WLS weighted average when one uses the standardized effect size (effect size divided by its standard error) as the dependent variable and precision (1/SE) as the independent variable with no intercept. Nothing else is needed.

[8] PET (or PEESE) is the estimated intercept from a simple regression with effect size as the dependent variable and SE (or SE^2) as the single independent variable, using 1/ SE^2 as the WLS weights.

Stanley and Doucouliagos (2014) approximate by a restricted polynomial function of the standard error (PEESE). When the true effect is zero, it can also be shown mathematically that this complex function collapses to a linear relation with standard error, giving PET (Stanley & Doucouliagos, 2014). A series of statistical simulation studies documents how PET-PEESE often greatly reduces selective reporting bias and is preferable to conventional meta-analysis methods and to the "trim-and-fill" publication bias correction algorithm (Moreno et al., 2009; Stanley, 2008, 2017; Stanley & Doucouliagos, 2014, 2017; Stanley et al., 2017). PET-PEESE provides a more aggressive approach to selective reporting bias than any simple weighted average, but it too has limitations, overcorrecting for publication bias in some cases (Stanley, 2017).[9]

To recap, we calculate power and bias in a robust, yet conservative, manner by employing three proxies of true average effect size: (a) the WLS unrestricted weighted average, with point estimates equivalent to the fixed-effect, (b) the weighted average of the adequately powered (WAAP-WLS), and (c) the PET-PEESE meta-regression reduction of selective reporting bias. Two of these approaches (WLS and WAAP-WLS) are known to overestimate the true effect if there is any type of selective reporting bias. They are conservative in the sense that they give the benefit of doubt to the psychological research record as reported and are likely to overestimate psychological research's power, on average. PET-PEESE, on the other hand, more aggressively attempts to identify and filter out selective reporting bias; thus, it is possible to underestimate true effect and thereby underestimate statistical power in some cases (Stanley, 2017). Because our survey is descriptive, we focus on the median powers across reported effect sizes, research topics, and also across these three estimation approaches to the average true effect size. The median of these three will thus tend to overestimate the quality of psychological research. Our overall survey results do not depend on the accuracy or validity of PET-PEESE. We include PET-PEESE only for the sake of robustness and to see what if any difference might result when a more aggressive approach to reducing selective reporting bias is used. Below, we find that it makes little difference.

Assessing Adequate Power

With an estimate of the true effect for a given meta-analysis (WLS, WAAP-WLS, or PET-PEESE), adequate power is easy to assess. We assume null hypotheses are two-tailed with a 5% significance level, and we accept Cohen's 80% as the definition of adequate power. These conventions for Type I and Type II errors imply that the true effect needs to be equal to or greater than 2.8 standard errors, in absolute magnitude, if power is to reach 80%. This value of 2.8 is the sum of 1.96 and 0.84, where 1.96 is the minimum number of standard errors from zero that an observed effect must fall to be rejected with a 5% significance level and 0.84 is the number of additional standard errors that the true effect must fall from zero such that 80% of the distribution of the observed effect is in the rejection region (see Figure 1). Hence, for a study to have adequate power, its standard error needs to be smaller than the absolute value of the underlying mean true effect divided by 2.8. All that remains to assess adequate power, retrospectively, are (a) the values of the standard error and (b) an estimate (or estimates) of the true effect. If the standard error of a study is less than the absolute value of an estimated true effect (from WLS, WAAP-WLS, or PET-PEESE) divided by 2.8, we know that this study is adequately powered to detect a true effect equal or greater than this estimate. Median power for a given area of research can then be calculated as one minus the cumulative normal probability of the difference between 1.96 and the absolute value of an estimate of the true effect divided by the median standard error. This probability (median power) would look much like the 80% adequate power displayed in Figure 1, except the relevant Z-value is now the absolute value of one of these proxies for true effect (WLS, WAAP-WLS, or PET-PEESE) divided by median standard error, rather than 2.8 as displayed in Figure 1. Because our survey is descriptive, we focus on the median powers across: studies within a meta-analysis, areas of research, and across these three estimation approaches to the average true effect size.

Assessing Residual Selective Reporting Bias

If an area of research is selectively reporting effect size to be statistically significant in a direction consistent with the prevailing psychological theory, then the average reported effect will, on average, have a larger magnitude than the true effect (whether or not prevailing psychological theory suggests a direct or an inverse association). As before, we can use these meta-averages: WLS, WAAP-WLS, and PET-PEESE as proxies for true effect and then compare them to average reported effect for an assessment of residual reporting bias. Needless to say, each reported estimate is subject to random sampling error, and will be sometimes larger and sometimes smaller than the mean true effect. Such differences cannot be regarded as bias, but merely as sampling or estimation errors. However, when there is a systematic trend for the simple average (i.e., the unweighted mean) to be larger than a meta-average known to be less biased when there is selective reporting bias, we can regard the average difference when it persists over hundreds of separate areas of research as a lower limit of residual reporting bias.[10] This average difference is calculated as the median absolute value of the simple average found among these 200 meta-analyses minus the median absolute value of WLS, WAAP-WLS or PET-PEESE. Below, we report this average residual reporting bias as a percent of the median absolute value of the meta-estimate (WLS, WAAP-WLS or PET-PEESE) that serves as a proxy for true effect. Ioannidis et al. (2017) find that the typical reported estimate in economics is twice as large, or larger, than either WAAP or PET-PEESE.

Summary

To recap the process that we used to calculate these statistics and how the below results were obtained:

[9] We do not use Simonsohn, Nelson, and Simmons's (2014) "*p*-curve" correction for "*p*-hacking." Recently, several papers establish that the *p*-curve approach is biased and unreliable when there is either heterogeneity, misspecification biases, or when some non-significant studies are reported (Bruns & Ioannidis, 2016; McShane, Böckenholt, & Hansen, 2016; van Aert et al., 2016). Such conditions are ubiquitous in the social sciences. For example, we find that the typical heterogeneity variance among the 200 meta-analyses that we survey is nearly 3 times larger than the corresponding random sampling variance. That is, 74% of the observed variation of reported research results from study to study is typically due to actual differences in the true effect (heterogeneity) or to differential bias, in either case overwhelming the *p*-curve's assumed pattern of p-values from sampling errors alone.

[10] It is a lower limit because all weighted averages are known to be biased in the same direction as the simple average when there is some selective reporting.

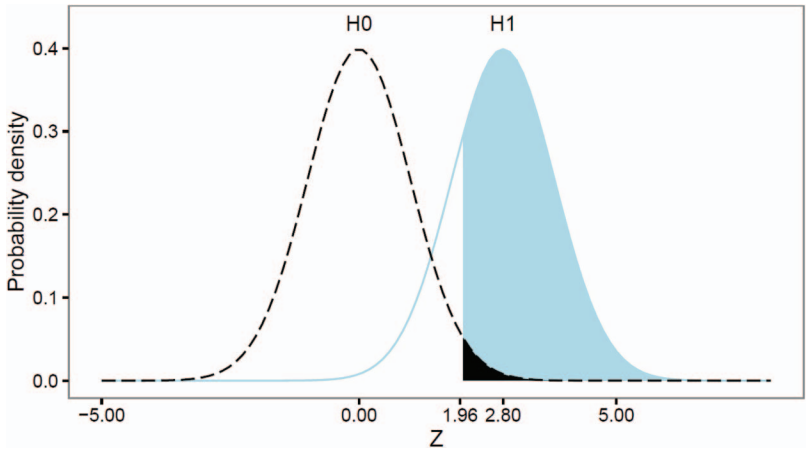

Figure 1. Obtaining adequate power (80%). Z is distributed as a standard normal distribution, N(0,1). 1.96 is the critical value for testing against 0 at $\alpha = .05$. Figure 1 illustrates how the standardized mean of the sampling distribution (H_1) needs to 2.8 standard errors away from 0 (the mean of H_0) for an experiment to have adequate power—Cohen's 80%, which is represented by the shaded area of H_1. Because the normal distribution is symmetric, psychological phenomena that are inversely related or negative correlated will work exactly as depicted above when the absolute value of the mean of the true effects distribution, or its proxy, is employed. See the online article for the color version of this figure.

1. First, we gathered data on effect sizes and their standard errors from the 200 most recent *Psychological Bulletin* meta-analyses where the required information is available.

2. For each of these 200 meta-datasets, we calculate several summary statistics—average effect size, median effect size, and median standard error (*SE*)—and perform several meta-regression analyses. From these meta-regression analyses, two weighted averages, WLS and WAAP-WLS, are produced. These meta-regression analyses also calculate: the precision-effect test (PET), the PET-PEESE corrected estimate, the funnel-asymmetry test (FAT), and I^2. All of these statistics are calculated with the effects sizes as reported, whether positive or negative. From these statistics and the methods described above, the proportion of studies that are adequately powered and the median power are each calculated using three different proxies (WLS, WAAP-WLS and PET-PEESE) for the mean of the true effect distribution. These 13 statistics plus two statistical tests (FAT and PET) are computed, separately, for each of our 200 meta-datasets.

3. Lastly, descriptive statistics, largely medians, are computed across an aggregate data file containing 200 rows, each one of which records all of the statistics mentioned in Step 2, above. In total, 600 meta-regression analyses and more than a dozen meta-meta-analyses are conducted, jointly producing thousands of statistics.

Results

Among these 200 meta-analyses containing 12,065 estimated effects, the average effect size is 0.389, expressed as the median of average standardized mean differences, or 0.191 as a correlation coefficient. This overall effect size is nearly the same as the average of the first 100 years of social psychology ($r = .21$) uncovered by Richard et al. (2003). The typical standard error is 0.104, expressed as a correlation; 0.21 when represented by a standardized mean difference (SMD). Contrary to recent concerns about publication bias, questionable research practices and null hypothesis significance testing, we find that the typical psychological research study is statistically nonsignificant at the conventional .05 level of significance.

Table 1 reports the median absolute value of the average reported effect size from these 200 meta-analyses. Here, all effect sizes are first converted to SMD to be comparable. However, an interesting pattern emerges when these 200 meta-analyses are divided by the types of effect sizes that are commonly reported: correlation versus SMD. The typical effect size found among the 108 "correlation-based" meta-analyses in our survey (0.458, in SMD units) is 57% larger than those meta-analyses measured by SMDs (0.291).[11]

Power

The median of the percent of reported effects that are adequately powered across these 200 meta-analyses are (a) 7.7% when the unrestricted WLS (or fixed effect) weighted average is used to represent the mean of the true effects distribution, (b) 7.1% when WAAP-WLS proxies for true effect, and (c) 9.1% if PET-PEESE substitutes for true effect. Figure 2 displays the distributions of the proportion of studies that are adequately powered across these 200

[11] We thank Frank Schmidt for pointing out that, technically, only point-biserial correlations can be converted to Cohen's d. *Ceteris paribus*, other correlations will be larger than the point-biserial correlation due to the latter's restricted range. Thus, a small part of the larger average effect size of correlation-based meta-analyses might be due to the conversion of all correlations to Cohen's d. Because these 108 "correlation-based" meta-analyses often contain an undisclosed mix of correlation types, we cannot fully correct this small bias. However, as we discussed above, none of our calculations of power, heterogeneity or bias depend in any way on the transformation of correlations to standardized mean differences.

Table 1
Median Statistical Power and Average Effect Sizes

Type of effect	Mean	Proportion with adequate power			Median power		
	Absolute effect sizes	WLS	WAAP-WLS	PET-PEESE	WLS	WAAP-WLS	PET-PEESE
Overall (m = 200)	.389	.077	.071	.091	.364	.331	.360
Correlations (m = 108)	.458	.243	.328	.272	.577	.610	.607
SMDs (m = 92)	.291	.013	.000	.011	.230	.171	.170

Note. Table entries are medians. Mean absolute effect sizes are reported in this table in units of standardized mean differences (SMD), regardless of whether they were reported in the meta-analysis as correlations or as SMD. WLS is the unrestricted weighted least squares weighted average. WAAP-WLS = the weighted average of adequately powered effect sizes (WAAP) or weighted least squares (WLS) when there are no adequately powered studies; PET-PEESE = the conditional precision-effect test-precision-effect estimate with standard error meta-regression correction for publication bias; m = the number of meta-analyses. Adequate power is defined as 80%, following Cohen (1977).

areas of research. Clearly, underpowered studies dominate psychological research. But how underpowered are they?

We also calculate the median powers for each of these 200 meta-analyses. The typical power of psychological research is around 36%: 36.4% based on WLS, 33.1% based on WAAP-WLS, and 36% based on PET-PEESE. Figure 3 shows the distribution of median powers across these 200 areas of research. Note their striking shapes. The two most frequent categories are the lowest (0–10%) and the highest (over 90%). Even though typical areas of research are quite inadequately powered, as measured by median power, approximately one fifth of these 200 areas of psychological research are quite highly powered. Between 19% and 23% have an average statistical power of 90% or higher (see Figure 3). It should not be surprising that some areas of research have high power. Aside from sample size, statistical power depends on the underlying true effect size, and some psychological phenomena have large effects. When WAAP-WLS is used to estimate true effect, one third of these 200 areas of psychological research (32%) have large or medium effect sizes; defined as |WAAP-WLS| > 0.5 SMD. Even rather modest sample sizes will estimate large effects powerfully.

As before, we find a striking difference between correlation-based meta-analyses and SMD-based meta-analyses. Table 1 breaks down the median proportion of studies that are adequately powered and the median power by type of effect size reported in *Psychological Bulletin* meta-analyses. Those areas of psychological research that predominately report standardized mean differences (SMDs) are highly underpowered; typically, 99% are underpowered compared to 67% to 76% for correlation-based meta-analyses. The median statistical power in SMD meta-analyses is between 17% to 23%. For correlation-based research, typical power is nearly three times higher—58% to 61%.

Residual Selective Reporting Bias

Recall that residual reporting bias may be calculated as the difference between the median absolute value of the simple unweighted mean reported effect and the median absolute value of one of our less vulnerable proxies for true effect—WLS, WAAP-WLS or PET-PEESE. We find only small amounts of residual reporting bias in these 200 meta-analyses of psychological research: 8% based on WLS, 12% based on WAAP-WLS, and 15% based on PET-PEESE. Thus, our survey identifies only a small systematic exaggeration, overall.

There are some important qualifications to make about this finding. First, all these estimates are themselves biased and two of them consistently underestimate residual bias when there is selective reporting (Stanley, 2017; Stanley & Doucouliagos, 2014,

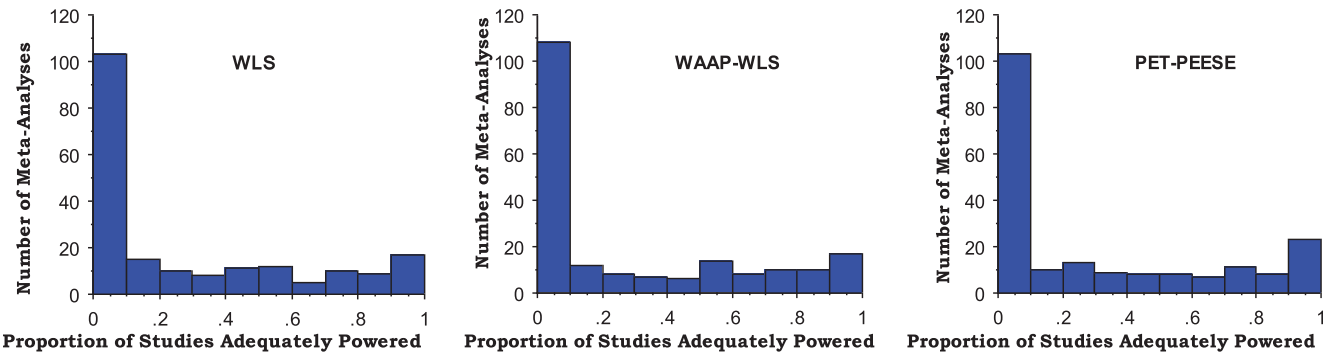

Figure 2. Histograms of adequately powered estimates from 200 areas of research. Weighted least squares (WLS) is the unrestricted weighted least squares weighted average. WAAP-WLS is the weighted average of adequately powered effect sizes (WAAP) or WLS when there are no adequately powered studies. PET-PEESE is the conditional precision-effect test-precision-effect estimate with standard error meta-regression correction for publication bias. Adequate power is defined as 80%, following Cohen (1977). See the online article for the color version of this figure.

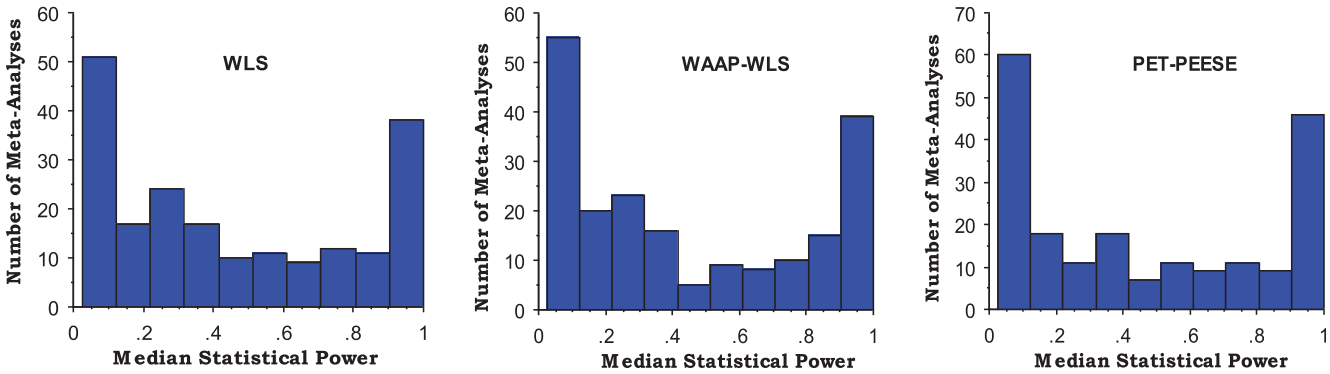

Figure 3. Histograms of median statistical power from 200 areas of research. Weighted least squares (WLS) is the unrestricted weighted least squares weighted average. WAAP-WLS is the weighted average of adequately powered effect sizes (WAAP) or WLS when there are no adequately powered studies. PET-PEESE is the conditional precision-effect test-precision-effect estimate with standard error meta-regression correction for publication bias. Adequate power is defined as 80%, following Cohen (1977). See the online article for the color version of this figure.

2015; Stanley et al., 2017). Second, a notable proportion of psychology might still be affected by selective reporting bias, even if the median amount of exaggeration is relatively small. 27.5% (or 55 areas of research) find evidence of some type of selective reporting or small-sample bias using the Egger test for funnel asymmetry (FAT), and this is likely to be an underestimate of the incidence of these biases because the Egger test is known to have low power (Egger, Davey Smith, Schneider, & Minder, 1997; Stanley, 2008). Third, we again find notable differences between types of effect sizes reported by meta-analysts. When using WLS as a proxy for the true effect, we find that the simple unweighted mean of reported SMDs is now exaggerated by 13%, on average, by 20% if WAAP-WLS substitutes for true effect, and by 30% relative to the median absolute value of PET-PEESE.

Heterogeneity

The median percent of the observed variation of reported effect sizes within a given area of research that is attributed to heterogeneity (I^2) is 74%, which means that the variance among true effects is nearly 3 times larger than the reported sampling variance. According to Pigott's (2012) guidelines for small (25%), medium (50%) and large (75%) heterogeneity, typical areas of research have nearly "large" excess heterogeneity. Yet, this level of heterogeneity appears to be the norm for research in psychology. For example, van Erp, Verhagen, Grasman, and Wagenmakers (2017) extracted estimates of heterogeneity from 705 meta-analyses published in *Psychological Bulletin* between 1990 and 2013 and found that the median reported $I^2 = 70.62\%$ (interquartile range: [33.24%, 87.41%]). Figure 4 displays the distribution of I2 in our survey of 200 Psychological Bulletin meta-analyses.

However, it is important to remember that I^2 is a relative measure of heterogeneity and does not reflect the variation in true effects as measured in units of SMDs. When our median I^2 is applied to the typical area of research, the standard deviation among *true* effects is 0.354 SMD,[12] and the standard deviation from one study to the next due to both heterogeneity and sampling error becomes 0.412, larger than the typical reported effect size, 0.389. In practical terms, this observed level of heterogeneity is huge.

Experimental Versus Observational Research

As a post hoc secondary analysis,[13] we examined whether the systematic differences between SMD-based and correlation-based meta-analyses are due to experimental design: experimental versus observational. Unfortunately, this differentiation is not perfect. Many meta-analyses contain a mix of experimental and observational research designs at the study level. For example, Williams and Tiedens (2016) meta-analysis of the effects of gender and dominance behavior includes 97 experimental studies (85%) where dominance behavior was somehow manipulated and 17 purely observational studies.

Because a substantial percent (42.4%) of those meta-analyses that report effect sizes in terms of SMD are observational, and 31.4% of correlation-based meta-analyses are experimental, there is only small correlation ($\phi = 0.263$; $p < .001$) between experimental design (1 if primarily experimental; 0 elsewhere) and effect type (1 if SMD; 0 for correlation). However, we do see some interesting differences in power and heterogeneity by experimental design. First, there is a difference between heterogeneity as measured by I^2 ($p < .01$): The median I^2 for experimental research is

[12] Anonymous reviewers expressed concern that the choice of the level of aggregation (to report one overall meta-analysis with all effects sizes included or to report only subdivided meta-analyses) might influence the findings. This decision by *Psychological Bulletin* authors does not affect our survey's findings in any noteworthy way. For those 23 papers that report a single aggregated meta-analysis with all effects sizes included, the typical mean effect is 0.401 vs 0.387 for those that subdivide; the median SE for these 23 aggregated meta-analyses is 0.181 vs 0.21, the typical proportion with adequate power is 7.1% vs 7.7%, median power is 27.4% vs 33.1% (WAAP-WLS), the median I^2 is 78.1% vs 72.4%, and the typical standard deviation among true effects, as discussed in text above, is 0.342 vs 0.340.

[13] Investigating these differences was not part of our pre-analysis plan. Anonymous reviewers asked that we code for experimental design and report the differences.

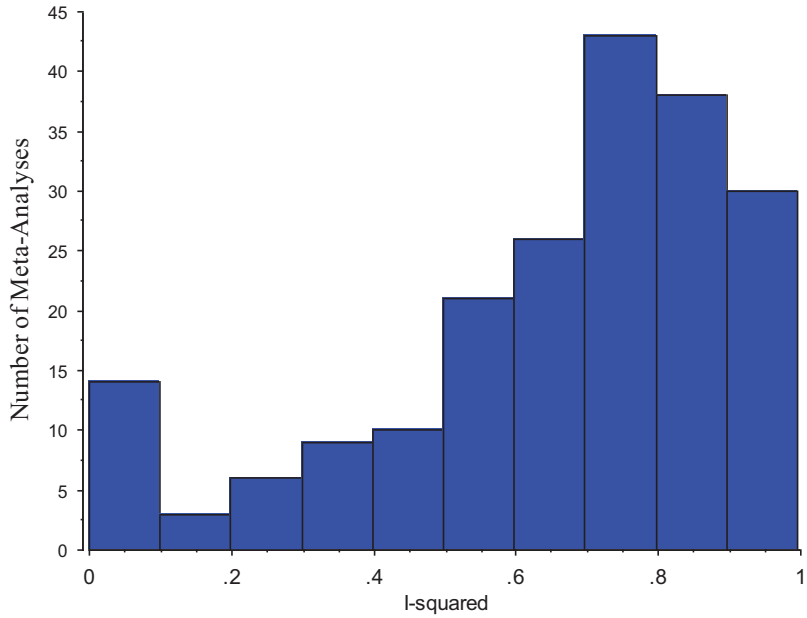

Figure 4. Histograms of I^2 from 200 areas of research. I^2 is the proportion of observed variation among reported effect sizes that cannot be explained by the calculated standard errors associated with these reported effect sizes. It is a relative measure of heterogeneity. See the online article for the color version of this figure.

68% versus 76% for observational designs. But I^2 is a relative measure of heterogeneity with a nonsymmetric distribution. To correct for I^2's nonstandard distribution, we used Abramowitz and Stegun's (1964) normal approximation for the χ^2 distribution applied to the Cochran Q test for heterogeneity. Doing so causes this difference in relative heterogeneity to be only marginally larger than statistical noise ($p = .045$). In addition, experimental research designs have larger sampling errors and lower power. Typical sampling errors are 0.26 versus 0.19, measured as median SEs in units of SMD. Table 2 reports the median proportion of studies that are adequately powered and the median of median powers by type of research design. Even though there is only a small association between experimental design and effect type, we find a similar pattern among the typical levels of power for experimental design that we see for effect type, confirming the concerned expressed by dozens of researchers over the years that scarcely any experimental studies are adequately powered. All of these results about experimental versus observational research designs should be interpreted with caution because, as mentioned, they are conducted post hoc.

Research Domains

Because anonymous reviewers requested the breakdown of power and heterogeneity by research domains, we provide a second post hoc set of comparisons that were not part of our preanalysis plan. Cognitive psychology is the most frequently represented research domain, 27.5%, with nearly as many of these 200 *Psychological Bulletin* meta-analyses as the next two domains (clinical and social) combined (see Figure 5).

Figure 6 displays how average power and heterogeneity are distributed across these domains. Because our survey is descriptive, we make no inference about what might underlie these distributions of power and heterogeneity. Nonetheless, a few descriptive patterns seem clear. In terms of both the percent of studies that are adequately powered and median power, behavioral

Table 2
Median Statistical Power by Experimental Design

	Proportion with adequate power			Median power		
Experimental design	WLS	WAAP-WLS	PET-PEESE	WLS	WAAP-WLS	PET-PEESE
Observational ($m = 113$)	.278	.259	.268	.621	.613	.585
Experimental ($m = 87$)	.032	.000	.053	.247	.232	.236

Note. Table entries are medians. WLS = the unrestricted weighted least squares weighted average; WAAP-WLS = the weighted average of adequately powered effect sizes (WAAP) or WLS when there are no adequately powered studies; PET-PEESE = the conditional precision-effect test-precision-effect estimate with standard error meta-regression correction for publication bias; m = the number of meta-analyses. Adequate power is defined as 80%, following Cohen (1977).

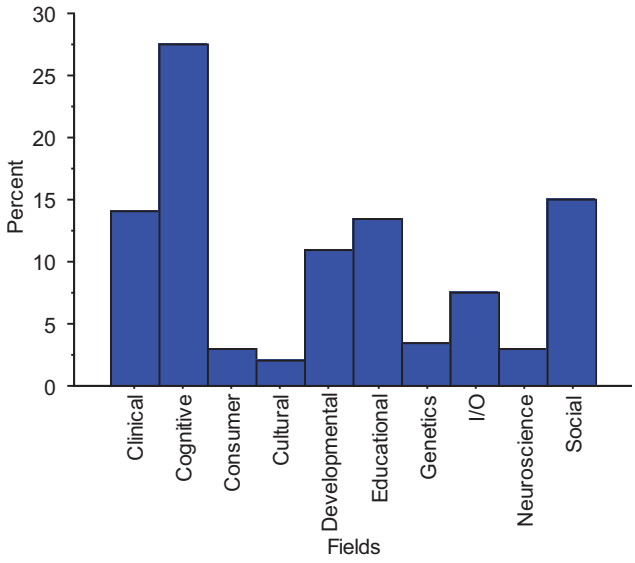

Figure 5. Histograms of 200 meta-analyses by research domain. "Genetics" is short for behavioral genetics as "I/O" represents industrial/organizational. See the online article for the color version of this figure.

genetics (labeled *genetics* in Figure 6) stands out. On average, the median power of behavioral genetics studies is 86.4% (when WAAP-WLS serves as the proxy for true effect), and 71.5% are adequately powered. The primary reason that power is so much higher than what is typically found across these psychological domains is that effect sizes are also much larger for behavioral genetics—average |WAAP-WLS| $>$ 1 SMD. Unfortunately, behavioral genetics also has the highest average heterogeneity (I^2 = 80%)—see Figure 6. Because replication depends on the combination of I^2 and power (see the Discussion section below), behavioral genetics will also be challenged to closely replicate a previously reported effect size.

After behavioral genetics, educational psychology and industrial/organizational (labeled I/O in Figure 6) also have high levels of power, where median powers are approximately 60% and over 40% are adequately powered. It is interesting to note descriptively that those areas of greatest power are also areas of research that are highly observational. All behavioral genetics and educational psychology meta-analyses and 87% of industrial/organizational meta-analyses are primarily observational. At the other end of the power distribution, those areas which have the highest concentration of experimental studies (cognitive, consumer, and neuroscience) are among the least powered domains. Further research is needed to understand the complex interactions of domain, design, sample size, effect size, and heterogeneity, reliably.

Last, note the apparent flatter distribution of heterogeneity across research domains in Figure 6. Although behavioral genetics is again the highest, it is not notably more heterogeneous than several other domains. The only stand out is neuroscience, where average heterogeneity is about half that of behavioral genetics. Unfortunately, neuroscience tends to be at the lower end of the power distribution. In summary, no area of psychology represented by these 200 *Psychological Bulletin* meta-analyses has that desirable combination of high statistical power and low heterogeneity; hence, all psychological domains will find replication a challenge under typical conditions. Again, these breakdowns by research domains should be interpreted with caution because they are conducted post hoc.

Discussion

Our survey of 12,065 estimated effects sizes from nearly 8,000 articles in 200 meta-analyses reveals that the typical effect size is 0.389 SMD with a median standard error of 0.21. We also find low statistical power, small residual selection bias, and high levels of relative heterogeneity. The central purpose of our review is to assess the replicability of psychological research through meta-analysis and thereby better understand recent failures to replicate. Our findings implicate low statistical power and high levels of heterogeneity as the primary causes of failed replications, however defined.

We find that only a small proportion of psychological studies as reported in *Psychological Bulletin* are adequately powered, approximately 8%, and the median power is about 36%, across three proxies for the mean true effect.[14] This median power is somewhat less than what is reported in a recent survey. Fraley and Vazire (2014) find that the median power to detect a correlation of 0.2 is 49% among top social-personality journals for the years 2006–2010. But then, Fraley and Vazire (2014) calculate prospective, rather than retrospective, power and their sampling frame is different than ours. Thus, we would expect some differences, especially when considering that our median power calculations reflect the observed effect sizes of each area of research and the distribution of statistical power within each of these areas of research.

What does our survey imply about replication? A median power of 36% means that the typical replication study that uses the typical care in duplicating research protocols, conditions, and methods and typical sample sizes will have only a 36% chance of finding a statistically significant effect in the expected direction. Coincidentally, this value is exactly the same percent of replications found to be statistically significant in the same direction by the Open Science Collaboration (Open Science Collaboration, 2015, Table 1). Thus, when replication is viewed in terms of sign and statistical significance, it is no wonder that rates of replication are considered low in psychology research.

Improving replication, as noted by others (e.g., Maxwell, Lau, & Howard, 2015), would seem to be a matter of: conducting both initial studies and replications with larger samples, reducing sources of nonsampling error (e.g., measurement error; Stanley & Spence, 2014), and focusing on larger effects. Because researchers work with limited resources and knowledge, these obvious recommendations are extremely difficult to implement in most cases.

More practical recommendations have centered on redefining replication success and adjusting researchers' expectations about nonsignificant replications. For example, Patil et al. (2016) examined the data from the Reproducibility Project: Psychology (Open Science Collaboration, 2015) in terms of prediction intervals—confidence intervals that account for variability in both the original and replication study—and found that 77% of the replication attempts were consistent with the original findings. This finding

[14] To be more precise, 36% is the median among the medians of medians.

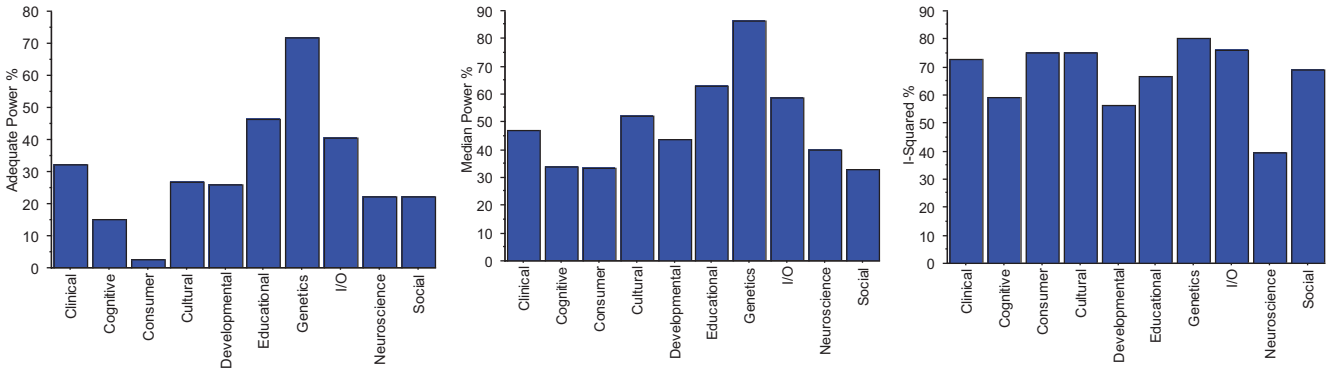

Figure 6. Breakdown of power and heterogeneity by research domain. Adequate and median power are calculated using WAAP-WLS as the proxy for the mean of the true effect distribution. WAAP is the weighted average of adequately powered effect sizes (WAAP) or WLS when there are no adequately powered studies. Adequate power is defined as 80%, following Cohen (1977). I^2 is the proportion of observed variation among reported effect sizes that cannot be explained by the calculated standard errors associated with these reported effect sizes. It is a relative measure of heterogeneity. "Genetics" is short for behavioral genetics as "I/O" represents industrial/organizational. See the online article for the color version of this figure.

seems to be in apparent contrast to the highly publicized result that only 36% of the replications in those data were statistically significant, but completely in accord with the less well-publicized result that meta-analytically combining the replications with the original studies resulted in 70% statistical significance (Open Science Collaboration, 2015). Along similar lines, several authors have argued for assessing replication primarily within the context of meta-analysis (e.g., Braver et al., 2014; Fabrigar & Wegener, 2016; Stanley & Spence, 2014). Doing so could shift the evaluation of success versus failure of replications to a judgment about the degree of information that new data add to our knowledge base (Patil et al., 2016).

Consistent with a previous survey of heterogeneity (van Erp et al., 2017) and several findings from recent multisite replication attempts (Eerland et al., 2016; Hagger et al., 2016; Klein et al., 2015), our survey reveals clear evidence for high levels of heterogeneity in research reported in *Psychological Bulletin*. We find that heterogeneity accounts for nearly three fourths of the observed variation among reported effects (i.e., median $I^2 = 74\%$). When applied to the median reported standard error (measured in terms of SMD), this high I^2 implies that the typical standard deviation of heterogeneity is equal to 0.354 (again, in units of SMDs). Importantly, even a replication study with millions of subjects will remain vulnerable to heterogeneity among true effects. When we apply our survey's typical heterogeneity to its typical effect size, it is unlikely that any replication will be successful. For example, if our median effect size, 0.389, is the mean of the distribution of true effects, then there is a 29.8% probability that the largest possible replication study will find a negligible, zero, or opposite-signed effect. The probability is 32.4% that this ideal replication (i.e., $n \to \infty$) finds a small effect, and it is 25.5% for a medium-sized effect—using Cohen's guidelines of 0.2, 0.5, and 0.8. Even though 0.389 is considered a small effect by these guidelines, the probability that a very large replication finds a large effect remains non-negligible at 12.3%. Thus, it is quite likely (68%) that an ideal replication will not reproduce a small effect when the mean of the distribution of true effects is equal to our median average effect size. The wide distribution of true effect sizes that our survey finds is also similar to what Open Science Collaboration observed when attempting to replicate 100 psychological experiments—see Figure 3 in the Open Science Collaboration (2015).

No matter what the mean of the distribution of true effect sizes might be, or how large the replication study is, it is unlikely that a replication will find an effect size close to what was found previously when there is this much heterogeneity. If a successful replication is defined as finding an effect size similar to what some previous study or studies have found (e.g., to within ±0.1 or within ±0.2 SMD), then there will always be a sizable chance of unsuccessful replication, no matter how well conducted or how large any of the studies are. For example, suppose that two studies are conducted with nearly infinite sample sizes and therefore infinitesimal sampling error. Heterogeneity of the size that our survey finds implies that the probability that these two ideal studies find effect sizes that are within ±0.1 from one another is 15.8% and 31.1% to within ±0.2. Indeed, there remains a 50% or greater probability of a failed replication whenever the acceptable difference for a successful replication is set at less than 0.35 SMD. Needless to say, if we have less-than-ideal sample sizes, the resulting added sampling error will reduce the probability of a successful replication further. Levels of heterogeneity this high further explain why the Open Science Collaboration (2015) found that "(n)o single indicator sufficiently describes replication success" (p. 943).

Because our survey is descriptive, we make no inference about the sources or explanation of this high level of heterogeneity nor about research that does not happen to be included in *Psychological Bulletin* meta-analyses. We only hope that our study serves as a stimulus for other researchers to investigate systematically the sources of heterogeneity among psychological research results and the implications that such heterogeneity might have on the practice and credibility of psychology.

Limitations

There are important caveats to these calculations of replication success that need to be mentioned. First, there is a wide distribu-

tion of observed heterogeneity among these areas of psychological research (see Figure 4); 22% have I^2s that are less than 50%, 47% have I^2 values 75% or higher and one out of seven have excess heterogeneity of 90% or more. Thus, successful replication will be more likely in some areas and domains of research but much less likely in others. Second, these calculations assume that the typical heterogeneity observed in this research record is entirely heterogeneity among true effects; that is, variation beyond the control of researchers. Because exact replication is rare in psychology, some of this observed heterogeneity will be due to variation in experimental conditions, measures, methods and the characteristics of the population from which the samples are drawn. Indeed, meta-analyses often attempt to account for systematic variation by further employing meta-regression analyses or subgroup comparisons. For example, North and Fiske (2015) found that such factors as geographical region and cultural individualism help to predict attitudes toward older adults, and Williams and Tiedens (2016) found that studies where dominance was explicit exhibited greater backlash to women's dominance behavior. Williams and Tiedens (2016) also conducted an exploratory investigation of eight additional moderator variables, including gender of first author of the study, article source, design type, sample type, dominance medium, target of the dominance, study location, and sample nationality.

In any case, the studies summarized by these 200 meta-analyses are a mix of direct and conceptual replications. Thus, a careful exact replication study could avoid a notable amount of this variation in expected effect by carefully duplicating all of the controllable features of the experimental design, execution, and evaluation. However, those large-scale efforts to control these controllable research dimensions still find that notable heterogeneity remains (Eerland et al., 2016; Hagger et al., 2016; Klein et al., 2015). Moreover, if half the observed heterogeneity variance is due to differences that are under the control of the researcher,[15] then our estimate of the typical standard deviation due this uncontrollable heterogeneity among true effects would be reduced to 0.25. In which case, the probability that the largest possible replication study will find a negligible, zero, or opposite-signed effect declines to 22%, 28% for a medium effect, and it will find a large effect only about 5% of the time. Even if half of the observed heterogeneity among reported effects is under the researcher's control, the probability that any replication effort successfully replicates a small effect (recall 0.389 is our survey's median effect size) remains less than 50%.

Third, we do not calculate random effects, because random effects is widely known to be more biased if there is any type of selective reporting bias (Henmi & Copas, 2010; Poole & Greenland, 1999; Stanley, 2017; Stanley & Doucouliagos, 2014; Stanley & Doucouliagos, 2015; Stanley et al., 2017; Sutton et al., 2000). As a result, we cannot compute random-effects' measure of heterogeneity, τ^2, either. Perhaps, τ^2 would provide a somewhat different assessment of the effect of typical heterogeneity on potential replication success than our use of the medians of I^2 and *SE*? However, it is important to note that either the low power or the high heterogeneity that our survey finds to be typical, independently, explain the low levels of replication success recently reported. Thus, even if one were to find that heterogeneity is not so severe, successful replication remains unlikely when combined with the typical powers found in this survey.

A further limitation is that because the *Psychological Bulletin* is our sampling frame, we cannot be sure that what we find in our survey is more broadly representative of psychological research. From the outset, our intention has been to be conservative in our choice of methods and approaches by erring of the side of the credibility of psychology, so although we fully accept this limitation, it comes from a set of principled choices. *Psychological Bulletin* is a leading academic journal in psychology. It tends to publish the highest quality meta-analyses on well-respected areas of research that are likely to be, if anything, somewhat more mature and credible. If so, our sample frame would tend to shine a more favorable light on psychological research. Last, recall that we are sampling meta-analyses, not studies, so these data sets have potentially had some degree of bias removed through the meta-analysts' choices of inclusion criteria. Each meta-analysis published in *Psychological Bulletin* goes to great effort to identify and include all relevant research studies (60, on average), regardless of where they might be published and often including unpublished research.

In contrast, an anonymous reviewer suggested that what is published in *Psychological Bulletin* might be systematically different than psychology more broadly defined. This argument further suggests that small meta-analyses are rarely seen in *Psychological Bulletin*, and they tend to be conducted on areas of research literature that are well established. If this is the case, might one expect *Psychological Bulletin* meta-analyses to be somehow different? In a broad survey of research across many scientific disciplines, Fanelli et al. (2017) found partial support for the presence of a decline effect; where "The earliest studies to report an effect might overestimate its magnitude relative to later studies, due to a decreasing field-specific publication bias over time or to differences in study design between earlier and later studies" (p. 3714). Stanley, Doucouliagos, and Jarrell (2008) suggested that there might be a "research cycle," generated by a preference for novelty by editors and reviewers. A novel and interesting hypothesis tends to be initially confirmed in the first phase of a new line of research until yet another confirmation is no longer seen to be novel or as a sufficient contribution. Eventually, studies with findings very different from the seminal paper(s) that began a line of research will be viewed as novel and sufficiently interesting to be published. In either case, the large effects observed in early, small literatures would tend to be reversed or contradicted over the evolution of a given line of research, thus generating heterogeneity. If either of these dynamic views of research maturation were correct, even partially, larger fields of research might have greater heterogeneity than small, less mature, ones. This, then, might provide a reason for higher levels of heterogeneity in *Psychological Bulletin* meta-analyses. But would psychological research that contained a higher frequency of small and thereby less mature lines of inquiry provide a better representation of psychological research?

Regardless, the smaller meta-analyses in our sample are not that much more likely to be successfully replicated. If we look only at

[15] The typical heterogeneity found by two large replication studies that attempted to be as exact as possible is larger than half of the median heterogeneity that our survey finds (Eerland et al., 2016; Hagger et al., 2016).

those meta-analyses that contain 10 or fewer research papers, representing 20% of our sample, we find that the typical effect size is nearly the same as before (0.382), typical power is lower (0.0% and 28.5% for the median percent with adequate power and median power, respectively), and the standard deviation of heterogeneity is 20% smaller (0.286). To be clear, 65% of these 40 small meta-analyses do not contain a single study with adequate power, and, as an anonymous reviewer and the above dynamic views of research suggests, heterogeneity is somewhat lower. Nonetheless, using the same methods as discussed above, the probability of successfully replicating a small-sized effect remains less than 40% (39.8% vs. 32.4%) even if the replication involves millions of subjects.

Conclusions

Our survey of 12,065 estimated effect sizes from nearly as many studies and 8,000 research articles finds that failed replications are to be expected at a rate consistent with the widely publicized Reproducibility Project: Psychology (Open Science Collaboration, 2015). Like many other researchers, we find that statistical power is chronically low. Unlike previous research, however, our survey reveals a more severe challenge for replication. High levels of heterogeneity are evidently the norm in psychology (median I^2 = 74%), at least as represented by studies reported in *Psychological Bulletin.* As discussed above, heterogeneity this large makes successful replication difficult, whether defined in terms of hypothesis testing or estimation precision. This high heterogeneity will, of course, be due in part to methodological differences between the different studies in each of the 200 meta-analyses that we survey. However, data from multisite registered replication efforts further suggest that when obvious methodological differences are removed (i.e., when exact replications are conducted) heterogeneity is not reduced to negligible levels. In other words, it is unrealistic to believe that variance in psychological phenomena studied can always be tightly controlled or reduced to practically insignificant levels.

Perhaps more surprising is our finding that there is relatively small exaggeration or overall selective reporting bias in recent psychological research. This is in contrast to the view that publication bias and questionable research practices are the primary cause of failed replications. Of course, as mentioned above, there are good reasons to be cautious in applying this hopeful result to any individual area of research. In particular, 27.5% of these areas of research produce a significant Egger's test (FAT) for publication bias (or small-sample bias), and this test is known to have low power. Our survey implies that the effects of selective reporting bias (i.e., publication bias, small-sample biases, p-hacking, and/or questionable research practices) have a less clear pattern than is maybe assumed by some researchers and by what has been observed in other fields (e.g., economics, Ioannidis et al., 2017).

What then are the more practical implications of our survey of psychological research? First, we recommend that future meta-analyses be expected to report median power as another indicator of research quality, along with conventional meta-analysis statistics and measures of heterogeneity that include I^2 and τ^2, or their equivalents. Second, readers must be especially circumspect about any summary of effect size (random effects, fixed-effect, WLS or WAAP-WLS) when heterogeneity is high and median power is low, regardless of what the *CI* or significance test might indicate.

To illustrate these issues, we return to our previous examples. Consider the rather strong conclusion expressed in by Williams and Tiedens (2016) regarding the backlash to women's dominance behavior, "we have demonstrated that the tendency for women to be penalized more than men for expressing dominance emerges *reliably* across a heterogeneous sample of studies" (p. 180; emphasis added). Although we commend these authors for conducting moderator analyses and qualifying meta-analyses with discussions of heterogeneity and the differences between explicit and implicit dominance in their text, this conclusion seems rather strong when one further considers power and heterogeneity. First, in only one of the six overall meta-analyses (Dominance × Target Gender interaction effect) is median power higher than 10% and with any of the studies being adequately powered (where power is based on the generous WLS proxy for the true effect). Second, there is high heterogeneity, I^2 = 84.9%, among these reported effects. Heterogeneity this high indicates little beyond true effects are either small, medium, or large, in spite of the highly significant, medium-sized, reported random-effects estimate (−0.58 SMD; $p < .0005$, p. 178). True positive effects cannot be entirely ruled out (8%). Given that only half of these 6 meta-analyses produce statistically significant findings, all but one are highly underpowered, and that those that are statistically significant have high heterogeneity, greater caution about the conclusion that women are penalized for dominance behavior is warranted based on the current research base. Readers should greatly discount any unqualified conclusion from a meta-analysis where there is both quite high heterogeneity and very little power.

In contrast, our second example of a *Psychological Bulletin* paper, North and Fiske (2015), is an exemplar of employing high heterogeneity to soften a conclusion. Recall that North and Fiske's (2015) review concerned attitudes toward older adults, and they combined all of these observational studies into a single meta-analysis. Even though the overall effect is both statistically and practically significant (*SMD* = −0.31; 95% confidence interval = −0.41, −0.20), they concluded, "the current analysis found evidence for a reverse overall pattern—albeit one with high heterogeneity, suggesting significant moderators, and a story warranting more qualification than broad, East-versus-West categories" (p. 1016). A qualification such as this one in the conclusion is exactly what our survey suggests should become the norm for any meta-analysis that finds high levels of heterogeneity. North and Fiske's (2015) meta-analysis of attitudes toward older adults reports quite high heterogeneity (I^2 = 92%), but statistical power is correspondingly high as well. Sixty percent are adequately powered, and median power is 89%. Although North and Fiske's (2015) conclusion appears much weaker, its research base warrants a stronger conclusion than does Williams and Tiedens (2016). In the case of attitudes toward older adults, we have some confidence that the typical design is adequate to make inferences on the topic. With backlash to women's dominance behavior, we can be confident only that the typical study design is nearly powerless to address many of the central issues. None of the studies have adequate power to study the three main effects to women's dominance behavior, and median powers are very low (4.8%, 29.7%, 4.3%). Cautious readers should take both power and heterogeneity into

account fully before forming strong beliefs about what an area of psychological research reveals.

In light of the typically high heterogeneity that we observed, how should replications in psychology be interpreted? To begin, it seems extremely likely that the *p*-value from any single, nonpreregistered, nonexact replication study will have nil informative value. In the face of high heterogeneity, this is true regardless of the sample size of the replication, because it is quite likely that the replication study will correctly reflect a substantially different true effect than the original study. However, individual nonpreregistered, nonexact replication studies may still contribute to our collective knowledge when added to a carefully conducted meta-analysis or meta-regression analysis. Meta-analysis moves the focus away from the statistical significance of the single replication study, increases statistical power by pooling across studies, allows one to accommodate and reduce selective reporting bias, and meta-regression analysis can use nonexact replications to help isolate and quantify methodological heterogeneity (e.g., Braver et al., 2014; Fabrigar & Wegener, 2016; Stanley & Doucouliagos, 2012; Stanley & Spence, 2014).

Replication findings from multisite, preregistered replications clearly offer a different class of evidence from the single, nonpreregistered, nonexact replication. Statistical power can be maximized by the pooling of resources across research sites, preregistration is likely to greatly reduce some types of selective reporting bias and heterogeneity due to obvious methodological characteristics can be minimized through tightly controlled study protocols. Multisite, preregistered replication also allows researchers to directly isolate and quantify the specific heterogeneity among true effects which remains after methodological heterogeneity is minimized. Such information can provide useful insights about the psychological phenomenon in question, thereby helping to guide and redirect future research and meta-regression analyses.

The central implication of our survey is that the typical under powered study, individually or when simply combined into a meta-analysis, offers little information about the magnitude or significance of the phenomenon it examines. Our core recommendation, then, is that researchers change their expectations about the ease with which convincing evidence, either for or against an effect, can be claimed.

References

References marked with an asterisk indicate *Psychological Bulletin* articles used in this survey.

Abramowitz, M., & Stegun, I. A. (Eds.). (1964). *Handbook of mathematical functions with formulas, graphs and mathematical tables*. Washington, DC: U.S. Department of Commerce.

American Psychological Association. (2010). *Manual of the american psychological association* (6th ed.). Washington, DC: Author.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. http://dx.doi.org/10.1002/per.1919

*Baas, M., Nijstad, B. A., Boot, N. C., & De Dreu, C. K. W. (2016). Mad genius revisited: Vulnerability to psychopathology, biobehavioral approach-avoidance, and creativity. *Psychological Bulletin, 142,* 668–692. http://dx.doi.org/10.1037/bul0000049

Baker, M. (2015, August 27). Over half of psychology studies fail reproducibility test. *Nature: International Weekly Journal of Science*. Advance online publication. http://dx.doi.org/10.1038/nature.2015.18248

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27,* 1069–1077. http://dx.doi.org/10.1177/0956797616647519

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

*Balliet, D., Li, N. P., Macfarlan, S. J., & Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological Bulletin, 137,* 881–909. http://dx.doi.org/10.1037/a0025354

*Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137,* 594–615. http://dx.doi.org/10.1037/a0023489

*Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140,* 1556–1581. http://dx.doi.org/10.1037/a0037737

Bohannon, J. (2015). Many psychology papers fail replication test. *Science, 349,* 910–911. http://dx.doi.org/10.1126/science.349.6251.910

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9,* 333–342. http://dx.doi.org/10.1177/1745691614529796

*Brewin, C. R. (2014). Episodic memory, perceptual memory, and their interaction: Foundations for a theory of posttraumatic stress disorder. *Psychological Bulletin, 140,* 69–97. http://dx.doi.org/10.1037/a0033722

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

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

*Byron, K., & Khazanchi, S. (2012). Rewards and creative performance: A meta-analytic test of theoretically derived hypotheses. *Psychological Bulletin, 138,* 809–830. http://dx.doi.org/10.1037/a0027652

*Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin, 140,* 980–1008. http://dx.doi.org/10.1037/a0035661

*Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotion expression in children: A meta-analytic review. *Psychological Bulletin, 139,* 735–765. http://dx.doi.org/10.1037/a0030737

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65,* 145–153. http://dx.doi.org/10.1037/h0045186

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York, NY: McGraw-Hill.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *Handbook of research synthesis*. New York, NY: Russell Sage Foundation.

*Defoe, I. N., Dubas, J. S., Figner, B., & van Aken, M. A. (2015). A meta-analysis on age differences in risky decision making: Adolescents versus children and adults. *Psychological Bulletin, 141,* 48–84. http://dx.doi.org/10.1037/a0038088

*Degner, J., & Dalege, J. (2013). The apple does not fall far from the tree, or does it? A meta-analysis of parent-child similarity in intergroup attitudes. *Psychological Bulletin, 139,* 1270–1304. http://dx.doi.org/10.1037/a0031436

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., . . . Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 15343–15347. http://dx.doi.org/10.1073/pnas.1516179112

*Eastwick, P. W., Luchies, L. B., Finkel, E. J., & Hunt, L. L. (2014). The predictive validity of ideal partner preferences: A review and meta-analysis. *Psychological Bulletin, 140,* 623–665. http://dx.doi.org/10.1037/a0032432

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P. A., . . . Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11,* 158–171. http://dx.doi.org/10.1177/1745691615605826

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634. http://dx.doi.org/10.1136/bmj.315.7109.629

*Else-Quest, N. M., Higgins, A., Allison, C., & Morton, L. C. (2012). Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological Bulletin, 138,* 947–981. http://dx.doi.org/10.1037/a0027930

Erica, C. Y., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review, 21,* 268–282. http://dx.doi.org/10.3758/s13423-013-0495-z

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology, 66,* 68–80. http://dx.doi.org/10.1016/j.jesp.2015.07.009

*Fairbairn, C. E., & Sayette, M. A. (2014). A social-attributional analysis of alcohol response. *Psychological Bulletin, 140,* 1361–1382. http://dx.doi.org/10.1037/a0037563

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5,* e10068. http://dx.doi.org/10.1371/journal.pone.0010068

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America, 201618569, 114,* 3714–3719. http://dx.doi.org/10.1073/pnas.1618569114

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

*Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., . . . Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137,* 517–537. http://dx.doi.org/10.1037/a0023304

*Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137,* 316–344. http://dx.doi.org/10.1037/a0021663

*Fox, N. A., Bakermans-Kranenburg, M. J., Yoo, K. H., Bowman, L. C., Cannon, E. N., Vanderwert, R. E., . . . van IJzendoorn, M. H. (2016). Assessing human mirror activity with EEG mu rhythm: A meta-analysis. *Psychological Bulletin, 142,* 291–313. http://dx.doi.org/10.1037/bul0000031

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9,* e109019. http://dx.doi.org/10.1371/journal.pone.0109019

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological & Personality Science, 7,* 8–12. http://dx.doi.org/10.1177/1948550615598377

*Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138,* 296–321. http://dx.doi.org/10.1037/a0026556

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9,* 641–651. http://dx.doi.org/10.1177/1745691614551642

*Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P. D., Robins, R. W., & Yan, T. (2015). Gender differences in narcissism: A meta-analytic review. *Psychological Bulletin, 141,* 261–310. http://dx.doi.org/10.1037/a0038231

*Haedt-Matt, A. A., & Keel, P. K. (2011). Revisiting the affect regulation model of binge eating: A meta-analysis of studies using ecological momentary assessment. *Psychological Bulletin, 137,* 660–681. http://dx.doi.org/10.1037/a0023660

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multi-lab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11,* 546–573. http://dx.doi.org/10.1177/1745691616652873

*Harkin, B., Webb, T. L., Chang, B. P., Prestwich, A., Conner, M., Kellar, I., . . . Sheeran, P. (2016). Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. *Psychological Bulletin, 142,* 198–229. http://dx.doi.org/10.1037/bul0000025

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Henmi, M., & Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine, 29,* 2969–2983. http://dx.doi.org/10.1002/sim.4029

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558. http://dx.doi.org/10.1002/sim.1186

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24. http://dx.doi.org/10.1198/000313001300339897

*Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin, 141,* 901–930. http://dx.doi.org/10.1037/a0038822

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Hozo, I., & Djulbegovic, B. (2013). Optimal type I and type II error pairs when the available sample size is fixed. *Journal of Clinical Epidemiology, 66,* 903–910.e2. http://dx.doi.org/10.1016/j.jclinepi.2013.03.002

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, C. (2017). The power of bias in economics research. *Economic Journal, 127,* F236–F265. http://dx.doi.org/10.1111/ecoj.12461

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

*Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin, 141,* 747–768. http://dx.doi.org/10.1037/bul0000015

*Karlin, B., Zinger, J. F., & Ford, R. (2015). The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin, 141,* 1205–1227. http://dx.doi.org/10.1037/a0039650

*Kim, S., Thibodeau, R., & Jorgensen, R. S. (2011). Shame, guilt, and depressive symptoms: A meta-analytic review. *Psychological Bulletin, 137,* 68–96. http://dx.doi.org/10.1037/a0021466

*Klahr, A. M., & Burt, S. A. (2014). Elucidating the etiology of individual differences in parenting: A meta-analysis of behavioral genetic research. *Psychological Bulletin, 140,* 544–586. http://dx.doi.org/10.1037/a0034205

Klein, R. A., Vianello, M., Hasselman, F., Alper, S., Aveyard, M., Axt, J. R., & Nosek, B. A. (2015). *Many Labs 2: Investigating variation in replicability across sample and setting*. Retrieved from http://projectimplicit.net/nosek/ML2protocol.pdf

*Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137,* 616–642. http://dx.doi.org/10.1037/a0023557

*Kredlow, M. A., Unger, L. D., & Otto, M. W. (2016). Harnessing reconsolidation to weaken fear and appetitive memories: A meta-analysis of post-retrieval extinction effects. *Psychological Bulletin, 142,* 314–336. http://dx.doi.org/10.1037/bul0000034

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE, 9*(9), e105825. http://dx.doi.org/10.1371/journal.pone.0105825

*Kuykendall, L., Tay, L., & Ng, V. (2015). Leisure engagement and subjective well-being: A meta-analysis. *Psychological Bulletin, 141,* 364–403. http://dx.doi.org/10.1037/a0038508

*Landau, M. J., Kay, A. C., & Whitson, J. A. (2015). Compensatory control and the appeal of a structured world. *Psychological Bulletin, 141,* 694–722. http://dx.doi.org/10.1037/a0038703

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8,* 424–432. http://dx.doi.org/10.1177/1745691613491437

LeBel, E. P., Vanpaemel, W., McCarthy, R. J., Earp, B. D., & Elson, M. (2017). A unified framework to quantify the trustworthiness of empirical research. *Advances in Methods and Practices in Psychological Science*. Retrieved from https://osf.io/preprints/psyarxiv/uwmr88/23/2017

*Lee, E.-S., Park, T.-Y., & Koo, B. (2015). Identifying organizational identification as a basis for attitudes and behaviors: A meta-analytic review. *Psychological Bulletin, 141,* 1049–1080. http://dx.doi.org/10.1037/bul0000012

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26,* 1827–1832. http://dx.doi.org/10.1177/0956797615616374

*Lui, P. P. (2015). Intergenerational cultural conflict, mental health, and educational outcomes among Asian and Latino/a Americans: Qualitative and meta-analytic review. *Psychological Bulletin, 141,* 404–446. http://dx.doi.org/10.1037/a0038449

*Lull, R. B., & Bushman, B. J. (2015). Do sex and violence sell? A meta-analytic review of the effects of sexual and violent media and ad content on memory, attitudes, and buying intentions. *Psychological Bulletin, 141,* 1022–1048. http://dx.doi.org/10.1037/bul0000018

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7,* 537–542. http://dx.doi.org/10.1177/1745691612460688

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

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70,* 487–498. http://dx.doi.org/10.1037/a0039400

*Mazei, J., Hüffmeier, J., Freund, P. A., Stuhlmacher, A. F., Bilke, L., & Hertel, G. (2015). A meta-analysis on gender differences in negotiation outcomes and their moderators. *Psychological Bulletin, 141,* 85–104. http://dx.doi.org/10.1037/a0038184

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21,* 47–60. http://dx.doi.org/10.1037/met0000036

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11,* 730–749. http://dx.doi.org/10.1177/1745691616662243

*Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin, 140,* 409–433. http://dx.doi.org/10.1037/a0033890

*Melby-Lervåg, M., Lyster, S. A., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin, 138,* 322–352. http://dx.doi.org/10.1037/a0026744

*Mendelson, J. L., Gates, J. A., & Lerner, M. D. (2016). Friendship in school-age boys with autism spectrum disorders: A meta-analytic summary and developmental, process-based model. *Psychological Bulletin, 142,* 601–622. http://dx.doi.org/10.1037/bul0000041

*Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137,* 267–296. http://dx.doi.org/10.1037/a0021890

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology, 9,* 2. http://dx.doi.org/10.1186/1471-2288-9-2

*North, M. S., & Fiske, S. T. (2015). Modern attitudes toward older adults in the aging world: A cross-cultural meta-analysis. *Psychological Bulletin, 141,* 993–1021. http://dx.doi.org/10.1037/a0039469

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716–aac4716. http://dx.doi.org/10.1126/science.aac4716

*Orquin, J. L., & Kurzban, R. (2016). A meta-analysis of blood glucose effects on human decision making. *Psychological Bulletin, 142,* 546–567. http://dx.doi.org/10.1037/bul0000035

*Ottaviani, C., Thayer, J. F., Verkuil, B., Lonigro, A., Medea, B., Couyoumdjian, A., & Brosschot, J. F. (2016). Physiological concomitants of perseverative cognition: A systematic review and meta-analysis. *Psychological Bulletin, 142,* 231–259. http://dx.doi.org/10.1037/bul0000036

*Pahlke, E., Hyde, J. S., & Allison, C. M. (2014). The effects of single-sex compared with coeducational schooling on students' performance and attitudes: A meta-analysis. *Psychological Bulletin, 140,* 1042–1072. http://dx.doi.org/10.1037/a0035740

*Pan, S. C., & Rickard, T. C. (2015). Sleep and motor learning: Is there room for consolidation? *Psychological Bulletin, 141,* 812–834. http://dx.doi.org/10.1037/bul0000009

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536. http://dx.doi.org/10.1177/1745691612463401

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530. http://dx.doi.org/10.1177/1745691612465253

Patil, P., & Leek, J. T. (2015). *Reporting of 36% of studies replicate in the media*. Retrieved from https://github.com/jtleek/replication_paper/blob/gh-pages/in_the_media.md

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11,* 539–544. http://dx.doi.org/10.1177/1745691616646366

*Phillips, W. J., Fletcher, J. M., Marks, A. D. G., & Hine, D. W. (2016). Thinking styles and decision making: A meta-analysis. *Psychological Bulletin, 142,* 260–290. http://dx.doi.org/10.1037/bul0000027

Pigott, T. (2012). *Advances in meta-analysis*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2278-5

*Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin, 142,* 79–106. http://dx.doi.org/10.1037/bul0000026

Poole, C., & Greenland, S. (1999). Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology, 150,* 469–475. http://dx.doi.org/10.1093/oxfordjournals.aje.a010035

Popper, K. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Psychonomic Society. (2012). *New statistical guidelines for journals of the psychonomic society*. Retrieved from http://www.psychonomic.org/page/statisticalguideline

*Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin, 140,* 1411–1431. http://dx.doi.org/10.1037/a0037428

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

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656. http://dx.doi.org/10.1037/0022-006X.58.5.646

*Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin, 141,* 213–235. http://dx.doi.org/10.1037/bul0000006

Scargle, J. D. (2000). Publication bias: The "File-Drawer" problem in scientific inference. *Journal of Scientific Exploration, 14,* 91–106.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Los Angeles, CA: Sage. http://dx.doi.org/10.4135/9781483398105

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or something else? *Archives of Scientific Psychology, 4,* 32–37. http://dx.doi.org/10.1037/arc0000029

*Sedlmeier, P., Eberth, J., Schwarz, M., Zimmermann, D., Haarig, F., Jaeger, S., & Kunze, S. (2012). The psychological effects of meditation: A meta-analysis. *Psychological Bulletin, 138,* 1139–1171. http://dx.doi.org/10.1037/a0028168

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316. http://dx.doi.org/10.1037/0033-2909.105.2.309

*Sheeran, P., Harris, P. R., & Epton, T. (2014). Does heightening risk appraisals change people's intentions and behavior? A meta-analysis of experimental studies. *Psychological Bulletin, 140,* 511–543. http://dx.doi.org/10.1037/a0033065

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

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

*Smith, S. F., & Lilienfeld, S. O. (2015). The response modulation hypothesis of psychopathy: A meta-analytic and narrative analysis. *Psychological Bulletin, 141,* 1145–1177. http://dx.doi.org/10.1037/bul0000024

*Soderberg, C. K., Callahan, S. P., Kochersberger, A. O., Amit, E., & Ledgerwood, A. (2015). The effects of psychological distance on abstraction: Two meta-analyses. *Psychological Bulletin, 141,* 525–548. http://dx.doi.org/10.1037/bul0000005

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys, 19,* 309–347.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics, 70,* 103–127.

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8,* 581–591. http://dx.doi.org/10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. Oxford, United Kingdom: Routledge.

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

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine, 34,* 2116–2127. http://dx.doi.org/10.1002/sim.6481

Stanley, T. D., & Doucouliagos, H. (2017). Neither fixed nor random: Weighted least squares meta-regression. *Research Synthesis Methods, 8,* 19–42. http://dx.doi.org/10.1002/jrsm.1211

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in Medicine, 36,* 1580–1598.

Stanley, T. D., Doucouliagos, H., & Jarrell, S. B. (2008). Meta-regression analysis as the socio-economics of economics research. *Journal of Socio-Economics, 37,* 276–292. http://dx.doi.org/10.1016/j.socec.2006.12.030

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science, 9,* 305–318. http://dx.doi.org/10.1177/1745691614528518

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

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49,* 108–112.

Sutton, A. J., Song, F., Gilbody, S. M., & Abrams, K. R. (2000). Modelling publication bias in meta-analysis: A review. *Statistical Methods in Medical Research, 9,* 421–445. http://dx.doi.org/10.1177/096228020000900503

*Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracín, D. (2015). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin, 141,* 1178–1204. http://dx.doi.org/10.1037/a0039729

*Toosi, N. R., Babbitt, L. G., Ambady, N., & Sommers, S. R. (2012). Dyadic interracial interactions: A meta-analysis. *Psychological Bulletin, 138,* 1–27. http://dx.doi.org/10.1037/a0025767

Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology, 6,* 726. http://dx.doi.org/10.3389/fpsyg.2015.00726

*Vachon, D. D., Lynam, D. R., & Johnson, J. A. (2014). The (non)relation between empathy and aggression: Surprising results from a meta-analysis. *Psychological Bulletin, 140,* 751–773. http://dx.doi.org/10.1037/a0035236

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11,* 713–729. http://dx.doi.org/10.1177/1745691616650874

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 113,* 6454–6459. http://dx.doi.org/10.1073/pnas.1521897113

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in psychological bulletin from 1990–2013. *Journal of Open Psychology Data, 5,* 4. http://dx.doi.org/10.5334/jopd.33

*Verhage, M. L., Schuengel, C., Madigan, S., Fearon, R. M. P., Oosterman, M., Cassibba, R., . . . van IJzendoorn, M. H. (2016). Narrowing the transmission gap: A synthesis of three decades of research on intergenerational transmission of attachment. *Psychological Bulletin, 142,* 337–366. http://dx.doi.org/10.1037/bul0000038

*von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin, 139,* 841–869. http://dx.doi.org/10.1037/a0030746

*Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140,* 1174–1204. http://dx.doi.org/10.1037/a0036620

*Vukasović, T., & Bratko, D. (2015). Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin, 141,* 769–785. http://dx.doi.org/10.1037/bul0000017

*Wanberg, C. R., Kanfer, R., Hamann, D. J., & Zhang, Z. (2016). Age and reemployment success after job loss: An integrative model and meta-analysis. *Psychological Bulletin, 142,* 400–426. http://dx.doi.org/10.1037/bul0000019

*Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin, 142,* 472–497. http://dx.doi.org/10.1037/bul0000030

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology, 7,* 1832. http://dx.doi.org/10.3389/fpsyg.2016.01832

*Williams, M. J., & Tiedens, L. Z. (2016). The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin, 142,* 165–197. http://dx.doi.org/10.1037/bul0000039

*Winer, E. S., & Salem, T. (2016). Reward devaluation: Dot-probe meta-analytic evidence of avoidance of positive information in depressed persons. *Psychological Bulletin, 142,* 18–78. http://dx.doi.org/10.1037/bul0000022

Yuan, K., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30,* 141–167. http://dx.doi.org/10.3102/10769986030002141

Appendix

Distribution of Survey Estimates

Year	Number of meta-analyses published	Number of meta-analyses sampled	% of published meta-studies	Number of estimates	% of sample
2016	17	14	82%	46	22%
2015	22	18	82%	40	19%
2014	25	12	48%	46	22%
2013	17	3	18%	16	8%
2012	15	6	40%	34	16%
2011	19	8	42%	31	15%

Received May 4, 2017
Revision received July 2, 2018
Accepted July 21, 2018 ■