The Annals of Applied Statistics
2021, Vol. 15, No. 1, 208–229
https://doi.org/10.1214/20-AOAS1387

AN EVALUATION OF STATISTICAL METHODS FOR AGGREGATE PATTERNS OF REPLICATION FAILURE

BY JACOB M. SCHAUER[1], KAITLYN G. FITZGERALD[2,*], SARAH PEKO-SPICER[2,†], MENA C. R. WHALEN[2,‡], RRITA ZEJNULLAHI[2,§] AND LARRY V. HEDGES[2,¶]

[1]*Feinberg School of Medicine, Northwestern University, jms@northwestern.edu*
[2]*Department of Statistics, Northwestern University, *kgfitzgerald@u.northwestern.edu; †SarahPekoSpicer2020@u.northwestern.edu; ‡menawhalen2020@u.northwestern.edu; §RritaZejnullahi2020@u.northwestern.edu; ¶l-hedges@northwestern.edu*

Several programs of research have sought to assess the replicability of scientific findings in different fields, including economics and psychology. These programs attempt to replicate several findings and use the results to say something about large-scale patterns of replicability in a field. However, little work has been done to understand the analytic methods used to do this, including what they are assessing and what their statistical properties are. This article examines several methods that have been used to study patterns of replicability in the social sciences. We describe in concrete terms how each method operationalizes the idea of "replication" and examine various statistical properties, including bias, precision and statistical power. We find that some analytic methods rely on an operational definition of replication that can be misleading. Other methods involve more sound definitions of replication, but most of these have limitations, such as large bias and uncertainty or low power. The findings suggest that we should use caution interpreting the results of such analyses and that work on more accurate methods may be useful to future replication research efforts.

1. Introduction. The replication crisis in science, particularly in psychology, has involved efforts to empirically replicate scientific findings. Though not the first such programs, the Replication Project: Psychology (RPP) (Open Science Collaboration (2015)) and the Replication Project: Economics (RPE) (Camerer et al. (2016)) have been among the most prominent in this discussion. Both of these took a set of findings and attempted a single replication of each: the RPE involved 18 findings of different phenomena, while the RPP attempted to replicate 100 findings. These programs were influential in shaping how we think about replicability, as various research programs have likewise attempted to replicate multiple findings (e.g., Camerer et al. (2018); Klein et al. (2018)). The results of such programs remain among the most commonly cited evidence of a crisis. For example, the results of the RPP have widely been interpreted to indicate that 61% of their replication attempts failed by both the academic literature and popular press (e.g., Yong (2016); Wood and Randall (2018)). Similarly concerning are reports that replication studies in the social sciences tend to find effects that are between 46% and 64% smaller than original studies (Camerer et al. (2018)).

Yet, figures like that, 61% failure rate or 64% decrease in effect sizes, are without context: they arise from statistical analyses, and their interpretation must take into account at least two aspects of those analyses. First, any analysis method for replication depends on a precise operational definition of what it means for a finding (or several findings) to "replicate." While the idea of "replication" might seem intuitive, it is often difficult to define it precisely (see Bollen et al. (2015); Shapin and Shaffer (1985)). Perhaps because of this, researchers seldom

Received March 2020; revised August 2020.
Key words and phrases. Replication, meta-analysis, power, bias.

specify a concrete definition, and often it is up to the reader to discern it. Moreover, different analysis methods can rely on different or even conflicting definitions of replication. Second, the results of statistical analyses are subject to error. Null hypothesis tests, for instance, can produce type I and type II errors. Estimates may have bias, and, even if they are unbiased, they still must be viewed in light of their statistical uncertainty. Thus, in order to understand a statistic like that 61% failure rate, we need to know: (1) what it means for replications to fail, and (2) how accurate the methods are that produced that statistic.

When researchers attempt replications of findings for several different phenomena, there are at least two ways to talk about analysis methods. The first involves determining whether a specific replication study failed (i.e., for a single finding). We call these *pairwise* analyses, and programs like the RPP and RPE have used a variety of such methods. The most common pairwise analysis concludes that a replication failed if it disagrees with the original study in sign or statistical significance (e.g., the original study is significant, but the replication is not). Various researchers have challenged pairwise analysis methods common in replication research (e.g., Etz and Vandekerckhove (2016); Hartgerink, Wicherts and van Assen (2017); van Aert and van Assen (2017); Hedges and Schauer (2019a, 2019b)). However, such challenges largely focus on proposing alternative methods rather than clarifying the properties of existing ones.

The second class of methods is less concerned with individual replication studies but rather on the entire group of findings. We call these *groupwise* methods, and they quantify the extent to which a series of different findings were successfully replicated. For instance, inferences are often framed in terms of whether original and replication studies involve similar effects on average (see Open Science Collaboration (2015); Camerer et al. (2016)). Very little attention has been paid to groupwise analysis methods. This is important, because the results of these methods are often used to characterize the replication crisis, but their statistical properties are seldom understood.

This article examines groupwise analysis methods that researchers have used to assess the replicability of several findings. Our goal is to shed light on the properties of these methods so that we can better understand the results of empirical research. We focus primarily on six methods used in RPP, RPE and the Replication Project: *Science* and *Nature* (RPSN) (Camerer et al. (2018)). We also consider one other method that was proposed to address some of the shortcomings of these methods which was also used by the RPE and RPSN (Patil, Peng and Leek (2016)). While these are not the only relevant groupwise analysis methods, they have been used to support some prominent claims about the replication crisis in science. Though we are less concerned with proposing a litany of alternative methods, we do discuss potential corrections (where possible). The following sections outline the types of replication research relevant to groupwise analysis methods and describe a relevant statistical model to formalize analyses of replication. Then, for a variety of analytic methods that replication research programs have used, we examine what they are attempting to assess and delineate some of their statistical properties. These methods, displayed in Table 1, include estimating the proportion of failed replications as well as comparisons of effect sizes and p-values from original and replication studies. For each method we highlight its properties under plausible sets of conditions, including with data from the RPE, RPP and RPSN. In some cases we find that the methods focus on conceptions of "replication" that may be misleading and that many tend to have poor statistical properties.

2. Data. It has been increasingly common for researchers to attempt replications of several findings as part of the same program of research. The RPE, RPP and RPSN did this, as have several other major replication research programs (e.g., Klein et al. (2014, 2019); Schweinsberg et al. (2016)). Such programs have used a variety of group- and pairwise methods to assess replicability. The groupwise analyses involve reporting the mean relative effect

size (i.e., the average ratio of the replication effects to the original study effects), examining the correlation and average difference between effects from the original and replication studies and comparing differences in the p-values from original and replication studies. These methods, listed in Table 1, are described in detail throughout this article.

The properties of many analysis methods discussed here will depend on how precisely effects are estimated in each experiment which will, in turn, depend on how large those studies are. In order to demonstrate these properties under realistic conditions, we use data from the RPE, RPP and RPSN. These research programs each used at least one (and often more than one) of the analysis methods discussed in this article. These programs also have publicly available datasets at the Open Science Framework (see https://osf.io). From these data we have extracted relevant information related to how precise each study was (on the scale of Cohen's d). We use these not to conduct any reanalysis of the RPE, RPP or RPSN but rather to demonstrate the properties of the analysis methods they used. All of our data and code are available as a supplement to this article and are available at online (Schauer et al. (2021)).

Although RPE, RPP and RPSN conducted their analyses in the metric of effect sizes transformed into correlation coefficients, we conducted our analyses in the metric of standardized mean differences for two reasons. First, most of the data arises from between-group experiments, for which the standardized mean difference seems to be a more direct and mathematically natural effect size than the correlation coefficient. Second, the sampling distribution of the standardized mean difference, when transformed to the metric of the correlation coefficient, is not the same as that of a directly computed correlation coefficient (see Borenstein et al. (2009), pages 48–49). Therefore, the salutary properties of the Fisher z-transform (normalization and variance stabilization) do not hold for these transformed "correlations."

3. Model and notation. Analyses of replication can be understood within the framework of meta-analysis which is the statistical methodology for combining information from multiple (i.e., two or more) studies (see Borenstein et al. (2009)). The models commonly used in meta-analysis can help clarify important aspects of analyses of replication (Hedges and Schauer (2019b); Schauer (2018); Valentine et al. (2011)).

Suppose we are interested in the replicability of a population of N findings and that a subset of $m \leq N$ findings are selected to be replicated. The analyses considered here assume that there are $k = 2$ studies per finding, an original study and a replication study. In this article, "finding" refers to a specific phenomenon under investigation, and "study" refers to experiments used to investigate a finding; for instance, the RPP had $m = 100$ findings each with $k = 2$ studies (the original and replication studies). When multiple replication studies are conducted for a finding, their results are often aggregated into a single result, such as with the Many Labs Replication Projects (Klein et al. (2014, 2018)).

3.1. *Parameters of interest.* Let θ_{ij} be the effect in study $i = 1, 2$ for finding $j = 1, \ldots, N$. We assume that θ_{ij} is on the scale of one of the standard effect sizes used in meta-analysis, such as standardized mean differences or z-transformed correlations (see Cooper, Hedges and Valentine (2009)). The effect θ_{ij} is what would be observed in study i of finding j in the absence of any estimation error, such as from the sampling of experimental units.

The θ_{ij} are the scientific estimands of interest in each study, and so replication should be defined as some function of the θ_{ij} (see Hedges and Schauer (2019a, 2019b)). For a single finding j, replication *failure* typically involves effects that disagree in size (i.e., $\theta_{1j} \neq \theta_{2j}$) or in sign (e.g., $\theta_{1j} > 0$ but $\theta_{2j} \leq 0$) (see Bollen et al. (2015)). It stands to reason that aggregate definitions of replication (across N findings) ought to be somewhat compatible with these pairwise definitions.

Precisely defining replication across a series of N findings requires at least one additional consideration about the θ_{ij}: are they fixed or random? One reason to treat the $[\theta_{1j}, \theta_{2j}]$ as random is if the m findings to be replicated are randomly selected from a population of N findings. If we treat the $[\theta_{1j}, \theta_{2j}]$ as random, one appropriate model is the multivariate random effects model used in meta-analysis (see Hedges and Olkin (1985); Olkin and Gleser (1994); Raudenbush et al. (1988)). This assumes $[\theta_{1j}, \theta_{2j}]$ are exchangeable draws from a distribution with mean $[\mu_1, \mu_2]$, marginal variances τ_1^2 and τ_2^2 and correlation ρ. Note that μ_i, τ_i and ρ are attributes of the population of N findings and vectors $[\theta_{1j}, \theta_{2j}]$ from which the selected findings are a sample, and so inferences about replication pertain to that population.

In practice, it will often be difficult to generalize from the sample of m findings to the entire population of N findings. Research programs seldom sample findings randomly but rather select them because their findings are of interest or the source of skepticism. Even when programs identify findings to replicate using quasi-probability sampling (e.g., Open Science Collaboration (2015)), there are reasons to suspect that these samples are not necessarily representative of an entire field (see Gilbert et al. (2016)).

Instead, one may treat the m findings, for which replications are conducted as the entire population of interest, so that $N = m$. This is equivalent to treating the θ_{ij} as fixed but unknown constants, and inferences pertain only to the m findings. However, we can use similar notation as the random effects model, denoting the mean of $[\theta_{1j}, \theta_{2j}]$ as $[\mu_1, \mu_2]$, the marginal variances τ_1^2 and τ_2^2 and their correlation ρ. Here, the mean and variance are not properties of random variables but rather are descriptive statistics of the m vectors $[\theta_{1j}, \theta_{2j}]$. In this article we will mostly treat the θ_{ij} as fixed. While this leads to different conceptions of analyses, Hedges and Schauer (2019b) argue that fixed- and random-effects replication analyses tend to have relatively similar properties and that the parameters are analogous between the fixed- and random-effects models. Note that the variance components τ_i^2 represent variation in true effect sizes across findings that measure fundamentally different effects. This differs from variance components, usually encountered in meta-analysis, where experiments are (at some level of generality) estimating the same effects. Thus, the size of the τ_i^2 values in this paper depends on how findings subject to replication attempts are selected.

In this article we show that common groupwise analyses of replication often, but not always, frame replication as a function of the θ_{ij}. An increasingly common metric for quantifying replication success or failure is the mean relative effect size (MRES) which can be expressed as

$$\eta = \sum_{j=1}^{m} \frac{\theta_{2j}/\theta_{1j}}{m}. \tag{1}$$

Researchers have also examined quantities related to the distribution of θ_{ij}. For instance, researchers appear interested in the correlation ρ between θ_{1j} and θ_{2j} and have examined the differences between effects $\delta_j = \theta_{1j} - \theta_{2j}$, including the average difference $\mu_\delta = \mu_1 - \mu_2$.

3.2. *Statistical model and estimates.* While the parameters above are used in framing the definition of "replication," what makes analyses difficult is that we do not actually observe θ_{ij} directly but, instead, must estimate them; analyses of replication must also rely on these estimates. Let T_{ij} be the estimate of θ_{ij}. A useful assumption is that T_{ij} is unbiased and normally distributed with known variance v_{ij},

$$T_{ij}|\theta_{ij} \sim N(\theta_{ij}, v_{ij}). \tag{2}$$

This is an accurate approximation for most effect sizes (see Cooper, Hedges and Valentine (2009)), including standardized mean differences (Cohen's d) which is the scale we use to

report results in this article. Note that δ_j is often estimated by $D_j = T_{1j} - T_{2j}$, under the model $D_j|\delta_j \sim N(\delta_j, v_{1j} + v_{2j})$.

Researchers have also assessed replication based on p-values from original and replication studies. These p-values typically arise from a test that $\theta_{ij} = 0$. In this article we assume two-sided p-values, so that under the model, the p-value for study ij is given by

$$p_{ij} = 2\left[1 - \Phi\left(\frac{|T_{ij}|}{\sqrt{v_{ij}}}\right)\right], \tag{3}$$

where $\Phi(x)$ is the standard normal distribution function.

The probability of a statistically significant result in study ij is given by

$$1 - \beta_{ij} = 1 - \Phi\left(c_{1-\alpha/2} - \frac{\theta_{ij}}{\sqrt{v_{ij}}}\right) + \Phi\left(c_{\alpha/2} - \frac{\theta_{ij}}{\sqrt{v_{ij}}}\right), \tag{4}$$

where c_x is the xth percentile of the standard normal distribution and α is the significance level, which we assume is $\alpha = 0.05$ throughout this article. Note that when $\theta_{ij} \neq 0$, then $1 - \beta_{ij}$ is the power of the test, and when $\theta_{ij} = 0$, then $1 - \beta_{ij}$ is the significance level of the test.

From equation (4) we can see that the power of any one study will depend on $|\theta_{ij}|/\sqrt{v_{ij}}$, and thus on θ_{ij}^2/v_{ij}. This is because the distribution of p_{ij} will depend on θ_{ij}^2/v_{ij}. Bahadur (1960) and Lambert and Hall (1982) show that p-values are asymptotically log-normal when the null hypothesis is false (i.e., when $\theta_{ij} \neq 0$). Based on their results, it can be shown that when $\theta_{ij} \neq 0$, then $-2\log(p_{ij})$ has an asymptotic distribution that is normal,

$$-2\log(p_{ij}) \sim \mathrm{AN}\left(\frac{\theta_{ij}^2}{v_{ij}}, 2\frac{\theta_{ij}^2}{v_{ij}}\right). \tag{5}$$

Some methods involve averages of p-values which will depend on the averages of the θ_{ij}^2/v_{ij}. For $j = 1, \ldots, m$, denote the average of the θ_{1j}^2/v_{1j} as λ_1 and the average of the θ_{2j}^2/v_{2j} as λ_2.

Table 1 highlights the methods that this article examines. These are methods that research programs, including those whose data are used in this article, have used to assess replication. The table describes each method, highlights how it defines replication in terms of the parameters discussed in this section, how those definitions are assessed and any glaring strengths or limitations (which are discussed throughout this article).

There are two different approaches to defining replication used in the methods we discuss. One approach aggregates comparisons (e.g., ratios or differences) among θ_{1j} and θ_{2j} values. The other approach involves comparisons of the collection of θ_{1j} values with those of the θ_{2j} values (e.g., comparisons of average properties of θ_{1j} values with those of θ_{2j} values, p-values or the correlation between the θ_{1j} and θ_{2j}). Methods that aggregate the comparisons between θ_{1j} and θ_{2j} do not depend on the variation between the θ_{1j}'s or the variation between the θ_{2j}'s and thus have exactly the same properties whether we consider the θ_{ij}'s fixed or varying randomly across values of j. On the other hand, the properties of methods that involve the variation of the θ_{ij}'s across values of j do depend on the distribution of the θ_{ij}'s, a point we try to clarify in our discussion.

4. Mean relative effect size. Replication research programs have reported the mean relative effect size which is used to show how much larger or smaller effects in the replication studies are, on average, relative to the effects in the original studies. This method frames replication in terms of the mean of the θ_{2j}/θ_{1j} and uses the mean of the T_{2j}/T_{1j} to estimate it. Note that this analysis depends on ratios which can be difficult to work with statistically.

TABLE 1

This table summarizes the methods examined in this paper. For each method the table highlights the type of replication pattern it is attempting to assess and lists any important limitations

Method	Definition	Estimator	Primary limitation
Mean relative effect size (MRES) *Estimate the average ratio of replication study effects to original study effects*	$\eta = \sum_{j=1}^{m} \frac{\theta_{2j}/\theta_{1j}}{m}$	$H = \sum_{j=1}^{m} \frac{T_{2j}/T_{1j}}{m}$	**Large uncertainty:** H can be close to 0 when replications succeed and $\eta = 1$.
Correlation between effects *Determine if replication studies and original studies produce effects that are correlated.*	$\rho = \text{Cor}(\theta_{1j}, \theta_{2j})$	$r = \text{Cor}(T_{1j}, T_{2j})$	**Inconsistent definition of replication:** High correlation between effect parameters does not mean that they are similar in size **Bias:** Reported correlation will be downwardly biased
Paired tests of effects *Determine if replication studies and original studies produce different effect sizes on average.*	$E[\theta_{1j} - \theta_{2j}] = \mu_\delta = 0$	$\sum_{j=1}^{m} \frac{T_{1j} - T_{2j}}{m}$	**Inconsistent definition of replication:** All findings can fail to replicate, but averaging across findings ignores this.
Prediction interval coverage *Determine the proportion of replication studies in the 95% prediction interval of the original study.*	$\pi = P[\frac{\lvert T_{1j} - T_{2j}\rvert}{\sqrt{v_{1j} + v_{2j}}} < 1.96]$	$p = \sum_{j=1}^{m} \frac{\mathbb{1}[\frac{\lvert T_{1j} - T_{2j}\rvert}{\sqrt{v_{1j} + v_{2j}}} < 1.96]}{m}$	**Not sensitive to replication failures:** Wide range of $\lvert\theta_{1j} - \theta_{2j}\rvert$ values lead to large values of p
Fisher's method *Assess if nonsignificant replication studies are actually false negatives.*	$\theta_{1j} \neq 0 \wedge \theta_{2j} \neq 0$	$X_F^2 = \sum_{j:T_{2j}\text{ null}} -2\log(p_{2j})$	**Low Power:** Requires many false negatives that are each highly powered in order to achieve adequate power
McNemar's test *Determine if replication studies are significant at a different rate than original studies.*	$\beta_1 = \beta_2$	$X_M^2 = \frac{(m_{10} - m_{01})^2}{m_{10} + m_{01}}$	**Inconsistent definition of replication:** Requires the power of the original and replication studies to be equal, but not the effects
Tests of p-value means *Determine if p-values for original studies and replications have the same mean.*	$\beta_1 = \beta_2$	$t_p = \sum_{j=1}^{m} \frac{p_{1j} - p_{2j}}{m}$	**Inconsistent definition of replication:** Requires the power of the original and replication studies to be equal, but not the effects

This section shows that, because of this, the reported mean relative effect size (i.e., the mean of the T_{2j}/T_{1j}) can be highly inaccurate.

This analysis concerns the mean relative effect size, which refers to the average of θ_{2j}/θ_{1j}, denoted as η in equation (1). Two caveats are worth noting here. First, if any of the $\theta_{1j} = 0$, then η will not be defined. Second, η is different than μ_2/μ_1. But, assuming $\theta_{1j} \neq 0$, then η provides an intuitive scale on which to quantify replication. When the replications (mostly) succeed so that $\theta_{2j} = \theta_{1j}$, then η would be near 1.0. When the original study produces a much larger effect than the replication (i.e., $\theta_{1j} \gg \theta_{2j}$), then η will be closer to 0. The quantity η is a summary statistic of the θ_{2j}/θ_{1j}, and the θ_{2j}/θ_{1j} may vary for each finding j. Therefore, it is possible for their mean to be 1 even if all of the θ_{2j}/θ_{1j} are themselves quite different from 1.

The reported mean relative effect size has been used as an estimate of η,

$$H = \sum_{j=1}^{m} \frac{T_{2j}/T_{1j}}{m}. \tag{6}$$

A key point is that H is an estimator of the actual mean relative effect size η, and so it must be interpreted in light of its accuracy and precision. When studies largely replicate, so that $\eta = 1$, then we would want H to be close to 1 with high probability. But if H were, say, very small (e.g., less than 0.1) or very large (e.g., greater than 2) with high probability, then we would worry about the accuracy of H as an estimator for η because it would indicate that the studies, mostly, failed to replicate. Similarly, if the effects in original studies were typically much larger than the effects in replication studies, so that $\eta = 0$, then we would want H to be near zero with high probability; values of H that were near 1 (i.e., greater than 0.9) would be inaccurate in this case because that would indicate that the replications were largely successful when they were not.

The distribution of H is not known; however it will depend on the T_{2j}/T_{1j}. If T_{2j}/T_{1j} are poor estimates of θ_{2j}/θ_{1j}, then H will likely be a poor estimator of η. Under the model, T_{2j}/T_{1j} is a ratio of normal random variables which has been studied thoroughly in the statistical literature. The exact distribution of T_{2j}/T_{1j}, which is quite complex, was derived and studied by various researchers, including Geary (1930) and Fieller (1932). The shape of this distribution, which largely depends on $\theta_{1j}/\sqrt{v_{1j}}$, can be unimodal or bimodal and asymmetric or symmetric (see Diaz-Frances and Rubio (2013)).

An important aspect of the distribution of T_{2j}/T_{1j} is that its moments (i.e., its mean and variance) do not exist since T_{1j} has a nonzero probability of being zero. Because the mean of T_{2j}/T_{1j} does not exist, neither does the mean of H which means that H cannot be an unbiased estimator of η. Under certain conditions, T_{2j}/T_{1j} approximately follows a normal distribution with mean θ_{2j}/θ_{1j} (see Diaz-Frances and Rubio (2013); Geary (1930); Hayya et al. (1975); Marsaglia (2006)). Given those results, it can be shown that H is asymptotically normal with mean η and a variance that depends on the θ_{2j}/θ_{1j} and each $\theta_{1j}/\sqrt{v_{1j}}$ and $\theta_{2j}/\sqrt{v_{2j}}$. While this would seem to imply unbiasedness, at least up to an asymptotic approximation, simulations have found that this approximation is only accurate when both $\theta_{1j}/\sqrt{v_{1j}}$ and $\theta_{2j}/\sqrt{v_{2j}}$ are large so that both studies have exceptionally high power. Diaz-Frances and Rubio (2013) found that the approximation was only "good" when each study had over 95% power. Since studies in the social sciences are seldom that high-powered (see Maxwell (2004); Vankov, Bowers and Munafò (2014)), this approximation will likely not be accurate when applied to replication studies.

In addition to bias, another issue is the variability of H. Because the variance of T_{2j}/T_{1j} does not exist, neither does the variance of H, and hence its standard error is undefined. But just because it is undefined does not mean that uncertainty in H can be ignored. A well-known property of ratio distributions is that they are notoriously heavy-tailed. Because

of this, even if $\theta_{2j}/\theta_{1j} = 1$, very large (i.e., greater than 2.0) or very small (i.e., less than 0.1) values of T_{2j}/T_{1j} can occur with surprisingly high probability. As an example, suppose $\theta_{1j} = \theta_{2j} = 0.2$ in Cohen's d units so that their ratio is 1.0 and that both studies had a sample size of 80 so that $v_{1j} = v_{2j} \approx 4/80$. Then, the probability that $T_{2j}/T_{1j} < 0.1$ is about 33%, and the probability that $T_{2j}/T_{1j} > 2.0$ is about 18%. In other words, there is higher than a 50% chance that the value of T_{2j}/T_{1j} implies that these studies fail to replicate when they successfully replicate.

Often in statistics, including in meta-analysis, averages of noisy estimates tend to have better precision than the estimates themselves; however this is not necessarily the case with ratios. Notably, when $\theta_{1j} = \theta_{2j} = 0$ for each j, then, T_{2j}/T_{1j} follows a Cauchy distribution. The average of m Cauchy random variables is itself Cauchy. This means that H follows the same distribution as T_{2j}/T_{1j} and averaging does nothing to reduce noise. Thus, not only is the distribution of each T_{2j}/T_{1j} heavy-tailed but taking their average does not necessarily result in a less variable statistic.

Though the distribution of H is not known, we studied it with Monte Carlo simulations. These simulations involve drawing m pairs of (T_{1j}, T_{2j}) at random and computing H as in (6). Following the model, we drew T_{ij} from normal distributions. To help tie these simulations to empirical research, we use the estimation variances v_{ij} from studies in the RPE, RPP and RPSN. Thus, these simulations proceed by specifying a θ_{ij} for each study in each program. We then draw $T_{ij} \sim N(\theta_{ij}, v_{ij})$ for each of the $j = 1, \ldots, m$ findings in a program and compute H. This constitutes one draw of H from its distribution for that program (and assuming the θ_{ij} values). We repeat this procedure 100,000 times for different configurations of θ_{ij} and estimated various quantities involving the distribution of H.

Our first simulation concerns how likely very large or very small values of H are to occur when $\eta = 1$. Previous work (Marsaglia (2006); Diaz-Frances and Rubio (2013)) suggests that the results of these simulations will be sensitive to the size of θ_{ij}, particularly the effect parameters of the original studies θ_{1j}. Thus, these simulations set $\theta_{1j} = \theta_{2j} = 0.2, 0.5$ and 0.8 which correspond with conventions of small, medium and large effects in the social sciences.

Figure 1 shows the distribution of H when $\eta = 1$ and $\theta_{1j} = \theta_{2j} = 0.2$, 0.5, or 0.8. Each panel corresponds to a value of θ_{ij}, and each colored region shows the density of H for a given research program. In the left panel, where effects are small ($\theta = 0.2$), the distribution of H is highly variable, and the probability that H is less than 0.1 is about 33% for each program in that plot; the probability H is greater than 2.0 is about 16% for each program.

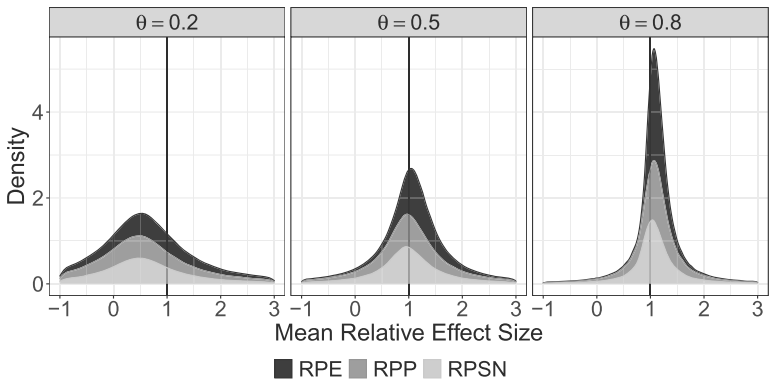

FIG. 1. *This plot shows the density of the estimated mean relative effect size H when the true mean relative effect size is $\eta = 1$. Each panel shows the distribution of H when effects in each pair of studies are the same size and are small ($\theta = 0.2$), medium ($\theta = 0.5$) or large ($\theta = 0.8$). Within each panel the colors correspond to the distribution of H for each replication research program. Note that an accurate estimate would correspond to $H \approx 1$.*

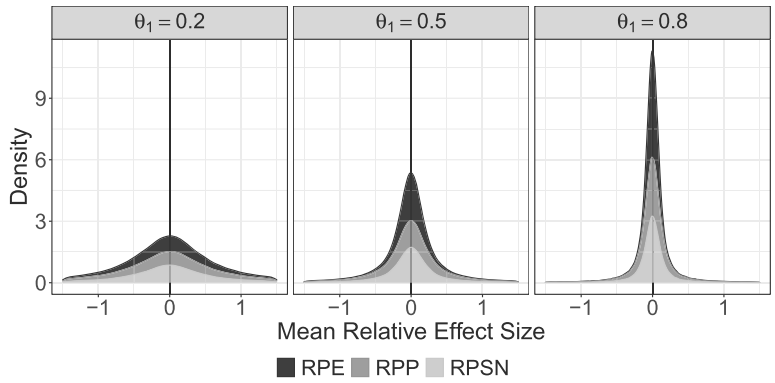

FIG. 2. *This plot shows the density of the estimated mean relative effect size H when the true mean relative effect size is $\eta = 0$. Each panel shows the distribution of H when effects in the replication study are zero (i.e., $\theta_{2j} = 0$) and effects in the original studies are small ($\theta_1 = 0.2$), medium ($\theta_1 = 0.5$) or large ($\theta_1 = 0.8$). Within each panel the colors correspond to the distribution of H for each replication research program. Note that an accurate estimate would correspond to $H \approx 0$, and an inaccurate estimate would correspond to $H \approx 1$ or -1.*

Not only that, when effects are small ($\theta = 0.2$), the mode of the distribution of H is smaller than 1.0 for these research programs. For medium-sized effects ($\theta = 0.5$, middle panel), the distribution of H is less variable, though the probability that H is less than 0.1 is still about 15% for each program. Only when effects are large ($\theta = 0.8$, right panel) do extreme values of H become less probable. In other words, when all of the studies replicate exactly, so that $\theta_{1j} = \theta_{2j}$ and $\eta = 1$, this method will only estimate η accurately when the effects for each finding are large. But if the effects are small, the probability that H is near zero can be very large (over 30%).

Our second simulation involves a scenario where $\theta_{2j} = 0$ and $\theta_{1j} = 0.2$, 0.5, and 0.8. This would correspond to an effect being positive in the original study and zero in the replication study. In that case, $\eta = 0$, and we would want H to be near zero. Figure 2 plots the density of H when $\theta_{2j} = 0$ and $\theta_{1j} = 0.2$, 0.5 and 0.8 which means that $\eta = 0$. Each panel corresponds to a value of θ_{1j} (0.2, 0.5 and 0.8), and each region corresponds to the distribution of H for a given research program. In the right panel, when the effect of the original study is small ($\theta_{1j} = 0.2$) and the replication study effect is zero, H will be particularly variable; the probability that H is less than -0.9 is about 16%, and the probability that H is greater than 0.9 is about 16% which means that the probability that $|H - \eta|$ is greater than 0.9 is nearly 33%. However, when the original study effect parameters are large, so that $\theta_{1j} = 0.8$, then the distribution of H is less likely to be substantially different from zero.

In sum, because ratios of random variables are so noisy when effects are not large, the following can happen: When the studies largely replicate (i.e., $\eta = 1$), H can be near zero with high probability and when the studies largely fail to replicate (i.e., $\eta = 0$), H can be far from zero with high probability. This means that, unless the original study effect parameters are large for the findings considered by each research program, it will be almost impossible to say anything conclusive about η on the basis of the reported mean relative effect size H.

5. Analyses of differences in effects. Though the ratio of two effect estimates can be difficult to work with (see previous section), their difference is often much less noisy. Research programs, like the RPE and RPP (though not the RPSN), have applied paired t- and Wilcoxon tests to the effect-size estimates for the original and replication studies. In this section we argue that this can provide a well-powered test of whether the original studies have the same average effect as the replication studies (i.e., $\mu_1 = \mu_2$). However, we also point out that the focus on means can be a little misleading; a group of original and replication studies

can have the same average effect, even if each original-replication study pair obtains very different effects (i.e., $\mu_1 = \mu_2$, but $\theta_{1j} \neq \theta_{2j}$ for all j).

Paired tests of effect sizes can be understood in terms of the parameters $\delta_j = \theta_{1j} - \theta_{2j}$ and their estimates $D_j = T_{1j} - T_{2j}$; note $D_j \sim N(\delta_j, v_{1j} + v_{2j})$. As discussed in the notation section, we can think of the δ_j as having some distribution with mean μ_δ, which means that paired tests of effect sizes are formally testing,

$$H_0 : \mu_\delta = 0.$$

Rejecting H_0 is taken as a sign of poor replicability. If the δ_j are approximately normally distributed, then one can just compute the paired t-statistic ignoring the v_{ij}, as was done by the RPP. A more powerful version of that test uses a precision-weighted mean of the differences rather than the unweighted mean (see Hedges and Olkin (1985)). Alternatively, if the δ_j are not normally distributed, one can use the Wilcoxon test.

Hedges and Pigott (2001) derive the power of the test that uses a precision-weighted average which will be at least as powerful as the other tests that have been used. They find that with large numbers of studies (or smaller numbers of large studies) that the power of this test will be high. Using their results, the RPP would have had 80% power to detect a difference of $\mu_\delta = 0.05$, and the RPE would have had 80% power to detect a difference of about $\mu_\delta = 0.15$.

While this test helps pool information across studies, it only provides part of the picture. This is because the mean difference between original and replication effects μ_δ is just one summary statistic of an entire distribution; $\mu_\delta = 0$ does not imply that any of the studies replicate successfully. It is possible for $\mu_\delta = 0$, even if $|\delta_j|$ is large for all j: that is, it is possible for all of the replications to have failed dramatically but for the mean difference between effect parameters of original and replication studies to be zero. Moreover, if the distribution of the δ_j has a large variance τ_δ^2, then, even if $\mu_\delta = 0$, large values of $|\delta_j|$ may be probable which would be a sign of poor replicability. Perhaps a more complete analysis would examine the full distribution of δ_j. If the θ_{ij} are treated as random, then so are the δ_j, and hence common methods used with random-effects meta-analyses can provide inference for the mean μ_δ and variance τ_δ^2 or produce prediction intervals for the distribution of the δ_j (see, e.g., Borenstein et al. (2009); Cooper, Hedges and Valentine (2009); Hedges and Vevea (1998); Riley, Higgins and Deeks (2011); Veroniki et al. (2016)). Not only would this provide a more complete understanding about replication across findings, it may prove to be a more statistically precise approach than examining ratios of effect estimates for the reasons described in the previous section.

6. Prediction intervals. A different strategy for comparing original and replicated studies is to evaluate the proportion π of effect sizes of the replicated studies are contained in the $100 \times (1-\alpha)\%$ prediction interval of the original study. A prediction interval, as proposed by Patil, Peng and Leek (2016), is $T_{1j} \pm c_{(1-\alpha/2)}\sqrt{v_{1j} + v_{2j}}$ where $c_{(1-\alpha/2)}$ is the $1-\alpha/2$ percentile of the standard normal distribution. Most prediction interval analyses involve a 95% prediction interval which would mean $\alpha = 0.05$ and $c_{(1-\alpha/2)} \approx 1.96$. "Successful" replication occurs when T_{2j} is contained in that interval. A groupwise aggregation of this approach is equivalent to asking how frequently the difference between T_{1j} and T_{2j} is statistically significant, and the proportion π is the average acceptance rate for the tests (across the m pairs). This strategy has the virtue that, when $\theta_{1j} = \theta_{2j}$ for all $j = 1, \ldots, m$, exactly 95% of the T_{2j} values will lie in the prediction interval.

The weakness of this method is that the acceptance rate of tests between two effect sizes can be relatively large when $\theta_{1j} \neq \theta_{2j}$, even if $\theta_{1j} - \theta_{2j}$ is not negligible. When $\theta_{1j} \neq \theta_{2j}$, the acceptance rate is one minus the power of the tests, and the power of the test for differences

between effects is often rather small, unless the studies have unusually large sample sizes (see Hedges and Schauer (2019a)). For example, suppose that that both studies had a sample size of 80 so that $v_{1j} = v_{2j} \approx 4/80$. Then, if $\theta_{1j} = \theta_{2j}$ for all j, then the probability that T_{2j} is in the 95% prediction interval of T_{1j} is 95%, but if $\theta_{1j} - \theta_{2j} = 0.2$ for all j, then the probability that T_{2j} is in the 95% prediction interval of T_{1j} is 90%, and if $\theta_{1j} - \theta_{2j} = 0.4$ for all j, then the probability that T_{2j} is in the 95% prediction interval of T_{1j} is 76%. This latter figure matches closely the 77% of T_{2j} values that were in the 95% prediction interval based on T_{1j} that Patil, Peng and Leek (2016) computed in their analysis of the RPP studies. Thus, the analysis they conducted is consistent with differences between effect sizes of as much as 0.4 for every finding—a difference that is closer to Cohen's benchmark for a "medium-sized" effect ($d = 0.5$) than a "small effect" ($d = 0.2$). Despite a 77% coverage probability, it seems unlikely that researchers would characterize a difference between a pair of effects as large as 0.4 as a successful replication, let alone a difference that size between every pair of effects.

7. Correlation between effects. Replication has been assessed in terms of the linear relationship between effect estimates T_{1j} and T_{2j}, including numerically with the Pearson or Spearman correlation as well as visually with scatterplots of (T_{1j}, T_{2j}) (e.g., Open Science Collaboration (2015)). This can be seen as assessing replication via the correlation between effect parameters $\rho = \text{Cor}(\theta_{1j}, \theta_{2j})$ which is estimated with the correlation of the effect estimates $r = \text{Cor}(T_{1j}, T_{2j})$. The idea behind this is that if pairs of studies successfully replicate, their effects should be similar, and hence their correlation should be close to 1.0. However, there are two limitations to such analyses. First, even if $\rho = 1$, this does not necessarily mean that $\theta_{1j} = \theta_{2j}$; for instance, if $\theta_{1j} = 100 \times \theta_{2j}$, so that each original effect is 100 times larger than the replication effect, the correlation is still $\rho = 1$. Second, as detailed below, the sample correlation r can have a substantial downward bias.

Because the estimation errors of T_{1j} and T_{2j} are independent, r will tend to underestimate ρ. When the θ_{ij} are treated as fixed, the expectation of r can be written as

$$E[r|\theta_{ij}] \approx \rho \frac{\tau_1 \tau_2}{\sqrt{(\tau_1^2 + \bar{v}_1)(\tau_2^2 + \bar{v}_2)}} < \rho, \tag{7}$$

where $\bar{v}_i = \sum_{j=1}^{m} v_{ij}/m$ is the mean within study variance for the original ($i = 1$) and replication ($i = 2$) studies and τ_i^2 is the variance of the effect parameters for the original ($i = 1$) and replication ($i = 2$) studies, as described in the notation section. Given equation (7), we would expect r to be smaller than ρ, and its bias will increase as a function of v_{ij}/τ_i^2. Figure 3 shows the expected value of r on the y-axis as a function of ρ (x-axis) for each research program. The expected values in the figure are computed using meta-analytic estimates of τ_i^2 for each program and the reported estimation error variance v_{ij}. The light gray line in the figure indicates an unbiased estimate of ρ. Figure 3 shows that, even if all of the studies replicated exactly, these programs would be expected to report a correlation of r less than 0.8, and possibly even below 0.6.

To gain some intuition about the bias of r, suppose that $\tau_i^2 = a\bar{v}_i$ for some constant a. Then, the bias of r can be written as

$$\text{Bias}(r|\theta_{ij}) \approx \frac{-\rho}{1+a}. \tag{8}$$

When a is very small (i.e., near zero), then the bias will be $-\rho$; that is, when $\tau_i^2 \ll v_{ij}$, we would expect r to be near zero, regardless of the value of ρ. However, the bias decreases as a increases which means that, when $\tau_i^2 \gg v_{ij}$, the bias will be smaller. For instance, if $a > 20$, then the bias will be less than 0.05. This is consistent with the fixed-effects logic: if the studies are really large so that $v_{ij} \to 0$ and $a \to \infty$, then we would observe the θ_{ij} with

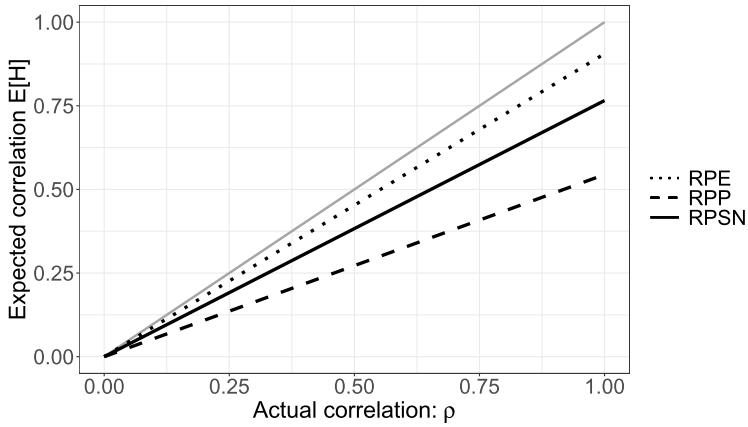

FIG. 3. *This figure shows the approximate expectation of the correlation between effect estimates r (y-axis) as a function of the correlation between effect parameters ρ (x-axis) for the RPE, RPP and RPSN. The light gray line indicates an unbiased estimate, and values below that line mean that r understates ρ.*

almost no error and hence be able to compute ρ without bias. In other words, what is driving the bias of the fixed-effects correlation estimate is that the θ_{ij} are estimated with error by the T_{ij}. Viewed this way, (7) is analogous to the attenuation formula for measurement error and correcting this attenuation has long been studied in the statistical literature (e.g., Muchinsky (1996); Spearman (1904, 1910)).

If the θ_{ij} are treated as random, there is an additional source of bias. It has long been known that, unless the correlation between two random variables is -1, 0, or 1, the sample correlation will be a downward-biased estimate of their true correlation (see Fisher (1915, 1921)). If the θ_{ij} are normally distributed, this bias is largely a function of sample size and will be negligible if a larger number of findings (i.e., $m > 20$) are replicated. However, if the θ_{ij} are not normally distributed, the bias can be larger (Bishara and Hittner (2015)). This means that, even if $v_{ij} \to 0$, the sample correlation r would still be a downward-biased estimate of ρ in the random-effects model, particularly when m is small or the θ_{ij} are not normally distributed. Thus, there is bias due to the fact that we estimate θ_{ij} with error (as described in the previous paragraphs) and also from using the sample of m findings to estimate ρ in the population. We would note that it is possible to estimate ρ without bias, including methods described by Olkin and Pratt (1958) or Garren (1998).

8. Fisher's method. The use of Fisher's method in replication research is tied to the idea that pairwise "replication failure" is often concluded when an original study has a statistically significant effect but the replication study does not. Various researchers, including the RPP, have pointed out that a null result in a replication study is *not* evidence that $\theta_{2j} = 0$, and so some "replication failures" may arise from "false negatives," replication studies that failed to detect a true nonzero effect due to low power. To evaluate the existence of false negatives, the RPP applied a post hoc adaptation of Fisher's method that was later formalized by Hartgerink, Wicherts and van Assen (2017), who concluded that this could generally be seen as a well-powered test. Here, we reconsider these findings with asymptotic statistical theory and simulations and determine that this method is unlikely to have high power. Moreover, even when it has high power, it cannot tell which or how many of the null replication studies are false negatives.

Suppose that for findings $j = 1, \ldots, s \leq m$ that T_{1j} is statistically significant, but T_{2j} is not. This means that $p_{2j} > \alpha$ for $j = 1, \ldots, s$; in this article we assume $\alpha = 0.05$. This method tests the null hypothesis

$$H_0 : \theta_{21} = \cdots = \theta_{2s} = 0. \tag{9}$$

Traditionally, Fisher's method would use the test statistic

$$-2\sum_{j=1}^{s}\log(p_{2j}). \tag{10}$$

However, this conditional application of Fisher's method uses p-values that are necessarily on the interval [0.05, 1] and hence adapts this statistic as follows:

$$X_F^2 = \sum_{j=1}^{s} -2\log\left(\frac{p_{2j}-\alpha}{1-\alpha}\right). \tag{11}$$

Under the null hypothesis, X_F^2 will have a chi-squared distribution with $2s$ degrees of freedom. Thus, we reject H_0 in (9) when X_F^2 exceeds $c_{(1-\alpha)}(s)$, the $1-\alpha$ percentile of that distribution.

There are two key limitations to this procedure. First, this test is relatively uninformative. Failure to reject H_0 is inherently ambiguous, and, even if we do reject H_0, that does not tell us which or how many θ_{2j} are nonzero or whether those nonzero effects are positive or negative. Second, contrary to the reporting by Hartgerink et al., this method is likely to be underpowered to detect false negatives. This is because the power will depend on how many of the θ_{2j} are nonzero and how large they are (see below), and power will only be high if several of θ_{2j} are nonzero and large.

The power of this test will depend on the nonnull sampling distribution of X_F^2 which, in turn, depends on the distribution of the p_{2j} when $\theta_{2j} \neq 0$. Equation (5) gives the unconditional asymptotic distribution of p_{ij} (i.e., $p_{ij} \in [0, 1]$), but the p-values used by this method are conditional: they are only used if $p_{2j} > \alpha$. The relevant asymptotic distribution of the conditional p-value is much more complex which means that the asymptotic distribution of X_F^2 in (11) is not known exactly. However, it will be closely related to the traditional test statistic (11), and the properties of the conditional test will be similar to that of the unconditional test.

Given the result in (5), it follows that the asymptotic power of the unconditional test is

$$1-\Phi\left(\frac{c_{(1-\alpha)}(s)-\sum_{j=1}^{s}\theta_{2j}^2/v_{2j}}{\sqrt{2\sum_{j=1}^{s}\theta_{2j}^2/v_{2j}}}\right). \tag{12}$$

To gain some intuition about (12), suppose that $u \leq s$ of the studies involve $\theta_{2j} \neq 0$ and that they all have roughly the same power so that $\theta_{2j}^2/v_{2j} = \lambda$ for those u studies. Then, (12) reduces to

$$1-\Phi\left(\frac{c_{(1-\alpha)}(s)-s\lambda\frac{u}{s}}{\sqrt{2s\lambda\frac{u}{s}}}\right). \tag{13}$$

From (13) we can see that the power of the unconditional Fisher's method will increase with: (a) the number of null replication studies s, (b) the proportion that are false negatives u/s and (c) how powerful those false negative studies were λ. For reference, if $u = 30$ of $s = 100$ null replications were false negatives and they each had 80% power to detect $\theta_{2j} \neq 0$ (so that $\lambda \approx 7.85$), the power of Fisher's method would be about 50%.

Similar factors would seem to govern the power of the conditional test. Though the distribution of X_F^2 is not known exactly, we can approximate it with simulations based on the model in equation (2) (described further in the Appendix). In these simulations, sets of effect estimates T_j are drawn from normal distributions with mean θ_j and variance v_j, and their two-sided p-values are computed as in (3). Only statistically significant ($\alpha = 0.05$) values of

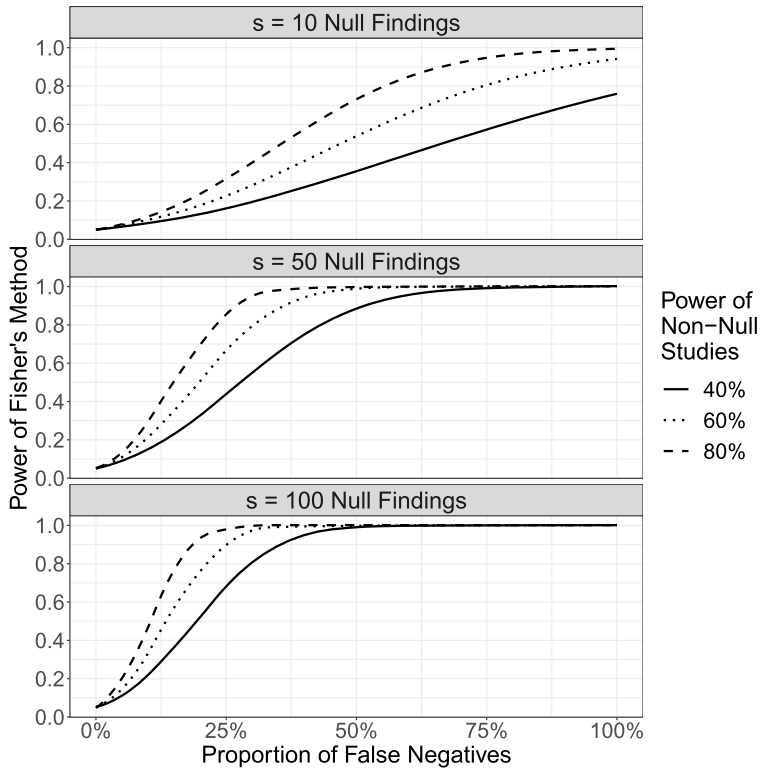

FIG. 4. *This figure shows the power of the conditional application Fisher's method to detect at false negatives as a function of the number of null findings* (s)*, the proportion of null findings that are false negatives* (x*-axis*) *and the power of those studies involving false negatives* (*linetype*).

T_j are retained in the sample, and their p-values are then used to compute X_F^2. Each simulation involves specifying different values of θ_j, v_j and s and hence different values of s, u/s and λ.

The first set of simulations uses values of θ_j and v_j so that the false negatives have a given statistical power. Figure 4 shows the results of these simulations: it plots the power of this test as a function of s, u/s and λ. For instance, the first panel shows the power of Fisher's method when there are $s = 10$ experiments with nonsignificant results: the x-axis corresponds to the proportion of those findings that are nonzero u/s, and the linetype corresponds to the power of the u nonnull experiments (which depends on λ). These graphs show that the power is only high when s, u/s and λ are large. For example, in the second panel we see that, for $s = 50$ nonsignificant findings, the conditional test would only have high power if nearly a quarter ($u \geq 12$) of those studies were false negatives that all had 80% power.

This presents something of a paradox. In order for this test to have high power, there would need to be a large number of false negatives, each with high power. However, the higher the power of each individual study, the less likely it is that they all fail to detect an effect. Thus, it would seem that Fisher's method is unlikely to have high power. For the $s = 50$ example, the probability that 12 studies each powered at 80% all fail to detect an effect is less than 10^{-8}.

Empirically, we can get a sense of the best-case scenarios for the power of the conditional test based on data from the RPE, RPP and RPSN. From (13), it is clear that the most powerful this test could be would involve a scenario where all of the false negatives (i.e., $\theta_{2j} \neq 0$) were from the largest and hence, potentially, more powerful experiments. In simulations this means ordering studies from smallest v to largest and then iteratively setting effects for the first $u < s$ effects to be nonzero.

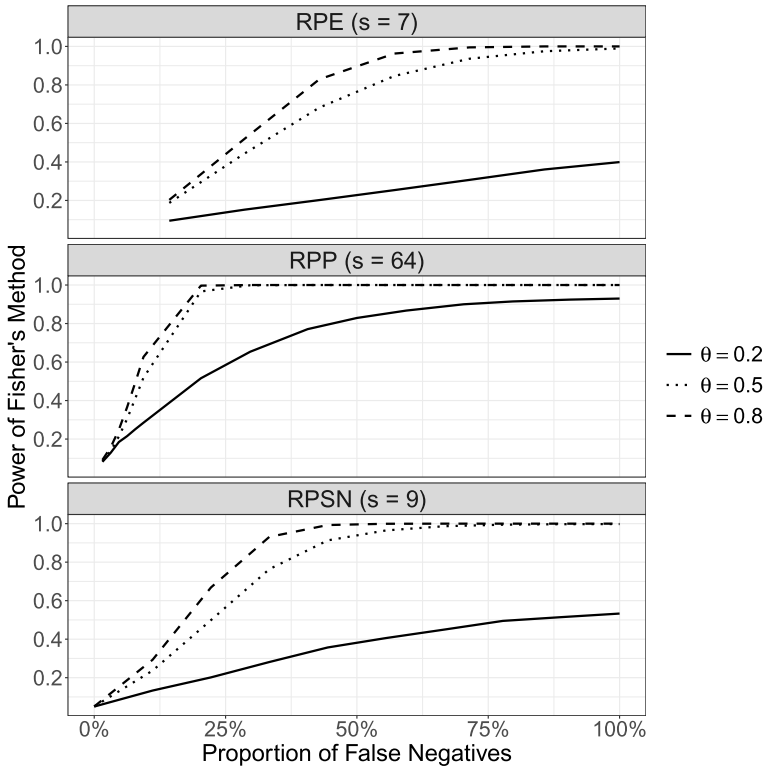

FIG. 5. *This figure shows the power of Fisher's method using the estimation error variances v_{2j} of the RPE, RPP, and RPSN experiments. The x-axis indicates the proportion of the s null findings that are false negatives, and the linetype corresponds to various magnitudes of the true effect sizes (θ) for those false negatives.*

Figure 5 shows the best-case power of Fisher's method for the RPE, RPP and RPSN. For the RPE the conditional test will only have high power if at least half of the null findings involved moderate (0.5) or large effects (0.8). For the RPP the test will be well powered if about 10 of the null findings involved moderate or large effects or if most (>65%) involved small effects. For the RPSN this test would only have high power if more than four of the nine null findings actually had medium or large effects. While it may be worth conducting this test post hoc, the fact that it will only be well powered in certain (unlikely) scenarios means that any failure to reject H_0 in (9) should be interpreted with caution.

9. Analyses of significance patterns and *p*-values. Comparisons of p-values have played a prominent role in groupwise analyses of replication, including the RPP's use of a few different paired tests of p-values. In this section we show how such tests implicitly define replication in a way that is misleading. Specifically, we show that these tests concern whether the within-study power to detect a nonnull effect is about the same for original and replication studies. However, substantial pairwise differences in effects (i.e., $\theta_{1j} \neq \theta_{2j}, \forall j$) can exist even when original and replication studies have the same power. Conversely, differences in power between the original and replication studies can mask the fact that effects are actually quite similar. In other words, because these analyses rely on a misleading definition of replication, their conclusions about replication are difficult to interpret.

Groupwise analyses of p-values have compared the distributions of p-values from the original studies p_{1j} and replication studies p_{2j}. The RPP used McNemar's test to conclude that original effect estimates were more likely to be statistically significant than replication effect estimates ($p < 0.001$), and they also tested whether the original and replication studies had the same average p-values.

McNemar's test of p-values concerns the proportion of original and replication studies that are statistically significant ($p_{ij} < 0.05$). This can be summarized in the following 2×2 table:

	Replication Studies	
Original Studies	Significant	Nonsignificant
Significant	π_{11}	π_{10}
Nonsignificant	π_{01}	π_{00}

where π_{kl} are marginal probabilities of significant patterns: $k = 1$ indicates a significant original study, and $l = 1$ indicates a significant replication study (e.g., π_{11} is the probability that both the original and replication studies are significant). The null hypothesis of McNemar's test is

$$H_0 : \pi_{11} + \pi_{01} = \pi_{10} + \pi_{11}. \tag{14}$$

In addition, the RPP conducted paired t- and Wilcoxon tests of p-values. This formally tests whether the average of the original study p-values μ_{p1} is equal to the average of the replication study p-values μ_{p2}; it can be written

$$H_0 : \mu_{p1} = \mu_{p2}. \tag{15}$$

For the sake of simplicity, we focus on the t-tests.

Note that H_0 in both (14) and (15) concern the distribution of p-values, and by equations (5) and (4) can be expressed in terms the θ_{ij}^2/v_{ij}. For (14) the probability that a study chosen at random from a group of studies results in a statistically significant effect is just the average of $1 - \beta_{ij}$ of the studies in that group (i.e., their average power). Thus, (14) will only be true when the average of the $-\beta_{1j}$ is equal to the average of the $1 - \beta_{2j}$ which, in turn, will be true when λ_1 (the mean of the θ_{1j}^2/v_{1j}) is equal to λ_2 (the mean of the θ_{2j}^2/v_{2j}). For (15) the arguments of equation (5) show that the mean of a single p-value is determined by θ_{ij}^2/v_{ij}, and so the mean of a group of p-values is determined by λ_i. Thus, both (14) and (15) can be written as

$$H_0 : \lambda_1 = \lambda_2. \tag{16}$$

Viewed this way, both tests are a comparison of the within-study power between a set of original and replication studies.

The null hypotheses in (14)–(15) can be a misleading definition of replication because they focus on statistical power and not effects. Similarity in power does not imply that effect parameters are the same size or direction. For instance, suppose $\theta_{1j} = -\theta_{2j}$ and $v_{1j} = v_{2j}$ so that each replication got the opposite effect as the original study. However, this implies that $\theta_{1j}^2/v_{1j} = \theta_{2j}^2/v_{2j}$ and $\lambda_1 = \lambda_2$. Thus, for these tests, scenarios where all studies disagree qualitatively can correspond to H_0 being true, and the probability of rejecting it is only $\alpha = 0.05$. Conversely, suppose $\theta_{1j} = \theta_{2j}$ so that all of the replication attempts succeeded. If $v_{1j} \neq v_{2j}$, then the power of the original studies will be different than the power of the replications. This means that the null hypothesis will be false, and both tests will be more likely to reject it. In other words, there are conditions under which studies clearly fail to replicate that these tests would be unlikely to detect as well as conditions under which studies successfully replicate, but these tests would be more likely to indicate otherwise.

Understanding how probable these tests are to result in misleading conclusions requires some knowledge of the nonnull sampling distributions of their test statistics X_M^2 and t_p, respectively. Both X_M^2 and t_p depend on θ_{ij}^2/v_{ij}. The exact nonnull sampling distributions

TABLE 2

This table shows the rejection rate for McNemar's test and the t-test of p-values as a function of the power of the original and replication studies and the number of findings m. Cells report the simulated probability and Monte Carlo standard error. This table assumes that all findings successfully replicate, so that $\theta_{1j} = \theta_{2j}$, which means that rejecting the null hypothesis and concluding replication failure is an error

		Rejection rate		
	Test	Power 40%/60%	Power 60%/80%	Power 40%/80%
$m = 25$	McNemar	0.267 (0.015)	0.304 (0.01)	0.819 (0.013)
	t-test	0.289 (0.012)	0.304 (0.01)	0.822 (0.011)
$m = 50$	McNemar	0.522 (0.016)	0.599 (0.016)	0.988 (0.003)
	t-test	0.529 (0.016)	0.561 (0.015)	0.985 (0.004)
$m = 100$	McNemar	0.801 (0.01)	0.873 (0.01)	0.999 (0.0003)
	t-test	0.849 (0.01)	0.825 (0.01)	0.999 (0.0002)

are not known, but we can use Monte Carlo simulations to closely approximate them. These simulations, discussed in further detail in the Appendix, follow the same approach as the simulations for the mean relative effect sizes. Since the distributions of t_p and X^2_M depend on the power of each individual study (via θ^2_{ij}/v_{ij}) as well as the number of findings subject to replication attempts m, our simulations involved different numbers of findings ($m = 25$, 50, 100), and different power levels for the original and replication studies (40%, 60%, 80%). For simplicity, these simulations assume that each original study has power $1 - \beta_1$, and each replication has power $1 - \beta_2$.

Suppose that all of the studies replicate exactly so that $\theta_{1j} = \theta_{2j}, \forall j$, but, because the original and replication studies have different sample sizes, they have different average power. Table 2 shows the rejection rate of both methods under this assumption. The rejection rate is shown for different discrepancies between the original and replication study power ($1 - \beta_1$ vs. $1 - \beta_2$) and numbers of findings m. In this table, rejecting the null hypothesis and concluding that the replications failed would be an error. The table shows that the power of both tests increases as a function of the discrepancy in within-study power $\beta_2 - \beta_1$ and the number of findings m. For instance, if all of the original studies have 40% power and the replications have 60% power, then McNemar's test will have a rejection rate of 27% when there are only $m = 25$ findings and 80% when $m = 100$. However, for larger power discrepancies, such as when original studies have 40% power and the replications have 80% power, McNemar's test will reject the null hypothesis with over 98% probability for $m = 50$ findings.

This highlights the importance of basing analyses on valid definitions of replication. When there are large differences between the within-study power of original and replication studies, these tests have high power. On its own, this sounds like a desirable feature, but, because of the way these tests define replication, their high power means that they are very likely to conclude studies fail to replicate, even when all of them replicate successfully.

This dynamic can be demonstrated on the RPE, RPP and RPSN; each of which designed replication studies to be larger than the original studies. Suppose that $\theta_{1j} = \theta_{2j} = 0.5$ (in Cohen's d units) for $j = 1, \ldots, m$ so that all the studies successfully replicated. Given the v_{1j} and v_{2j} in each program, we would expect McNemar's test to reject H_0 with nearly 100% probability for all three programs. The t-test would reject H_0 with probability greater than 41% in the RPE and over 98% for the RPP and RPSN. Thus, if all of the replications in these programs succeeded, these tests would be almost certain to indicate otherwise.

10. Conclusions. This article has examined the properties of groupwise analysis methods that have been used to assess replication and found that most methods we considered had serious limitations. The mean relative effect size can have substantial uncertainty which can lead to misleading conclusions with surprising frequency. Estimates of the correlation between original and replication studies can greatly understate the actual correlation between effects in those studies. Fisher's method, which has been used to detect false negative replication studies, is bound to have low power in this context. Finally, comparisons of p-values frame replication as a comparison of power between original and replication studies which can be a misleading definition of replication.

Because many of these analysis methods have poor statistical properties under seemingly plausible conditions, it is, therefore, difficult to interpret the results of such methods with much confidence. For instance, a reported mean relative effect size near zero may imply that the actual ratio of replication study effects to original study effects is near zero, but this also has a reasonable chance of happening even when their ratio is one. Our focus here is not to criticize the results of prior replication research but to emphasize that the methods producing those results (and the results of future efforts) have statistical properties that must be considered.

Perhaps the most important consideration for assessing replication is its operational definition. Methods that rely on a flawed definition of replication will necessarily be flawed, and, in some sense, discussion of their properties becomes somewhat irrelevant: if such an analysis method has good statistical properties, it will simply be more certain about the wrong thing. Greater effort should be devoted to ensuring that any proposed analysis method aligns with clear and justifiable definitions of replication. We have argued that such definitions should depend on effect parameters. As scientific and statistical fields increasingly emphasize the interpretation of experiments in terms of effect sizes, it seems only natural to extend this emphasis to interpretations of replication (see Wasserstein and Lazar (2016); Cooper (2011)). Further, patterns used to describe replication across multiple findings should be somewhat consistent with the definitions used to define replication for a single finding.

Additional work is also needed on design and analysis methods. Estimators of important quantities pertaining to replication should be accurate; large and unpredictable biases should be avoided, as should tests with uncontrolled or poor error rates. However, the properties of analysis methods are closely tied to design. The same principles used to design a single study to ensure high power or precision can be adapted for ensembles of studies. In this way we can ensure that the results of replication studies are accurate and conclusive.

Finally, throughout this article we have advocated for a meta-analytic framework for assessing replication. While it is not the only way to think about replication, we find that meta-analysis offers a few important advantages. The model that underpins most meta-analyses distinguishes between the effect parameters and estimation errors and allows for a more clear-cut approach to defining replication. Inferential procedures based on this model have been studied in the meta-analytic literature for decades, and such procedures may possibly be adapted to the study of replication (see Hedges and Schauer (2019b); Schauer (2018)). In particular, standard meta-analytic methods can be used to explore the distribution of differences between original and replication study results, which we described in this article.

APPENDIX: SIMULATIONS

This article used simulations to approximate the sampling distribution of three different test statistics. The first is the test statistic for Fisher's method, X_F^2 given in equation (11). The second is the test statistic for McNemar's test of the null hypothesis defined in (14),

$$X_M^2 = \frac{(m_{10} - m_{01})^2}{m_{10} + m_{01}}, \tag{17}$$

where m_{kl} is the number of experiments that exhibit a given statistical significance pattern, $k = 1$ indicates a significant original study and $l = 1$ indicates a significant replication study; for instance, m_{11} is the number of experiments for which both the original and replication studies are significant.

The third is the test statistic for tests of p-values for null hypothesis (15). The test statistic for this test is

$$t_p = \frac{\bar{p}_d}{\mathrm{SD}[\bar{p}_d]/\sqrt{m}}, \tag{18}$$

where $\bar{p}_d$ is the average of the $p_{1j} - p_{2j}$ and $\mathrm{SD}[p]$ is their standard deviation,

$$\bar{p}_d = \frac{1}{m}\sum_{j=1}^{m}(p_{1j} - p_{2j}); \qquad \mathrm{SD}[p] = \frac{1}{\sqrt{m-1}}\sum_{j=1}^{m}\big[(p_{1j} - p_{2j}) - \bar{p}_d\big]^2. \tag{19}$$

All tables and graphics in this paper are based on 100,000 simulations which are described in greater detail below. Each of these sampling distributions depend on θ_{ij} and v_{ij}; thus, a given simulation consisted of drawing m pairs of studies from normal distributions as in equation (2) and computing the relevant test statistic. What varied between simulations were the values of θ_{ij} and v_{ij}.

For the simulations of X^2_M and t_p in Table 2, the sampling distribution of each statistic depends on the power of each null hypothesis test described by equations (14)–(15). The power of these tests depends on the value θ^2_{ij}/v_{ij}. In Table 2, we assumed that studies had 40%, 60% and 80% power. Thus, we set the values of θ^2_{ij}/v_{ij} to be 2.91, 4.90 and 7.85. For instance, in the "Power = 40%/60%" column, we set $\theta^2_{1j}/v_{1j} = 2.91$ and $\theta^2_{2j}/v_{2j} = 4.90$ for all j. Further, to obtain the potential error rates in the RPP, RPE and RPSN data, we set $\theta_{1j} = \theta_{2j} = 0.5$ and used the v_{ij} from the RPP, RPE and RPSN data that were converted to be on the scale of Cohen's d.

For Fisher's method the nonnull sampling distribution of X^2_F also depends on θ^2_{ij}/v_{ij}, and hence we used the same values as above. To obtain an upper bound of the power for the RPP, RPE and RPSN, we assumed the largest studies were the false negatives. To do this, we sorted the s nonsignificant replication studies in each program by v_{2j} in ascending order. We then iteratively set the first u effect parameters θ_{2j} to be equal to a given value (0.2, 0.5 and 0.8) and ran the simulations.

Acknowledgments. The authors would like to thank the referees and Editors for their constructive comments that improved the quality of this paper.

This work was supported by Institute of Education Sciences (IES) Grant R305B140042 as well as National Science Foundation (NSF) Grant DRL-1841075.

SUPPLEMENTARY MATERIAL

Aggregate patterns of replication code and data (DOI: 10.1214/20-AOAS1387SUPP; .zip). This supplement contains a repository of data, code, and output (including graphics) used to obtain the results reported in this article. This repository contains a directory for data (including raw and cleaned data). It also contains a directory of analysis and simulation scripts, as well as results from simulations. Finally, the graphics directory contains the graphics produced for this article.

REFERENCES

BAHADUR, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Stat.* **31** 276–295. https://doi.org/10.1214/aoms/1177705894

BISHARA, A. J. and HITTNER, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educ. Psychol. Meas.* **75** 785–804. https://doi.org/10.1177/0013164414557639

BOLLEN, K., CACIOPPO, J. T., KAPLAN, R. M., KROSNICK, J. A. and OLDS, J. L. (2015). Reproducibility, replicability, and generalization in the social, behavioral, and economic sciences. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. National Science Foundation, Arlington, VA.

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T. and ROTHSTEIN, H. R. (2009). *Introduction to Meta-Analysis*. Wiley-Blackwell, Oxford.

BRANDT, M. J., IJZERMAN, H., DIJKSTERHUIS, A., FARACH, F. J., GELLER, J., GINER-SOROLLA, R., GRANGE, J. A., PERUGINI, M., SPIES, J. R. et al. (2014). The replication recipe: What makes for a convincing replication? *J. Exp. Soc. Psychol.* **50** 217–224.

CAMERER, C. F., DREBER, A., FORSELL, E., HO, T.-H., HÜBER, J., JOHANNESSON, M., KIRCHLER, M., ALMENBERG, J., ALTMEJD, A. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* **351** 1433–1436.

CAMERER, C. F., DREBER, A., HOLZMEISTER, F., HO, T.-H., HÜBER, J., JOHANNESSON, M., KIRCHLER, N. G., NOSEK, B. A. et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* **2** 637–644.

COOPER, H. M. (2011). *Reporting Research in Psychology*: *How to Meet Journal Article Reporting Standards*. APA Books, Washington, DC.

COOPER, H. M., HEDGES, L. V. and VALENTINE, J. (2009). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. The Russell Sage Foundation, New York.

CUMMING, G., FIDLER, F., KALINOWSKI, P. and LAI, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Aust. J. Psychol.* **64** 138–146.

DERSIMONIAN, R. and LAIRD, N. M. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7**.

DIAZ-FRANCES, E. and RUBIO, F. J. (2013). On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. *Statist. Papers* **54** 309–323. https://doi.org/10.1007/s00362-012-0429-2

ETZ, A. and VANDEKERCKHOVE, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE* **11** e0149794. https://doi.org/10.1371/journal.pone.0149794

FIELLER, E. C. (1932). The distribution of the index in a bivariate normal distribution. *Biometrika* **24** 3–4. https://doi.org/10.1093/biomet/24.3-4.428

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521.

FISHER, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

GARREN, S. T. (1998). Maximum likelihood estimation of the correlation coefficient in a bivariate normal model with missing data. *Statist. Probab. Lett.* **38** 281–288. MR1629923 https://doi.org/10.1016/S0167-7152(98)00035-2

GEARY, R. C. (1930). The frequency distribution of the quotient of two normal variates. *J. R. Stat. Soc.* **93** 442–446. https://doi.org/10.2307/2342070

GILBERT, D. T., KING, G., PETTIGREW, S. and WILSON, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science* **351** 1037–1037.

GLESER, L. J. and OLKIN, I. (1994). Stochastically dependent effect sizes. *The handbook of research synthesis*. H. Cooper & L. V. Hedges (Eds.). Russell Sage Foundation 339–355.

HARTGERINK, C. H. J., WICHERTS, J. M. and VAN ASSEN, M. A. L. M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra*: *Psychology* **3** 9.

HAYYA, J., ARMSTRONG, D. and GRESSIS, N. (1975). A note on the ratio of two normally distributed variables. *Manage. Sci.* **21** 1338–1341. https://doi.org/10.1287/mnsc.21.11.1338

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL. MR0798597

HEDGES, L. V. and PIGOTT, T. D. (2001). The power of statistical tests in meta-analysis. *Psychol. Methods* **6** 203–217. https://doi.org/10.1037/1082-989x.6.3.203

HEDGES, L. V. and SCHAUER, J. M. (2019a). More than one replication study is needed for unambiguous tests of replication. *J. Educ. Behav. Stat.* **44** 543–570.

HEDGES, L. V. and SCHAUER, J. M. (2019b). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychol. Methods* **24** 557–570.

HEDGES, L. V. and VEVEA, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychol. Methods* **3** 486–504. https://doi.org/10.1037/1082-989X.3.4.486

HSUEH, H.-M., CHEN, J. J. and KODELL, R. L. (2003). Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J. Biopharm. Statist.* **13** 675–689.

HUNG, H. M. J., O'NEILL, R. T., BAUER, P. and KÖHNE, K. (1997). The behavior of the *P*-value when the alternative hypothesis is true. *Biometrics* **53** 11–22. MR1450180 https://doi.org/10.2307/2533093

KALAIAN, H. K.and RAUDENBUSH, S. W. (1986). A multivariate mixed linear model for meta-analysis. *Psychol. Methods* **1** 227–235. https://doi.org/10.1037/1082-989X.1.3.227

KLEIN, R. A., RATLIFF, K. A., VIANELLO, M., ADAMS, R. B., BANHÍK, Š., BERNSTEIN, M. J., BOCIAN, K., BRANDT, M. J., BROOKS, B. et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Soc. Psychol.* **45** 142–152. https://doi.org/10.1027/1864-9335/a000178

KLEIN, R. A., VIANELLO, M., HASSELMAN, F., ADAMS, B. G., ADAMS, R. B., ALPER, S. et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1** 443–490.

KLEIN, R. A., COOK, C. L., EBERSOLE, C. R., VITIELLO, C. A., NOSEK, B. A., CHARTIER, C. R., CHRISTOPHERSON, C. D. et al. (2019). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. Available at https://psyarxiv.com/vef2c.

LAMBERT, D. and HALL, W. J. (1982). Asymptotic lognormality of *P*-values. *Ann. Statist.* **10** 44–64. MR0642718

MARSAGLIA, G. (2006). Ratios of normal variables. *J. Stat. Softw.* **16** 1–10. https://doi.org/10.18637/jss.v016.i04

MAXWELL, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychol. Methods* **9** 147–163.

MUCHINSKY, P. M. (1996). The correction for attenuation. *Educ. Psychol. Meas.* **56** 63–75.

OLKIN, I. and PRATT, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Ann. Math. Stat.* **29** 201–211. MR0093854 https://doi.org/10.1214/aoms/1177706717

OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science* **349** aac4716.

PATIL, P., PENG, R. D. and LEEK, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* **11** 539–544. https://doi.org/10.1177/1745691616646366

RILEY, R. D., HIGGINS, J. P. T. and DEEKS, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ* **342**. d549.

SCHAUER, J. M. (2018). Statistical Methods for Assessing Replication: A Meta-Analytic Framework. Doctoral Dissertation, Northwestern Univ., Evanston, IL.

SCHAUER, J. M., FITZGERALD, K. G., PEKO-SPICER, S., WHALEN, M. C. R., ZEJNULLAHI, R. and HEDGES, L. V. (2021). Supplement to "An evaluation of statistical methods for aggregate patterns of replication failure." https://doi.org/10.1214/20-AOAS1387SUPP

SCHWEINSBERG, M., MADAN, N., VIANELLO, M., SOMMER, S. A., JORDAN, J., TIERNEY, W., AWTREY, E., ZHU, L. L., DIERMEIER, D. et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *J. Exp. Soc. Psychol.* **66** 55–67.

SHAPIN, S. and SCHAFFER, S. (1985). *Leviathan and the Air-Pump*: *Hobbes*, *Boyle*, *and the Experimental Life*. Princeton Univ. Press, Princeton, NJ.

SIMONSOHN, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26** 559–569.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* **15** 72–101.

SPEARMAN, C. (1910). Correlation calculated from faulty data. *Br. J. Psychol.* **3** 271–295.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

TAMHANE, A. C. and SHI, J. (2009). Parametric mixture models for estimating the proportion of true null hypotheses and adaptive control of FDR. In *Optimality. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **57** 304–325. IMS, Beachwood, OH. MR2681678 https://doi.org/10.1214/09-LNMS5718

VALENTINE, J. C., BIGLAN, A., BORUCH, R. F., CASTRO, F. G., COLLINS, L. M., FLAY, B. R., KELLAM, S., MOŚCICKI, E. K. and SCHINKE, S. P. (2011). Replication in prevention science. *Prev. Sci.* **12** 103–117.

VAN AERT, R. C. M. and VAN ASSEN, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS ONE* **12** e0175302.

VANKOV, I., BOWERS, J. and MUNAFÒ, M. R. (2014). On the persistence of low power in psychological science. *Q. J. Exp. Psychol.* **67** 1037–1040.

VERONIKI, A. A., JACKSON, D., VIECHTBAUER, W., BENDER, R., BOWDEN, J., KNAPP, G., KUSS, O., HIGGINS, J. P. T., LANGAN, D. et al. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synth. Methods* **7** 55–79.

WASSERSTEIN, R. L. and LAZAR, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose [Editorial]. *Amer. Statist*. **70** 129–133. MR3511040 https://doi.org/10.1080/00031305.2016.1154108

WOOD, P. and RANDALL, D. (2018). How bad is the government's science? *Wall St. J*. Available at https://www.wsj.com/articles/how-bad-is-the-governments-science-1523915765.

YONG, E. (2016). The inevitable evolution of bad science. The Atlantic. Available at https://www.theatlantic.com/science/archive/2016/09/the-inevitable-evolution-of-bad-science/500609/.