Reconsidering Statistical Methods for Assessing Replication

J. M. Schauer
Institute for Policy Research
Northwestern University

Abstract

Recent empirical evaluations of replication in psychology have reported startlingly few successful replication attempts. At the same time, they have noted that the proper way to analyze replication studies is far from a settled matter and have thus analyzed their data in several different ways. This presents two challenges to interpreting the results of these programs. First, different analysis methods assess different operational definitions of replication. Second, the properties of these methods are not necessarily common knowledge; it is possible for a successful replication to be deemed a failure by nearly all of the metrics used, and it is not always immediately clear how likely such errors are to occur. In this article, we describe the methods commonly used in replication research and how they imply specific operational definitions of replication. We then compute the probability of false failure (i.e., a successful replication is concluded to have failed) and false success determinations. These are shown to be high (often over 50%) and in many cases uncontrolled. We then demonstrate that errors are probable in the data to which they have been applied in the literature. We show that the probability that some conclusions in the literature about replication are incorrect can be as high as 75-80%.

The Replication Crisis

The emergence of a replication crisis over the past decade has cast doubt on just how much we should trust the findings of scientific research (see Ioannidis, 2005; Pashler & Harris, 2012; Baker, 2015). One facet of this crisis, which has been noted by many researchers, is that replication attempts (successful or unsuccessful) seldom appear in the scientific literature (e.g., Moonesinghe, 2007; Freese, 2007). At the same time, high-profile efforts to empirically evaluate the replicability of findings have reported alarmingly few successes (e.g., Open Science Collaboration, 2015; Camerer et al., 2016).

Efforts to systematically replicate experiments have become a centerpiece of the replication crisis. They remain the source of some of the best evidence about the crisis, and their publication has generated much discussion. For instance, the Replication Project: Psychology (RPP) attempted to independently replicate 100 distinct experiments in social and behavioral psychology (Open Science Collaboration, 2015). The RPP concluded based on “subjective determinations” that 61 of their replication attempts failed, a statistic that has been echoed in several academic papers (e.g., Baker, 2016), as well as in the popular press (e.g., Connor, 2015; Yong, 2016).

However, the analyses of such programs have been called into question (Etz & Vandeckerckhove, 2016; Hartgerink et al., 2017; Hedges & Schauer, 2019b). Indeed, these programs have often stated that there is no single standard metric for assessing replication (Open Science Collaboration, 2015; Camerer et al., 2016; Schweinsberg et al., 2016). Absent a standard analysis method, it has become common for replication research programs to analyze their data using several different methods, which can (and do) give different signals about whether a replication was successful.

The focus on analysis is important for at least three reasons. First, different analysis methods would appear to imply different definitions of replication. On its face, replication seems as though it has an obvious definition: successful replication studies get the same results. However, as noted by a recent National Science Foundation (NSF) Subcommittee on Replication in Science, "Although we have an intuitive sense of what it means for results to replicate, the meaning becomes less clear the more closely we look," (Bollen et al., 2015). The various analysis methods used so far operationalize "results" in different ways (e.g., *p*-values, effects, statistical significance patterns), and use different notions of what it means for those results to be the same. Thus, each analysis method may say something different about what replication means, and if an attempt was successful.

A second reason is that the properties of some of the analysis methods are not necessarily widely known. Analyses of replication research have sought to make determinations about whether a finding was replicated. However, it is possible that these determinations are incorrect; they may indicate that a replication succeeded when it failed, or vice versa. Understanding how likely these methods are to make errors is necessary to contextualize the results of these programs. Hypothetically, if a decision procedure falsely concludes that a replication failed when it actually succeeded 60% of the time, then results like those of the RPP could reflect less about replicability of scientific findings, and more about the analysis methods they used.

Third, because there is no settled analysis method, and the properties of the ones used so far are not common knowledge, it seems unlikely that replication research efforts were designed to ensure sensitive analyses. Efficient analyses and careful designs can limit the likelihood of drawing erroneous conclusions about the replicability of a finding (for discussion, see Hedges & Schauer, 2019a). This is part of the reason that major scientific funding agencies often require research proposals to include statistical power analyses. Particularly for replication research, the value of which has been questioned, it would seem imperative that resources are not wasted on research designs and analyses unlikely to net sufficiently precise conclusions.

This paper reviews several methods that researchers have used to determine whether a replication succeeded or failed. We formalize the distinction between analysis methods and potential definitions of replication, and identify some common definitions used so far. We then examine analysis methods by formalizing them as statistical procedures: this includes determining what they might say about the definition of replication, describing the types of errors they may make, and computing how frequently they might make those errors. We then demonstrate how likely these types of errors could be made in empirical research to provide context to published findings about replication.

Data: Empirical Evaluations of Replication

Several programs of research about replication have been conducted in the social sciences, and we examine some of these programs. The RPP, described above, attempted a single direct replication of each of 100 distinct experiments in social and behavioral psychology. They used various methods to analyze their data and determined that between 64% and 53% of their attempts failed, though the 61% statistic appears to be the one that has resonated (Open Science Collaboration, 2015). The RPP note that a subset of 73 study pairs would be appropriate for meta-analytic methods, and that 46 (63%) of those were deemed to have failed.

The RPE followed a similar design as the RPP, attempting to replicate 18 distinct experiments in economics (Camerer et al., 2016). Their analysis methods were nearly identical to those of the RPP, and led to the RPE concluding that between 4 and 7 of the 18 replication attempts failed (22–39%). While their reported results involve a higher replication rate than the RPP, the authors were still pessimistic about the state of replication in behavioral economics.

The Many Labs Project, whose results were published in the year prior to the RPP, took a slightly different approach to studying replication (Klein et al., 2014). Many Labs selected 16 experiments that were each replicated not a single time, but several times. The effort recruited 36 labs, that each conducted all 16 experiments. The effect estimates from the replication studies (i.e., studies 2, …, *k*) were aggregated via a meta-analysis and that was compared to the original finding using similar methods as those used by the RPP and RPE. Many Labs determined that two of the 16 replications failed (12%).

Finally, the Pipeline Project, here referred to as PPIR, followed a similar approach as Many Labs (Schweinsberg et al., 2016). They selected 10 experiments that were "in the pipeline" and had not yet been published. They then recruited 25 laboratories to conduct subsets of these experiments. Thus, each experiment was conducted between 12 and 18 times. As with Many Labs, the results of the replication studies were aggregated via a meta-analysis and compared to the findings of the initial experiments using many of the same metrics as RPP and RPE. The PPIR ultimately concluded that four of the 10 experiments failed to replicate.

Data on effect sizes and sample sizes from each of these programs are available on the Open Science Framework (https://osf.io/). They were collected, often using the original investigator's own code, and transformed so that all study results were on the scale of Cohen's *d*. In addition, the sampling variances of each effect size estimate was computed. Where information was not available, we contacted the original authors, who responded with important clarifications. It should be noted that only the meta-analytic subset of studies from the RPP were collected. These data will be used to demonstrate the properties of the analysis methods of these programs, which will help to better contextualize the results of these and similar research programs.

Analyses, Definitions, and Their Implications

In this paper, we make a distinction between an *analysis* of replication and an operational *definition* of replication. This may seem like a trivial distinction, since these are two sides of the same coin: an analysis of replication inherently relies on some definition of what replication *means*. However, the difference between them appears particularly muddied in the replication research literature (Hedges & Schauer, 2019a,b; Schauer, 2018). Here, we try to clarify these concepts, and show why both need to be considered to interpret the results of research.

The distinction between definition and analysis is akin to the difference between a parameter and an estimator in statistics. The definition of replication, much like a parameter, precisely operationalizes what is meant by "replication." This requires careful consideration about the scientific quantities of interest in experiments and how they should relate to each other. For example, suppose two identical pharmaceutical trials examine if certain drug reduces mortality. One definition of a successful replication is that the drug reduces mortality by the same amount in each study. An alternative definition would be that the drug is effective in both studies, so that it reduces mortality in each study, but by potentially different amounts.

An analysis is a procedure used to assess a given definition of replication amid statistical uncertainty. The proper analysis method for one definition of replication may be different from that of another definition. Frequently in empirical work, analyses take the form of decision rules; researchers are interested in yes/no decisions about whether an experiment replicated successfully. However, alternative analysis methods, such as sensitivity analyses or point estimation have also been proposed in the context of replication (e.g., Simonsohn, 2015; Etz & Vandekerkhove, 2016; Schauer, 2018).

This article is concerned with four analysis methods that have been prevalent in the replication research literature: the significance criterion, confidence interval overall criterion, prediction interval, and meta-analytic average. These have been used by many empirical evaluations of replication. For each method, we want to know (a) what definition of replication are they assessing, and (b) how accurately are they assessing it.

Error Rates

There are two types of errors that we explore in this paper. For a given procedure, a *false failure determination* would involve the analysis concluding that the studies failed to replicate when they actually successfully replicated according to a given definition. We refer to the probability that a procedure makes this type of error as the *false failure rate* (FFR). The second type of error is a *false success determination*, where the analysis concludes that the studies successfully replicated when they did not (for a given definition of replication). We call the probability that this type of error occurs the *false success rate* (FSR).

While the FFR and FSR can be thought of as decision theoretic properties of a general decision procedure, most of the analysis methods discussed in this article either are or can be thought of as a test of the null hypothesis that the studies replicate successfully. In this context, the FFR is analogous to the type I error rate α, and the FSR is like the type II error rate β. However, not all of the procedures discussed here share some of the standard properties of hypothesis tests. First, the type I error rate is typically *controlled* in a hypothesis test, meaning that regardless of sample size, type I errors will only occur with some pre-specified probability

α. However, the following sections will document how some procedures used to assess replication have *uncontrolled* type I error rates (i.e., uncontrolled FFRs) that will depend on how the original and replication studies are designed.

Second, when thought of hypothesis tests, a false success error would occur if the studies did not replicate, but the test failed to reject the null hypothesis. However, this cannot necessarily be interpreted as a successful replication; failure to reject the null hypothesis does not prove that it is true, and interpreting that otherwise is a logical fallacy. Thus, the terms *false success* and *FSR* are somewhat misleading for these methods. We use them here because interpretations of the analyses of replication often make this fallacy (see, e.g., Klein et al., 2014; Open Science Collaboration, 2015). This article shows why doing so can greatly distort conclusions.

Statistical Model

While it has been argued that there is no standard approach to assessing replication, it is possible to describe the data from replication studies in terms of a statistical model. Perhaps the most relevant literature is that of meta-analysis, which has been used to clarify subjective notions about replication, and to propose methods to assess it (Hedges & Schauer, 2019b).

Suppose $k \geq 2$ replication studies are conducted. If $k = 2$, this would refer to an original study and replication study. Denote θ_i as the effect of study *i*. Typically, the θ_i are one of the standard effect sizes used in meta-analysis, such as the standardized mean difference (Cohen's *d*) or Fisher-transformed correlation coefficient (see Cooper, Hedges, & Valentine, 2009). The parameter θ_i is the scientific estimand of interest in study *i*, and is what would be observed in the absence of any estimation error (e.g., due to sampling experimental units).

In practice, we do not observe the θ_i directly, but instead must estimate them. Let T_i be the estimate of θ_i, and let v_i be its estimation error variance. It is common in meta-analysis to assume that T_i is unbiased and normally distributed, and that v_i is known:

$$T_i \sim N(\theta_i, v_i) \tag{1}$$

This is exactly true for some effect sizes and is a good large-sample approximation for others (Cooper, Hedges, & Valentine, 2009).

The analyses discussed in this article are appropriate for $k = 2$ studies. We index the original study with $i = 1$ and the replication study with $i = 2$. However, not all research programs conduct only a single replication study, such as the Many Labs Project, which conducted $k - 1 = 36$ replication studies. In such cases, effect estimates are aggregated into a single effect estimate via a meta-analysis, which we refer to with index $i = 2$.

Often the value of θ_i is assessed in scientific research using a null hypothesis test, where the null hypothesis may be written as H_{0i}: $\theta_i = \theta_{0i}$. It is common in the social sciences to test a null hypothesis that the effect of a manipulation is zero, so that $\theta_{0i} = 0$. Assuming that H_{0i} is tested against a two-sided alternative, the p-value of this test is given by

$$p_i = 2\left[1 - \Phi\left(\frac{|T_i|}{\sqrt{v_i}}\right)\right] \tag{2}$$

where Φ is the standard normal distribution function.

Some of the analyses involve averaging the original and replication studies via:

$$T. = \frac{\frac{T_1}{v_1}+\frac{T_2}{v_2}}{\frac{1}{v_1}+\frac{1}{v_2}} \tag{3}$$

Under the model, $T.$ has a normal distribution with mean $\theta.$ and variance $v.$ given by:

$$\theta. = \frac{\frac{\theta_1}{v_1}+\frac{\theta_2}{v_2}}{\frac{1}{v_1}+\frac{1}{v_2}}, \qquad v. = \frac{1}{\frac{1}{v_1}+\frac{1}{v_2}} \tag{4}$$

Various researchers have pointed out that publication bias can affect analyses of replication (Hedges & Schauer, 2019; Etz & Vandeckerckhove, 2016). Replication studies often attempt to re-create a published finding (such as in the RPP). However, there is evidence that such findings are subject to a selection process that favors the publication of significant results. This selection can lead to bias in the effect estimate T_1 that can be quite large, as much as 250% in extreme cases (see Hedges, 1984). Several methods exist to correct that bias, including a maximum likelihood approach introduced by Hedges (1984), and subsequent refinements (see McShane et al., 2016; Rothstein et al., 2005 for discussion). These methods effectively result in a corrected estimate T_1^* that has variance v_1^*. Since these corrections typically require the estimation of additional parameters associated with selection, the corrected estimates tend to have a larger variance, so that $v_1^* > v_1$. Moreover, adjustments that rely on maximum likelihood methods, such as Hedges (1984), will result in effect estimates that are asymptotically normal. While we consider the properties of analysis methods without publication bias (with T_1 and v_1), analogous arguments follow from these results (with T_1^* and v_1^*) if there bias.

Definitions of Replication

Since the θ_i are the quantities of interest in a single study, definitions of replication should be framed in terms of the θ_i. The methods examined in this article would seem to suggest two approaches to doing so. *Definition 1* requires effect parameters to be the same value. This typically means effect parameters are identical (i.e., $\theta_1 = \theta_2$), referred to as *exact replication*. However, it Definition 1 can be expanded to accommodate notions of *approximate replication* where effects are very similar but not identical (see Hedges & Schauer, in press). Systematic reviews of replications in hard sciences, such as particle physics, suggest that even in the face of strong theory and sound scientific practice, minor differences may be expected among seemingly perfect replications (Olive, 2014; Hedges, 1987; Rosenfeld, 1975). Moreover, we might regard very small differences between effects as unimportant or negligible. In this article, the methods considered rely on ideas of exact replication.

Definition 2 can be seen as requiring qualitative agreement among effect parameters. A replication is successful if one of the following is true:

(a) $\theta_1 = \theta_2 = 0$

(b) $\theta_1 > 0$ and $\theta_2 > 0$

(c) $\theta_1 < 0$ and $\theta_2 < 0$

Relative to Definition 1, this is a somewhat looser notion of replication. In fact, under this definition, a successful replication can involve a difference between two effect parameters $\theta_1 - \theta_2$ that is unbounded, so long as both θ_1 and θ_2 are positive (or both are negative).

Statistical Methods for Determining Replication Success

In this section we examine methods used to analyze replication research programs. Each method has its own subsection, within which we describe the procedure, highlight the operational definition of replication it assesses, and compute the probability that it makes different types of errors. We show that the probability that a procedure results in an error often depends on the estimation error variances v_i, the effects parameters θ_i, or both. We then assess how likely these methods are to have made errors in empirical replication research based on the design of those programs.

Prediction Interval

Patil, Peng, and Leek (2017) proposed the prediction interval analysis, which has been used by such programs as the RPE and Many Labs 2 (Klein etl al., 2018). The prediction interval amounts to a two-sided z test between the original and replication effect parameters, and it is equivalent to the meta-analytic Q-test when there are only $k = 2$ effects (Hedges & Schauer, 2019b).

The prediction interval concludes that a replication failed when T_1 is not contained in a "prediction interval" for θ_2, which occurs when the following is false: $-1.96\sqrt{v_1 + v_2} < T_1 - T_2 < 1.96\sqrt{v_1 + v_2}$. This procedure is actually a test is that the underlying effects are identical:

$$H_0\text{: } \theta_1 = \theta_2 \tag{5}$$

This is tested by computing

$$Z = (T_1 - T_2)/\sqrt{v_1 + v_2} \tag{6}$$

Under the null hypothesis, Z has a standard normal distribution, and we reject the null hypothesis if $|Z|$ exceeds $c_{(1-\alpha/2)}$, the $(1 - \alpha/2)$ percentile of that distribution:

$$c_{(1-\alpha/2)} = \Phi^{-1}(1 - \alpha/2)$$

Users of this procedure have set $\alpha = 0.05$ and $c_{(1-\alpha/2)} = 1.96$.

Based on the null hypothesis, this procedure assesses Definition 1 of replication, that $\theta_1 = \theta_2$. A false failure error, then, would involve the test concluding that $\theta_1 \neq \theta_2$ when in fact $\theta_1 = \theta_2$. This scenario would correspond to a type I error for this test, which occurs with probability α. Note that for this test, this error rate is controlled: we can set some value of α and the test is guaranteed to have a type I error rate no larger than α.

A false success error occurs when this test does not detect a failed replication (i.e., we maintain H_0 when $\theta_1 \neq \theta_2$). This would correspond to a type II error for the test, and the probability of a type II error occurs will depend on the non-null sampling distribution of Z. When

$\theta_1 \neq \theta_2$, Z follows normal distribution with unit variance and mean $(\theta_1 - \theta_2)/\sqrt{v_1 + v_2}$. Therefore, the FFR is given by

$$1 - \Phi\left(c_{(1-\alpha/2)} - \frac{\theta_1 - \theta_2}{\sqrt{v_1 + v_2}}\right) + \Phi\left(-c_{(1-\alpha/2)} - \frac{\theta_1 - \theta_2}{\sqrt{v_1 + v_2}}\right) \tag{7}$$

Note that equation (7) implies the FSR depends on $|\theta_1 - \theta_2|$ and $v_1 + v_2$. Figure 1 shows the FSR as a function of $|\theta_1 - \theta_2|$, v_1, and v_2 (assuming $\alpha = 0.05$). Each colored line corresponds to a given value of $|\theta_1 - \theta_2|$ where both are on the scale of Cohen's *d*, so that the three colors correspond to differences between a zero effect and a small effect (0.2), medium effect (0.5), or large effect (0.8). Solid lines correspond to replications where $v_1 = v_2$, and dashed lines correspond to scenarios where $v_1 = 4v_2$, so that study 2 is four times as large as study 1. One way to intuit the *x*-axis is that on this scale, $v_1 \approx 4/n$ where *n* is the total sample size of study 1; thus 0.02 would involve study 1 having 200 participants. What we see from this plot is that if the difference between parameters is 0.2, then this test will fail to detect that over 75% of the time. When $n > 200$, this procedure will be more likely to detect differences greater than $d = 0.5$.

In the empirical evaluations of replication, the $\alpha = 0.05$ level test would be unlikely to detect small differences between effect parameters for most experiments. Figure 2 shows the FSR (type II error rate of the test) for studies in each program that failed to reject the null hypothesis of replication. Error rates are shown as a function of $|\theta_1 - \theta_2|$, where each boxplot corresponds to a given value of $|\theta_1 - \theta_2|$. In Figure 2, we see that if all of the replications failed such that $|\theta_1 - \theta_2| = 0.2$ (in Cohen's *d* units) then the probability that the analysis would fail to detect that is mostly over 80%! Moreover, even if $|\theta_1 - \theta_2| = 0.5$ for each pair of studies, error rates would still be largely above 50%.

Because the FFR is controlled at level α, we can potentially decrease the FSR by increasing α. For instance, suppose we were interested in detecting replication failures such that $|\theta_1 - \theta_2| = 0.5$. If we set $\alpha = 0.1$, both Many Labs and PPIR would involve replication studies where this method would have over 75% power to detect a difference this large, so the FSR would be smaller than 25%. However, even with $\alpha = 0.1$ the FSR for the RPP and RPE would still be over 50% for most studies; for $\alpha = 0.2$, the FSR would still be over 50% for the RPP.

That this test has low power to detect differences between effects has been known to statisticians for some time (see Hedges & Pigott, 2001). However, it is the likelihood ratio test, and thus will be the most powerful test to detect failed replications under the model (and according to Definition 1). This means that it will be more powerful than any other test for exact replication that has a controlled type I error rate. The low power is less an issue of the method (since alternative methods will either have low power or will have uncontrolled error rates), but rather a limitation of the information about replication that can be gleaned from only $k = 2$ studies (Hedges & Schauer, 2019a). The meta-analytic Q test naturally generalizes this procedure to $k > 2$ studies, and Hedges & Schauer (2019b) show that when multiple ($k > 2$) replications are conducted, the Q test can have much greater power to detect meaningful differences between effect parameters.

Confidence Interval Overlap

An approach similar to the prediction interval examines a measure of confidence interval overlap. This method was seemingly proposed by Brandt et al. (2014) as an important analysis method in their "recipe" for replications. It was used by the RPP, RPE, and PPIR as part of their confirmatory analyses. Moreover, while the Many Labs Project did not report the results of this analysis, they did use a version of it to argue whether a replication was successful.

In this method, a replication is said to have failed if T_1 is *not* contained in a 95% confidence interval for θ_2, which occurs if the following is false: $-1.96\sqrt{v_2} < T_1 - T_2 < 1.96\sqrt{v_2}$. This is equivalent to testing the null hypothesis (assuming $\alpha = 0.05$) in equation (5) by comparing the test statistic

$$S = (T_1 - T_2)/\sqrt{v_2} \tag{8}$$

to the standard normal distribution. Note that this procedure implies replication follows Definition 1, that $\theta_1 = \theta_2$.

The probability that this procedure determines the studies fail to replicate when $\theta_1 = \theta_2$ (i.e., a false failure) can be quite high, and more importantly is not controlled. This is because while it compares S to the standard normal distribution, under the null hypothesis, *S actually* follows a normal distribution with mean zero and variance $1 + v_1/v_2$. Thus, the FFR is given by

$$1 - \Phi\left(\frac{1.96}{\sqrt{1+v_1/v_2}}\right) + \Phi\left(\frac{-1.96}{\sqrt{1+v_1/v_2}}\right) \tag{9}$$

The FFR is an increasing function of v_1/v_2, and hence will vary depending on the relative sample sizes in each study; it is not controlled. Figure 3(A) shows the FFR as a function of v_1/v_2. The smallest the FFR can be is 5%, which occurs when study 1 is infinitely larger than study 2, so that $v_1/v_2 \rightarrow 0$. However, the FFR will increase as v_1/v_2 gets larger. If both studies are the same size so that $v_1 = v_2$, then the FFR is 16.6%; when study 2 is twice the size of study 1, so that $v_1 = 2v_2$, then the FFR is 25.8%. In other words, even if studies replicate exactly, this test may be likely to determine that they do not.

This type of error will be particularly common for two reasons. First, replications are often designed to be larger than the original study, so we might expect $v_1/v_2 > 1$. Moreover, if T_1

must be corrected for publication selection, then the estimation error variance of this bias-corrected estimate v_1^* will be larger than v_1. Hedges (1984) shows that there are cases where corrections for publication selection increase estimation error variance by over 150%, so that $v_1^* > 1.5v_1$. Thus, if study 1 and study 2 are the same size, but the analysis must correct T_1 for publication selection, the FFR can be as high as 22%.

Second, false failures will be quite common when this procedure is applied to research designs that conduct multiple replications. Because analyses of such programs combine the results of the replications into a single effect estimate (see Klein et al., 2014; or Schweinsberg et al., 2016), that combined estimate will have a much smaller estimation error variance than if only a single replication had been conducted. This means that v_1/v_2 will almost certainly be large. In the case of Many Labs, this ratio is on average about 60!

A false success error would involve the test failing to reject H_0 when $\theta_1 \neq \theta_2$. When $\theta_1 \neq \theta_2$, S follows a normal distribution with mean and variance:

$$E[S] = \frac{\theta_1 - \theta_2}{\sqrt{v_2}}, \quad V[S] = 1 + \frac{v_1}{v_2} \tag{10}$$

Thus, the probability it makes this type of error is

$$\Phi\left(\frac{1.96}{\sqrt{1+v_1/v_2}} - \frac{\theta_1-\theta_2}{\sqrt{v_1+v_2}}\right) - \Phi\left(\frac{-1.96}{\sqrt{1+v_1/v_2}} - \frac{\theta_1-\theta_2}{\sqrt{v_1+v_2}}\right) \tag{11}$$

From equation (11), it can be shown that errors are less likely when either (a) the difference between effect parameters is large, or (b) when v_1 and v_2 are small such that v_2/v_1 is also small. This presents something of a paradox. In order to reduce the false failure rate, designs should have values of v_1/v_2 that are quite small, but those are the same designs that will increase false success rate.

Figure 3(B) shows the FSR as a function of v_1, v_2, and $|\theta_1 - \theta_2|$, where the θ_i are on the scale of Cohen's *d*. Each color corresponds to a value of $|\theta_1 - \theta_2|$ and each line type corresponds to a value of v_2 (relative to v_1). For reference $v_1 = 0.04$ would correspond to a study with about 100 subjects. What can be seen in this plot is that unless the studies are quite large (greater than 400 subjects or $v_1 < 0.01$) or the difference between effects is large (greater than 0.5), then the FSR will be high. Comparing Figure 3(A) and 3(B), we also see that designs that lead to low false failure probabilities will have $v_1/v_2 < 1$. Yet, unless study 1 is quite large (greater than 150 or 200 individuals, for example), then this procedure will be unlikely to detect even moderate differences between effect parameters.

The high FFR and FSR are borne out in the replication data. Figure 4(A) shows the probability of false failures among studies for which this method determined the replication failed. The *x*-axis reports the fraction of studies that were deemed failures by this criterion for each program, and each boxplot shows the distribution the FFR for the designs of the studies in that program. In the plot we see that all of the Many Labs replications failed according to this criterion, but that the probability that those are in error is greater than 75%. Error rates are similar among the PPIR studies. However, for both RPP and RPE, these are still above 25%.

Figure 4(B) shows the FSR among studies for which this test did not rule out replication. For each program, there are three boxplots that correspond to the FSR for a given value of $|\theta_1 - \theta_2|$. None of these studies are likely to detect differences between effects on the order of 0.2. However, the PPIR designs were much less likely to miss larger differences. Contrast that with RPP and RPE, where most of the studies would have been unlikely to detect differences that are about 0.5 in size, which means even if study 1 had a small negative effect and study 2 had a small positive effect, this test would result in a success between 30% and 60% of the time for those studies.

To put this in perspective, suppose all of the studies failed to replicate, so that $|\theta_1 - \theta_2| = 0.5$, then we would expect this test to fail to detect that 6 of the RPE studies, and 34 of the RPP studies; 40 of the 117 findings would be falsely labeled as successful replications. Conversely, suppose all of the studies in each program replicated successfully, and that there were no failures. Then we would expect this test to determine that the replications failed for about 11 of the Many Labs studies, six of the PPIR studies, 5 of the RPE studies, and 17 of the RPP studies, so that 41 of the 117 findings would be incorrectly labeled as failed replications.

Correspondence in Sign and Statistical Significance

Perhaps the most common method to determine if studies failed to replicate is if the original and replication study do not correspond in sign and statistical significance. Heuristically, this would seem logical, and indeed, Fisher (1935) noted that "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results."

The definition of replication implied by this method follows directly from the individual hypothesis tests that generate the *p*-values used in the analysis. Thus, the analysis implies that replication simply means effects are in the same direction (or both are null), and hence relies on Definition 2.

When $\theta_1 = \theta_2 = 0$, so that both studies involve null effects, then this procedure's FFR is $2\alpha(1 - \alpha) + \alpha^2/2$, where α is the level of the null hypothesis tests in the two studies. When $\alpha =$

0.05, then this rate is 9.6%. If both effects are nonzero and in the same direction, which would characterize a successful replication in this analysis, then the FFR is given by:

$$\left[1-\ \Phi\left(1.96-\tfrac{\theta_1}{\sqrt{v_1}}\right)\right]\ \Phi\left(1.96-\tfrac{\theta_2}{\sqrt{v_2}}\right)\ +\Phi\left(1.96-\tfrac{\theta_1}{\sqrt{v_1}}\right)\left[1-\ \Phi\left(1.96-\tfrac{\theta_2}{\sqrt{v_2}}\right)\right]+$$

$$\Phi\left(-1.96-\tfrac{\theta_1}{\sqrt{v_1}}\right)\left[1-\ \Phi\left(-1.96-\tfrac{\theta_2}{\sqrt{v_2}}\right)\right]+\left[1-\ \Phi\left(-1.96-\tfrac{\theta_1}{\sqrt{v_1}}\right)\right]\Phi\left(-\ 1.96-\tfrac{\theta_2}{\sqrt{v_2}}\right) \qquad (12)$$

The first line of equation (12) gives the probability that one of the estimates (T_1 or T_2) is positive and significant and the other is not, and the second line gives the probability that one is negative and significant and the other is not. This error rate is a function of $\theta_1/\sqrt{v_1}$ and $\theta_2/\sqrt{v_2}$ which determine the power of the test of no effect in study 1 and study 2, respectively. If both studies have high power, then both T_1 and T_2 are likely to be significant, and false failures will be less probable. Likewise, if both studies have low power, then T_1 and T_2 are likely to be nonsignificant, and the procedure will be less likely to determine that the studies failed to replicate.

Figure 5(A) shows the shows the FFR as a function of the power of study 1 and study 2, assuming that θ_1 and θ_2 are both nonzero and in the same direction. For instance, if both studies had 80% power to detect a nonzero effect, then the FFR will be just over 30%. In the figure we see that the FFR is largest when the power of the studies is more unequal (i.e., if one study has high power and the other low power). Moreover, unless both studies have very high power to detect a nonzero effect (and assuming the effects are nonzero), then the FFR will be higher than 30-40%.

Part of the reason the FFR is so high for this method is that concluding that a replication failed often involves a logical fallacy. A common scenario for which this method would conclude that studies failed to replicate is when one study (say the original one) has a significant effect, but the other (the replication) does not. This is implicitly taken to mean that the effect is nonzero in the first study, and the effect is zero in the second. However, just because an estimate is not statistically significant does not mean that the effect is null, and assuming that it does runs counter to the logic of null hypothesis tests.

A false success will occur when the effect estimates correspond in sign and statistical significance, but the effect parameters do not qualitatively agree. This would involve scenarios where (a) $\theta_1 > 0$ and $\theta_2 \leq 0$, (b) $\theta_1 < 0$ and $\theta_2 \geq 0$, or (c) $\theta_1 = 0$ and $\theta_2 \neq 0$. The probability that this test makes this type of error is given by:

$$\left[1 - \Phi\left(1.96 - \frac{\theta_1}{\sqrt{v_1}}\right)\right]\left[1 - \Phi\left(1.96 - \frac{\theta_2}{\sqrt{v_2}}\right)\right] +$$

$$\left[\Phi\left(1.96 - \frac{\theta_1}{\sqrt{v_1}}\right) - \Phi\left(-1.96 - \frac{\theta_1}{\sqrt{v_1}}\right)\right]\left[\Phi\left(1.96 - \frac{\theta_2}{\sqrt{v_2}}\right) - \Phi\left(-1.96 - \frac{\theta_2}{\sqrt{v_2}}\right)\right] +$$

$$\Phi\left(-1.96 - \frac{\theta_1}{\sqrt{v_1}}\right)\Phi\left(-1.96 - \frac{\theta_2}{\sqrt{v_2}}\right) \qquad (13)$$

where θ_1 and θ_2 satisfy (a), (b), or (c) above. The top and bottom rows of this equation give the probability both estimates are significant and positive or negative, respectively, and the middle row is the probability that both are nonsignificant. Note that as with the FFR, the FSR depends on the power of individual studies to detect an effect via $\theta_1/\sqrt{v_1}$ and $\theta_2/\sqrt{v_2}$.

Figure 5(B) shows the FSR as a function of the power of study 1 and study 2 to detect nonzero effects. The black line corresponds to a scenario where one effect is zero and the other is not. The other colored lines correspond to scenarios where effects are in opposite directions. It turns out that the FSR will be highest when one of the effect parameters is zero and the other is not (black line). However, even if effects are in opposite directions, unless both studies have high power to detect nonzero effects, the FSR can be larger than 20%.

Among the replication studies in the data, there are scenarios where error rates are low, and scenarios where they are high. Figure 6(A) shows the false failure rate among studies for which this method concluded the studies failed to replicate. Each research program has four box plots, and the boxes are colored according to one of four scenarios of successful replication: the effect parameters are identical and equal to 0.2, 0.5, and 0.8, as well as a scenario where study 1 has a large effect and study 2 has a small effect in the same direction. What we see is that in most scenarios that error rates are well above 20-25%. Moreover, when both effects are moderate (gold boxes) or if study 1 has a large effect and study 2 has a small effect, the RPP and RPE analyses would be very likely to determine that the replications failed.

A similar story emerges in Figure 6(B), which shows the FSR among studies deemed to have successfully replicated by this method. Each research program has four box plots that correspond to different scenarios of failed replication: three involve study 2 having an effect of zero, but study 1 having a positive effect, and the fourth involves both studies having small effects in the opposite direction. For instance, if study 1 involved a small effect ($\theta_1 = 0.2$) and study 2 had an effect of zero, which correspond to the red boxes, the FSR would be above 75% for nearly all of the studies in the data for which this metric determined the replication was successful. Moreover, even if $\theta_1 = 0.5$ and $\theta_2 = 0$, this test would frequently determine that these studies replicated. Somewhat shockingly, for the RPP and RPE programs, if study 1 and study 2 involved small effects in opposite directions (dark blue boxes) this test would be more likely to determine that these studies replicated than to determine that they did not.

Taken together, there are plausible scenarios under which this method is highly likely to lead to an error. For reference, suppose that all of the studies failed to replicate, such that $\theta_1 = 0.2$ and $\theta_2 = -0.2$, so that the effects are small but in opposite directions for each study. Then we would expect this method to determine that 12 of the RPE and 53 of the RPP replicated successfully. Conversely, suppose all of the studies in each program replicated exactly, so that $\theta_1 = \theta_2 = 0.5$. Then we would expect this method would conclude that replication failed for four Many Labs studies, four PPIR studies, eight RPE studies, and 26 RPP studies (42 out of 117 experiments)!

Meta-analytic Averages

Another type of analysis method involves a precision-weighted average of the effect estimates, often referred to as a fixed-effects meta-analysis. Determinations about replication focus whether that average is statistically significant. Interpretation of this analysis has varied across research programs. While significant averages are largely interpreted as "evidence for an effect," the RPE and PPIR further examined whether the original study and meta-analytic average corresponded in sign and statistical significance.

Using this method, the RPP determined that 68 of 100 findings involved precision-weighted averages that were statistically significant. The RPE also conducted this analysis and determined that 14 of 18 (78%) of these averages were statistically significant and in the same direction as the original study. Meanwhile PPIR conducted a meta-analysis of all studies in a given ensemble (i.e., the original study and the k – 1 replications), and determined that the average effect estimate corresponded with sign and statistical significance with the original study.

The procedure involves computing the weighted average of effects given by $T_{\cdot}$:

$$T_{\cdot} = \frac{T_1/v_1 + T_2/v_2}{1/v_1 + 1/v_2} \qquad (14)$$

Under the model, $T_{\cdot}$ has a normal distribution with mean $\theta_{\cdot}$ and variance $v_{\cdot}$:

$$\theta_{\cdot} = \frac{\theta_1/v_1+\theta_2/v_2}{1/v_1+1/v_2}, \quad v_{\cdot} = \frac{v_1 v_2}{v_1+v_2} \tag{15}$$

The analysis involves testing

$$\mathrm{H}_0\text{: } \theta_{\cdot} = 0 \tag{16}$$

by computing $T_{\cdot}/\sqrt{v_{\cdot}}$. We reject H_0 if $|T_{\cdot}|/\sqrt{v_{\cdot}}$ exceeds $1 - \alpha/2$ percentile of the standard normal distribution $c_{(1-\alpha/2)}$ The power of this test is given by:

$$1 - \Phi\left(c_{(1-\alpha/2)} - \frac{\theta_{\cdot}}{\sqrt{v_{\cdot}}}\right) + \Phi\left(-c_{(1-\alpha/2)} - \frac{\theta_{\cdot}}{\sqrt{v_{\cdot}}}\right) \tag{16}$$

Note that the power is increasing as a function of $\theta_{\cdot}/\sqrt{v_{\cdot}}$, and has been studied by various authors (Hedges & Pigott, 2001).

Because the parameter of interest for this analysis is $\theta_{\cdot}$, it cannot identify if a given replication succeeded or failed according to either Definition 1 or 2. However, it can be a useful analysis when $\theta_1 \approx \theta_2 \approx \theta_{\cdot}$, so that both studies involve the same effect size, because it provides a powerful test than the one conducted individually in studies 1 or 2. In this sense, this method is less an analysis of replication, and more an analysis *assuming a successful replication*.

However, when $\theta_1 \neq \theta_2$, it is not clear how to interpret $\theta_{\cdot}$ and the results of this analysis can be misleading. This is because determination about "an effect" does not make much sense if there are two distinct effect parameters that disagree. For instance, if $\theta_1 = 0.8$, $\theta_2 = -0.8$, and $v_1 = v_2$, then the average effect $\theta_{\cdot} = 0$. Thus, any interpretation of this test would not identify the fact that both individual studies involve large effects and that those large effects are very different from each other. However, it would only conclude the effect is nonzero with probability α. An important corollary to this issue is that the results of the previous sections show that tests to determine if θ_1 and θ_2 disagree (in size or sign) will almost always be underpowered, and hence will be unlikely to shed light on whether θ_1 and θ_2 are similar.

It is further worth noting that effects from both studies are statistically significant and in the same direction, then their weighted average $T.$ will also be statistically significant and in that same direction. When effect parameters are very different (e.g., $\theta_1 >> \theta_2 > 0$), then both the statistical significance criterion *and* the meta-analytic average would indicate successful replication. When two criteria point a successful replication, one might be more inclined to conclude that the replication was indeed successful, even if θ_1 and θ_2 differ by orders of magnitude.

Limitations of Existing Designs and Analyses

This article has shown that in practice, many analysis methods for replication have suboptimal properties, such as uncontrolled and large error rates. Other approaches that have controlled error rates, or that are unbiased tests of replication, tend to have low power. This has been noted by various researchers, including Maxwell, Lau, and Howard (2015). Hedges and Schauer (2019a) show that even the most powerful unbiased test to assess Definition 1 of replication (the prediction interval/Q test) will almost never be large with only two studies. This has serious implications not just for interpreting the analyses of prior replication efforts, but also for the design of future studies. Put another way, how should replication research proceed if even the most powerful test will almost always be underpowered no matter how large the replication study is?

One answer is to conduct more studies. Hedges and Schauer (2018) show that with a greater number of studies, the Q test can detect smaller differences between effect parameters, and it is possible to design multiple replications so that the Q test has higher power. This approach changes the focus of analysis slightly: it assesses how consistent the body of evidence for a finding is. The underlying definition of replication remains unchanged, namely that effects should be similar in size, but the aim is less on rooting out false publications and more about assessing how reliably an effect can be recreated.

Another approach is to shift the question of replication to one of sensitivity. This is part of what underpins the "small telescopes" analysis method (Simonsohn, 2015). Small telescopes does not seek to determine if effects are similar under either of the potential definitions described in this paper. Rather, it asks whether the effect in the replication could have been detected by the original experimental design. In doing so, replication studies can be designed to ensure high power for the small telescopes test. However, it will not be conclusive about potentially important definitions of replication, nor is it clear how this method accounts for multiple and potentially heterogeneous replications (such as with Many Labs).

A third approach is to quantify the uncertainty inherent in analyses and present some gradient of how conclusive they are. Bayesian approaches to this have been proposed in the literature. For instance, Etz and Vandekerckhove (2016) use Bayes factors to qualify evidence about replication as strong, weak, and inconclusive. However, given that standard errors of effect estimates will often be large relative to potentially meaningful differences between studies, such metrics may frequently categorize evidence as "inconclusive" in the absence of very strong priors. In fact, in their re-analysis of the RPP data, Etz and Vandekerckhove's metric was inconclusive for 64% of findings. This is consistent with work by Hedges and Schauer (2019a) who show that both Bayesian and frequentist estimation methods will have low precision with only $k = 2$ studies.

Conclusions

This paper attempted to clarify two important aspects about methods for assessing replication: (1) the operational definition of replication that they test, and (2) how frequently they can lead to erroneous conclusions. We found that there are really two main definitions of replication in use. One involves effect parameters that are similar sizes. The other requires effects to agree qualitatively (i.e., have the same sign) though does not require them to be the same size. We also found that some common metrics used to determine if a replication is successful (confidence interval overlap and correspondence in statistical significance) do not have controlled error rates, which can lead to scenarios where these procedures are very likely to conclude that studies failed to replicate when they actually replicated successfully.

This is borne out empirically. For much of the data from replication research programs, determinations that studies failed to replicate based on some of these criteria are likely to be in error. Conversely, while the prediction interval test controls the type I error rate, it will often have low power when applied to only two studies. This is not just a limitation of that test, but also a limitation of design. Any assessment of replication (as it seems to be commonly defined) must contend with the fact often effects are estimated with enough noise to overwhelm signals of successful or unsuccessful replication (Hedges & Schauer, 2019a).

These results suggest two main issues with designing and analyzing replication studies. The first is that it is clear that certain analysis methods have very poor statistical properties, including uncontrolled error rates (i.e., confidence overlap criterion, statistical significance criterion). These ought to be avoided, particularly because optimal analysis methods exist for Definition 1 of replication (Hedges & Schauer, 2019a,b) and there is relevant work in meta-analysis relevant to Definition 2 (see, e.g., Piantadosi & Gail, 1993).

The second is that even using optimal analysis methods, there may be no feasible design that renders analyses sufficiently sensitive. The prediction interval (and the Q test), which is the UMP test, is bound to have low power when only $k = 2$ studies are involved, particularly if the original study has been published, and it certainly appears to have low power in the empirical research programs considered here. We have argued that there are two potential ways to proceed from this problem. One involves changing the definition of replication to something less meaningful and less consistent with standard scientific ideas about replication (e.g., small telescopes). The other involves conducting additional studies (i.e., $k > 2$) and studying variation across all k studies, rather than privileging the original study.

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1), 1–12. https://doi.org/10.1037/met0000051

Bahr, H. M., Caplow, T., & Chadwick, B. A. (1983). Middletown III: Problems of replication, longitudinal measurement, and triangulation. *Annual Review of Sociology*, 9(1), 243–264. https://doi.org/10.1146/ annurev.so.09.080183.001331

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*, 452-454.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & 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. Arlington, VA: National Science Foundation.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217-224.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Cassella, G. & Berger, R. L. (2002). *Statistical Inference (2nd edition)*. Ontario: Thompson Learning.

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

Collins, H. M. (1992). *Changing order: Replication and induction in scientific practice*. Chicago: University of Chicago Press.

Connor, S. (2015). Study reveals that a lot of psychology research really is just “psycho-babble”. *The In- dependent*. Retrieved from http://www.independent.co.uk/news/science/study-reveals-that-a-lot-of- psychology-research-really-is-just-psycho-babble-10474646.html

Cooper, H. M. (2011). *Reporting research in psychology: How to meet Journal Article Reporting Standards*. Washington, DC: APA Books.

Cooper, H. M., Hedges, L. V. & Valentine, J. (2009). *The handbook of research synthesis and meta-analysis (2nd edition)*. New York: The Russell Sage Foundation.

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. *Australian Journal of Psychology*, 64: 138-146.

Etz, A. & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11(2).* e0149794. doi:10.1371/journal.pone.0149794.

Fisher, R. A. (1935) *The Design of Experiments*. Oxford: Oliver & Boyd.

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research*, 36(2), 153–172.

Hartgerink, C. H. J., Wicherts, J. M., & van Assen, M. A. L. M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology, 3(1*), 9.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61-85.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science?: The empirical cumulativeness of research. *American Psychologist, 42*, 443-455.

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

Hedges, L. V. & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*, 203-217.

Hedges, L. V., & Schauer, J. M. (2018). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*.

Hedges, L. V., & Schauer, J. M. (in press). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294(2)*, 218-228.

Klein, R. A. et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142-152. doi:10.1027/1864-9335/a000178.

Klein, R. A. et al. (2018). Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. Retrieved from https://www.psychologicalscience.org/redesign/wp-content/uploads/2018/11/ManyLabs2.pdf.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151-159.

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published research findings are false, but a little replication goes a long way. *PLOS Medicine*, 4(2), e28. https://doi.org/10.1371/ journal.pmed.0040028

Olive, K. A. et al. (2014). Review of particle properties. *Chinese Physics Journal C, 38*, 090001. http://iopscience.iop.org/issue/1674-1137/38/9.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716.

Pashler, H. & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Psychological Science, 7*, 531-536.

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*(4), 539-44.

Rosenfeld, A. (1975). The particle data group: Growth and operations. *Annual Review of Nuclear Science,* 555-559.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90-100.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, 66, 55–67.

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

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., ... Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, 12(2), 103–117. https://doi.org/10.1007/s11121-011-0217-6

Viechtbauer, W. & Cheung, M. W. (2010), Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2), 112-125.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2): 129-133. doi:10.1080/00031305.2016.1154108

Yong, E. (2016). The inevitable evolution of bad science. *The Atlantic*. Retrieved from https://www.theatlantic.com/science/archive/2016/09/the-inevitable-evolution-of-bad-science/500609/

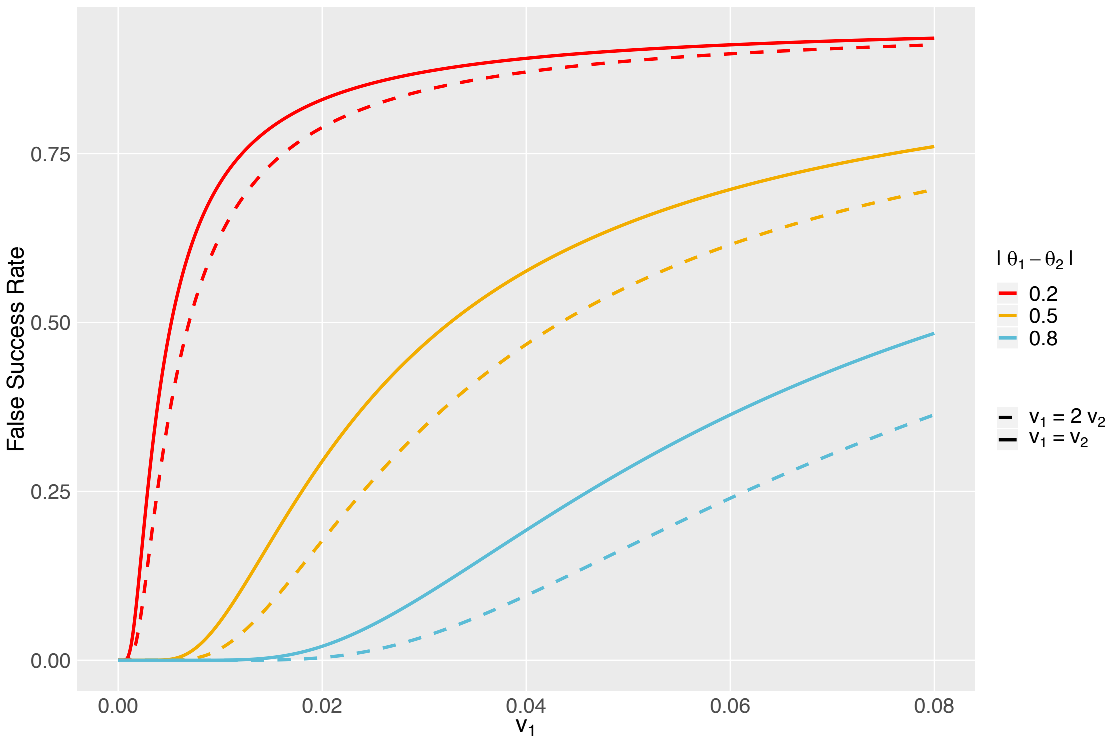

Figure 1: False Success Rate of the *Q* Test for Exact Replication. This plot shows the FSR of the *Q* test as a function of the difference between effects $|\theta_1 - \theta_2|$ (i.e., how badly the replication failed), v_1, and the ratio v_1/v_2.

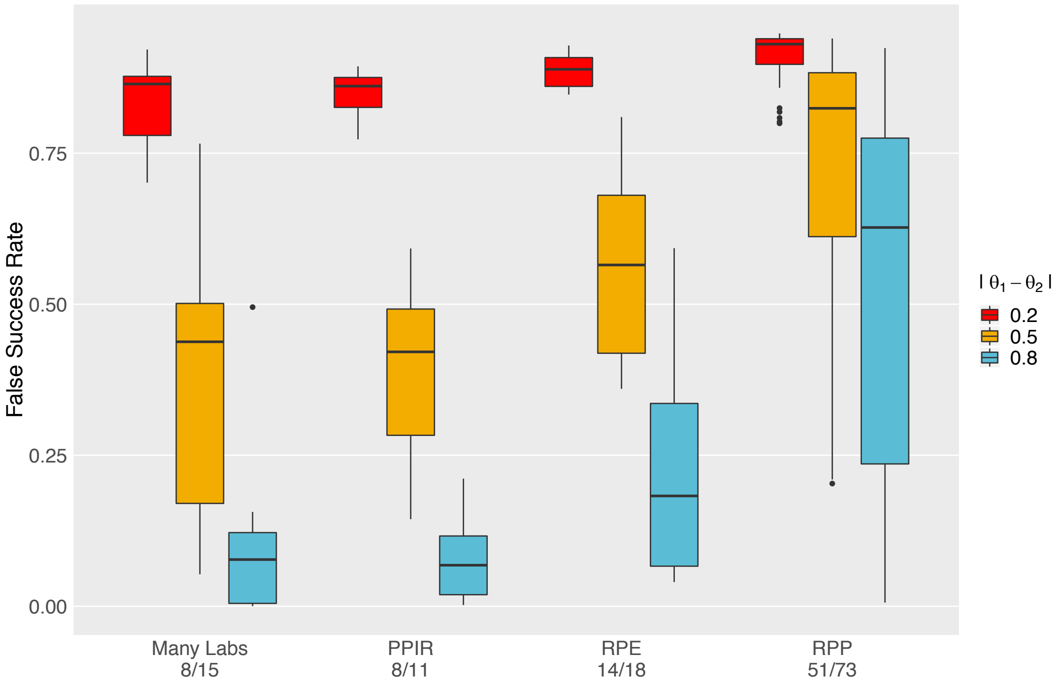

Figure 2: False Success Rates of Q Test for Empirical Replication Research Programs. This plot shows the empirical false success rate of the Q test among studies for which Q is not significant. Each box corresponds to the distribution of failure rates for a given value of $|\theta_1 - \theta_2|$.

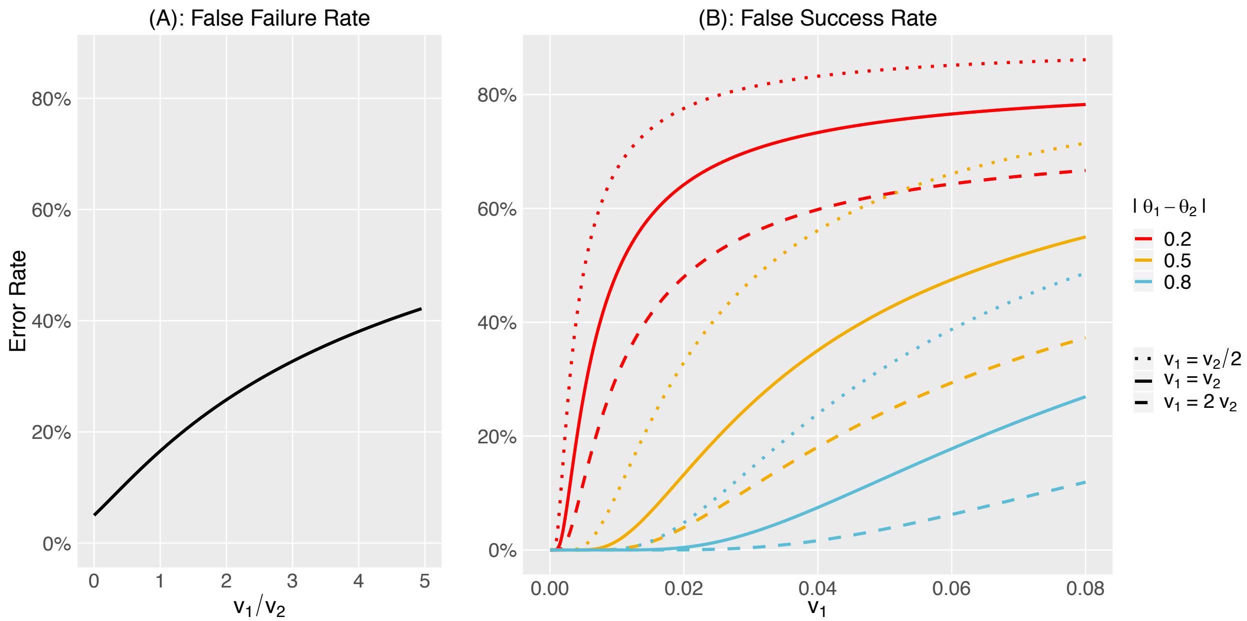

Figure 3: Error Rates of Confidence Interval Overlap Procedure. This plot show the FFR (A) and FSR (B) for the CI overlap criterion. For the FFR, the error rate increases as a function of v_1/v_2. For the FSR, the error rate depends on $|\theta_1 - \theta_2|$, v_1, and v_1/v_2.

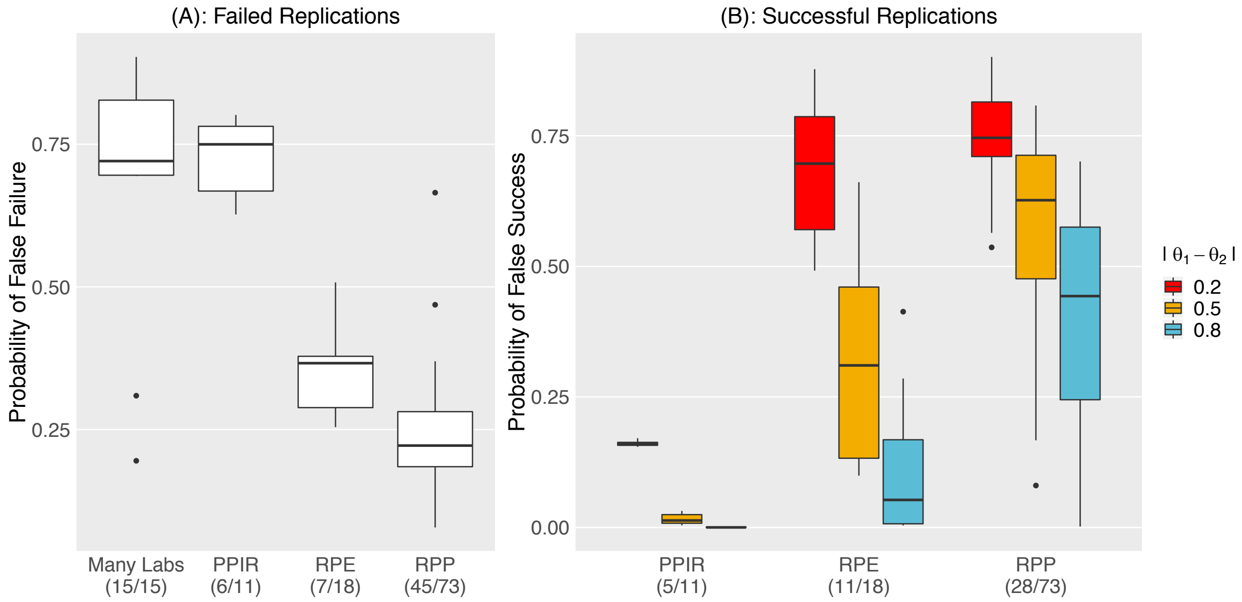

Figure 4: Error Rates For Confidence Interval Overlap Among Empirical Research Designs. These plots show the distribution of FFRs (A) and FSRs (B) of the CI overlap method for studies determined to have failed (A) and succeeded (B) according to that criterion. Note the FSR depends on $|\theta_1 - \theta_2|$.

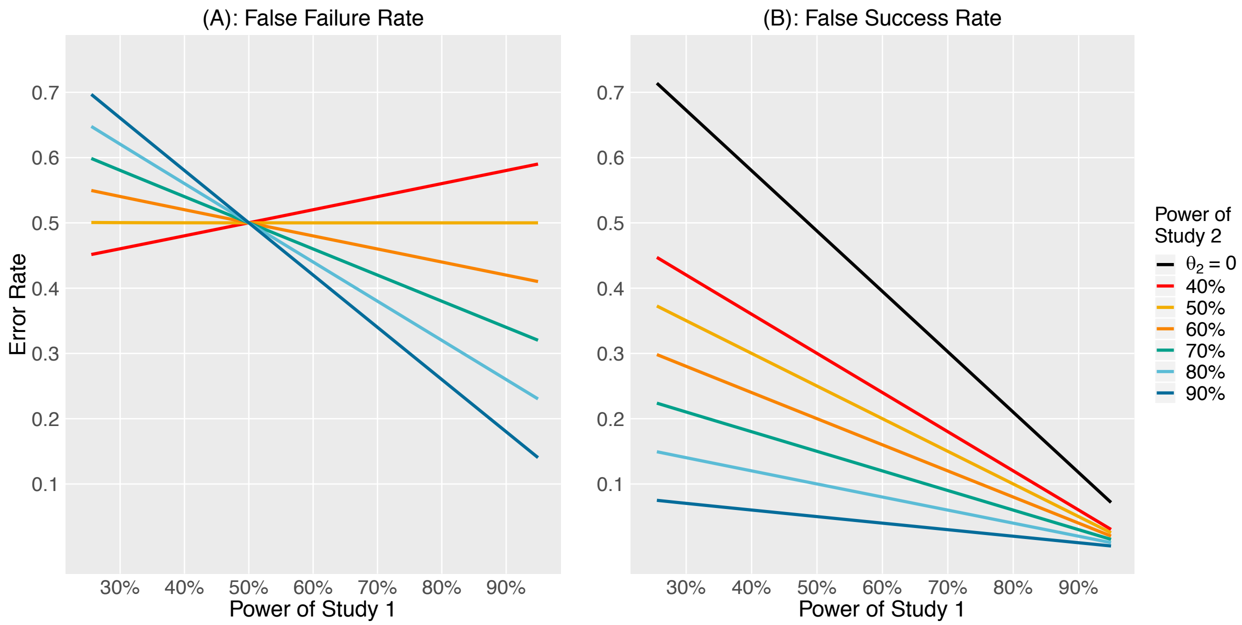

Figure 5: Error Rates for Sign and Statistical Significance Comparisons. These plots show the FFR (A) and FSR (B) of the statistical significance criterion as a function of the power of study 1 and study 2.

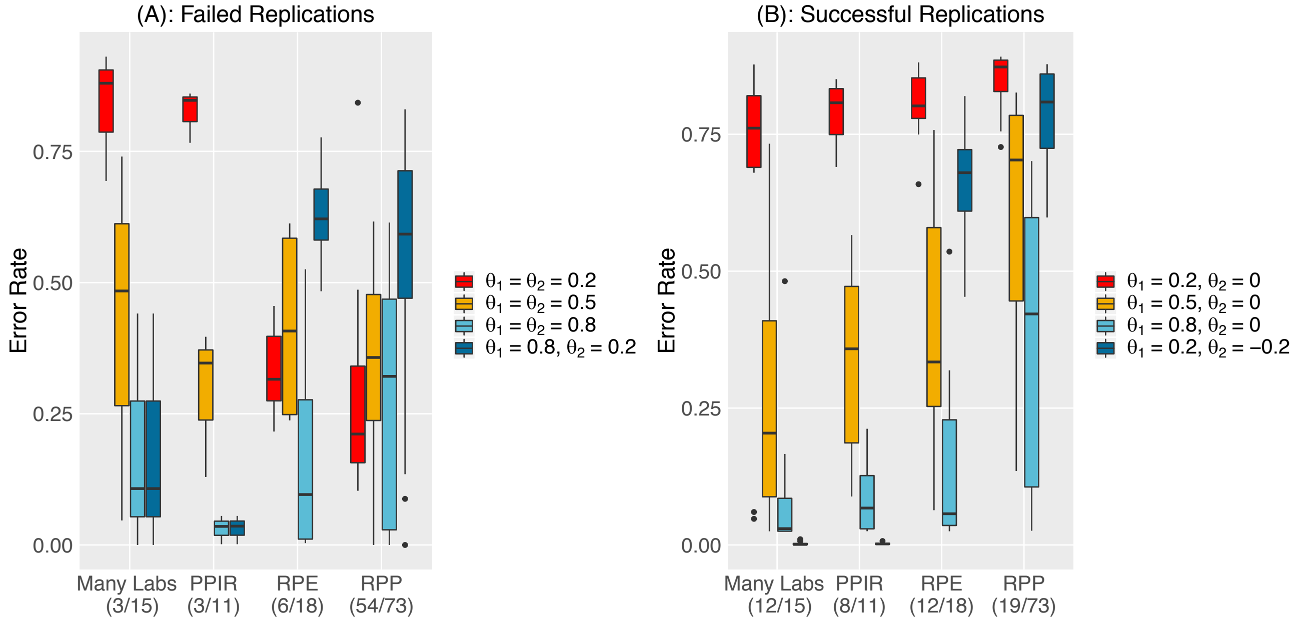

Figure 6: Error Rates for Sign and Statistical Significance Comparisons in Empirical Research. These plots show the FFR (A) and FSR (B) for studies that failed (A) and passed (B) the statistical significance criterion. For each plot, the error rates depend on the value of θ_1 and θ_2.