American Psychological Association

Psychological Methods

© 2018 American Psychological Association
1082-989X/19/$12.00

2019, Vol. 24, No. 5, 557–570
http://dx.doi.org/10.1037/met0000189

Statistical Analyses for Studying Replication: Meta-Analytic Perspectives

Larry V. Hedges and Jacob M. Schauer
Northwestern University

Abstract

Formal empirical assessments of replication have recently become more prominent in several areas of science, including psychology. These assessments have used different statistical approaches to determine if a finding has been replicated. The purpose of this article is to provide several alternative conceptual frameworks that lead to different statistical analyses to test hypotheses about replication. All of these analyses are based on statistical methods used in meta-analysis. The differences among the methods described involve whether the burden of proof is placed on replication or nonreplication, whether replication is exact or allows for a small amount of "negligible heterogeneity," and whether the studies observed are assumed to be fixed (constituting the entire body of relevant evidence) or are a sample from a universe of possibly relevant studies. The statistical power of each of these tests is computed and shown to be low in many cases, raising issues of the interpretability of tests for replication.

Translational Abstract

The idea that a finding can be replicated is fundamental to scientific progress. However, several recent studies have called into question the replicability of findings in different fields, including psychology. These studies have garnered attention both in academia and in the popular press, and have become important evidence of a crisis in science. On its face, replication seems like a straightforward idea: just repeat an experiment and check that you get the same results. However, authors of replication studies have noted that it is not that simple. Indeed, analyses of these studies have revealed that we might mean several different (and conflicting) things when we refer to "results" being "the same." This article attempts to clarify some of this ambiguity. It describes a way to precisely define when study results are the same. It also provides analyses that test whether data from replicate studies are consistent with that definition. In general, we find that defining replication and properly framing the analysis requires serious effort, and that unless several studies are conducted, the results of analyses about replication may be inconclusive.

Keywords: replication, meta-analysis, heterogeneity, replicability, reproducibility

Supplemental materials: http://dx.doi.org/10.1037/met0000189.supp

The idea that scientific studies can be replicated is a fundamental aspect of the rhetoric of the scientific method, and is part of the logic supporting the claim that science is self-correcting, because replication attempts will identify findings that are incorrect (see, e.g., McNutt, 2014). The replicability of scientific findings in medicine has recently been called into question by empirical analyses (e.g., Collins & Tabak, 2014; Ioannidis, 2005; Perrin, 2014). Similar challenges have also emerged in psychology (e.g., Open Science Collaboration, 2015) and economics (e.g., Camerer et al., 2016). There is substantial evidence that scientists themselves are concerned about replicability in many disciplines (e.g., Baker, 2016; Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015), including psychology (e.g., Pashler & Harris, 2012). Such concerns are not limited to academic journals but have also appeared in the popular press, including *Newsweek* and *The Economist*.

There seems little doubt that empirical evaluation of the replicability of research will continue. Yet there appear to be important differences in the methods used to determine whether replication has occurred. Ioannidis (2005) used agreement between "the final interpretation of the results by the authors" (p. 219) as a criterion for replication. The Open Science Collaboration (2015) used several methods, including comparing *p* values of original and replication studies, and by assessing whether the effect size of the original study was in the confidence interval of the replication, but Gilbert, King, Pettigrew, and Wilson (2016) challenged their analysis methods. It is worth noting that the discussion of how to assess replication has persisted for many years (see, e.g., Humphreys, 1980).

Because the concept of replication is so central to the logic and rhetoric of science, it would be reasonable to expect a substantial literature on the methodology of replication, including guidelines for designing replication studies and analyzing ensembles of stud-

This article was published Online First August 2, 2018.

Larry V. Hedges and Jacob M. Schauer, Department of Statistics, Northwestern University.

This research was supported in part by two grants from the U.S. Institute of Education Sciences: R305D140045 and R305B140042.

Correspondence concerning this article should be addressed to Larry V. Hedges, Department of Statistics, Northwestern University, Evanston, IL 60208. E-mail: l-hedges@northwestern.edu

ies that are replications. But as Schmidt (2009) has pointed out, "the opposite is true" (p. 90). This is not to say that there is no literature on replication; however, much of it (e.g., Lykken, 1968) focuses on the definition and functions of replication, not on the analysis of replications (see Schmidt, 2009). Note that the kind of replication that is the focus of this article is what Schmidt would call direct replication, which involves the "replication of an experimental procedure" (p. 91) as opposed to what he calls "conceptual replication," which involves the "repetition . . . of earlier research work with different methods" (p. 91). Although it is not always easy to distinguish between these two types of replication, the intent of researchers is sometimes reasonably clear (as it is in the programs of preregistered replications such as that considered in our example).

One reason that findings may fail to replicate is because studies are more likely to appear in the published literature if their results are statistically significant, a phenomenon often called "publication bias" (see., e.g., Rothstein, Sutton, & Borenstein, 2005). There is considerable evidence that publication bias exists in the biomedical and social sciences (see, e.g., Dickersin, 2006). Such selection can lead to biases in effect size estimates as much as 200% in extreme cases (Hedges, 1984). Because a replication is not necessarily subject to the same publication selection as the original study (particularly in programs of replication that have registered protocols in advance), effect sizes in replications are expected to be smaller in absolute magnitude.

Consequently, some scholars have emphasized the need to adjust for the effects of publication bias in original published studies before comparing their results with replications. Several methods of adjustment for publication bias have been suggested, including maximum likelihood estimation of effect sizes under selection models specified a priori or estimated from the data (e.g., Hedges & Vevea, 2005), or Bayesian methods (e.g., Guan & Vandekerckhove, 2016). It is therefore natural to adjust for the effects of publication bias in original published studies before comparing their results with replications. A Bayesian method for doing so was presented by Etz and Vandekerckhove (2016), and a hybrid model was presented by van Aert and van Assen (2017). A frequentist approach is presented by Hartgerink, Wicherts, and van Assen (2017).

Publication bias is only one of many possible causes that might lead to failures to replicate. A variety of questionable research practices (such as modifying samples or analyses until the results are statistically significant or erroneously rounding p values until they are significant), often called p-hacking, reduce the replicability of research (see, e.g., Head, Holman, Lanfear, Kahn, & Jennions, 2015). There is empirical evidence that p-hacking occurs in published research (see, e.g., Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016). Moreover, surveys of researchers suggest that many psychologists admit to practices that constitute p-hacking (see John, Loewenstein, & Prelec, 2012, or Fiedler & Schwarz, 2016).

The purpose of this article is to offer several approaches to testing hypotheses about replication that are consistent with meta-analysis as a method of summarizing research findings. Approaches to the analysis of replication efforts that emphasize comparison of effect sizes (rather than p values or qualitative interpretations of findings) are not the only way to study replication empirically, but they are one way that is clearly in the same spirit as meta-analysis and contemporary research synthesis. The approaches we present do not include attempts to adjust for publication bias or p-hacking. Thus, they are suitable for situations in which these considerations should be minimal, such as evaluating studies that are part of designed programs of replication or other situations in which study protocols have been registered in advance. We present methods to illustrate the sensitivity of these tests (including power calculations), as well as examples of how to conduct them using data from the Open Science Collaboration. The approach we propose uses the Q-statistic that was introduced to study replication in physics by Birge (1932). It was independently introduced for studying heterogeneity of effects in some analysis of variance models by Cochran (1954) and independently introduced again by Hedges (1982) for studying heterogeneity of effect sizes in psychology, who also gave its formal asymptotic distribution when the effect sizes being combined were standardized mean differences. The power function of the Q test was given by Hedges and Pigott (2001). We introduce no new statistical theory here.

Theoretical Considerations

This article explores replication somewhat differently than other empirical evaluations or proposed methods. A common approach in the literature compares a target study (typically the initial finding) with one or several replications (e.g., Etz & Vandekerckhove, 2016; Klein et al., 2014). The general framework of "target study versus replication" answers important questions but has some limitations. For example, comparing an initial study with an aggregate finding from replications may not address lack of agreement among the replications. Likewise, methods that do not evaluate replication in terms of effect parameters (such as comparing the sign and statistical significance of observed effects) may conclude that very different patterns of findings constitute replication.

The meta-analytic approach presented here addresses both of these issues. This approach considers the effect size parameters of the observed studies to be the underlying results of those studies, unaffected by estimation error. The meta-analytic approach assesses replication not in terms of observed estimates (which are affected by estimation error), but in terms of the effect parameters they estimate. Then, rather than singling out a privileged study for comparison, this approach characterizes the overall heterogeneity of effects across all studies. This allows one to answer questions about differences throughout a body of evidence instead of validating one result.

As proposed in this article, statistical approaches to studying replication depend on three considerations that are largely conceptual: how replication is defined (as exact or approximate replication), how the hypothesis test is structured (whether the burden of proof lies with replication or nonreplication), and whether the studies are conceived as the only studies relevant to the replication question or as a random sample from a universe of studies relevant to evaluating replication. Each of these considerations has important consequences for the properties of tests for replication.

The Definition of Replication: Exact Replication or Approximate Replication?

Perhaps the most important consideration is the precise *definition* of replication. One possible definition of replication is that all studies have exactly the same effect parameter. Although this is logically appealing, it may be too strict to be useful in scientific practice. Even in strong sciences like physics, there is awareness that even the most careful experiments cannot eliminate all biases (see, e.g., Hedges, 1987; Olive et al., 2014; Rosenfeld, 1975). Therefore, *some* variation in effects across attempted replication studies might be expected as a consequence of good scientific practice. There is also an issue of how much precision is implicit in the interpretation of a result. Small differences in the magnitude of effects (associated with slight variations in instrumentation and procedures) may not lead to different interpretations of a finding.

Alternatively, replication may correspond to practical equivalence, in which effects are "almost the same" across studies, such that "almost the same" is defined precisely. Later in this article, we offer conventions used in three scientific areas to quantify this notion. However, we regard the decision of what specific convention is appropriate in any field to be a matter of scientific judgment that might well differ across fields.

How the Hypotheses Test Is Structured: Is the Burden of Proof on Nonreplication or on Replication?

Another important consideration is exactly how the null hypothesis is conceived. One possibility is to structure the null hypothesis to correspond with replication (by some definition) and the alternative hypothesis with nonreplication. In this conception, we maintain the hypothesis that the studies replicate unless the evidence supports its explicit rejection. Thus, the burden of proof is on nonreplication. In this setup, the finding of "replication" (failure to reject the null hypothesis) is inconclusive, but the finding of "nonreplication" (rejection of the null hypothesis) is conclusive.

Conversely, one may structure the null hypothesis so that it corresponds to nonreplication and the alternative hypothesis to replication. In this case, we conclude only that studies replicate when the evidence supports the explicit rejection of nonreplication. Thus, the burden of proof is on replication. This is similar to the framework for equivalence testing (see, e.g., Wellek, 2002). It frames the underlying effect parameters as being not necessarily identical but nearly so according to some indifference zone of limited heterogeneity. The null hypothesis is that some effect parameters are outside this indifference zone, so that rejection implies that all effect parameters lie within the indifference zone.

Do the Observed Studies Comprise the Population or a Sample?

Regardless of how replication is defined and how the hypothesis test is structured, the studies available can be considered in either of two different ways. If the observed studies constitute the entire population of studies relevant to assessing replication, then inferences about replication are inferences about the effect parameters in the studies actually observed. This is consistent with the fixed effects framework in meta-analysis (see, e.g., Hedges & Vevea, 1998). One might say that conclusions about replication in the fixed studies framework are conclusions about how well the effect parameters in the *observed* studies agree.

If the studies observed are a sample from a hypothetical population or universe of studies, then their effect parameters are a sample from a hypothetical universe of effect parameters. Inferences about replication are inferences about the universe of effect parameters from which the sample was taken. The observed studies and their effect parameters are of interest only in that they provide information about this hypothetical universe. This is consistent with the random effects framework in meta-analysis (see, e.g., Hedges & Vevea, 1998). One might say that conclusions about replication in the random studies framework are conclusions about how well effect parameters agree *in the universe* of studies, in which that universe is one that might have yielded the observed studies as a random sample.

The difference between these two frameworks may seem trivial. In fact, both use the same test statistic (the Q-statistic) from meta-analysis. When there is perfect agreement across studies (exact replication), Q has the same distribution regardless of whether the studies are treated as fixed or random (see, e.g., Hedges & Olkin, 1985). However, there are two important differences between these models. First, they answer slightly different questions. The fixed effects model addresses agreement between only the observed studies, whereas the random effects approach pertains also to an entire population of studies that might include potential future studies. Second, when there is not perfect agreement in effect parameters across studies, the Q-statistic has a slightly different sampling distribution when studies are considered fixed than when they are considered random. This has implications for statistical power and constructing tests for approximate replication. It should be noted that the choice of whether to treat the studies as fixed or random is sometimes contentious in meta-analysis (see, e.g., Hedges & Vevea, 1998). In fact, it is a special case of the general issue of conditional versus unconditional inference in statistics, which has been an important debate since the beginning of modern statistics early in the 20th century (see Camilli, 1990, for a review of the conditionality issue in conjunction with the analysis 2 × 2 contingency tables).

When the amount of heterogeneity is small or the number of studies is large, Q has approximately the same distribution in both the fixed and random effects analyses. To simplify presentation in this article, we give only the results of the fixed studies analyses. The corresponding random studies framework analyses exhibit similar power to the values in Tables 1 to 4, which are computed under the fixed studies assumption. In many practical scenarios, the power differs only in the second decimal place.

Statistical Models

Suppose that k studies are potential replicates of one another. Let $\theta_1, \ldots, \theta_k$ be the effect size parameters from the studies and let $T_1, \ldots, T_k$ be the effect size estimates with known estimation error variances $v_1, \ldots, v_k$. Assume that the effect size estimates are approximately normally distributed so that

$$T_i \sim \mathrm{N}(\theta_i,\ v_i).$$

A primary statistical tool used in this article is the Q-statistic, which is also used in testing for heterogeneity of effects across studies in meta-analysis, and is defined by

$$Q = \sum_{i=1}^{k} (T_i - T_\bullet)^2 / v_i, \tag{1}$$

where $T_\bullet$ is the inverse variance weighted mean of the T_i given by

$$T_{\bullet} = \frac{\sum_{i=1}^{k} T_i/v_i}{\sum_{i=1}^{k} 1/v_i}$$

(see, e.g., Hedges & Olkin, 1985).

When studies are conceived as fixed, but when

$$H_0\colon \theta_1 = \cdots = \theta_k$$

is false, then Q has the noncentral chi-squared distribution with $k - 1$ degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{k} \frac{(\theta_i - \theta_{\bullet})^2}{v_i}, \qquad (2)$$

where $\theta_{\bullet}$ is the weighted mean of the θ_i given by

$$\theta_{\bullet} = \frac{\sum_{i=1}^{k} \theta_i/v_i}{\sum_{i=1}^{k} 1/v_i}.$$

(see, e.g., Hedges & Pigott, 2001). Note that the distribution of Q when the null hypothesis of exact homogeneity is false is determined only by k, the number of studies, and the noncentrality parameter λ.

It is also worth noting here that we can define replication in this model entirely in terms of λ. This is not the only way to define replication, but it provides a natural measure of heterogeneity among the effect parameters. Exact replication, in which all of the θ_i are equal, corresponds to $\lambda = 0$. Small values of λ can be associated with approximate replication. For example, if $\lambda \leq \lambda_0$ for some "small enough" value λ_0, we might conclude that the studies approximately replicate. However, as described below, characterizing the magnitude of negligible differences in effects is an important consideration in assessments of replication.

Some authors have noted in the context of meta-analysis that the heterogeneity test based on the statistic Q has low power. It is important to recognize that this is the likelihood ratio test under the model considered here. Thus, the test based on Q is the uniformly most powerful unbiased test. This means that no other unbiased test (including those that have not been proposed yet) can have higher statistical power. Thus, the low power observed in some situations is not a fault of the test but a limitation of the information contained in the data in that situation, because no other unbiased test could have higher power.

It is also true that the distribution of Q depends on the fact that the effect size estimates are conditionally normally distributed given the effect size parameters with known variances. Although this is often a reasonable modeling assumption, for example, when the effect sizes are standardized mean differences or Fisher z-transformed correlations derived from normally distributed observations, substantial departures from normality of the effect size distribution would be a cause for concern (as it is in other parametric tests such as the F test in analysis of variance). Note that no distributional assumptions about the θ_i are required in the studies-fixed model considered in this article, because the derivation of the distribution of Q assumes that the θ_i are fixed, but unknown, constants.

How Should We Assess the Magnitude of Heterogeneity?

We argued previously that exact replication may be too stringent a definition of replication to be scientifically useful because some variation in results is expected even in strong sciences. Exact replication is well defined, but the definition of approximate replication requires some characterization of how much heterogeneity may be considered negligible. Even if the analysis is framed in terms of exact replication, some judgment about the magnitude of heterogeneity is required to carry out power analyses to evaluate the sensitivity of the test.

We offer two frameworks that might be useful in evaluating the magnitude of heterogeneity: one based on the variation of study results and one based on the (largest) difference between any two study's results. The first of these is more natural when studies are treated as having random effects, whereas the second conception is more natural when studies are conceived as having fixed effects. However, both frameworks can be seen to apply in a loose sense regardless of whether studies are conceived as fixed or random.

Assessing Heterogeneity by Variation of Study Results

One way to gain insight about the noncentrality parameter λ is to note that if all the estimation error variances are the same so that $v_1 = \ldots = v_k = v$, then λ can be seen as $(k - 1)/v$ times the "variance" of the θ_i values,

$$\lambda = \sum_{i=1}^{k} \frac{(\theta_i - \overline{\theta})^2}{v} = (k-1)\tau^2/v, \qquad (3)$$

where τ^2 is the "variance" of the θ_i values. (The concept of variance invoked here is as a descriptive statistic, not as a property of a random variable, but we offer it as crude way to gain intuition about λ.)

We offer three conventions that have arisen in different sciences for identifying a negligible value of λ. In high-energy physics, the Particle Data Group (which has been compiling meta-analyses of high energy physics experiments for over 50 years) concludes that $Q/(k - 1) \leq 1.25$ corresponds to negligible heterogeneity (see Olive et al., 2014). Because the expected value of Q (under the fixed effects model) is $k - 1 + \lambda$, this implies that $\lambda = (k - 1)/4$ would be a negligible value of λ.

In personnel psychology, Hunter and Schmidt (1990) propose a 75% rule, which says that when the estimation error variance v is at least 75% as large as the total variance of the effect size estimates $(v + \tau^2)$, then the variance of the effect size parameters τ^2 could be considered negligible. This implies that values of $\tau^2/v = 1/3$ and $\lambda = (k - 1)/3$ correspond to negligible amounts of heterogeneity in effect size parameters.

In medicine, a value of $I^2 = 100\% \times \tau^2/(v + \tau^2)$ of 40% or less is considered to be "not important" (see Higgins & Green, 2008, Section 9.5.2). This implies that $\tau^2/v = 2/3$ would be a negligible value of τ^2/v and that $\lambda = 2(k - 1)/3$ would be a negligible value of λ.

These three conventions provide a range of definitions of negligible heterogeneity from $\lambda = (k - 1)/4$ to $\lambda = 2(k - 1)/3$. They are certainly not the only way to characterize the magnitude of heterogeneity (see, e.g., Pigott, 2012, pp. 55–66). But they are conventions that are shared by large groups of researchers in each of three different sciences. We will use these three conventions to illustrate the methods suggested in this article, but the methods could be used with any standard for negligible heterogeneity that can be expressed as values of λ.

Relating Heterogeneity Parameters to Differences Between Studies

It is natural to think of heterogeneity in terms of the differences among the θ_i values. Here we mean a priori, theoretically justified differences that are independent of the data under consideration. When all the v_i are equal so that $v_1 = \ldots = v_k = v$, and we can describe the θ_i as normally distributed about a common mean with variance τ^2, then the average difference between any two θ_is is

$$E\{|\theta_i - \theta_j|\} = 2\tau/\sqrt{\pi}. \tag{4}$$

When the θ_i are mean differences, then $v = 2\sigma^2/n$, where σ^2 is the variance of the n observations in each treatment group, and the standardized difference $(\theta_i - \theta_j)/\sigma$ can be interpreted as the difference between two Cohen's d's: It describes the difference between study findings in units of the standard deviation of the observations in each study. Thus, because $\lambda = (k - 1)\tau^2/v$, the average absolute *standardized* difference between study effects is

$$E\left\{\left|\frac{\theta_i}{\sigma} - \frac{\theta_j}{\sigma}\right|\right\} = 2\sqrt{\frac{2\lambda}{n\pi(k-1)}}. \tag{5}$$

Because λ is a weighted sum of squares, it also provides a bound on the largest contrast among the θ_i. In particular, for any two values θ_i and θ_j

$$\lambda \geq (\theta_i - \theta_j)^2/2v$$

and therefore

$$|\theta_i - \theta_j| \leq \sqrt{2\lambda v} \tag{6}$$

for all $i, j = 1, \ldots, k$. When the observations are the differences between sample means of n observations with variance σ^2, then the largest possible standardized difference between mean differences becomes

$$\left|\frac{\theta_i}{\sigma} - \frac{\theta_j}{\sigma}\right| \leq \sqrt{\frac{2\lambda}{n}}. \tag{7}$$

Thus, we might define a negligible value of λ by starting with the largest average difference or largest possible difference between any two θ_i (or standardized θ_i's), and then solve either Equation 5 or Equation 7 to obtain a corresponding value of λ.

Conventional Heterogeneity Analysis in Meta-Analysis: Testing for Exact Replication With Burden of Proof on Nonreplication

Thus far, we have described how we might conceive of replication as the similarity of underlying effect parameters. The conventional heterogeneity test in meta-analysis uses the null hypothesis that effect size parameters are exactly the same and the alternate hypothesis that studies do not have identical effect size parameters. This perspective puts the burden of proof on the alternate hypothesis of nonreplication. This section details how to test null hypotheses structured to correspond to either exact or approximate replication when the studies are treated as fixed.

The k studies replicate exactly if $\theta_1 = \ldots = \theta_k$, so testing for replication corresponds to testing the null hypothesis

$$H_0: \theta_1 = \cdots = \theta_k \tag{8}$$

versus the alternative that at least one θ_i is different from the rest. Statistically, this is equivalent to the conventional test for heterogeneity of effect sizes based on the Q-statistic in meta-analysis (see, e.g., Hedges & Olkin, 1985, pp. 122–123). If the null hypothesis H_0 described in Equation 8 is true, that is, if the studies replicate exactly, the statistic Q given in Equation 1 has a chi-squared distribution with $k - 1$ degrees of freedom. Thus, the test consists of rejecting H_0 if the obtained value of Q exceeds c_α, the 100 × $(1 - \alpha)$ percentile of the chi-squared distribution with $k - 1$ degrees of freedom.

In this situation, the significance level α controls the probability of a Type I error (deciding that nonreplication has occurred when in fact the studies replicate exactly). The statistical power is the probability that the test would correctly decide that nonreplication had occurred when in fact the studies did not replicate exactly.

The interpretation of the statistical analysis should be influenced by the sensitivity of the statistical test. One way to characterize sensitivity is via the statistical power of the test. When the null hypothesis of exact replication is not true, Q has a noncentral chi-squared distribution with $k - 1$ degrees of freedom and noncentrality parameter λ given in Equation 2. Therefore, for any particular configuration of $\theta_1, \ldots, \theta_k$ the power of the α-level test is

$$P(\lambda) = 1 - F(c_\alpha | k-1, \lambda), \tag{9}$$

where $F(x \mid \nu, \lambda)$ is the cumulative distribution function of the noncentral chi-squared distribution with ν degrees of freedom and noncentrality parameter λ, and c_α is defined above.

Substantively meaningful power calculations using Equation 9 require that we input substantively meaningful values of λ. We illustrate power calculations using three conventions of negligible heterogeneity, which are discussed in the previous section. This provides insight into the test's sensitivity to even negligible differences in effect size parameters.

Table 1 gives the power of the test as a function of the number of studies k to detect heterogeneity relative to three definitions a negligible value for λ: $\lambda_0 = (k - 1)/4$, $\lambda_0 = (k - 1)/3$, and $\lambda_0 = 2(k - 1)/3$. These values of λ_0 correspond to conventions proposed in three different fields of science, as discussed in the previous section. The table is organized into four vertical panels, with the first three corresponding to different values of a negligible noncentrality parameter λ_0. The columns within each panel show the power when the true value of λ is equal to multiples of λ_0. The table shows that even the largest of the three conventional values $[\lambda_0 = 2v(k - 1)/3]$ would require between 20 and 30 studies to achieve statistical power approaching 80% to detect heterogeneity of $1.50\lambda_0$.

Conventional power analysis starts with a presumed degree of true heterogeneity and computes the probability that the test will detect that amount of heterogeneity by rejecting the null hypothesis. The power computed reflects the sensitivity of the test to detect the specified amount of heterogeneity. Another way to characterize sensitivity is to start with a desired statistical power (e.g., 80%) and determine the smallest amount of true heterogeneity required to achieve that power. The amount of heterogeneity required could be called the "minimum detectable heterogeneity"

(MDH; at a certain specified level of statistical power). The advantage of this characterization of sensitivity is that it does not require specification of a hypothetical level of heterogeneity but shows how much heterogeneity would be necessary for the test to have adequate sensitivity.

The MDH for significance level α and power value π can be computed by solving

$$\pi = 1 - \mathrm{F}(c_\alpha | k-1, \lambda), \tag{10}$$

for λ, where $\mathrm{F}(x \mid \nu, \lambda)$, ν, λ, and c_α are defined as in Equation 9, which requires an iterative procedure. The fourth vertical panel of Table 1 provides MDH values for $\alpha = .05$ and $\pi = 0.80$, expressed as the ratio τ^2/v. Displaying the MDH on this scale makes it easier to evaluate the sensitivity of the test with the λ_0 values that were offered to represent conventional values of heterogeneity (or any other λ_0 values). If the MDH is less than any particular λ value, then the test will have at least the power used to define the MDH (e.g., 80% in this table).

We see from the last column of Table 1 that the MDH is quite large unless the number of studies is large. Over 40 studies are needed to obtain an MDH of $\tau^2/v = 2/3$ (corresponding to our smallest convention of negligible heterogeneity), and over 200 studies are needed to obtain an MDH of $\tau^2/v = 1/4$, which represents the smallest convention of negligible heterogeneity.

Approximate Replication: Burden of Proof Is on Failure to Replicate

In the previous section we evaluated tests for replication for which the definition of replication was that the effect parameters were identical in all studies. In this section, we evaluate tests for when the definition of replication is that the effect size parameters $\theta_1, \ldots, \theta_k$ are "almost the same," that is, the studies replicate *approximately*. Of course, the concept of almost the same needs to be operationalized. One way to do that is in terms of the noncentrality parameter λ defined in Equation 2. In the previous section, we offered several ways to interpret λ in terms of the largest contrast among the θ_i, the typical deviation of θ_i from their mean or the variation of the θ_is compared with that of the estimation errors.

Testing for approximate replication requires choosing a value λ_0 of λ that corresponds to the operational definition of the θ_i being "almost the same" in the particular context. We then test whether there is greater heterogeneity in the estimates than would be expected had the actual amount been characterized by a noncentrality parameter of λ_0 or smaller. Specifically, to test the null hypothesis

$$H_0: \lambda \leq \lambda_0 \tag{11}$$

versus the alternative $\lambda > \lambda_0$ at significance level α, use the test statistic Q given in Equation 1 and reject the null hypothesis if Q exceeds the critical value c_α, where c_α is defined by

$$1 - F(c_\alpha | k-1, \lambda_0) = \alpha \tag{12}$$

or

$$c_\alpha = \mathrm{F}^{-1}(1 - \alpha | k-1,\ \lambda_0). \tag{13}$$

Note that c_α is a function of λ_0 and should properly be written $c_\alpha(\lambda_0)$.

In this situation, the statistical power is the probability that the test would correctly decide that nonreplication had occurred when in fact the studies did not replicate approximately (as operationally defined by λ_0). It can be computed using the noncentral chi-squared distribution. Thus, the power of this test to detect an amount of heterogeneity characterized by λ at significance level α is given by

$$P(\lambda) = 1 - \mathrm{F}(c_\alpha | k-1, \lambda). \tag{14}$$

Note that the difference between tests of the null hypotheses for exact replication (Equation 8) and approximate replication (Equation 11) is that the critical value for the former is based on the central distribution of Q (the distribution when $\lambda = 0$), whereas the critical value of the test for approximate replication is based on the

Table 1
Power of the Test of H_0: $\lambda = 0$

	$\lambda_0 = (k-1)/4$			$\lambda_0 = (k-1)/3$			$\lambda_0 = 2(k-1)/3$			
k	$\lambda = \lambda_0$	$\lambda = 1.5\lambda_0$	$\lambda = 2\lambda_0$	$\lambda = \lambda_0$	$\lambda = 1.5\lambda_0$	$\lambda = 2\lambda_0$	$\lambda = \lambda_0$	$\lambda = 1.5\lambda_0$	$\lambda = 2\lambda_0$	MDH (τ^2/v)
2	.08	.09	.11	.09	.11	.13	.13	.17	.21	7.85
3	.09	.11	.13	.10	.13	.16	.16	.23	.29	4.82
4	.10	.13	.15	.12	.15	.19	.19	.28	.36	3.63
5	.11	.14	.17	.13	.17	.22	.22	.32	.42	2.98
10	.14	.19	.25	.17	.25	.34	.34	.51	.66	1.74
20	.19	.29	.39	.25	.39	.53	.53	.76	.90	1.08
30	.24	.37	.50	.32	.50	.67	.67	.89	.97	.84
40	.28	.44	.60	.39	.60	.77	.77	.95	.99	.70
50	.32	.51	.68	.44	.68	.85	.85	.98	1.00	.61
100	.49	.75	.90	.67	.90	.98	.98	1.00	1.00	.41
200	.74	.94	.99	.90	.99	1.00	1.00	1.00	1.00	.28
300	.87	.99	1.00	.97	1.00	1.00	1.00	1.00	1.00	.22
500	.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.17

Note. This table displays the power of the test for exact replication to detect given values of λ as a function of the number of studies *k*. The table is divided into three vertical panels that correspond to conventions for negligible values of λ, and power is computed for multiples of these conventions. The final panel displays the Minimally Detectable Heterogeneity (MDH) that could be detected with 80% power for a given number of studies *k*. The MDH is displayed on the scale of τ^2/V; for reference, the conventions of negligible heterogeneity correspond to mdh values ranging from 1/4 to 2/3.

noncentral distribution of Q when $\lambda = \lambda_0$. Hence, the critical values for the test of approximate replication are larger than the corresponding critical values for exact replication. This means that it is more difficult to reject (it requires a larger value of Q to reject) the hypothesis of approximate replication than that of exact replication. Consequently, the test for approximate replication with $\lambda_0 > 0$ will be less powerful than the test for exact replication. For example, if $k = 10$, the $\alpha = .05$ level test for exact replication has critical value $c_{0.05} = 16.92$. The corresponding test for approximate replication with $\lambda_0 = (k - 1)/3$ has critical value $c_{0.05} = 22.17$. Therefore, in this instance, the power of test for exact replication to detect heterogeneity of $\lambda = 2(k - 1)/3$ (that is, $\lambda = 2\lambda_0$) is $P(2\lambda_0) = 0.34$, and the corresponding test for approximate replication with $\lambda_0 = (k - 1)/3$ has power $P(2\lambda_0) = 0.14$.

Table 2 gives the power of the test for approximate replication as a function of the number of studies k to detect heterogeneity beyond what might be considered negligible according to the three conventions mentioned earlier. The table is organized with three vertical panels that correspond to different values of λ_0. The first three columns within each panel give the statistical power when the true value of λ is equal to multiples of λ_0. For example, suppose the true heterogeneity in effect parameters was characterized by $\lambda = (k - 1)$. Using the largest of the three conventional values [$\lambda_0 = 2(k - 1)/3$, $\lambda = 1.5\lambda_0$], it would take nearly 300 studies to attain 80% power to conclude the studies do not approximately replicate. The statistical power increases and the required number of studies to obtain 80% power decreases as we decrease our choice of λ_0, so that if $\lambda_0 = (k - 1)/3$ (and $\lambda = 3\lambda_0$), we would need between 50 and 100 studies for the same power.

The MDH for significance level α and power value π can be computed by using an iterative procedure to solve

$$\pi = 1 - \mathrm{F}(c_\alpha | k-1,\ \lambda), \qquad (15)$$

for λ, where $\mathrm{F}(x \mid \nu, \lambda)$, ν, and λ, are defined as in Equation 14 and c_α is defined by Equation 13. The last column in each vertical panel of Table 2 provides the MDH for $\alpha = .05$ and $\pi = 0.80$ and the value of λ_0 for that panel, expressed as the ratio τ^2/v for easier comparison with the λ_0 values. We see from the last columns in each vertical panel of Table 2 that the MDH values are quite large unless the number of studies is also quite large.

Burden of Proof Is on Replication

In the previous sections, the methods for studying replication formulated the problem as testing the null hypothesis that the studies replicated (either exactly or approximately) with the alternative hypothesis being failure to replicate. The inherent problem with this formulation is that deciding that studies replicate involves accepting the null hypothesis, which is considered inconclusive in conventional hypothesis testing procedures. In this section, we describe tests that structure nonreplication as the null hypothesis and replication as the alternative hypothesis. Thus, by rejecting the null hypothesis of no replication, we may conclude that replication has occurred using a test with a known false-positive (false conclusion that replication has occurred) error rate.

Exact Replication

Exact replication among k studies implies that $\theta_1 = \ldots = \theta_k$, so testing for replication using a test based on the Q-statistic would involve testing the null hypothesis H_0: $\lambda > 0$ versus the alternative $\lambda = 0$. No such test is available, so we offer no test of exact replication when the alternative corresponds to the hyperplane in the parameter space defined by $\theta_1 = \ldots = \theta_k$.

Approximate Replication

The k studies replicate approximately if the effect size parameters $\theta_1, \ldots, \theta_k$ are "almost the same." In terms of the noncentrality parameter λ defined in Equation 2, we can test for approximate replication by choosing a value λ_0 of λ that corresponds to

Table 2
Power of the Test of H_0: $\lambda \leq \lambda_0$

	$\lambda_0 = (k - 1)/4$				$\lambda_0 = (k - 1)/3$				$\lambda_0 = 2(k - 1)/3$			
k	$\lambda = 1.5\lambda_0$	$\lambda = 2\lambda_0$	$\lambda = 3\lambda_0$	MDH (τ^2/v)	$\lambda = 1.5\lambda_0$	$\lambda = 2\lambda_0$	$\lambda = 3\lambda_0$	MDH (τ^2/v)	$\lambda = 1.5\lambda_0$	$\lambda = 2\lambda_0$	$\lambda = 3\lambda_0$	MDH (τ^2/v)
2	.06	.07	.10	9.14	.06	.08	.11	9.53	.07	.10	.15	10.94
3	.06	.08	.11	5.76	.07	.09	.13	6.04	.08	.11	.19	7.07
4	.07	.09	.13	4.43	.07	.10	.15	4.67	.09	.13	.24	5.55
5	.07	.09	.14	3.70	.07	.10	.17	3.91	.09	.15	.28	4.00
10	.08	.12	.20	2.28	.09	.14	.26	2.45	.12	.22	.46	3.08
20	.09	.15	.31	1.53	.11	.19	.41	1.67	.16	.34	.72	2.20
30	.11	.19	.41	1.24	.13	.24	.53	1.37	.20	.44	.86	1.86
40	.12	.22	.49	1.08	.14	.29	.63	1.20	.24	.54	.93	1.67
50	.13	.25	.56	.98	.16	.33	.72	1.10	.27	.62	.97	1.55
100	.18	.39	.81	.74	.22	.52	.93	.85	.43	.86	1.00	1.27
200	.26	.61	.97	.58	.35	.77	1.00	.68	.66	.99	1.00	1.08
300	.33	.75	1.00	.52	.45	.90	1.00	.62	.81	1.00	1.00	1.00
500	.47	.91	1.00	.45	.62	.98	1.00	.55	.94	1.00	1.00	.92

Note. This table displays the power of the test for approximate replication to detect given values of λ as a function of the number of studies k. The table is divided into three vertical panels that correspond to conventions for negligible values of λ_0, and power is computed for multiples of these conventions. The final column in each panel displays the Minimally Detectable Heterogeneity (MDH) that could be detected with 80% power for a null hypothesis H_0: $\lambda \leq \lambda_0$ and number of studies k. The MDH is displayed on the scale of τ^2/V; for reference, the conventions of negligible heterogeneity correspond to MDH values ranging from 1/4 to 2/3.

some operational definition of the θ_i being "almost the same." Testing for approximate replication, therefore, is testing whether there is less heterogeneity in the estimates than would be expected if the actual amount was characterized by a noncentrality parameter of λ_0 or larger.

Specifically, to test the null hypothesis

$$H_0 \colon \lambda \geq \lambda_0 \tag{16}$$

versus the alternative $\lambda < \lambda_0$ at significance level α, use the test statistic Q given in Equation 1 and reject the null hypothesis if Q is *smaller than* the critical value c_α, where c_α is defined by

$$\mathrm{F}(c_\alpha | k-1, \lambda_0) = \alpha$$

or

$$c_\alpha = \mathrm{F}^{-1}(\alpha | k-1, \lambda_0), \tag{17}$$

so that c_α is a function of λ_0 similar to that in Equation 13.

In this situation, the significance level α controls the probability of a Type I error (deciding that approximate replication has occurred when in fact the studies do not replicate approximately). The statistical power is the probability that the test would correctly decide that approximate replication had occurred when in fact the studies did approximately replicate.

Note that the null hypothesis for approximate replication when the burden of proof is on nonreplication (H_0: $\lambda \leq \lambda_0$) and when the burden of proof is on replication (H_0: $\lambda \geq \lambda_0$) are mutually exclusive. Thus, if the probability of rejecting the hypothesis $\lambda \leq \lambda_0$ is $\pi_1 = \mathrm{P}\{Q > c_\alpha\}$, and the probability of rejecting the hypothesis $\lambda \geq \lambda_0$ is $\pi_2 = \mathrm{P}\{Q < c_\alpha\}$, then $\pi_1 = 1 - \pi_2$.

The power of the test for approximate replication (as operationally defined by λ_0) can be computed using the noncentral chi-squared distribution. In the test described here, the power to detect an amount of heterogeneity characterized by λ at significance level α is given by

$$P(\lambda) = \mathrm{F}(c_\alpha | k-1, \lambda). \tag{18}$$

In contrast to tests that place the burden of proof on nonreplication, the power of this test is a *decreasing* function of λ and is maximum when $\lambda = 0$, that is, when there is no heterogeneity at all.

Table 3 gives the power of the test for approximate replication (with the burden of proof on replication) as a function of the number of studies k for three definitions of λ_0 corresponding to the conventions mentioned earlier. The table is organized into three vertical panels like Table 2, which correspond to different negligible values λ_0 of the noncentrality parameter. The first three columns within each panel correspond to power when the true value of λ corresponds to multiples of λ_0. In contrast to Tables 1 and 2, the values of λ considered in Table 3 are multiples of λ_0 by numbers less than one (1/2, 1/4, etc.). This is because, unlike when the burden of proof is on nonreplication, when the burden of proof is on replication, power increases as the noncentrality parameter λ decreases. The table shows that even the largest of the three conventional values [$\lambda_0 = 2(k - 1)/3$] would require between 40 and 50 studies to achieve statistical power approaching 80% when there is no heterogeneity at all (that is, when $\lambda = 0$). Unless λ_0 is large relative to λ, the power of this test is likely to be low.

In the previous sections, in which the burden of proof was on nonreplication, we defined the MDH to characterize the sensitivity of the test. When the burden of proof is on replication, power of the test is a decreasing function of λ (not an increasing function of λ, as when the burden of proof is on nonreplication). This distinction gives rise to a corresponding concept of maximum allowable heterogeneity (MAH) when the burden of proof is on replication. The MAH is the largest amount of true heterogeneity for which the test has some prespecified value of statistical power.

The MAH for significance level α and power value π can be computed by solving

$$\pi = \mathrm{F}(c_\alpha | k-1, \lambda), \tag{19}$$

for λ, where $F(x \mid \nu, \lambda)$ and λ are defined as in Equation 9 and Equation 2, and c_α is defined by Equation 17. Solving Equation 19 requires an iterative procedure, but it is straightforward to pro-

Table 3
Power of the Test of H_o: $\lambda \geq \lambda_o$

	$\lambda_0 = (k - 1)/4$				$\lambda_0 = (k - 1)/3$				$\lambda_0 = 2(k - 1)/3$			
k	$\lambda = \lambda_0/2$	$\lambda = \lambda_0/4$	$\lambda = 0$	MAH (τ^2/v)	$\lambda = \lambda_0/2$	$\lambda = \lambda_0/4$	$\lambda = 0$	MAH (τ^2/v)	$\lambda = \lambda_0/2$	$\lambda = \lambda_0/4$	$\lambda = 0$	MAH (τ^2/v)
2	.05	.06	.06	—	.05	.06	.06	—	.06	.06	.07	—
3	.06	.06	.06	—	.06	.06	.07	—	.07	.08	.09	—
4	.06	.07	.07	—	.06	.07	.08	—	.08	.10	.12	—
5	.06	.07	.08	—	.07	.08	.09	—	.09	.11	.15	—
10	.07	.09	.10	—	.08	.10	.13	—	.12	.18	.26	—
20	.09	.12	.15	—	.10	.15	.20	—	.17	.29	.46	—
30	.10	.14	.19	—	.12	.18	.27	—	.22	.39	.62	—
40	.11	.17	.24	—	.14	.22	.33	—	.26	.49	.74	—
50	.13	.19	.28	—	.16	.26	.40	—	.31	.57	.83	.02
100	.18	.30	.46	—	.24	.42	.64	—	.50	.83	.98	.19
200	.27	.49	.72	—	.38	.67	.90	.04	.76	.98	1.00	.32
300	.36	.63	.86	.02	.50	.82	.97	.09	.89	1.00	1.00	.38
500	.51	.83	.97	.07	.69	.96	1.00	.14	.98	1.00	1.00	.44

Note. This table displays the power of the test for approximate nonreplication to detect given values of λ as a function of the number of studies k. The table is divided into three vertical panels that correspond to conventions for negligible values of λ_0, and power is computed for fractions of these conventions. The final column in each panel displays the Maximally Allowable Heterogeneity (MAH) that could be detected with 80% power for a null hypothesis H_0: $\lambda \geq \lambda_0$ and number of studies k. The MAH is displayed on the scale of τ^2/V. When the mah is not reported, the test has a maximum power below 80%.

gram. Because the power of the test is a decreasing function of λ, the MAH shows how little heterogeneity there would have to be in order for the test to have a "high" probability (specifically the probability π) of detecting that the effects had replicated.

However, note that for some values of π, α, and λ_0, there may be no solution to Equation 19. This means that, even when studies replicate exactly so that $\lambda = 0$, the level α test may not have the desired power π. This is not just a theoretical possibility. The last column of each panel of Table 3 provides MAH values for $\alpha = .05$ and $\pi = 0.80$ on the scale of τ^2/v. We see that there are many cases in which no value is given in these columns, meaning that even if $\lambda = 0$ so that all studies involve exactly the same effect size parameter, the test placing the burden of proof on replication has less than 80% power. This includes all values of negligible heterogeneity λ_0 considered with less than 50 studies.

Greater sensitivity could be obtained by increasing the significance level of the test, but even with higher significance levels, the maximum power of tests to detect replication is not large for the conditions examined here. Table 4 shows the maximum power (that is, power when $\lambda = 0$) of tests to detect replication for $\alpha =$ .05, 0.10, 0.15, and 0.20 level tests. Increasing he significance level from 0.05 to 0.10 decreases the number of studies necessary from between 40 and 50 to between 30 and 40 to detect exact replication when $\lambda_0 = 2(k - 1)/3$. Even increasing the significance level to 0.20 only decreases the number of studies necessary for 80% power to between 20 and 30.

Examples

The Many Labs Replication Project provides several examples of experimental replications (Klein et al., 2014). The initial effort recruited 36 labs from around the world to conduct the same 13 experiments under similar conditions. Each lab was required, among other things, to recruit at least 80 participants, though most recruited many more. Thus, for each experiment, the Many Labs project amassed 36 effect size estimates and associated sampling variances. These data are useful for demonstrating not only how to conduct tests of replication but also how sensitive our ultimate conclusions about replication are to the considerations outlined in this article. Code for conducting these analyses are available as part of the online supplemental materials.

To illustrate the methods proposed in this article, we focus on a specific experiment, the Reverse Gambler's Fallacy. In the replications, participants were randomly assigned to one of two conditions and asked to imagine a man rolling dice at a casino. In one condition, they imagined seeing the man roll three 6s. In the other, they imagined him rolling two 6s and a 3. Participants were then asked how many times they thought the man had rolled the dice before they witnessed the result in their assigned condition. On average, participants who imagined seeing three 6s tended to estimate the man had rolled the dice more times than those who imagined seeing only two 6s.

The examples below use only the data from the Many Labs replications and not the original finding from Oppenheimer and Monin (2009). The reason for this is primarily related to publication bias. This is a common concern in many analyses of replications (e.g., Etz & Vandekerckhove, 2016, or van Aert & van Assen, 2017). Although the meta-analytic approach may be adapted to correct for publication bias, as proposed here, it does

Table 4
Maximum Power of the Test of H$_0$*:* $\lambda \geq \lambda_0$ *for Various Significance Levels* α

	$\alpha = .05$			$\alpha = .10$			$\alpha = .15$			$\alpha = .20$		
k	$\lambda_0 = (k-1)/4$	$\lambda_0 = (k-1)/3$	$\lambda_0 = 2(k-1)/3$	$\lambda_0 = (k-1)/4$	$\lambda_0 = (k-1)/3$	$\lambda_0 = 2(k-1)/3$	$\lambda_0 = (k-1)/4$	$\lambda_0 = (k-1)/3$	$\lambda_0 = 2(k-1)/3$	$\lambda_0 = (k-1)/4$	$\lambda_0 = (k-1)/3$	$\lambda_0 = 2(k-1)/3$
2	.06	.06	.07	.11	.12	.14	.17	.18	.21	.23	.24	.28
3	.06	.07	.09	.13	.14	.18	.19	.20	.27	.25	.27	.34
4	.07	.08	.12	.14	.15	.22	.20	.22	.31	.27	.29	.40
5	.08	.09	.14	.15	.17	.26	.22	.24	.36	.28	.31	.44
10	.10	.13	.26	.19	.23	.41	.27	.32	.52	.34	.39	.60
20	.15	.20	.46	.26	.33	.62	.35	.43	.72	.43	.51	.79
30	.19	.27	.62	.32	.41	.76	.41	.52	.84	.50	.60	.89
40	.24	.33	.74	.37	.48	.85	.47	.59	.91	.55	.67	.94
50	.28	.39	.83	.42	.55	.91	.52	.65	.95	.60	.72	.97
100	.46	.64	.98	.61	.77	.99	.70	.85	1.00	.77	.89	1.00
200	.72	.90	1.00	.83	.95	1.00	.89	.97	1.00	.92	.98	1.00
300	.86	.97	1.00	.93	.99	1.00	.96	1.00	1.00	.97	1.00	1.00
500	.97	1.00	1.00	.99	1.00	1.00	.99	1.00	1.00	1.00	1.00	1.00

Note. This table displays the maximum possible power for a test of approximate nonreplication for different levels α. This occurs in a test of H_0: $\lambda \geq \lambda_0$, where $\lambda_0 > 0$ and $\lambda = 0$. Note that as α increases, the maximum power also increases.

not. Therefore, we will exclude the original published study, which mitigates the effects of publication bias on the analyses presented in the following sections.

The data, published in the original article (Klein et al., 2014), is also available from the Open Science Framework, which provides the results of each replication, including treatment and control group sample sizes, means, and standard deviations. Using this, we calculate the bias-corrected standardized mean difference g between the three-6s and two-6s conditions, and associated sampling variance v_g. These are presented in Table 5. Note that the effect sizes and variances presented here differ slightly from those documented by the Many Labs project, as we use a different method to pool the variances of the two groups.

Conventional Meta-Analytic Test of Heterogeneity for Exact Replication

Using Equation 1, we calculate the test statistic $Q = 51.61$. Under the null hypothesis of exact replication, we compare the test statistic with a central chi-squared distribution with $k - 1 = 35$ degrees of freedom. For an $\alpha = .05$ level test, the critical value is $c_\alpha = 49.80$. Because $Q > 49.80$, we would reject the null hypothesis that the studies replicate exactly. The exact p value is $p = .03$.

Table 5
Effect Sizes and Variances of Replicated Studies on the Reverse Gambler's Fallacy

Site	Effect size	Variance
abington	.590	.051
brasilia	.355	.036
charles	.886	.063
conncoll	.622	.050
csun	.517	.046
help	.516	.043
ithaca	.782	.053
jmu	.715	.026
ku	.527	.039
laurier	.961	.042
lse	.645	.016
luc	.528	.029
mcdaniel	.510	.046
msvu	.340	.054
mturk	.620	.004
osu	.111	.038
oxy	1.188	.048
pi	.724	.004
psu	.605	.048
qccuny	.419	.044
qccuny2	.338	.050
sdsu	.616	.026
swps	.114	.050
swpson	.593	.027
tamu	.747	.024
tamuc	.749	.054
tamuon	.592	.020
tilburg	.687	.059
ufl	.378	.034
unipd	.765	.035
uva	1.108	.059
vcu	.712	.040
wisc	.785	.045
wku	.441	.044
wl	.072	.046
wpi	.978	.053

Note. This table displays the bias-corrected standardized mean differences and their variances of 36 replications of the Reverse Gambler's Fallacy Experiment from the Many Labs Replication Project. Effect sizes and variances were recomputed from the original data using standard meta-analysis methods. Source: Open Science Framework.

In a previous section ("Assessing Heterogeneity by Variation of Study Results") we proposed three conventions for what constitutes negligible heterogeneity: $\lambda = (k - 1)/4$, $(k - 1)/3$, and $2(k - 1)/3$. The power of this test to detect these levels of heterogeneity in this ensemble of studies is 0.26, 0.36, and 0.74, respectively. Note that the MDH at level $\alpha = .05$ and power of $\pi = 0.80$ is $\lambda = 0.75(k - 1)$. In other words, the minimally detectable heterogeneity between effect parameters is on the order of three quarters of the average within-study variance. For reference, standard indices of heterogeneity in meta-analysis, such as the I^2 and $\hat{\tau}^2$ statistics, indicate that the ratio τ^2/v for these data is likely between 1/3 and 2/3.

Approximate Replication, Burden of Proof Is on Failure to Replicate

The null hypothesis that the studies approximately replicate can be written in terms of the noncentrality parameter determining the distribution of Q as in Equation 2. In this case, we compare the test statistic, $Q = 51.61$, with the critical value given in Equation 9 obtained from the noncentral chi-squared distribution with $k - 1 = 35$ degrees of freedom, and noncentrality parameter λ_0. To carry out this test, we must select some value of λ_0 that reflects a tolerable amount of between-study variation. As an illustration, we will use the three conventions for tolerable levels of heterogeneity from various scientific disciplines presented above: $\lambda_0 = (k - 1)/4$, $(k - 1)/3$, and $2(k - 1)/3$.

First consider the most stringent margin of allowable heterogeneity: $\lambda_0 = (k - 1)/4 = 8.75$. For this choice of λ_0, the critical value for the $\alpha = 0.05$ level test based on the noncentral chi-squared distribution is $c_\alpha = 61.83$. Therefore, we fail to reject the null hypothesis that the studies approximately replicate ($p = .21$). Thus, allowing for this amount of tolerable between-study variation, the conclusion about replication is different than when exact replication is required. When the true $\lambda = 2\lambda_0 = (k - 1)/2$, the power of this test is 0.21. The MDH for this value of λ_0 at level $\alpha = .05$ and power of $\pi = 0.80$ is $\lambda = 1.14(k - 1)$.

Increasing the margin of allowable heterogeneity leads to an even larger p value. If $\lambda_0 = (k - 1)/3 = 11.67$, the critical value for the $\alpha = .05$ level test is $c_\alpha = 65.70$. Again, we fail to reject the null hypothesis that the studies approximately replicate ($p = .30$). The power of this test to detect heterogeneity characterized by $\lambda = 2\lambda_0 = 2(k - 1)/3$ is 0.27. The MDH for this value of λ_0 at level $\alpha = .05$ and power of $\pi = 0.80$ is $\lambda = 1.26(k - 1)$.

Finally, if $\lambda_0 = 2(k - 1)/3 = 23.33$, the critical value for the $\alpha = .05$ level test is $c_\alpha = 80.72$, and we fail to reject the null hypothesis that the studies approximately replicate ($p = .68$). Even if we assume $\lambda = 2\lambda_0 = (k - 1)$, the power to of this test is still only 0.50. The MDH for this value of λ_0 at level $\alpha = .05$ and power of $\pi = 0.80$ is $\lambda = 1.74(k - 1)$.

Two things are worth noting in this example. First, as we allow λ_0 to increase, our null hypothesis considers increasingly looser notions of "approximate" replication. Therefore, setting a large λ_0, as in the third example above, will require much greater variation between studies in order to reject the null hypothesis. This is also evident in the critical and *p* values, both of which increase with λ_0. Second, and related, tests of approximate replication exhibit decidedly less power than tests of exact replication, and neither are particularly well-powered for these data. Given the number of studies ($k = 36$), the minimally detectable heterogeneity for level $\alpha = .05$ and power $\pi = 0.8$ is $\lambda = 0.75(k - 1)$ for testing the null hypothesis for exact replication.

Burden of Proof Is on Replication

When the burden of proof is on replication, the null hypothesis becomes that studies do not replicate—the null hypothesis is given by Equation 16. We compare the test statistic, $Q = 51.61$, with a critical value obtained from the noncentral chi-squared distribution with $k – 1 = 35$ degrees of freedom and noncentrality parameter λ_0 as described in Equation 17. Using the same values of λ_0 that reflect a tolerable amount of between-study variation—$\lambda_0 = 8.75$, 11.67, and 23.33—the $\alpha = .05$ level critical values from Equation 17 are $c_\alpha = 28.32$, 30.37, and 38.87, respectively.

In contrast to the tests in the previous section we reject the null hypothesis if the test statistic is less than the critical value. Because the obtained value of Q (51.61) exceeds the critical values for all three choices of λ_0, we fail to reject the null hypothesis for any of them (the exact *p* values are 0.79, 0.70, and 0.32, respectively). Unlike the results in the previous section when we shift the burden of proof on to replication, we maintain the hypothesis that the studies do not even approximately replicate.

Finally, under this configuration, the power increases when either the true heterogeneity λ decreases or when the definition of negligible heterogeneity λ_0 increases. Thus, the maximum power of the tests presented here corresponds to the case where $\lambda_0 = 23.33$ and $\lambda = 0$. That is, when we use the loosest definition of negligible heterogeneity but the studies replicate exactly, the power of this test is 0.71. However, if we use stricter definitions of approximate replication, namely, $\lambda_0 = 8.75$ or 11.67, the power is below 0.31.

To gain further insight, consider the power to detect heterogeneity characterized by some fraction of λ_0. The power to detect $\lambda_0 = \lambda_0/4$ ranges from 0.45 when $\lambda_0 = 23.33$, to 0.16 when $\lambda_0 = 8.75$. This illustrates two key points. First, the power of this test can be sensitive to our choice λ_0. Second, it is only reasonably powered to detect *exact* replication given that we consider $\lambda_0 = 23.33$ to be a negligible amount of heterogeneity.

Conclusions About This Example

The computations in this example demonstrate that, regardless of whether ones chooses exact or approximate replication as the definition of replication and how one frames the hypothesis testing problem, the analysis of the effect chosen from the Many Labs Project for the example is underpowered. We also computed the power for each of the other replicated effects in the Many Labs Project and found them to be similarly underpowered. This is particularly important because these studies are serving as a prominent example (a de facto "gold standard") for evaluating replication. The results of this article suggest that efforts to study replication may need a larger number of studies (at least 50) to be adequately powered. Moreover, although we have suggested adequate sample sizes to obtain power of 80% (a somewhat conventional value), some might argue that the question of whether psychological studies can withstand attempts to replicate them is important enough to warrant a standard of even higher power (90% or even 95%). Achieving these higher levels of power would require even more studies.

Comparing a Single Study With a Series of Replications

This article defined replication in terms of heterogeneity of effects among a series of *k* studies in which no particular study is privileged as different from all the others. A different framing of the replication problem is one in which some study (e.g., the first study of its type) is privileged and the question is whether a series of $k \geq 1$ additional studies have effect sizes that are consistent with the privileged study. If that privileged study was not subject to publication bias (e.g., if it had a registered protocol or was the first of a designed ensemble of replications conducted cooperatively), then the methods in this article could be extended to deal with that framing of the replication problem.

The analysis would involve defining two groups of studies, the first group consisting of only the privileged study and the second group consisting of the *k* replications of the original study. The "analysis of variance" for effect sizes could be used to test whether the effect in the first group (the privileged study) was the same as the average effect in the second group (the replications; see Hedges & Olkin, 1985, Chapter 7). Because the test statistic Q_B for testing whether there are differences between mean effect sizes in the two groups has a sampling distribution similar to that of the Q-statistic discussed in this article (see Hedges & Pigott, 2004), the analyses discussed in this article could be naturally extended to this situation. In addition, if there were $k > 1$ replications, it would be possible to analyze the heterogeneity of those replications (alone) using the methods suggested in this article.

Recommendations

The purpose of this article is to describe alternative methods for the statistical analysis of replication so that a scientific consensus could be formed about them. Our example demonstrated that conclusions about replication can change depending on the methodological choices that are made. However, one of our reviewers pointed out that there is a danger in presenting too many alternatives. In the absence of a well-established consensus, the alternatives presented permit a researcher (knowingly or inadvertently) to make a choice of burden of proof, definition of replication (λ_0 value), and power criterion that will increase the chance of reaching whatever conclusion they desire.

Consequently, we reluctantly offer suggestions for conventions until a broader consensus can be achieved. First, we argue that the most reasonable structure for the test is to put the burden of proof on nonreplication. The reason for this choice is that, in many situations, the test that puts the burden of proof on replication is so insensitive (has so little statistical power) as to be ambiguous unless the number of studies is unrealistically large. Second, we argue that because exact replication is too strict a definition to be useful in strong sciences like physics, it is too strict a definition to be useful in psychology (or other social or behavioral sciences). This leaves the question of what value of λ_0 should be used.

We believe that the value of λ_0 defining approximate replication should be determined as a matter of consensus in each field. However, in absence of that consensus, we propose the convention $\lambda_0 = (k - 1)/4$ used in particle physics is a reasonable suggestion for defining approximate replication. It seems highly unlikely that any future consensus in psychology or the behavioral sciences would arrive at a value smaller than this one. For studies with a sample size of $n = 65$ per treatment group, Equation 5 implies that this choice corresponds roughly to an average absolute pairwise difference among standardized effects $|\theta_i - \theta_j|/\sigma$ of 0.1 or half the size of what Cohen calls a "small effect." Note that $n = 65$ is a little smaller than the average sample size of the Many Labs experiments, and it is approximately the sample size required to detect what Cohen calls a medium-sized effect ($d = 0.5$) with 80% power.

In the absence of a current convention in the social and behavioral sciences, we also propose that any ensemble of studies designed to test for replication should have an MDH of $\lambda = (k - 1)$. This means that it should have power of at least 80% to detect heterogeneity of a value of heterogeneity that corresponds roughly to an average absolute pairwise difference among standardized effects $|\theta_i - \theta_j|/\sigma$ of 0.2 or the size of what Cohen calls a "small effect." Such a test could, of course, detect smaller amounts of heterogeneity but would have somewhat lesser power to do so.

These conventions suggested are arbitrary but were not chosen capriciously. In many situations, it should be feasible to design ensembles of replications that can achieve the criterion of 80% power to detect heterogeneity of magnitude $\lambda = k - 1$ with $\lambda_0 = (k - 1)/4$. For example, although the example we used to illustrate the computations did not have high enough power to meet the requirement we specified here (the power to detect approximate heterogeneity was just under 70%), a slightly larger ensemble (e.g., 48 instead of 36 studies) would have had over 80% power.

Conclusions

The tests given here illustrate different statistical analyses that might be conducted to test for replication. All of them are valid statistical approaches that assess replication within the meta-analytic framework. Because they use different conceptual definitions of "replication" and place the burden of proof differently, these tests vary in their sensitivity. The example illustrates that the same data might reject replication (if exact replication is required), fail to confirm approximate replication (if the burden of proof is place on nonreplication), or fail to reject approximate nonreplication (if the burden of proof is on replication). This suggests that studies of replication cannot be unambiguous unless they are clear about how they frame their statistical analyses and clearly define the hypotheses they actually test. Researchers should also recognize that different frameworks for evaluating replication could lead to different conclusions from the same data.

The power computations offered in this article illustrate that it is likely to be difficult to obtain strong empirical tests for replication, a finding that is not unique to this approach (see, e.g., Maxwell, Lau, & Howard, 2015, p. 495). We have shown that large numbers of studies are likely necessary to obtain adequate statistical power to detect modest amounts of heterogeneity. Regardless of whether the conceptual framework places the burden of proof on replication or nonreplication, low-power tests cannot lead to strong conclusions. Conclusive analyses require carefully designed ensembles of replication studies and substantial investment of resources. More work needs to be done on the theory for designing programs of research on replication and developing feasible multistudy designs that may have greater sensitivity. In the interim, it seems important to scrutinize empirical tests of replication to determine if they are sensitive enough to warrant strong conclusions.

The low power we found in our example might be interpreted as a deficiency of hypothesis testing as an analytic strategy for evaluating replication. Indeed, we found low power even when using the theoretically most powerful unbiased test possible, which implies that no other unbiased test could have higher power. In such a situation, Bayesian methods can provide an alternative analysis, providing a posterior distribution that summarizes the information in the data (combined with a prior distribution). However, the low power of the uniformly most powerful unbiased test implies that the data have inadequate information to make sharp distinctions about heterogeneity. In this situation, Bayesian methods using uninformative priors are likely to provide an alternative summary of the situation in which the data cannot support sharp conclusions about heterogeneity.

This article has been concerned with testing replication across studies intended to be similar enough to obtain the same effects. If studies differ in important ways, then it would be possible to incorporate study-level covariates into a so-called metaregression model to evaluate the effects of these covariates of study effects (see, e.g., chapter 8 of Hedges & Olkin, 1985). One might then evaluate replication conditional on having the same covariate values. In this analysis, the test for replication would involve a statistic that is the weighted residual sum of squares (often called Q_E, e.g., in Hedges & Olkin, 1985). The framework for tests of residual heterogeneity is analogous to that for Q (in fact, Q_E reduces to Q when there are no covariates), so the ideas in this article generalize directly to that situation. Methods for conducting power analyses of the test based on Q_E were given in Hedges and Pigott (2004).

This article examined statistical analyses for replication within a framework that assumes study outcomes were measured via conventional effect size measures used in meta-analysis (e.g., the standardized mean difference). In some domains of research, outcomes may be measured on the same scale of measurement in all studies (e.g., earnings or systolic blood pressure) so that standardization of effect sizes is unnecessary. A similar conceptual framework for replication in studies of this type could be used in that context, but the statistical methods for analyses of heterogeneity are not exactly the same and the properties of those methods are somewhat different. Although it seems likely that the same general conclusions would hold for those kinds of analyses, research on the on the sensitivity of analyses based on those methods would be useful.

Finally, more work is needed to establish standards for levels of heterogeneity that might be considered scientifically negligible. We have argued that exact replication may be too strict a standard in psychology and the social sciences because it is considered too strict in strong sciences like particle physics and because there is evidence that it is not met in other areas of physical science (see Hedges, 1987). If this is true, it is important to develop standards for the amount of heterogeneity that are appropriate for judging replication in psychology and the social sciences. Although we have offered tentative standards for approximate replication, we believe that such standards should be a matter of social consensus among scientists. They may be resolved differently in different areas of science (as our three examples seem to indicate) and cannot be determined entirely by the technical framework of mathematical statistics. Mathematics can help facilitate judgments, for example, by relating heterogeneity parameters to other quantities (such as the largest difference among pairs of effects from different studies), but the function of the mathematics is only to characterize heterogeneity in ways that are easier to make judgments about, not to determine the correct judgments themselves.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533,* 452–454. http://dx.doi.org/10.1038/533452a

Birge, R. T. (1932). The calculation of error by the methods of least squares. *Physical Review, 40,* 207–227. http://dx.doi.org/10.1103/PhysRev.40.207

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.

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,* 1433–1436.

Camilli, G. (1990). The test of homogeneity for 2 × 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin, 108,* 135–145. http://dx.doi.org/10.1037/0033-2909.108.1.135

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129. http://dx.doi.org/10.2307/3001666

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature, 505,* 612–613. http://dx.doi.org/10.1038/505612a

Dickersin, K. (2006). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 9–33). Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470870168.ch2

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

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

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037–1038. http://dx.doi.org/10.1126/science.aad7243

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review, 23,* 74–86. http://dx.doi.org/10.3758/s13423-015-0868-6

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of *p*-values smaller than. 05 in psychology: What is going on? *PeerJ, 4,* e1935. http://dx.doi.org/10.7717/peerj.1935

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

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology, 13*(3), e1002106. http://dx.doi.org/10.1371/journal.pbio.1002106

Hedges, L. V. (1982). Estimating effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499. http://dx.doi.org/10.1037/0033-2909.92.2.490

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. http://dx.doi.org/10.3102/10769986009001061

Hedges, L. V. (1987). How hard is hard science, how soft is soft science?: The empirical cumulativeness of research. *American Psychologist, 42,* 443–455. http://dx.doi.org/10.1037/0003-066X.42.5.443

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203–217. http://dx.doi.org/10.1037/1082-989X.6.3.203

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445. http://dx.doi.org/10.1037/1082-989X.9.4.426

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3,* 486–504. http://dx.doi.org/10.1037/1082-989X.3.4.486

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 145–174). Chichester, UK: Wiley.

Higgins, J. P. T., & Green, S. (Eds.). (2008). *The Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/9780470712184

Humphreys, L. G. (1980). The statistics of failure to replicate: A comment on Buriels (1978) conclusions. *Journal of Educational Psychology, 72,* 71–75. http://dx.doi.org/10.1037/0022-0663.72.1.71

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294,* 218–228. http://dx.doi.org/10.1001/jama.294.2.218

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45,* 142–152. http://dx.doi.org/10.1027/1864-9335/a000178

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159. http://dx.doi.org/10.1037/h0026141

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

McNutt, M. (2014). Reproducibility. *Science, 343,* 229. http://dx.doi.org/10.1126/science.1250475

Olive, K. A., Agashe, K., Amsler, C., Antonelli, M., Arguin, J.-F., Asner, D. M., . . . Zyla, P. A. (2014). Review of particle properties. *Chinese Physics C, 38,* 090001. http://dx.doi.org/10.1088/1674-1137/38/9/090001

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

Oppenheimer, D. M., & Monin, B. (2009). Investigations in spontaneous discounting. *Memory & Cognition, 37,* 608–614. http://dx.doi.org/10.3758/MC.37.5.608

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

Perrin, S. (2014). Preclinical research: Make mouse studies work. *Nature, 507,* 423–425. http://dx.doi.org/10.1038/507423a

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

Rosenfeld, A. (1975). The particle data group: Growth and operations. *Annual Review of Nuclear Science, 25,* 555–598. http://dx.doi.org/10.1146/annurev.ns.25.120175.003011

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470870168

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. http://dx.doi.org/10.1037/a0015108

van Aert, R. C. M., & van Assen, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS ONE, 12*(4), e0175302. http://dx.doi.org/10.1371/journal.pone.0175302

Wellek, S. (2002). *Testing statistical hypotheses of equivalence*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/9781420035964

Received April 18, 2017
Revision received March 3, 2018
Accepted March 19, 2018 ■