Original Article

The Statistics of Replication

Larry V. Hedges

Department of Statistics, Northwestern University, Evanston, IL, USA

Abstract: The concept of replication is fundamental to the logic and rhetoric of science, including the argument that science is self-correcting. Yet there is very little literature on the methodology of replication. In this article, I argue that the definition of replication should not require underlying effects to be identical, but should permit some variation in true effects to be allowed. I note that different possible analyses could be used to determine whether studies replicate. Finally, I argue that a single replication study is almost never adequate to determine whether a result replicates. Thus, methodological work on the design of replication studies would be useful.

Keywords: replications, replication crisis, meta-analysis

The concept of replication is central to the logic and rhetoric of science. The principle that scientific studies can be replicated by other scientists is part of the logic that science is self-correcting, because attempted replications will identify findings that cannot be replicated and are thus incorrect (see e.g., McNutt, 2014). It is therefore surprising that empirical evidence has called the replicability of evidence in medical sciences (e.g., Ioannidis, 2005; Perrin, 2014; Prinz, Schlange, & Asadullah, 2011). Empirical evidence brings into question the replicability of research in psychology (e.g., Open Science Collaborative, 2016), economics (e.g., Camerer et al., 2016), the social sciences generally (Camerer et al., 2018), and machine learning (Hutson, 2018). Moreover, scientists in many disciplines seem to be concerned about replicability (e.g., Baker, 2016; Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015) including psychology (e.g., Pashler & Harris, 2012). Articles raising questions about the replicability of scientific findings have also begun to appear in the popular press, including *The Economist, Newsweek,* and *The New Yorker*.

Recent concerns about replicability have led to responses from the scientific research community intended to enhance replicability (see e.g., Collins & Tabak, 2014; McNutt, 2014). In the biomedical research community, the emergence of registration for clinical trials has been an important response (see e.g., International Committee of Medical Journal Editors, 2004), an approach that has also been advocated in the social sciences (see e.g., Hedges, 2018).

An important distinction is that between reproducibility and replicability. Reproducibility concerns whether another investigator can obtain the same results when given the first investigator's research report and their data (and possibly the computer code they used to analyze the data). Replicability concerns whether another investigator can obtain the same results when they obtain their own (new) data by attempting to repeat the study that was carried out by the first investigator. A key difference between reproducibility and replicability is that the former involves whether two investigators can obtain the same answers when given the same data, but replicability involves whether two investigators can obtain the same answers from two different datasets. Replicability is more demanding than reproducibility. Moreover, in many cases, replicability will be more ambiguous than reproducibility. If statistical analyses involve deterministic computations (e.g., computing a test statistic based on an algebraic formula), then reproducibility can rest on exact agreement of the results of computations. Results are either reproduced exactly or not. In the case of replicability, the results of analyses of two different datasets will each involve statistical uncertainty. For example, estimates of treatment effects will each have a standard error reflecting the statistical error of estimation. We cannot expect estimates to be identical even if the underlying treatment effect parameters are identical, so some statistical inference will be necessary, and with it the inherent ambiguity of such inferences.

Because the concept of replication is so central to the logic and rhetoric of science, we might expect an established body of work on the topic:

> "... one would expect there to be a large body of literature on replication providing clear-cut definitions on such matters as 'what exactly is a replication experiment?' or 'what exactly is a successful replication?' Furthermore, one would expect to find guidelines on how to conduct a replication or maybe some standard operating procedures on this issue. ... The opposite is true." (Schmidt, 2009, p. 90)

There is some literature on replication, but 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). While the concepts of direct and conceptual replications are clear, it is not always easy to distinguish between these two types of replication, as they are carried out in scientific practice. Sometimes, the intent of researchers is reasonably clear however, as it is in the programs of preregistered replications such as the many labs project (Klein et al., 2014), but other times it is not.

Part I: Background

Defining Replication in Terms of Study Conclusions

Some treatments of replication have defined replication in terms of the conclusions obtained by studies (e.g., did both studies conclude that the treatment effect positive, or not; Humphreys, 1980). This may appear to be a sensible definition of replication. It also is consistent with the ways in which scientists talk informally about replication. For example, one might say that investigator Smith found an effect (by which we mean that Smith obtained a statistically significant positive treatment effect) while investigator Jones failed to replicate (meaning that Jones did not obtain a statistically significant positive treatment effect). While this definition of replication may be in accord with common language usage, it is not useful as a scientific definition of replication for both conceptual and statistical reasons.

This definition says the conclusion about whether Study 1 and Study 2 replicate one another can be reached entirely on the basis of the significance tests in the two studies, which is inconsistent with current thinking about the use of statistics in science. The third principle in the American Statistical Association's Statement on Statistical Significance and *p*-values is that "Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold" (Wasserstein & Lazar, 2016, p. 132).

The ASA statement goes on to say that methods that emphasize estimation (e.g., effect sizes) that "more directly address the size of an effect (and its associated uncertainty)" are a good supplement or even replacement for *p*-values. A similar recommendation is given in the American Psychological Association's Task Force on Statistical Inference, who say that "Reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (Wilkinson & The Task Force on Statistical Inference, 1999, p. 599).

The same principle is embedded in the American Educational Research Association's standards for reporting on empirical social science, which says that "It is important to report the results of analyses that are critical for interpretation of findings in ways that capture the magnitude as well as the statistical significance of those results" (AERA, 2006, p. 37).

Putting aside issues of inference, that is, supposing that the conclusion drawn about the treatment effect in a study are always correct, a definition of replication that says Study 1 and Study 2 replicate one another if the treatment effect parameters are both positive (or both negative) would imply that a treatment effect of $\theta = 0.01$ and a treatment effect of $\theta = 1{,}000{,}000$ are considered to be "the same" result. A science that did not distinguish between effects that differ by eight orders of magnitude would seem to be very theoretically or empirically impoverished. It would have very limited practical applications because practical applications always involve at least implicit considerations of cost-effectiveness tradeoffs.

The inference properties of using individual study conclusions to draw conclusions about replication are problematic. Consider the situation in which there is a real effect of exactly the same magnitude $\theta > 0$ in each of two studies. Then if the power of the significance test in each study is η, the probability that the significance test in the two studies obtain the same result (both significant or both nonsignificant) is

$$p\{\text{Agreement}\} = \eta^2 + (1-\eta)^2.$$

Note that this function has a minimum of 0.50 at $\eta = 0.5$ and increases for both $\eta < 0.5$ and for $\eta > 0.5$. This means that when both studies have exactly the same effect parameter, agreement is high when the statistical power is low (since in that case both studies make the *incorrect* inference by failing to reject the null hypothesis), agreement *decreases* as power increases to $\eta = 0.5$, then increases again as increases above 0.5. However, for conventional levels of power, agreement is not exceptionally high. For example, when power of both studies is $\eta = 0.80$, a value often used a benchmark for adequate power (see Cohen, 1977), the probability of agreement is only 68% (.64 + .04).

Statistical Background

Because best scientific practice is that decisions about replication, like other decisions should be based on effect sizes, and because replication involves comparisons of study results that include estimation error, the assessment of replications necessarily involves statistical inference. Meta-analysis is the branch of statistics that addresses statistical inference from several studies simultaneously. Consequently, I apply ideas from meta-analysis to the analysis of replication.

Consider experimental studies that are analyzed by focusing on the effect of a treatment. Statistical inference in a single study is about the underlying treatment effect parameter. For example, it might estimate, create a confidence interval (CI) for, or test hypotheses about that treatment effect parameter. Because inference is about the treatment effect parameter, that parameter is conceptually is the result of the experiment: Although the exact value of the treatment effect parameter is not known, it is the object of inference. Of course, we do not observe the treatment effect parameter, only estimates of it, and that is why we need statistical inference about the results of the experiment.

In evaluating replication, it is crucial that the effect size estimate is estimating the same parameter in each study. Artifacts such as reliability, restriction of range, or measurement invalidity can influence some effect size measures (e.g., standardized mean differences or correlation coefficients). Similarly, choice of analytic strategy can also influence effect size measures (see e.g., McGaw & Glass, 1980). It is important that any evaluation of replication take any differential effects of these artifacts across studies into account.

Let $\theta_1, \ldots, \theta_k$ be the effect parameters, $T_1, \ldots, T_k$ be the effect estimates, and let $\nu_1, \ldots, \nu_k$ be the estimation error variances from k independent studies. Assume that the effect size estimates are approximately normally distributed with known variances so that $T_i \sim N(\theta_i, \nu_i)$.

When effects are identical (homogeneous across studies) $\theta_1 = \cdots = \theta_k$. The Q-statistic, which is used in testing for heterogeneity of effects across studies in meta-analysis, is defined by

$$Q = \sum_{i=1}^{k} (T_i - T)^2/\nu_i, \tag{1}$$

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

$$T_\bullet = \frac{\sum_{i=l}^{k}}{\sum_{i=l}^{k} 1/\nu_i},$$

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

$$H_0 : \theta_1 = \cdots = \theta_k$$

is true, Q has the chi-squared distribution with $k - 1$ degrees of freedom.

The assumption that the variance is known is often not exactly true, but it is often a useful modeling assumption in both the social and physical sciences. The impact of uncertainty in the variances on the distribution of the Q-statistic can be taken into account using higher order expansions to the distribution of Q, which approximate the distribution of Q as a chi-squared with reduced degrees of freedom (see Kulinskaya, Dollinger, & Bjorkestol, 2011). Nonparametric approaches to testing heterogeneity are also available (see Mahlzahn, Böhning, & Holling, 2000).

When studies are conceived as fixed, but when

$$H_0 : \theta_1 = \cdots = \theta_k,$$

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

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

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

$$\theta_\bullet = \frac{\sum_{i=1}^{k} \theta_i/\nu_i}{\sum_{i=1}^{k} 1/\nu_i} \tag{3}$$

(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 and that when $\lambda = 0$, the noncentral chi-squared distribution reduces to the usual (central) chi-squared distribution.

The noncentrality parameter λ is a natural way to characterize heterogeneity when studies are assumed to be fixed, but there are alternatives, particularly when the studies themselves are considered a random sample from a universe of studies – the so called random effects model for meta-analysis (see e.g., Hedges & Vevea, 1998). If studies are a random sample from a universe of studies, so that their effect parameters are also a sample from a universe of effect parameters with mean μ and variance τ^2, then τ^2 (the between-studies variance component of effects) is a natural way to characterize heterogeneity of effects. When $\tau^2 = 0$, it follows that $\lambda = 0$, but when $\tau^2 > 0$, these two characterizations seem rather different. However, when $\nu_1 = \cdots = \nu_k = \nu$ (as they are likely to

be, at least approximately, when studies are attempting to replicate one another) then

$$\lambda = \sum_{i=1}^{k} \frac{\left(\theta_i - \bar{\theta}\right)^2}{v} = (k-1)\sum_{i=1}^{k} \frac{\left(\theta_i - \bar{\theta}\right)^2}{(k-1)v} = (k-1)\hat{\tau}^2/v, \quad (4)$$

where the symbol $\hat{\tau}^2$ is used to emphasize that this quantity is an (unobservable) estimate of the variance τ^2 in the entire universe of studies from which the k observed studies are a sample. Thus, in a crude sense, $\lambda = (k-1)\tau^2/v$ when estimation error variances are similar. In general meta-analysis (not just analyses of studies intended to be direct replications), it would be unreasonable to assume that all the estimation error variances are equal, yet this assumption is the starting point for defining statistics such as I^2 which characterize the relative amount of heterogeneity.

Part II: How Should Replication Be Defined?

It is logical to think of defining replication across studies as corresponding to the case when all of the effect parameters are identical, that is, when $\theta_1 = \cdots = \theta_k$ or equivalently when $\lambda = 0$, or when $\tau^2 = 0$. This situation might be characterized as *exact replication*.

It is also possible to think of that if the θ_i are quite similar, but not identical then the results of the studies replicate "approximately." When the value of λ (or τ^2) is "small enough," that is, smaller than some negligible value λ_0 (or τ_0^2), we might conclude that the studies *approximately replicate*. Of course defining the magnitude of negligible differences in effects (λ_0 or τ_0^2) is an important consideration in assessments of replication.

Scientific Studies in Established Sciences Often Fail to Replicate Exactly

Because replication is a concern of essentially all sciences, it is possible to examine empirical evidence about replication in various sciences to provide a context for understanding replication in the social sciences. The example of physics is particularly illuminating because it is among the most respected sciences and because it has a long tradition of examining empirical evidence about replication (see e.g., Mohr, Newell, & Taylor, 2016; Rosenfeld, 1975). Interestingly, physicists developed some of the methods that are essentially the same as those developed independently for meta-analysis in the social sciences, including the Q-statistic (see Birge, 1932; Hedges, 1987).

Historical Values

The Speed of Light

Determining the values of the so-called fundamental constants of mathematical physics is a continuing interest in physics. Theory suggests that the speed of light in a vacuum is a universal constant and there has been a considerable amount of empirical work to determine its value. Figure 1 shows the values of the studies estimating the speed of light from 1870 to 1973. The year of the determination is on the horizontal axis and the value of the speed of light is on the vertical axis. Each determination is given by a dot surrounded by one standard error bars (68% CI) for the estimate. It is clear that, while many of the CIs overlap, the estimates differ by more than would be expected by chance due to their estimation error. In fact, the Q-statistic is $Q = 36.92$ with a p-value of less than .01. You might also observe that values appear to become consistent after about 1940, but this is an illusion of scale. The insert to Figure 1 shows values from 1945 to 1958 on a different scale, which shows that variation that continues to be large in comparison with the statistical uncertainty of the values.

You might notice a seeming time trend with the values of the speed of light up to about 1940. There were serious physicists suggesting that perhaps the speed of light was actually decreasing over time (DeBray, 1934a, 1934b), but newer values seemed to disconfirm that hypothesis (see Birge, 1941).

Five Other Fundamental Constants

One might imagine that the speed of light is an exception in physics, being very difficult to estimate, and that there are few other historical parallels. Figure 2 shows the values of estimates of five other fundamental constants of physics: the inverse of the fine structure constant, Plank's constant, the electron charge, the electron mass, and Avogadro's number. Although the uncertainty of all of the estimates is smaller among more recent estimates, the estimates of all of the constants exhibit variation that is substantially greater than would be consistent with their statistical uncertainty using tests based on the Q-statistic.

Contemporary Values

One might object that these historical comparisons are not fair because methods improve over time and that if we look at the most contemporary studies, the consistency would be much better. The next few examples involve relatively contemporary values (e.g., those used in determining contemporary values of the fundamental values). They show that contemporary values are also more variable than would be expected given their estimation errors.

The Mass of the Proton

Figure 3 shows the values of the mass of the proton obtained from four high accuracy experiments between

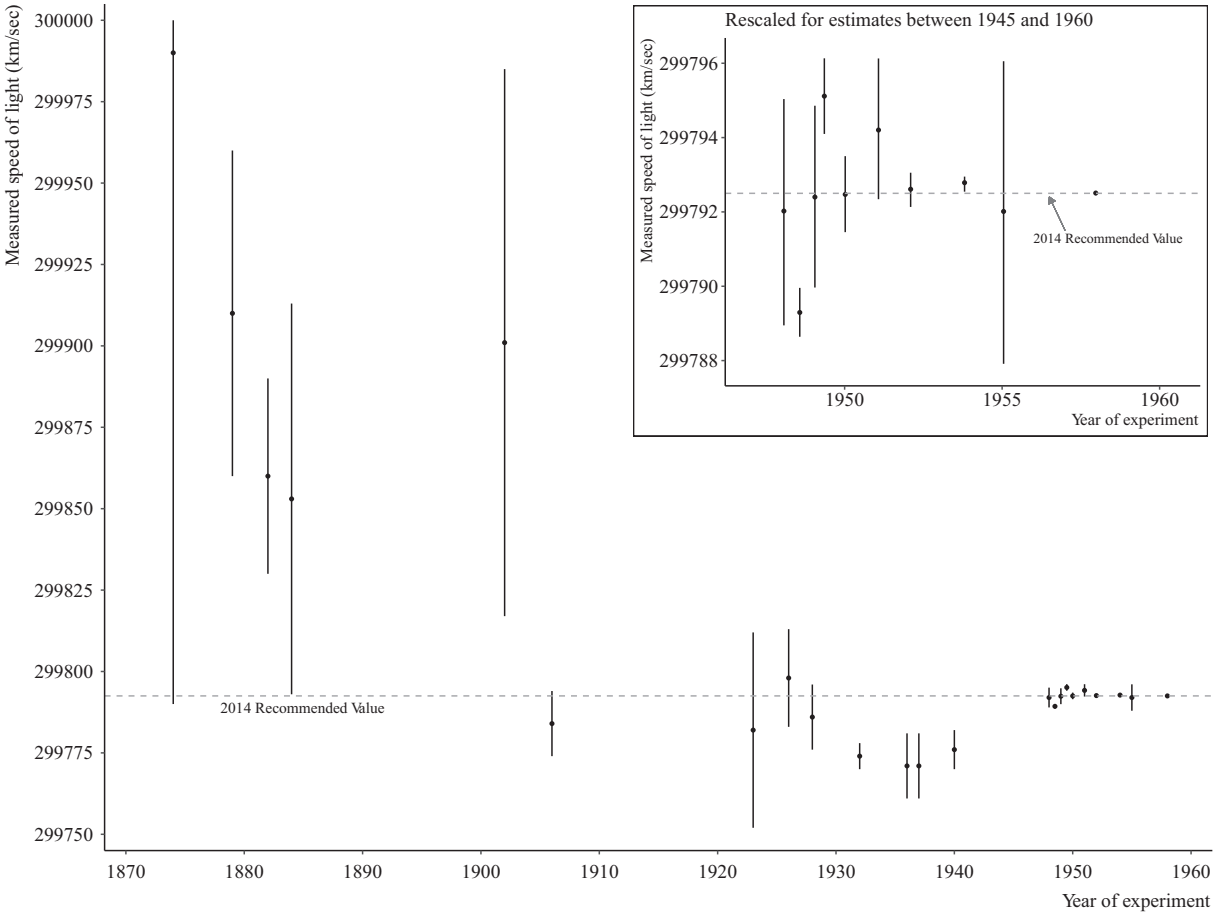

Figure 1. Historical values of the speed of light from 1875 to 1958 with 2014 horizontal line for the recommended value.

1960 and 1975. This example actually comes from the Particle Data Group and was used to illustrate the poor consistency of individual studies, even those intended to have high accuracy (Rosenfeld, 1975).

The Universal Gravitation Constant

Another important physical constant is the universal gravitation constant. The values that were used in a recent determination of this constant (Mohr et al., 2016) are shown in Figure 4, but not in chronological order and with the values of the estimates on the horizontal axis. These values exhibit much more variation than would be expected given their statistical uncertainties. In fact, $Q = 319$ with 13 degrees of freedom so that $p < .001$.

Plank's Constant

The values that were used in the most recent determination of Plank's constant (Mohr et al., 2016) are shown in Figure 5. They also exhibit inconsistency than would be expected on the basis of their estimation errors.

The Particle Data Group

The Particle Data Group is an international collaboration that carries out a program of meta-analyses (they just call them reviews) of all the high-energy physics experiments worldwide to estimate values of constants important to particle physics since 1957. Although they use several procedures to evaluate and harmonize data (including omitting over a third of the estimates), they routinely find estimates that vary by more than would be expected due to estimation errors (see e.g., Olive et al., 2014; Rosenfeld, 1975).

Other Physical Sciences

Although the examples above are all from physics, there are many examples from other physical sciences such as physical chemistry (Zwolinski & Chao, 1972), materials science (Touloukian, 1972), and thermodynamics (Ho, Powell, & Liley, 1972). For other examples see Hedges (1987) or Draper et al. (1993).

How Have Physicist Interpreted These Data?

As the historical data for individual constants suggests, scientists have understood the need to improve estimates and have generally sought to improve accuracy of their estimates. However, the general understanding is also that

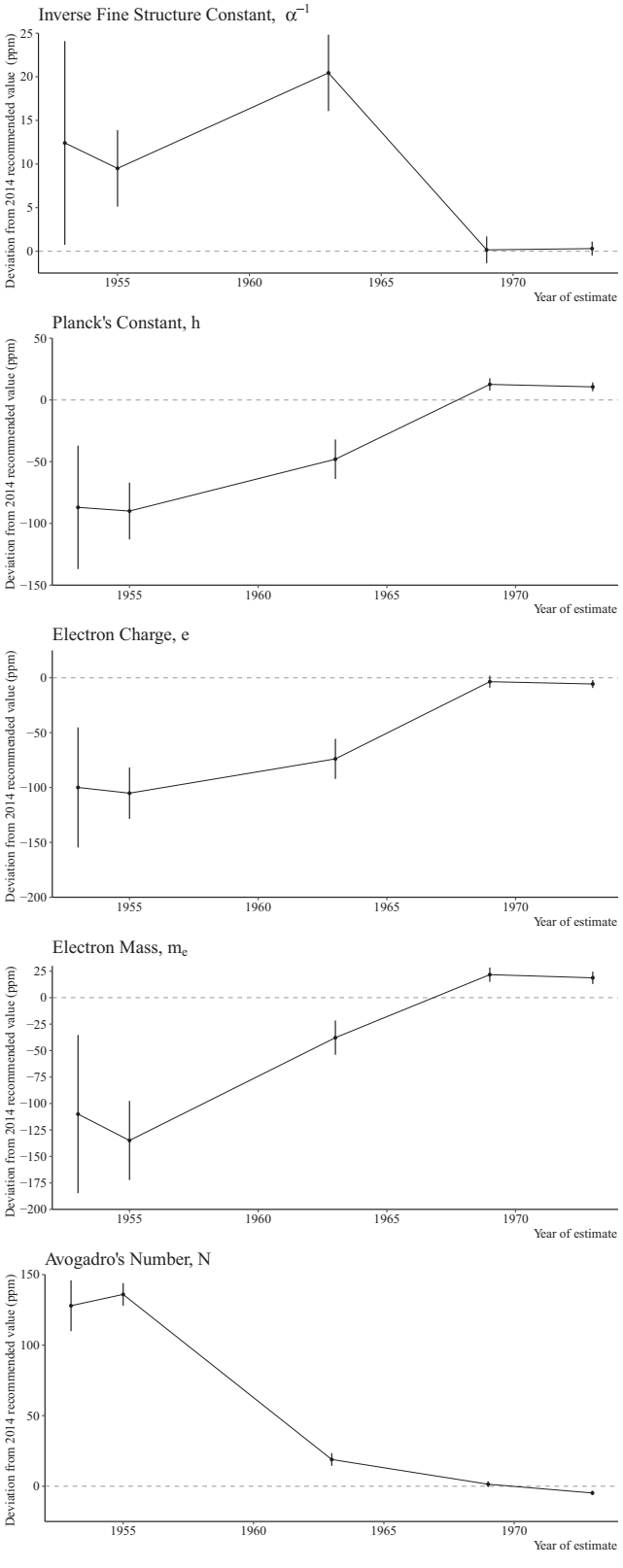

Figure 2. Historical values of the inverse of the fine structure constant, Plank's constant, the electron change, the electron mass, and Avogadro's number.

experimental determinations are difficult and that scientists probably underestimate the true uncertainty of their estimates. For example, Flowers and Pentley (2001), in commenting about the changes between recommended values of physical constants between 1973 and 1998 say that a "better (safer) measure of standard uncertainties would have been obtained by approximately doubling the estimated uncertainties" (p. 1240).

This has led to an understanding that reasonable scientific practice should be to tolerate a small amount of heterogeneity in estimates as negligible for scientific purposes. The Particle Data Group is quite explicit about this. They say that heterogeneity corresponding to $Q/(k-1) \leq 1.25$ is negligible heterogeneity regardless of the statistical significance (see Olive et al., 2014). Moreover, this criterion is applied *after* over a third of the studies are omitted from consideration (partially on the basis of the inconsistency of their findings with others). Note that the expected value of Q under the studies fixed model is

$$E\{Q\} = (k-1) + \lambda$$

and the expected value of Q under the studies random model is

$$E\{Q\} = (k-1)(1 + \tau^2/v).$$

Therefore, this convention for negligible heterogeneity corresponds to defining a negligible value of heterogeneity to be $\lambda_0 = (k-1)/4$ or $\tau_0{}^2/v = (k-1)/4$.

The question of what heterogeneity means scientifically is somewhat speculative. There is generally little question among physicists that the fundamental constants are indeed constant (except for brief periods like that in the 1930s when a genuine anomaly with respect to the speed of light seemed to be emerging). Apparent heterogeneity is usually attributed to underestimation of estimation error variances and/or the presence of uncontrolled biases in experiments, which would lead to apparent heterogeneity (see Rosenfeld, 1975). In a statistical sense, uncontrolled biases would be real heterogeneity inducing differences among the θ_i even if they were not a consequence of real differences in the physical world. For example, because the experiments that estimate the universal gravitation constant were conducted in different geographical locations, it is necessary to perform complex adjustments to the estimates to adjust for confounding effects of geography – adjustments that give rise to biases if they are not be quite right. In social and psychological experiments, such biases are clearly a major threat.

The principle used to arrive at an acceptable level of heterogeneity in physics is that competent experimenters attempt to reduce the biases in their experiments to a point

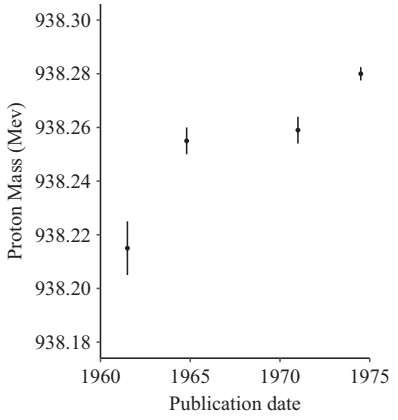

Figure 3. Recent values of estimates of the mass of the proton from Rosenfeld (1975).

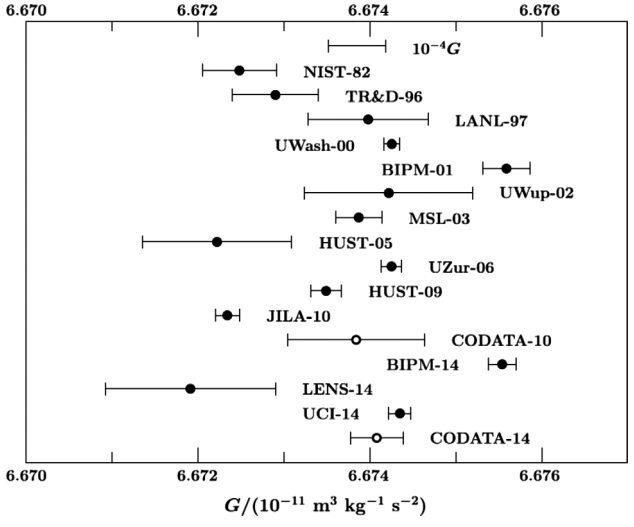

Figure 4. Recent estimates of the universal gravitation constant from Mohr et al. (2016).

that they are small compared to the estimation errors. Thus the variation due to biases (τ^2) is small in relation to variance due to estimation errors (ν). This principle suggests a criterion that is a multiple of τ^2/ν, but the judgment of which particular multiple is a matter of scientific consensus.

Other Sciences

There has been a considerable amount of experience with meta-analysis in medicine. In medicine, a value of $I^2 = 100\% \times \tau^2/(\nu + \tau^2)$ of 40% or less is considered to be "not important" (see Section 9.5.2 of Higgins & Green, 2008). This implies that a negligible amount of heterogeneity would be $\lambda_0 = 2(k-1)/3$ or $\tau_0{}^2/\nu = 2/3$.

In personnel psychology, Hunter and Schmidt (1990) proposed a "75% rule," which says that when the estimation error variance ν is at least 75% as large as the total

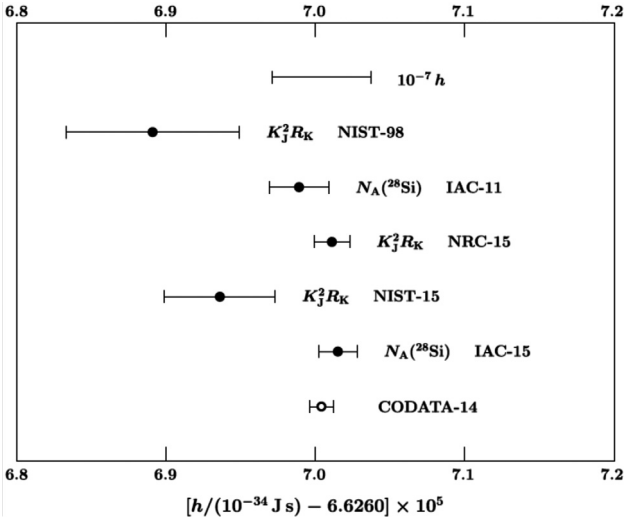

Figure 5. Recent estimates of the Planck's constant from Mohr et al. (2016).

variance of the effect size estimates ($\nu + \tau^2$), then the variance of the effect size parameters τ^2 could be considered negligible. This implies that $\tau_0{}^2/\nu = 1/3$ and $\lambda_0 = (k-1)/3$ correspond to negligible amounts of heterogeneity in effect parameters.

These three conventions in physics, medicine, and personnel psychology provide a range of definitions of negligible heterogeneity from $\lambda_0 = (k-1)/4$ to $\lambda_0 = 2(k-1)/3$ or alternatively, $\tau_0{}^2/\nu = 1/4$ to $\tau_0{}^2/\nu = 2/3$. Note that all of these definitions of negligible heterogeneity are social conventions among a group of scientists as all conventions for interpretation must be.

The conventions described above have the advantage that they are scale free. However, one weakness of conventions defined in terms of λ or τ^2/ν is that they are inversely proportional to ν, which is itself inversely proportional to the within-study sample size. This means that large values of these parameters can be obtained either if study sample size is large (so that the estimation error variance is small) or if the amount of effect size heterogeneity is large. For approaches to defining similar parameters (analogues to I^2) that are less sensitive to the absolute size of ν see Holling, Böhning, Masoudi, Böhning, and Sangnawakij (2019).

Conclusion About Defining Replication

The definition of replication is more complex that it first appears. While exact replication is logically appealing, it is too strict to be useful, even in well-established sciences like physics, chemistry, or medicine. Approximate replication has proven more scientifically useful in these sciences and in personnel psychology. However, it requires establishment of conventions of negligible heterogeneity among

groups of scientists. The fact that conventions have been established in these sciences shows that it is possible to do so. The fact that conventions of approximate replication have been established in terms of quantities like λ_0 or τ_0^2/ν suggests that there is value in doing so in ways that are also comparable to the established values in these sciences.

Part III: Statistical Analysis of Replication

The statistical test for heterogeneity typically used in meta-analysis using the *Q*-statistic given by (1) provides the basis for tests of replication in terms of effect size parameters. However, the details of the statistical approaches to studying replication depend on three considerations that are largely conceptual: How the hypotheses are structured (whether the burden of proof lies with replication or with failure to replicate), how replication is defined (as exact or approximate replication), and whether the studies are conceived as a fixed set or a random sample from a universe of studies (Hedges & Schauer, 2019).

Studies Fixed or Random

The Q statistic has the same sampling distribution when there is exact replication regardless of whether studies are fixed or random, but it has a different distribution when exact replication does not hold. Therefore, evaluation of the sensitivity (e.g., statistical power) of tests based on Q is somewhat different when studies are considered fixed than when they are considered random (Hedges & Schauer, 2019). Moreover, critical values for tests of approximate replication will be somewhat different when studies are considered fixed than when they are considered random.

If the studies are fixed, then inferences about replication are inferences about the effect parameters in the studies actually observed (see e.g., Hedges & Vevea, 1998). One might say that conclusions about replication in the fixed studies framework are conclusions about how well the *observed* studies agree.

If the studies are considered random, then the studies observed are a sample from a hypothetical universe of studies and 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 of effect parameters were taken. Thus, the observed studies and their effect parameters are of interest only in that they provide information about this hypothetical universe of study effects (see e.g., Hedges & Vevea, 1998). One might say that conclusions about replication in the random studies framework are conclusions about how well the studies agree *in a universe* of studies from which the observed studies are a random sample.

Definition of Replication

Replication among the results of k studies might be defined as exact (i.e., $\lambda = 0$ or $\tau^2 = 0$) or approximate (i.e., $\lambda < \lambda_0$ or $\tau^2 < \tau_0^2$ for some convention λ_0 or τ_0^2). Regardless of whether replication is defined as exact or approximate, the hypothesis test would have a similar form.

For exact replication regardless of whether studies are fixed or random, we reject H_0: $\lambda = 0$ (or $\tau^2 = 0$) at level α if Q exceeds the $100(1 - \alpha)$ percent point of the (central) chi-squared distribution with $k - 1$ degrees of freedom.

For approximate replication with studies fixed, we reject H_0: $\lambda < \lambda_0$ at level α if Q exceeds the $100(1 - \alpha)$ percent point of the noncentral chi-squared distribution with noncentrality parameter λ_0 and $k - 1$ degrees of freedom. For approximate replication with studies random, we reject H_0: $\tau^2 < \tau_0^2$ at level α if Q exceeds the $100(1 - \alpha)$ percent point of the distribution of Q when $\tau^2 = \tau_0^2$. If all of the studies have the same estimation error variance ν (e.g., they have the same sample size) then the noncentral distribution of Q is $(1 + \tau_0^2/\nu)$ times a (central) chi-squared distribution with $k - 1$ degrees of freedom. If the estimation error variances are not all equal then the noncentral distribution of Q has a more complex form (a mixture of central chi-squares) but it can be well approximated (see Hedges & Pigott, 2001).

Note that the distribution of Q when $\lambda > 0$ (or $\tau^2 > 0$) is stochastically larger (the distribution is shifted to the right). Thus tests of approximate replication have the same test statistic Q, but larger critical values than tests of exact replication and, therefore, have lower statistical power. For example, when $k = 2$, the test for exact replication uses a critical value of 3.94, but the test for approximate replication using $\lambda_0 = 1/4$ has a critical value of 4.76, that using $\lambda_0 = 1/3$ has a critical value of 5.03, and using $\lambda_0 = 2/3$ has a critical value of 6.06.

Burden of Proof

Note that whether studies are fixed or random and whether replication is exact or approximate, the form of the hypothesis test is the same. Rejection of the null hypothesis of exact or approximate replication is conclusive, but failure to reject is not. Thus the burden of proof is on failure to replicate. In such a situation, the results of the analysis can only be conclusive if the test has high power to detect meaningful amounts of heterogeneity. The inherent problem with this formulation is that concluding that

studies replicate involves accepting the null hypothesis, yet this is exactly the opposite of conventional hypothesis testing procedures.

A different way to structure the test is to alter the burden of proof so that it lies on replication (not failure to replicate). This is possible for tests of approximate replication. The test placing the burden of proof on replication uses the Q statistic but unlike the test placing the burden of proof on failure to replicate (which rejects for large values of Q), this test rejects for small values of Q. That is we structure the null hypotheses as:

$$H_0 : \lambda \geq \lambda_0 \left(\text{or } \tau^2 \geq \tau_0^2\right),$$

so that rejection of the null hypothesis leads to the (conclusive) decision that studies replicate.

If studies are fixed, we reject H_0: $\lambda \geq \lambda_0$ at level α if Q is less than the $100(1 - \alpha)$ percent point of the noncentral chi-squared distribution with noncentrality parameter λ_0 and $k - 1$ degrees of freedom. If studies are random we reject H_0: $\tau^2 \geq {\tau_0}^2$ at level α if Q exceeds the $100(1 - \alpha)$ percent point of the distribution of Q when $\tau^2 = {\tau_0}^2$.

Conclusion About Statistical Analyses for Replication

There are several possible statistical analyses of replication for any definition of replication. Each of these analyses is a valid way to explore a slightly different hypothesis about replication. The major conclusion about testing hypotheses about replication is that different tests are possible and the choice among them is not automatic, but a principled analytic decision that requires some care.

Part IV: Design of Replication Studies

The design of an ensemble of two or more studies to investigate replication might seem straightforward, but quite different designs have been used with little justification of why that design was appropriate. For example, the Open Science Collaborative (2016) and Camerer et al. (2018) chose to use a total of $k = 2$ studies (the original and one replication), while the Many Labs Project (Klein et al., 2014) used as many as $k = 36$ studies (the original and 35 replications) of each result. One might ask which, if either design is adequate and why.

While it may not be the only requirement of a sound design, a fundamental requirement of any research design is that it should lead to a statistical analysis that is sufficiently sensitive to detect the smallest effect deemed scientifically (or practically) important. If the analysis is a hypothesis test, then sensitivity could be measured by statistical power. Thus, a fundamental requirement of any design for studying replication using a hypothesis testing approach is that it leads to analyses that have sufficient power to detect the smallest amount of variation in results that is scientifically meaningful.

Evaluating Replication via a Single Replication Study

The simplest conception of the design to test whether Study 1 can be replicated is to simply repeat the study, so that the ensemble is two studies (Study 1 and Study 2). This is essentially the design used in Open Science Collaborative (2016) and Camerer et al. (2018). When $k = 2$, the Q-statistic becomes

$$Q = (T_1 - T_2)^2/(v_1 + v_2).$$

If we consider studies to be fixed, then the power of the test based on Q is determined by the noncentrality parameter

$$\lambda = (\theta_1 - \theta_2)^2/(v_1 + v_2).$$

Consider the power of the test for exact replication. To evaluate the statistical power we must identify smallest non-negligible value of λ for the replication test, which, will depend on $(\theta_1 - \theta_2)^2$. Call this smallest non-negligible value λ_R. Power is an increasing function of λ_R, so the bigger the λ_R, the higher the power. Thus the largest value that λ_R could reasonably take will correspond to the maximum power.

It seems reasonable that if the effects in the two studies had different signs, they could not be considered to replicate one another because the results would be qualitatively inconsistent. To guarantee that θ_2 has the same sign as θ_1, it must be true that $(\theta_1 - \theta_2)^2 < {\theta_1}^2$, since any larger value of $(\theta_1 - \theta_2)^2$ would be consistent with non-positive values of θ_2. Therefore,

$$\lambda_R < \theta_2^2/(v_1 + v_2).$$

However, a test of the null hypothesis that $\theta_1 = 0$ in Study 1 uses the test statistic ${T_1}^2/v_1$, which has the central chi-square distribution if $\theta_1 = 0$ and the noncentral chi-squared distribution if $\theta_1 \neq 0$. The power of the test is determined by the (noncentrality) parameter $\lambda_1 = {\theta_1}^2/v_1$. Comparing λ_1 to λ_R, we see that

$$\lambda_R < \theta_1^2/(v_1 + v_2) < \theta_1^2/v_1 = \lambda_1.$$

Thus the *maximum possible* power of the test for replication is smaller than the power of the original study's (Study 1's)

test of the null hypothesis of no effect. Moreover, the maximum possible power may be much lower in theory, and almost certainly will be lower in practice.

The theoretical limit requires perfect precision in the replication study (i.e., $\nu_2 = 0$ or an infinite sample size). If Study 2 has the same sample size as Study 1, so that, for example, $\nu_1 = \nu_2$, then the largest possible noncentrality parameter of the replication study becomes $\theta_1^2/2\nu_1 = \lambda_1/2$. In this case, if Study 1 had power of 80% (which occurs if $\lambda_1 = 7.85$), the power of the test for exact replication would be 51%. Even if Study 1 had power of 90%, that of the replication test would be only 63%. Study 1 would have to have a power of 98% for the replication test to have 80% power if $\nu_1 = \nu_2$. Such high power is unusual in most medical or social science contexts.

Note that the same logic applies regardless of which study is labeled as Study 1. The statistical power of a test for replication based on a total of $k = 2$ studies is limited by the study with the least statistical power. This means that it will be virtually impossible to achieve a high power test of replication unless both studies have very high power. Moreover, this analysis was based on a test for exact replication. Test for approximate replication have lower power than the corresponding test for exact replication, so they would have even lower power in this situation than a test for exact replication.

Evaluating Replication Using More Than Two Replication Studies

Statistical power of an analysis using the Q-statistic can be increased by using more than two studies. Methods for assessing the power of the test based on Q are available (see Hedges & Pigott, 2001). Adequate sensitivity can usually be achieved with enough replication studies. Extensive computations of statistical power of tests based on Q are given in Hedges and Schauer (2019). But important questions about design remain to be resolved. For example, in designing an ensemble of replication studies, how should one compromise between a greater number of studies and a larger sample size within each study? Rules of thumb and research on optimal allocations would be useful to aid in rational planning of replication studies.

Sometimes the object of the replication study is to determine whether an original study (already conducted) can be replicated. In that case, one design that might seem appealing is to use several replication studies (i.e., more than one replication of the original), and then to compare the results of replication studies (as a group) to the original study. Such analyses are often called subgroup analyses or fitting categorical models to effect sizes in meta-analysis (see Chapter 7 of Hedges & Olkin, 1985).

However, such a strategy is mathematically equivalent to combining the estimates from all of the replication studies into one "synthetic study" and computing an effect size estimate (and its variance) for that synthetic replication study. The analysis of the difference between the original study and the synthetic replication study is subject to exactly the same limitations of analyses comparing two studies that are described in this article. In other words, the sensitivity of that analysis is limited by the least sensitive of the two studies being compared (which will usually be the original study). Thus, no matter how many replication studies are conducted, it may be impossible to obtain a design of this type with adequate sensitivity.

This design is made even more problematic if the original study is from the published literature, was unregistered, and therefore was possibly subject to publication bias (see e.g., Dickersin, 2005). If the original study was subject to publication bias, the estimate of its effect would be biased upward in absolute magnitude (see Hedges, 1984), and the use of the uncorrected estimate would bias the test for replication by introducing heterogeneity as an artifact of publication bias. If the effect size estimate from the original study was corrected for publication bias (e.g., using maximum likelihood estimation under a selection model) then the estimate might be (approximately) unbiased, but variance of the corrected estimate would be even larger than that of uncorrected estimate (see Hedges, 1984), further reducing the statistical power of tests for replication.

Conclusion About Design of Replication Studies

Despite its obvious appeal, an ensemble of $k = 2$ studies (e.g., and original and a replication attempt) will almost never be an adequately sensitive design for an investigation of replication. More sensitive designs are possible, but the sensitivity of any design needs to be evaluated in conjunction with the choice of the definition of replication and the statistical analysis used to analyze the data. There has been rather little work on these kinds of design problems, but they are conceptually similar to the problem of designing studies of heterogeneity of treatment effects using multilevel models. Research on design and optimal design in those related models would be useful.

Part V: Overall Conclusions

Scientists have often approached the concept of replication as if its definition, appropriate analysis, and design of investigation of replication were straightforward. I have argued that none of these is the case and that methodological work is needed on all three issues.

Exact replication is logically appealing, but appears to be too strict a definition to be satisfied even in the most mature sciences like physics or medicine. Approximate replication is a more useful concept, but requires the development of social conventions in each area of science. Moreover, tests of approximate replication are less powerful than those of exact replication, leading to lower sensitivity in analyses of approximate replication.

For any particular definition of (exact or approximate) replication, several different, but perfectly valid, analyses are possible. They differ depending on whether a studies-fixed or studies-random framework is used and whether the burden of proof is imposed on failure to replicate (so that rejection of the null hypothesis leads to rejection of replication) or on replication (so that rejection of the null hypothesis leads to rejection of failure to replicate).

Finally there have been unappreciated problems in the design of a replication investigation (an ensemble of studies to study replication). The sensitivity of an ensemble of two studies is limited by the least sensitive of the studies, so that an ensemble of two studies will almost never be adequate to evaluate replication. Greater sensitivity can be obtained with more studies, but even then only with the appropriate analysis.

One might fault this article for arguing that interpretation of whether studies replicate should be based on effect sizes and then focusing on hypothesis testing and statistical power to evaluate replication designs. An alternative approach would be to focus on estimating a heterogeneity parameter (such as λ or τ^2), in which case sensitivity would be evaluated by the precision (e.g., variance or standard error) of the estimate. The standard error of estimates of λ and τ^2 from two studies will be larger than unity, large in comparison of values of interest (e.g., the conventional parameter values indicating negligible heterogeneity in physics, personnel psychology, or medicine are less than unity for $k = 2$). Thus, the conclusion that an ensemble of $k = 2$ studies is virtually always inadequately sensitive to evaluate replication would also hold if estimation was the focus of the analysis (see Hedges & Schauer, 2019). However, the details of design planning would be different if the object of the analysis were estimation. The details of how design recommendations might differ if the focus were on estimation, rather than testing, merits further research.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*, 33–40. https://doi.org/10.3102/0013189X035006033

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

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

Birge, R. T. (1941). The general physical constants: As of August 1941 with details on the velocity of light. *Reports of Progress in Physics, 8*, 90–134. https://doi.org/10.1088/0034-4885/8/1/307

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., … Heikensten, E. (2016). Evaluating the reproducibility of laboratory experiments in economics. *Science, 351*, 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., … Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and *Science* between 2010 and 2015. *Nature Human Behavior, 2*, 637–644. https://doi.org/10.1038/s41562-018-0399-z

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

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

DeBray, M. E. J. G. (1934a). The velocity of light. *Nature, 133*, 464. https://doi.org/10.1038/133464a0

DeBray, M. E. J. G. (1934b). The velocity of light. *Nature, 133*, 948–949. https://doi.org/10.1038/133948c0

Dickersin, K. (2005). 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. 11–33). Chichester, UK: Wiley. https://doi.org/10.1002/0470870168.ch2

Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., … Waternaux, C. (1993). *Combining information: Statistical issues and opportunities for research*. Washington, DC: American Statistical Association.

Flowers, J. L., & Pentley, B. W. (2001). Progress in our knowledge of the fundamental constants in physics. *Reports of Progress in Physics, 64*, 1191–1246. https://doi.org/10.1088/0034-4885/64/10/201

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. https://doi.org/10.2307/1164832

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

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness, 11*, 1–21. https://doi.org/10.1080/19345747.2017.1375583

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. https://doi.org/10.1037/1082-989X.6.3.203

Hedges, L. V., & Schauer, J. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods, 24*, 557–570. https://doi.org/10.1037/met0000189

Hedges, L. V., & Schauer, J. (in press). More than one replication is study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998619852953

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

Higgins, J. P. T., & Green, S. (2008). *The Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley.

Ho, C. Y., Powell, R. W., & Liley, P. E. (1972). Thermal conductivity of the elements. *Journal of Physical and Chemical Reference Data, 1*, 279–421. https://doi.org/10.1063/1.3253100

Holling, H., Böhning, W., Masoudi, E, Böhning, D., & Sangnawakij, P. (2019). Evaluation of a new version of I^2 with emphasis on diagnostic problems. *Communications in Statistics – Simulation and Computation*. https://doi.org/10.1080/03610918.2018.1489553

Humphreys, L. G. (1980). The statistics of failure to replicate: A comment on Buriels (1978) conclusions. *Journal of Educational Psychology, 72*, 71–75. https://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.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science, 359*, 725–726. https://doi.org/10.1126/science.359.6377.725

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

International Committee of Medical Journal Editors. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *The New England Journal of Medicine, 351*, 1250–1251. https://doi.org/10.1056/NEJMe048225

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

Kulinskaya, E., Dollinger, M. B., & Bjorkestol, K. (2011). Testing for homogeneity in meta-analysis I: The one parameter case: Standardized mean difference. *Biometrics, 67*, 203–212. https://doi.org/10.1111/j.1541-0420.2010.01442.x

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

Mahlzahn, U., Böhning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika, 87*, 619–632. https://doi.org/10.1093/biomet/87.3.619

McGaw, B., & Glass, G. V. (1980). Choice of metric for effect size in meta-analysis. *American Educational Research Journal, 17*, 325–337.

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

Mohr, P. J., Newell, D. B., & Taylor, B. N. (2016). CODATA recommended values of the fundamental physical constants: 2014. *Journal of Physical and Chemical Reference Data, 45*, 1–73. https://doi.org/10.1103/RevModPhys.88.035009

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

Open Science Collaborative. (2016). Estimating the reproducibility of psychological science. *Science, 349*, 943–951. https://doi.org/10.1126/science.aac4716

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

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

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Drug Discovery, 10*, 712–713. https://doi.org/10.1038/nrd3439-c1

Rosenfeld, A. H. (1975). The particle data group: Growth and operations – Eighteen years of particle physics. *Annual Review of Nuclear Science, 25*, 555–599. https://doi.org/10.1146/annurev.ns.25.120175.003011

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. https://doi.org/10.1037/a0015108

Touloukian, Y. S. (1972). Reference data on thermophysics. In H. A. Skinner (Ed.), *International review of science physical chemistry. Vol 10. Thermochemistry and thermodynamics* (pp. 119–146). London, UK; Butterworth. https://doi/org/10.1016/B978-0-08-019850-7.50040-X

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values. *The American Statistician, 70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wilkinson, L., The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Zwolinski, B. J., & Chao, J. (1972). Critically evaluated tables of thermodynamic data. In H. A. Skinner (Ed.), *International review of science physical chemistry. Vol 7. Thermochemistry and thermodynamics* (pp. 93–120). London, UK: Butterworth.

History
Received February 16, 2019
Revision received June 24, 2019
Accepted July 9, 2019
Published online November 1, 2019

Larry V. Hedges
Department of Statistics
Northwestern University
2040 N Sheridan Road
Evanston, IL 60208
USA
l-hedges@northwestern.edu

Larry V. Hedges is chair of the Department of Statistics at Northwestern University, where is also the Board of Trustees Professor if Statistics, Education and Social Policy, Psychology, and Medical Social Sciences. He is best known for his work in the development of statistical methods for meta-analysis.