Chapter 14
Replicability and Meta-Analysis

Jacob M. Schauer

Abstract In this chapter, I will discuss statistical considerations for studying replication. More specifically, I will approach replication from a framework based on meta-analysis. To do so, I will focus on *direct* replications, where studies are designed to be as similar as possible, as opposed to *conceptual* replications that (systematically or haphazardly) vary in at least one aspect of an experiment. The chapter starts with a brief description of recent research on replication in psychology and uses examples from that research to highlight relevant considerations in defining and parametrizing "replication." It then outlines different ways to frame analyses of replication and provides examples. Finally, it takes one possible definition of replication—that effects found across studies involving the same phenomenon are consistent—and describes relevant analyses and their properties.

Keywords Replicability · Meta-analysis · Replication research · Direct replication

Introduction

In the first two decades of the twenty-first century, multiple research programs called into question the replicability of scientific findings in several fields, including psychology (e.g., Ioannidis, 2005; Open Science Collaboration, 2015; Camerer et al., 2016; Klein et al., 2014). These findings would seem to have serious implications for the evidence behind evidence-based practices, particularly in clinical and behavioral psychology. In response, various scientific bodies, such as the Association for Psychological Science (APS) and National Institute of Health (NIH), as well as individual researchers called for steps to improve the transparency and reproducibility of scientific research (see Pashler & Harris, 2012; Collins & Tabak, 2014; Perrin, 2014; Bollen et al., 2015; Head et al., 2015).

J. M. Schauer (✉)
Department of Preventive Medicine, Northwestern University, Evanston, IL, USA
e-mail: jms@u.northwestern.edu

W. O'Donohue et al. (eds.), *Avoiding Questionable Research Practices in Applied Psychology*, https://doi.org/10.1007/978-3-031-04968-2_14

Though enhanced transparency can improve the face validity and potential replicability of research results, a critical step in establishing scientific evidence involves actually conducting replication studies. Yet, until the mid to late 2010s, literature on methods for designing and analyzing replication studies was limited (Schmidt, 2009; Hedges & Schauer, 2019b). Methods for studying replication would seem to be simple: Just conduct an additional study (or several studies) and examine whether results are *the same*. But empirical research on replication has demonstrated that replication is anything but simple (see Bollen et al., 2015). It can be extremely difficult and time-consuming to standardize procedures to ensure that relevant factors are controlled across multiple studies. Moreover, emerging work on the statistical aspects of studying replication has revealed several key challenges for researchers.

These statistical challenges intersect everything from the definition of replication to analytic methods to the design and sample size considerations for replication studies. Precise definitions of replication, which are seldom directly specified, are required in order to identify a relevant analytic method for replication. Schauer and Hedges (2021) argue that there are several possible definitions of replication, including agreement in the direction, interpretation, and magnitude of effects across studies. Moreover, a definitive analytic method must be specified in order to determine the sample size required of replication studies, or indeed, to determine how many studies need to be conducted.

In this chapter, I will discuss statistical considerations for studying replication. More specifically, I will approach replication from a framework based on meta-analysis. To do so, I will focus on *direct* replications, where studies are designed to be as similar as possible, as opposed to *conceptual* replications that (systematically or haphazardly) vary in at least one aspect of an experiment (for discussion, see Collins, 1992; Schmidt, 2009). The chapter starts with a brief description of recent research on replication in psychology and uses examples from that research to highlight relevant considerations in defining and parametrizing "replication." It then outlines different ways to frame analyses of replication and provides examples. Finally, it takes one possible definition of replication—that effects found across studies involving the same phenomenon are consistent—and describes relevant analyses and their properties.

What Does Research on Replication Look Like?

To understand what replication research looks like, it helps to look at how researchers have approached the study of replication, including those in psychology. Perhaps the most high-profile replication research projects were the Replication Project: Psychology (RPP; Open Science Collaboration, 2015) and the Replication Project: Economics (RPE; Camerer et al., 2016). Both of these projects took a series of scientific findings and ran a single replication of each: The RPE focused on 18 different experiments in behavioral economics, while the RPP looked at 100 social and behavioral psychology experiments, 73 of which they identified as a "meta-analytic

subset" for which meta-analysis methods would be appropriate. The lengths taken to standardize and register protocols, ensuring that the replications in question could be seen as direct replications (or as direct as possible), were documented by the RPP (Open Science Collaboration, 2012).

However, there is no reason that researchers need to stop at just one replication study. Depending on the finding in question, it may be possible to conduct multiple replication studies, as was the case with the Many Labs Replication Project (Klein et al., 2014). Many Labs recruited 36 laboratories to run the same set of experiments. In the same year Many Labs published their results, the APS announced a series on the Registered Replication Reports (Simons, Holcombe, & Spellman, 2014). These efforts have conducted several replications of a given finding, from as few as 13 to as many as 33 studies (see Alogna et al., 2014; Bouwmeester et al., 2017; Cheung et al., 2016; Eerland et al., 2016; Hagger et al., 2016; Wagenmakers et al., 2016). Subsequent projects, including various iterations of Many Labs (e.g., Ebersole et al., 2016; Klein et al., 2018, 2019) and the Pre-Publication Independent Replication (PPIR) project (Schweinsberg., 2016) have also approached the study of replication as one that relies on several replication studies. This approach has been adopted by the Psychological Science Accelerator, an international collaboration of over 500 laboratories across the globe dedicated to conducting simultaneous replications across several laboratories (see Moshontz et al., 2018). To date, large-scale programs devised to study replication have seldom attempted to replicate findings in clinical psychology.

To unpack what these programs imply about replication research, we can zoom in on a single experiment. For instance, the RPP (Open Science Collaboration, 2015) ran a replication of an experiment first described by Payne et al. (2008). The original study examined the correlation between time spent awake and participant's memory of negative objects or scenes. The experiment involved presenting participants with a series of negative and neutral images, randomizing participants to conditions that corresponded with different sleeping conditions, and then asking them to respond to a set of images similar to those they were shown previously. The RPP conducted a single replication of this experiment. In that sense, programs, such as the RPP and RPE, have taken an approach of studying replication by conducting a single replication study for a finding, typically with a larger sample size than the original study.

Contrast that with a program, such as Many Labs, which replicated experiments like the reverse gambler's fallacy (Oppenheimer & Monin, 2009). In the original experiment, participants were asked to imagine a man rolling dice at a casino. In the two arms of the study, participants imagined seeing the man roll three sixes versus seeing him rolling two sixes and a three. 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 sixes tended to estimate the man had rolled the dice more times than those who imagined seeing only two sixes. Many Labs ran this experiment 36 times across different laboratories at (roughly) the same time. Analyses used by Many Labs included comparing the original study to an average of the effects found in the replication studies,

as well as examining variation across the replication studies. Viewed this way, the study of replication may require several replication studies conducted simultaneously, and potentially in different laboratories or settings.

Model and Notation

As noted above in this chapter, I adopt a model and notation to describe the data generated by replication studies that is commonly used in meta-analysis. Meta-analysis is particularly germane to discussions of replication as it concerns the statistical methodology for studying the results of multiple (i.e., two or more) studies (Hedges & Olkin, 1985; Cooper et al., 2019). Suppose there are k studies conducted; replication research involves $k \geq 2$ studies. Often, one of these studies is published, though it is neither infeasible nor without precedent that multiple replication studies could be conducted prior to publishing any result (see e.g., Schweinsberg et al., 2016; Moshontz, et al., 2018): For the Payne/RPP sleep-memory study, $k = 2$ (i.e., an original and a replication study), for Many Labs' gambler's fallacy study, $k = 36$.

In any single study, the focus of statistical analysis is a quantity known as the *estimand.* An estimand is a quantity to be estimated or evaluated in a statistical analysis. The term is used to more clearly distinguish the target of inference (i.e., the *estimand*) from the method used to make inferences about that target (i.e., the estimator) and the specific value obtained from a given method and dataset (i.e., the estimate; for discussion in participant or patient outcomes, see Lawrance et al., 2020). An estimand could be a treatment effect in a randomized trial, such as a standardized mean difference or log odds ratio, a population parameter, or some parameter in a statistical model. For the sake of simplicity, the language in this chapter will refer to *effects* or *treatment effects* and assume that effects are one of the standard effect size indices commonly used in meta-analysis, such as the mean difference, standardized mean difference (Cohen's d), log odds ratio, risk ratio, and correlation coefficient (see Cooper et al., 2019). I will present results on the scale of standardized mean differences; however, the statistical results presented largely hold for other quantities.

Within study $i = 1, \ldots, k$, let θ_i be the effect or estimand of interest. Note that it may be possible (even probable) that $\theta_i \neq \theta_j$ even among direct replications if there are any uncontrolled sources of variation between studies (e.g., samples derived from different populations, potentially unknown deviations in protocols; Hedges & Schauer, 2019b). In later sections I will discuss an important way to conceive of the θ_i as either fixed but unknown quantities, or as random variables (referred to as the random effects model in meta-analysis). When the θ_i are treated as random variables, their distribution is assumed to have a mean μ and variance τ^2.

In practice, we do not observe θ_i directly, but instead must estimate it from data collected within study i. Denote T_i as the estimate of θ_i and let v_i be the estimation variance of T_i. Thus, from each study, we obtain an effect estimate T_i, and a variance

v_i or standard error $\sqrt{v_i}$. The statistical results in this chapter make three assumptions about T_i. First, that T_i is an unbiased estimator of θ_i. Second, that T_i is normally distributed. Third, that v_i is known (or estimated with very little uncertainty). Taken together, these assumptions imply

$$T_i \sim N\left(\theta_i, v_i\right)$$

where v_i is known. This will be exactly or approximately true for estimates of most effect size indices, including standardized mean differences (with reasonably large sample sizes), mean differences, log odds ratios, or z-transformed correlation coefficients (Cooper et al., 2019; Borenstein et al., 2009). Note that in a two-armed experiment, the variance v_i of the standardized mean difference can be expressed as

$$v_i = \frac{n_i^T + n_i^C}{n_i^T n_i^C} + \frac{\theta_i^2}{2\left(n_i^T + n_i^C\right)} \tag{14.1}$$

where n_i^T and n_i^C are the sample size of a treatment and control groups, respectively (Hedges, 1982). In a balanced experiment, where $n_i^T = n_i^C = n_i/2$ (i.e., n_i is the total sample size), so long as effects are relatively small and sample sizes within groups $n_i/2$ are reasonably large, we can write

$$v_i \approx \frac{4}{n_i} \tag{14.2}$$

Additional Notation

A key attribute of the statistical model above is that it distinguishes between an effect parameter θ_i and effect estimate T_i. Understanding this distinction will be sufficient for unpacking most of the key considerations for defining and evaluating replicability. For readers interested in more technical aspects of analyses for replication, this section provides some other useful values that arise in analysis methods discussed in this chapter. These serve as a reference for subsequent equations in this chapter.

- The precision weighted average of effect parameters. This is one way to average the effects across replication studies, wherein effects that are more precisely estimated receive more weight.

$$\bar{\theta}_{\cdot} = \sum_{i=1}^{k} \frac{\theta_i}{v_i} \tag{14.3}$$

- The unweighted average of effect parameters. This is an alternative to the weighted average in (14.3).

$$\bar{\theta} = \sum_{i=1}^{k} \frac{\theta_i}{k} \tag{14.4}$$

  - Note that when all v_i are equal so that $v_i = v$, then $\bar{\theta}$ is equivalent to $\bar{\theta}_.$. Unless there is substantial variation in sample sizes across studies, the averages in (14.3) and (14.4) will often be similar in value.

- The precision weighted average of effect estimates. This is typically used to summarize or average effect estimates in meta-analysis, and gives greater weight to studies with smaller variances (i.e., more weight is given to studies with bigger sample sizes).

$$\bar{T}_. = \left(\sum_{i=1}^{k} \frac{T_i}{v_i} \right) \Big/ \left(\sum_{i=1}^{k} \frac{1}{v_i} \right) \tag{14.5}$$

  - Note that when all v_i are equal so that $v_i = v$, then $\bar{T}_.$ is equivalent to the unweighted mean $\bar{T} = \sum_{k}^{i=1} T_i / k$.
  - Among the $k = 36$ effect estimates and variances reported by Many Labs' reverse gambler's fallacy experiments (see Table 14.3), the weighted mean of effects is 0.63 and the unweighted mean is 0.61.

- The Q statistic is used to test heterogeneity and estimate between-study variance:

$$Q = \sum_{i=1}^{k} \frac{\left(T_i - \bar{T}_. \right)^2}{v_i} \tag{14.6}$$

  - For the Many Labs' reverse gambler's fallacy example, the Q statistics is 51.61.
  - For $k = 2$ studies, the Q statistic reduces to $(T_1 - T_2)^2/(v_1 + v_2)$.

- A sum of (powers of) precisions S_j is used in computing various quantities related to variation between studies and standard errors of meta-analytic estimates:

$$S_j = \sum_{i=1}^{k} \frac{1}{v_i^j} \tag{14.7}$$

  - Note that when all v_i are equal so that $v_i = v$, $S_j = k/v^j$.

- The constant S is a function of the estimation error variances v_i used in common estimators of between-study variation:

$$S = S_1 - \frac{S_2}{S_1} = \sum_{i=1}^{k} \frac{1}{v_i} - \frac{\sum_{i=1}^{k} \frac{1}{v_i^2}}{\sum_{i=1}^{k} \frac{1}{v_i}} \tag{14.8}$$

  - Note that when all v_i are equal so that $v_i = v$, $S = (k - 1)/v$.

- An estimate of the variation between effect parameters is based on the Q statistic in (14.6) (DerSimonian & Laird, 1986):

$$\hat{\tau}^2_{DL} = \frac{Q - (k-1)}{S} \tag{14.9}$$

  - For the reverse gambler's fallacy example, the estimated between-study variance is $\hat{\tau}^2_{DL} = 0.01$.

- A random effects weighted average of effect estimates:

$$\bar{T}^*_{\cdot} = \left(\sum_{i=1}^{k} \frac{T_i}{v_i + \hat{\tau}^2_{DL}} \right) \Big/ \left(\sum_{i=1}^{k} \frac{1}{v_i + \hat{\tau}^2_{DL}} \right) \tag{14.10}$$

  - This is analogous to the weighted average in (14.5), except the weights in (14.10) involve the estimated between-study variation $\hat{\tau}^2_{DL}$. In the reverse gambler's fallacy example, the unweighted mean is $\bar{T}_{\cdot} = 0.61$, the precision weighted mean is $\bar{T}_{\cdot} = 0.63$, and the random effects weighted mean is $\bar{T}^*_{\cdot} = 0.61$.

What We Mean When We Say "Replication"

Conventional understanding of successful replications is that they get "the same" result or outcome. Yet, when it comes to statistical analyses, "the same" has proven to be tricky to characterize precisely (see Valentine et al., 2011; Bollen et al., 2015; Hedges & Schauer 2019b; Schauer & Hedges, 2021; Schauer et al., 2021). Doing so requires some decision-making about the studies involved and how they pertain to the finding under scrutiny. Because of this, there are *several* possible definitions of replication success or failure, and different ways to quantify these definitions (Schauer & Hedges, 2021). As one might expect, an analytic method for one definition of replication may be wholly inappropriate for a different definition of replication. Thus, before conducting any analysis of replication, it is critical to formalize the relevant definition. In this section, I will discuss various ways in which

definitions of replication can be structured, and why that can matter for making inferences about the replicability of scientific findings.

Definition of Replication Versus Analysis Methods

An important distinction to make in the context of replication research is between the underlying *definition* of replication and *analysis methods* for a given definition. A definition of replication ought to concern the effect parameters θ_i. The θ_i are the actual effects produced in an experiment or study, and hence reflect a study's true results (i.e., the true effect). An analysis method uses data (i.e., the effect estimates T_i and variances v_i) to infer something about the relationships between the θ_i. That is, an analysis method—which is a function of the T_i and v_i—concerns a specific formal definition of replication—which is a function of the θ_i.

To unpack this distinction, consider a research design with $k = 2$ studies: an original study (study 1) and a replication (study 2). There appear to be two commonly accepted definitions of replication success for two studies. First, effects could agree in sign/direction, so that both effects are positive or negative (e.g., a treatment improves outcomes in both studies). Second, effects could agree in magnitude, so that effects are the same size in each study. The first column of Table 14.1 shows a mathematical formalization of these two definitions. Note that the sign () and 1{} functions are as follows:

$$sign(x) = \begin{cases} 1 \text{if } x > 0 \\ 0 \text{if } x = 0 \\ -1 \text{if } x < 0 \end{cases}, \qquad 1\{x\} = \begin{cases} 1 \text{if } x \text{ is TRUE} \\ 0 \text{if } x \text{ is FALSE} \end{cases} \tag{14.11}$$

Common analysis methods used to determine if study 2 failed to replicate study 1 include the statistical significance criterion, the confidence interval overlap (CIO) procedure (Brandt et al., 2014), and the prediction interval (PI) procedure (Patil, Peng, & Leek, 2016). Table 14.1 describes these approaches as statistical procedures.

Table 14.1 Some definitions of replication and commonly used methods to assess those definitions

Definition of replication	Some proposed analyses
Agreement in sign/*direction* of effects sign (θ_1) = sign (θ_2)	*Significance criterion:* sign (T_1) = sign (T_2) AND both are significant (or both nonsignificant)
Agreement in *magnitude* of effects $\theta_1 = \theta_2$	*Confidence interval overlap:* $1\{(T_1 - T_2)/\sqrt{v_2} > 1.96\}$
	Prediction interval: $1\{(T_1 - T_2)/\sqrt{v_1 + v_2} > 1.96\}$

- The statistical significance procedure concludes study 2 failed to replicate if it disagrees in sign or statistical significance compared to study 1 (e.g., T_1 and T_2 are both positive, but T_1 is statistically significant and T_2 is not). Typically, statistical significance is set to the $\alpha = 0.05$ level for this procedure and two-sided tests are used where appropriate.
- The CIO method concludes a study failed to replicate if T_1 is not contained in a 95% confidence interval for θ_2 in study 2. Note that the confidence interval for study 2 only accounts for estimation error variance in study 2, but not in study 1.
- To adjust for that, the PI approach concludes a study failed to replicate if T_1 is not contained in a 95% *prediction* interval for θ_2, where the prediction interval takes into account estimation error variance in both study 1 and study 2. This is equivalent to concluding study 2 failed to replicate study 1 if 95% confidence intervals from the two studies do not overlap.

Types of Agreement

An important consideration for defining replication is what we mean by "the same" results; that is, what type of agreement do we expect or desire out of our replication studies? The example above described two possible types of agreement: agreement in sign/direction and agreement in magnitude. These appear to be among the most commonly accepted types of agreement in replication research. However, they are not the only possible type of agreement. For instance, we might consider studies to agree qualitatively if their effects are all large enough to be considered clinically relevant, so that $\theta_i > q$ for some threshold value q that corresponds to a clinically relevant effect (see Mathur & VanderWeele, 2020).

In this chapter, I will focus on agreement in magnitude, which can be seen as a finer, more restrictive definition of replication. For instance, if $\theta_1 = 0.2$ in Cohen's d units and $\theta_2 = 20$, this would characterize a "successful" replication if our preferred definition involved the direction of effects, yet few social scientists would consider these to be similar in size given that they differ by two orders of magnitude. In that sense, agreement in direction is a coarser definition of replication. In a clinical setting, agreement in magnitude can provide greater confidence about the stability and predictability of effects and can potentially better inform decisions about implementing an intervention that must weigh potential benefits against anticipated costs or side effects.

Exact Versus Approximate Replication

Agreement in magnitude of effects can also be specified in different ways. An obvious way would be to require effects to be identical, so that $\theta_1 = \ldots = \theta_k$, a scenario referred to as *exact replication* (Hedges & Schauer, 2019b). However, one might

expect findings to vary slightly across repeated studies due to sampling subjects from slightly different populations or minor—often unknown—deviations in study implementation. If such differences produce small, but negligible variation between effects, that could still be seen as successful replication so long as the resulting effect parameters would warrant the same scientific or clinical interpretation. For instance, if $\theta_1 = 0.2$ (Cohen's d) and $\theta_2 = 0.201$, many social and psychological researchers might consider those to be about the same. Thus, we may also define *approximate replication* as when differences between effect parameters are negligibly small (for further discussion, see Schauer, 2018; Hedges & Schauer, 2019a,b). I will formalize ways to operationalize "negligibly small" in subsequent sections.

Falsification Versus Consistency

When $k > 2$ studies are involved in replication research (e.g., an original study and multiple replications), there are at least two ways to orient an analysis of replication. First, the focus could be on singling out a single study (or group of studies) and comparing it to the others. In such analyses, the original study is typically compared to subsequent replication studies as a means of falsifying the original study. If multiple replication studies have been conducted, analyses of replication may aggregate their results, including via a meta-analysis, so that the analysis compares the effect from the original study and the average effect found in the replication studies. We refer to this type of orientation as a *falsification* approach. Note that analyses based on a *falsification* approach to replication need not result in yes/no conclusions about the original study and could instead focus on continuous metrics, such as the size of the difference between the original study and subsequent replications.

Because falsification definitions contrast the original effect parameter θ_1 to an average of the replication study estimates, it can be seen as treating multiple replication studies (study 2, …, study k) as a single large study. Hence, analyses of falsification definitions of replication are statistically analogous to analyses for $k = 2$ studies, even when $k > 2$ studies (i.e., multiple replication studies) have been conducted.

Rather than singling out one specific study, a definition can focus on whether there is agreement across all studies. In this definition, the focus is on variation across all effects, rather than a comparison of one study versus an average of several others. If there is little or no variation among effects, then we might conclude that the finding is relatively consistent across all studies, and hence we refer to this framing as a *consistency* approach.

When there are only $k = 2$ studies, consistency and falsification analyses are identical. A comparison between an original study and a single replication is equivalent to an analysis that examines differences (i.e., variation) between the two study effects.

Fixed Versus Random Studies

Another consideration in defining and making inferences about replication is whether the studies and their resulting effect parameters are fixed or random (Hedges & Schauer, 2019b). The *fixed effects* model assumes that the studies and effect parameters are the only studies of interest. Inferences will pertain only to the studies that we observe and their corresponding effect parameters θ_i (i.e., did these specific studies successfully replicate?). The *random effects* model assumes that the studies observed are only a sample from a population of replication studies that could be observed. In this model, the θ_i are treated as draws from some distribution or putative population of effect parameters. Inferences about replication pertain to the population of studies, including those not observed. The distinction between fixed and random studies models is analogous to the fixed and random effects models in meta-analysis (Laird & Mosteller, 1990; Hedges & Vevea, 1998).

Putting It All Together: Defining Replication

All of the considerations above are necessary for defining "replication" (i.e., defining results "being the same") as a quantity on which we can conduct inference. Table 14.2 shows different ways we might define replication when we have different views of these considerations. These are not the only possible ways to define replication and other quantities related to replication may possibly be of interest to researchers.

Table 14.2 also demonstrates that the estimand that corresponds to "replication" will depend heavily on the considerations listed in this section. In practice, the distinction between fixed and random effects definitions and analyses is minor, and leads to parameters that differ in their precise statistical interpretation, but have roughly the same scale (see Schauer, 2018; Hedges & Schauer, 2019b). Yet, the framing of the definition of replication (consistency versus falsification) and the type of agreement (magnitude versus direction) can lead to markedly different parameters corresponding to "replication." It is therefore imperative that researchers identify the relevant framing and agreement type in advance.

Once again, in this chapter, I highlight definitions of replication that correspond to consistency across all effects in magnitude. This definition of replication seeks to identify if (and to what degree) effects of an intervention may change over repeated trials. Defining replication in this way can be seen as consistent with moves toward evidence-based practices, as well as with conventional notions regarding the role of replication in the scientific method. Furthermore, it can provide researchers and practitioners with a clearer picture of how stable an intervention's impact is, and potential conditions under which it may change.

Table 14.2 Some possible definitions/parametrizations of replication

		Studies fixed	Studies random
Falsification	Agreement in magnitude	Difference between original study and replication: $\lvert\theta_1 - \theta_2\rvert$ for $k = 2$ $\ddagger\ \theta_1 - \bar{\theta}_{\#}^{\#}$ for $k > 2$	Comparison of original study to distribution of effects in replications: $P_{\text{orig}} = P[\theta_1 > \theta_i]$, $i > 1$ $d_{\text{orig}} = (\theta_1 - \mu)/\tau$
	Agreement in direction	Replication effect parameters are in the same direction as the original: sign (θ_1) = sign (θ_2) for $k = 2$ $sign(\theta_1) = sign(\bar{\theta}_{\cdot})$ for $k > 2$	Probability that replication effects from population are in same direction as original: $P[\text{sign}\ (\theta_i) = \text{sign}\ (\theta_1)]$, $i > 1$
Consistency	Agreement in magnitude	Variation across effects from observed studies: $\lvert\theta_1 - \theta_2\rvert$ for $k = 2$ $\tau_F^2 = \sum_{i=1}^{k} \frac{\left(\theta_i - \bar{\theta}_{\cdot}\right)^2}{k-1}$ $\lambda = \sum_{i=1}^{k} \frac{\left(\theta_i - \bar{\theta}_{\cdot}\right)^2}{v_i}$	Variation across population of effect parameters: $Var[\theta_i] = \tau^2$
	Agreement in sign	Proportion of effects that are positive: $\sum_{i=1}^{k} \frac{1\{\theta_i > 0\}}{k}$	Probability effects are positive: $P_{>0} = P[\theta_i > 0]$

Considerations for Analyses of Replication

When we can precisely state a definition of replication, our next challenge involves formalizing analyses for replication. We can conduct statistical analyses in at least two ways: analyses that lead to binary conclusions about replication and analyses that quantify continuous metrics that correspond with a definition for replication. In addition, any analysis of replication that includes an extant published study must consider the potential impact of publication bias on the analysis (see below). In this section, I will describe some potential choices about how to frame an analysis for replication and discuss possible publication bias adjustments.

Categorical Decisions About Replication: Did the Finding Fail to or Successfully Replicate?

One class of statistical analyses is one that supports qualitative conclusions about replication (e.g., "the replication(s) failed") via some decision procedure. The analysis methods discussed above (Significance, CIO, PI) can be seen as part of this class, as they all result in a success/failure conclusion. More broadly, this type of analysis is common in the null hypothesis test (NHT) framework, wherein we test a null hypothesis about replication and draw conclusions about replication success or failure based on a test of that null hypothesis. For example, each of the Significance, CIO, and PI methods can be seen as tests of a null hypothesis that the replication succeeded, and we would reject that null hypothesis and conclude a finding failed to replicate if a criterion was not met (see Schauer & Hedges, 2021; Schauer et al., 2021).

The Burden of Proof

If using NHTs, an important consideration is the burden of proof, which dictates how to form the null hypothesis. The burden of proof for an NHT about replication can either be on replication or nonreplication (i.e., replication success or failure, respectively). If the burden of proof is on replication, then we would form a null hypothesis that corresponds to replication failure, and we would require evidence to reject that hypothesis and conclude replication success. Conversely, if the burden of proof is on nonreplication, the null hypothesis should correspond with replication success, and we would require evidence to reject that.

As an example, suppose we were interested in exact replication for $k = 2$ studies (i.e., $\theta_1 = \theta_2$). If the burden of proof was on nonreplication, we would form the null hypothesis H_0: $\theta_1 = \theta_2$ and we would only conclude replication failure if our analysis rejected H_0 (e.g., if the PI method indicated replication failure). Conversely, if the burden of proof is on replication, then we would need to form H_0: $\theta_1 \neq \theta_2$ and only conclude that $\theta_1 = \theta_2$ if we reject H_0. Forming a null hypothesis of replication failure that is testable can be done in a manner analogous to equivalence testing, which involves setting H_0: $|\theta_1 - \theta_2| > \varepsilon$ for some constant $\varepsilon > 0$ (see Wellak et al., 2002; Hedges & Schauer, 2019a,b). This null hypothesis contends that the replication failed and the difference between θ_1 and θ_2 is at least as big as some nonzero value ε that characterizes the smallest non-negligible difference between effects consistent with replication failure. We would reject that hypothesis and conclude replication success—that the difference in effects is less than the smallest non-negligible difference between effects ε (i.e., the difference between effects is negligible)—if the data provided evidence to do so.

Decision-Theoretic Properties/Error Rates

As with any NHT, analytic methods that produce qualitative inferences about replication can result in erroneous conclusions about replication. The rate at which these types of errors occur are crucial for interpreting their results. For example, if the burden of proof is on nonreplication, then a Type I error indicates that we conclude replication failure when the replication(s) was successful. If an analytic method has a high Type I error rate, then it has a high probability of labeling successful replications as failures.

The meaning of Type I and Type II errors depends on the burden of proof. A Type I error when the burden of proof is on nonreplication is the same as a Type II error when the burden of proof is on replication; in both cases successful replications are labeled as failures (or at the very least lacking evidence of success). To unify this nomenclature, we use the terms *false failure* and *false success* determinations (Schauer & Hedges, 2021). A false failure occurs when the replication succeeded but the analytic method does not indicate success. Conversely, a false success occurs when a replication failed but the analytic method does not indicate failure (for further discussion, see Schauer & Hedges, 2021).

Continuous Measures and Estimation

Rather than resulting in success/failure decisions about replication, analyses can involve estimating relevant quantities and their related uncertainty. Typically, uncertainty would include standard errors of estimators or confidence or credible intervals. As an example, for $k = 2$ studies, analyses that focus on agreement in magnitude may estimate $\theta_1 - \theta_2$ and report a standard error for that difference. If there are $k > 2$ studies that are treated as random, analyses could involve estimating τ^2, the between-study variation (discussed in subsequent sections) and its standard error. Conversely, if agreement in direction is the preferred definition of replication, there are a variety of alternatives. For example, Mathur and VanderWeele (2020) propose estimating the proportion of effect parameters that exceed some value q, which they denote $P_{>q}$. For agreement in direction, we can specify $q = 0$, and estimate $P_{>0}$ as

$$P_{>0} = 1 - \Phi\left(\frac{\bar{T}_{\cdot}^{*}}{\hat{\tau}_{DL}}\right) \tag{14.12}$$

where $\bar{T}_{\cdot}^{*}$ and $\hat{\tau}_{DL}^{2}$ are given in (14.10) and (14.9), respectively, and Φ is the distribution function for the standard normal distribution. This has an estimated standard error of:

$$\phi\left(\frac{-\bar{T}_{\cdot}^{*}}{\hat{\tau}_{DL}}\right)\sqrt{\frac{Var\left[\bar{T}_{\cdot}^{*}\right]}{\hat{\tau}_{DL}^{2}}+\frac{SE\left[\hat{\tau}_{DL}^{2}\right]^{2}\bar{T}_{\cdot}^{*2}}{4\hat{\tau}_{DL}^{6}}} \tag{14.13}$$

where ϕ is the standard normal density function, the variance of $\hat{\tau}_{DL}^{2}$, $SE\left[\hat{\tau}_{DL}^{2}\right]$, is described in (14.26) in later sections, and the variance $Var\left[\bar{T}_{\cdot}^{*}\right]=\left(\sum_{i=1}^{k}\frac{1}{v_i+\hat{\tau}_{DL}^{2}}\right)^{-1}$.

An alternative approach proposed by Etz and Vandekerckhove (2016) involves examining Bayes factors of original and replication studies. Bayes factors are analogous to hypothesis testing in that they evaluate the relative likelihood of competing hypotheses. In replication research, this often takes the form of a ratio of the likelihood of replication success relative to the likelihood of replication failure. This approach, which is appropriate for $k = 2$ studies, can be seen as examining the evidence provided by the replication study of a nonzero effect under competing models: that the effect in the replication study is $\theta_2 = 0$, and that the effect is equivalent to that estimated in study 1. Though the Etz and Vandekerckhove discuss these as continuous metrics, they also use their value to make qualitative inferences about replication success for failure. For instance, a Bayes factor at least as large as 10 is seen as strong support of a nonzero effect while a Bayes factor of 1/10 or less is seen as strong evidence of a null effect. If such inferences are made, then these methods can be seen as producing qualitative assessments about replication, and their properties should be discussed in terms of false failure and false success error rates rather than standard errors.

Publication Bias

Both empirical and theoretical researches suggest that published findings are subject to a selection process that favors the publication of statistically significant results (see Dickersin, 2005; Rothstein et al., 2005; Francis, 2012). If the probability that a finding is published depends on its statistical significance, this can induce bias in the effect size estimate T_i, and can impact the sampling distribution of T_i so that T_i is no longer normally distributed (see Hedges, 1984; Guan & Vandekerckhove, 2016). In the context of replication research wherein researchers conduct replications of a published study, there may be concern that the estimates reported by studies that were published prior to conducting replication studies may be affected by this process and could therefore be biased. This in turn can impact statistical analyses of replication and their properties.

Analyses can adjust for publication bias if it is suspected (see Rothstein et al., 2005; McShane et al., 2016). Adjustments should be focused on effect estimates for which researchers have good reason to suspect publication bias. This will likely include only a subset of relevant studies in replication research. Because many

replications are pre-registered and have not yet been published, it is unlikely that publication selection will bias effect estimates from those studies. However, if extant published findings are to be included in analyses of replication, it would seem more likely that those effect estimates would be biased due to selection.

There are several possible relevant adjustments for publication bias. For example, Hedges (1984) provides a maximum likelihood estimator for unbiased estimation in the face of publication selection. This was one of the first approaches based on selection modeling, in which the process by which findings are selected for publication is based on their effect estimates T_i, variances v_i, or p-values. At their most basic, selection models assume that we only observe a T_i conditional on it being published, which in turn depends on its p-value. For instance, we might expect "statistically significant" T_i with $p_i < 0.05$ to be published with high probability (i.e., near 100%), but "nonsignificant" T_i with $p_i \geq 0.05$ to be published with a lower probability (e.g., near 40–50%). Thus, a published T_i has a conditional distribution affected by the probability of selection. To back out its unconditional distribution (and reduce or eliminate bias), we need to model the probability that T_i is published given its p-value. Selection models typically involve estimating the probability that T_i is published given its p-value and making relevant adjustments based on that probability, but such estimates typically require large numbers of studies subject to publication bias (Hedges & Vevea, 1996).

These models have since been extended to account for increasingly complex relationships between estimators T_i, variance v_i, and the probability of publication (Hedges & Vevea, 2005). Vevea and Woods (2005) propose an adaptation to selection model approaches that would seem appropriate for cases where only one or two effect estimates are biased due to publication selection. This approach assumes that the probability that a significant T_i gets published and a nonsignificant T_i goes unpublished are known a priori and need not be estimated from the data. A Bayesian method that makes (more or less) the same set of assumptions was applied to replication studies by Etz and Vandekerckhove (2016), and a hybrid model was presented by van Aert and van Assen (2017).

Finally, for analyses of $k > 2$ studies that focus on consistency, if published effect estimates are suspected to have severe publication bias, they can be omitted from analyses. Omitting biased effect estimates may make sense if the assumptions made by publication bias adjustments (model specification and parameter values) are untenable or difficult to justify. However, excluding the original study limits the scope and sample size of the analysis.

Perhaps a more principled approach would be to conduct a series of analyses each based on different publication bias adjustments (including no adjustment at all). Results of each analysis would then be presented and interpreted in light of the plausibility and strength of relevant assumptions.

Some Limitations of Statistical Significance and Confidence Interval Overlap

At the time of this writing, conducting a single replication study ($k = 2$ designs) remains a popular approach to studying replication (see Camerer et al., 2018). Moreover, the Significance, CIO, and PI approaches are still in common use. However, this approach to studying replication and these methods have some serious flaws. First, all three analysis methods are really only appropriate for $k = 2$ studies: the original study (study 1) and a replication study (study 2; Schauer, 2018; Schauer & Hedges, 2021). When multiple replication studies are conducted, these methods proceed by aggregating their effect estimates via meta-analysis. Thus, these methods are limited to designs with $k = 2$ studies, or if the framing of replication is falsifiability. Note that this is true of many proposed Bayesian analyses (for discussion see Hedges & Schauer, 2019a; Schauer & Hedges, 2021).

In addition, the statistical properties of these approaches can result in erroneous conclusions about replication with high probability. Previous research examined the false failure and false success rates of the Significance, CIO, and PI methods (Schauer & Hedges, 2021). For example, Schauer and Hedges (2021) found that the error rates of the Significance criterion depend on the power of study 1 and study 2 to detect effects θ_1 and θ_2, respectively. Unless both studies have very high power (i.e., >90% power), the false failure rate can range from 30% to over 70%, while the false success rate is likely between 15 and 30%.

The error rates of the CIO method are largely a function of the ratio of v_1/v_2 (Schauer & Hedges, 2021; Schauer et al., 2021). When v_1/v_2 is high (which occurs if study 2 has a larger sample size than study 1), the false failure rate can be as large as 20–40%. However, when v_1/v_2 is small, this can inflate the false success rate, which could be as large as 80%. In short, depending on the sample sizes and effect sizes of the studies involved, both Significance and CIO may be more likely to result in an error than in an accurate conclusion about replication.

It is worth noting that in addition to potentially high error rates, neither CIO nor the Significance criteria control error rates. In traditional NHTs, the procedures used should (and often do) limit the probability of a Type I error to be no greater than some a priori threshold α. The benefit of controlling the Type I error rate is that (so long as assumptions are met) the probability of a Type I error is (more or less) known and independent of other factors, such as sample size. Because of this, rejection of the null hypothesis can be seen as conclusive, since the probability that it is rejected in error is known (or at least bounded). However, as shown by Schauer and Hedges (2021), the false failure and false success rates of the CIO and Significance methods are functions of the v_i, and hence functions of sample size, as well as the θ_i. In sum, these methods control neither the false failure nor false success rates. Conclusions about replication generated by these procedures may be false with unknown (and possibly large) probability, and therefore it is difficult to view the results of these methods as particularly conclusive in many settings. On a related note, by contrast, Schauer and Hedges (2021) show that the PI criterion is

equivalent to z-test for a difference in means, and therefore controls the false failure rate, a result we will point out in later sections of this chapter.

Defining Replication as Consistency of Effects

As argued above, agreement in magnitude may be a more informative definition of replication when it comes to clinical decision-making. To that end, there may be interest in the clinical psychology research community to emphasize definitions of replication that correspond to consistency of effects across studies. Table 14.2 characterizes ways we can parametrize such definitions for the fixed and random studies framework. If we treat the studies as random, we can quantify their consistency in terms of the variance of the distribution from which they were drawn, denoted τ^2. If $\tau^2 = 0$, then all of the effect parameters drawn from that distribution will be identical, and replication will be exact. If $\tau^2 > 0$ but is small, then effect parameters drawn from that distribution will be similar in size and may be seen to replicate approximately (Hedges & Schauer, 2019b).

When the studies are fixed, there are at least two ways to define agreement in magnitude for consistency analyses. One is with their "variance" τ_F^2:

$$\tau_F^2 = \sum_{i=1}^{k} \frac{\left(\theta_i - \bar{\theta}\right)^2}{k-1} \tag{14.14}$$

Here, $\bar{\theta}$ is the mean of the θ_i given in Eq. (14.4). The parameter τ_F^2 is a summary statistic of the θ_i akin to a sample variance, as opposed to a property of a distribution like τ^2 (Schauer, 2018).

Alternatively, Hedges and Schauer (2019b) suggest that replication be parametrized by

$$\lambda = \sum_{i=1}^{k} \frac{\left(\theta_i - \bar{\theta}_\cdot\right)^2}{v_i} \tag{14.15}$$

where $\bar{\theta}_\cdot$ is the precision weighted mean of the θ_i given in Eq. (14.3). The parameter λ is analogous to τ_F^2, but differs in that it accounts for the within-study estimation error variance v_i. When all of the studies have the same estimation error variance (e.g., the same sample size), then $v_1 = \ldots = v_k = v$, and the expression for λ reduces to

$$\lambda = \left(k-1\right)\frac{\tau_F^2}{v} \tag{14.16}$$

Thus, the primary difference between τ_F^2 and λ is that τ_F^2 is on the scale of the individual θ_i, while λ is a ratio of between-to-within-study variance, a scale commonly

used in meta-analysis (Schauer, 2018). This becomes evident when $k = 2$, so that τ_F^2 and λ reduce to expressions that depend on the magnitude of the difference $|\theta_1 - \theta_2|$:

$$\tau_F^2 = \frac{|\theta_1 - \theta_2|^2}{2}; \lambda = \frac{|\theta_1 - \theta_2|^2}{2\overline{v}} \tag{14.17}$$

where $\overline{v} = (v_1 + v_2)/2$ is the mean within-study variance.

The difference between τ_F^2 and λ is analogous to methods for quantifying heterogeneity in a random effects meta-analysis. The parameter τ^2 is on the scale of the individual θ_i. However, meta-analysts more often make judgments about between-study heterogeneity on the scale of between-to-within study variance τ^2/v, where v is some "typical" estimation error variance of the studies observed. Common meta-analytic statistics, such as I^2 or H^2, can be seen as depending on the scale of τ^2/v (see Higgins & Thompson, 2002).

Taken together, regardless of whether the studies are treated as fixed or random, a definition of replication that focuses on consistency of effects can be described as the variation between effect parameters. This variation can be computed on the scale of the θ_i, as with τ_F^2 and τ^2. Alternatively, it can be quantified on the scale of between- versus-within study variance τ^2/v (for random effects models) or λ (for fixed effects models). We note that while parameters like τ_F^2 and τ^2 refer to different quantities, in practice their scales can be interpreted in largely the same way (for discussion, see Hedges & Schauer, 2019b).

Evaluating the Amount of Heterogeneity Among "Consistent" Effects

Specifying an amount of heterogeneity among effects that corresponds with replication success or failure requires we set specific values of τ^2, τ^2/v, τ_F^2, or λ. Moreover, the following sections show that properties of analyses for replication will often depend on these values. Thus, to understand definitions of and analyses for replication, we need to understand the different scales of heterogeneity described above. What is a small or negligible value of λ or τ^2/v? What is a large value of τ^2 or τ^2/v?

The answer to such questions will be subject to scientific and clinical judgment. However, most researchers are used to intuiting the scale of individual effects, rather than variation across effects. In this section, we provide some insight into approaches for quantifying heterogeneity, as well as some conventions for negligible heterogeneity used in various scientific fields.

Hedges and Schauer (2019b) provide several ways to interpret τ^2 or τ_F^2 as a function of differences between pairs of effect parameters $\theta_i - \theta_j$ for $i \neq j$. Since $2\tau^2$ and $2\tau_F^2$ are equal to the mean pairwise squared difference between effects $E[(\theta_i - \theta_j)^2]$, it may be easier to describe replication or replication in terms of a meaningful value of $\theta_i - \theta_j$ and back out a value of τ_F^2 or τ^2. As an example, if the θ_i are standardized

mean differences, so that we consider a difference of $|\theta_i - \theta_j| > 0.2$ to be non-negligible, this suggests that $\tau^2 < 0.02$ could be seen as negligible. Alternatively, if the studies are treated as random, we might specify a form of the distribution of the θ_i (e.g., normal) and identify a value of τ^2 that renders large pairwise differences unlikely:

$$\mathrm{P}\left[\left|\theta_i - \theta_j\right| < \varepsilon\right] < \gamma$$

That is, the probability of a large difference between two effect parameters occurs with probability less than some desired level γ.

Using these approaches, Schauer (2018) argues that values of $\tau^2 \leq 0.035$ would result in effect parameters on the scale of Cohen's d that could be characterized as roughly the same size; it would characterize a distribution of parameters such that deviations from the mean of that distribution greater than $d = 0.2$ occur with probability less than 20%. See Hedges and Schauer (2019a,b) and Schauer (2018) for further detail.

As a matter of sound statistical practice, analyses for replication should focus on the parameters τ^2 or $\tau_F{}^2$ rather than on τ^2/v or λ. Since τ^2/v and λ depend on the within-study variance v_i, and hence the sample size within studies n_i, they are not (strictly speaking) parameters. However, the scale of τ^2/v has been easier to work with, with traditional metrics of between-study variation in meta-analysis depending on that scale. For illustration, suppose that the k studies have roughly the same sample size so that $v_1 \approx \ldots v_k \approx v$. Meta-analytic metrics typically used to quantify heterogeneity, such as I^2 or H^2, depend onτ^2/v (Higgins & Thompson, 2002). If the v_i are not similar in value, Higgins and Thompson (2002) provide an expression for the "typical" value v of the v_i (see Eq. 14.9 in their article), and variation can be described on the same τ^2/v scale. When studies are treated as fixed, the parameter λ can be thought of in the same terms, as seen in Eq. (14.16). Whether we treat the studies as fixed or random, a potentially useful scale for heterogeneity is τ^2/v, so long as it is clear what a typical or normative value of v is.

Since τ^2/v is a common scale in meta-analysis, it can be easier to work with. Several different fields that conduct meta-analyses have generated conventions for negligible heterogeneity on that scale. Such conventions may be of use when approaching design and analysis of replication studies. Hedges and Schauer (2019a,b) note that in high-energy physics, the Particle Data Group characterizes minor or unimportant variation in ways that suggest $\tau^2/v < 1/4$ would be seen as negligible (see Olive, et al., 2014). In personnel psychology, Hunter and Schmidt (1990) describe and propose $v/(\tau^2 + v) < 0.75$ as negligible, which means $\tau^2/v < 1/3$. In medicine, a value of $I^2 = 100\% \times \tau^2/(v + \tau^2) < 40\%$ or $\tau^2/v < 2/3$ is seen as "not important" (see Higgins & Green, 2008). These are far from the *only* conventions of negligible heterogeneity (see Pigott, 2012), but they reflect ideas about heterogeneity that guide and inform research in these fields.

Since $\tau^2/v \in [1/4, 2/3]$ could be seen as a range of negligible between-study variation, researchers would likely want to detect differences between studies at

least as large as these values; possibly two or three times larger. This would suggest that meaningful values of variation worth studying might range from $\tau^2/v = 1/4$ to $\tau^2/v = 3$ ($I^2 = 0.2$–0.75).

The benefit of using the scale of τ^2/v is that it does not depend on the scale of the θ_i. However, if we have a good idea of the scale of the θ_i, it makes more sense to focus on τ^2 or ${\tau_F}^2$. Based on the results above, if θ_i are on the scale of Cohen's d, we may consider values of $\tau^2 \in [0, 0.035]$ to be negligible and may be interested in studying values in the range of [0.005, 0.1] (see Schauer, 2018).

A General Approach for Studying Replication: Magnitude of Effects

In this section, I will outline a framework for studying replication when the preferred definition is agreement in magnitude, and the preferred framing involves the consistency of effects. Though this is far from the only way to define replication, it is consistent with research that seeks to understand the conditions under which effects are stable and is in line with the type of knowledge refinement prioritized in various scientific fields including physics, chemistry, and medicine.

I first present results for fixed studies when $k = 2$, and then assume studies are random for $k \geq 3$. The fixed effects analog of the random effects analyses presented here can be found in Hedges and Schauer (2019b). The key distinction between the properties of the fixed and random effects analyses (beyond their scope of inference) is that the random effects analyses will be slightly less powerful and efficient than the fixed effects studies, though the difference in power is relatively small.

A Note aAbout Publication Bias

The analysis methods and their properties that are presented in this section do not include explicit adjustments for publication bias in published effect estimates. Such adjustments could be included in these methods, which would presumably impact their sensitivity. To understand the extent to which they do, note the following two aspects about publication bias adjustments. First, adjusting effect estimate T_i for bias can result in a corrected effect estimate T_i^* with variance v_i^*, where $v_i^* \geq v_i$ (i.e., corrected estimates tend to have greater variance). Second, corrections such as Hedges' (1984) maximum likelihood approach can result in adjusted effect estimates that are asymptotically normally distributed. Because the statistical results that follow depend on the normality of effect estimates, analyses can proceed with T_i and v_i if publication bias is not suspected, and T_i^* and v_i^* if publication bias is likely. Subsequent sections will detail that the sensitivity (statistical power or standard errors) will be worse when the v_i are larger. As a result, one impact of

publication bias adjustments is that they reduce the power of tests for replication or increase standard errors or estimates for relevant quantities.

Fixed Effects for Two Studies

When there are $k = 2$ studies, the focus of analysis is on θ_1 and θ_2 being (about) the same value. Thus, analyses of replication can be viewed in terms of the difference $\theta_1 - \theta_2$. When $\theta_1 - \theta_2$ is large in magnitude, then there is a large difference between study results, but when $\theta_1 - \theta_2$ is small, then study results are more similar. We can directly estimate $\theta_1 - \theta_2$ with $T_1 - T_2$. Under the model it is the estimator with the smallest variance, and that variance is simply $v_1 + v_2$. When effect sizes are on the scale of standardized mean differences, the variance of the estimated difference can be written as approximately $4/n_1 + 4/n_2$, where n_i is the total sample size of study i; the standard error can be written as approximately $2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Example. Recall the Payne et al. study replicated by the RPP. The original study estimated an effect of $T_1 = 0.75$ (Cohen's d) with a variance $v_1 = 0.066$. The RPP replication study found an estimated effect of $T_2 = 0.30$ with variance of $v_2 = 0.23$. The estimated difference between effects is $T_1 - T_2 = 0.45$, which has a standard error of $\sqrt{v_1 + v_2} = 0.30$. A 95% confidence interval for the difference in effects is [–0.14, 1.04].

Alternatively, there are two different ways we can test null hypotheses about replication for $k = 2$ studies.

Tests When the Burden of Proof is on Nonreplication

If the burden of proof is on nonreplication, then we can structure a null hypothesis

$$\mathrm{H}_0 : \left|\theta_1 - \theta_2\right| \leq \varepsilon \tag{14.18}$$

for some $\varepsilon \geq 0$. Here, ε corresponds to the largest difference between effect parameters that would be considered negligible. When $\varepsilon = 0$, H_0 corresponds to exact replication, but when $\varepsilon > 0$ (but is still small), H_0 corresponds to approximate replication. To test H_0 in Eq. (14.18), we compute the test statistic

$$Q_2 = \frac{\left(T_1 - T_2\right)^2}{v_1 + v_2} \tag{14.19}$$

Note that Q_2 is equivalent to the Q statistic in Eq. (14.6) for $k = 2$ studies. Under H_0, Q_2 follows a chi-square distribution with one degree of freedom and noncentrality parameter

$$\lambda_{02} = \frac{\varepsilon^2}{v_1 + v_2} \tag{14.20}$$

An α-level test would involve rejecting H_0 when $Q_2 > c_{1-\alpha}(1, \lambda_{02})$ where $c_{1-\alpha}(\nu, \lambda)$ is the $1 - \alpha$ percentile of the chi-square distribution with ν degrees of freedom and noncentrality parameter λ. Note that when $\varepsilon = 0$, so that the test concerns exact replication, $\lambda_{02} = 0$ and Q_2 follows a central chi-square distribution under H_0.

The power of this test to detect a difference $|\theta_1 - \theta_2| > \varepsilon$ is given by

$$1 - F\left(c_{1-\alpha}\left(1, \lambda_{02}\right) | , 1 | , \lambda_{12}\right) \tag{14.21}$$

where $c_{1-\alpha}(1, \lambda_{02})$ is defined as above, and $F(x\,|\nu, \lambda)$ is the distribution function of a chi-square random variable with ν degrees of freedom and noncentrality parameter λ. The power depends on a number of quantities:

- It is decreasing in ε, so that tests of exact replication will be more powerful than tests of approximate replication. Tests of increasingly looser notions of approximate replications with larger ε (and hence larger differences between effects that are seen as negligible) will be less powerful.
- It is increasing as a function of the true difference between effects $|\theta_1 - \theta_2|$. If $|\theta_1 - \theta_2|$ is larger, then tests will have more power.
- It is increasing as a function of the variance of each effect estimate $v_1 + v_2$. If v_1 and v_2 are smaller (so that sample sizes in each study are larger), the test will have greater power.
- In practice, the power of this test for $k = 2$ studies is bound to be low unless both studies have uncommonly large sample size. Assuming effects on the scale of Cohen's d and an $\alpha = 0.05$ level test for exact replication, to detect a difference $|\theta_1 - \theta_2| = 0.2$ (Cohen's d) with 80% power would imply $v_1 + v_2 \leq 0.0013$, which is consistent with both studies having sample sizes of at least 1569 given Eq. (14.2). Detecting a difference of $|\theta_1 - \theta_2| = 0.5$ with 80% power would require $v_1 + v_2 \leq 0.008$, which is consistent with both studies having sample sizes of at least 251. Detecting a difference of $|\theta_1 - \theta_2| = 0.5$ with 80% power would require $v_1 + v_2 \leq 0.02$, which is consistent with both studies having sample sizes of at least 98.

It is worth noting that when $\varepsilon = 0$, so that the test concerns exact replication, then this procedure is statistically equivalent to the PI criterion. Because of this, the PI criterion can be seen as a test of the null hypothesis that the studies replicated exactly. In that case, the false failure rate is simply α, and is controlled.

Example. Consider the Payne et al.'s memory studies referenced above ($T_1 = 0.753$, $v_1 = 0.0662$, $T_2 = 0.304$, $v_2 = 0.0229$, converted to Cohen's d). A test for

exact replication ($\varepsilon = 0$) would fail to reject the null hypothesis that the studies replicated. The power of this test to detect a difference as large as $|\theta_1 - \theta_2| = 0.5$ is 38%. If instead we consider a difference of $|\theta_1 - \theta_2| < 0.1$ to be negligible, then we might test for approximate replication ($\varepsilon = 0.1$). In doing so we would again fail to reject the null hypothesis that the studies replicated. The power of this test to detect a difference as large as $|\theta_1 - \theta_2| = 0.5$ is 35%. Note that the test for approximate replication is less powerful than the test of exact replication.

Tests When the Burden of Proof Is on Replication

If the burden of proof is on replication, then the null hypothesis can be formed to correspond to nonreplication. As discussed previously in this chapter, forming a testable null hypothesis of replication can be done via approaches used in equivalence testing. Concretely, let ε denote the smallest difference between effects that would be considered non-negligible. Then we form a null hypothesis

$$\mathrm{H}_0 : \left|\theta_1 - \theta_2\right| \geq \varepsilon \tag{14.22}$$

To test H_0, we compute Q_2. Under the null hypothesis, Q_2 follows a chi-square distribution with one degree of freedom and noncentrality parameter λ_{02} in Eq. (14.20). Since we need conclusive evidence that the studies successfully replicate, an α-level test involves rejecting H_0 if Q_2 is less than $c_\alpha(1, \lambda_{02})$, the α-percentile of the chi-square distribution with one degree of freedom and noncentrality parameter λ_{02}.

The power of this test is given by

$$F\left(c_\alpha\left(1,\lambda_{02}\right)\mathrm{l},1\,\mathrm{l},\lambda_{12}\right) \tag{14.23}$$

where F is as in Eq. (14.21) and $\lambda_{12} = |\theta_1 - \theta_2|^2/(v_1 + v_2) \leq \lambda_{02}$. Note that if the studies replicate exactly, then $\lambda_{12} = 0$. However, if the studies replicate approximately, so that $|\theta_1 - \theta_2| < \varepsilon$, then $0 < \lambda_{12} < \lambda_{02}$.

The power of this test depends on a few quantities:

- It is increasing as function of ε. The bigger the difference between effects that is considered negligible, the greater the power of the test.
- It is increasing as a function of $v_1 + v_2$, so that when the variances for each study decrease (sample sizes within studies increase), the power of the test for replication will increase.
- It is decreasing as a function of $|\theta_1 - \theta_2|$. The smaller the actual difference between effects is, the greater the power. In fact, the power is greatest when the studies replicate exactly, so that $\theta_1 = \theta_2$, and $\lambda_{12} = 0$.
- In practice, unless we consider extremely large differences between effects to be negligible, the test when the burden of proof is on replication is lower than the power of the test when the burden of proof is on nonreplication. For example,

consider effects on the scale of Cohen's d and a design such that $v_1 + v_2 = 0.02$ (i.e., sample size of 98 per study). The power of the test of exact replication when the burden of proof is on nonreplication (H_0: $\theta_1 = \theta_2$) has 80% power to detect a difference of $|\theta_1 - \theta_2| = 0.8$. However, a test for nonreplication assuming $|\theta_1 - \theta_2|$ is at least as large as 0.8 (H_0: $|\theta_1 - \theta_2| \geq 0.8$) has maximum power of 75%, which occurs when the studies replicate exactly (i.e., when $\theta_1 = \theta_2$).

- Note that neither of these tests will be particularly well powered with $k = 2$ studies, as argued in the following section.

Example. Suppose we deem $\varepsilon = 0.2$ to be the smallest non-negligible difference between effects. Then with the replication of Payne et al.'s memory study, we fail to reject the null hypothesis that the studies failed to replicate. The power of this test will be greatest if $\theta_1 = \theta_2$, so that the studies replicate exactly. In that case, the power would be 30%.

More than One Study Is Likely Necessary for Conclusive Results About Replication

Two key questions about the design of replication studies involve how many replications should be conducted and how large the sample size should be for each study in order to ensure sufficiently powerful analyses. If the design a priori sets $k = 2$, as has been common in some social science research, and if the original study has already been conducted and published, then the question of design involves how large the sample size ought to be in the replication study in order to ensure sufficiently powerful analyses.

Hedges and Schauer (2019a) showed that a design with $k = 2$ studies where the original study had already been conducted will almost never support sufficiently sensitive analyses, and in fact the power of tests for replication will typically be bounded by the power of the original study to detect an effect. To see this, note that the power of the original study (study 1) to detect an effect as large as θ_1 is given by

$$1 - F\left(c_{1-\alpha}\left(1,0\right)|\theta_1^2 / v_1\right) \tag{14.24}$$

where F and $c_{1-\alpha}$ are given in Eq. (14.21). The power of this test depends largely on θ_1^2/v_1.

Now consider the test for replication when the burden of proof is on nonreplication. This test has power given in Eq. (14.21). Note that the power for both the test for an effect in study 1, Eq. (14.24), and the test for replication in Eq. (14.21) depend on the chi-square distribution function with one degree of freedom, and that both are decreasing as the critical value $c_{1-\alpha}$ increases. Further, $c_{1-\alpha}(1, \lambda_{02}) \geq c_{1-\alpha}(1, 0)$ with equality holding only if $\lambda_{02} = 0$, so that the test involves exact replication.

The test for replication is an increasing function in $|\theta_1 - \theta_2|$. Yet, an upper bound on differences between effects we might want to detect likely occurs when

$|\theta_1 - \theta_2| \leq \theta_1$. The reasoning behind this is that if $|\theta_1 - \theta_2| = \theta_1$, this would involve a scenario where θ_1 is in one direction (e.g., $\theta_1 > 0$) and θ_2 is in another direction ($\theta_2 \leq 0$), which would constitute qualitative disagreement in effects (e.g., the effect in study 1 helps patients, the effect in study 2 does nothing for them) and run contrary to agreement of magnitude or direction of effects. Thus, the difference we may wish to detect in a test of replication is bounded above by $|\theta_1|$.

Finally, the power of the test for replication increases as v_2 decreases. The smallest v_2 could possibly be is 0, which would occur if study 2 had an infinite sample size. Putting these pieces together, it follows that the power of study 1 to detect an effect is an upper bound for the test of replication:

$$1 - F\left(c_{1-\alpha}(1,0), \frac{\theta_1^2}{v_1}\right) \geq 1 - F\left(c_{1-\alpha}(1,0), \frac{|\theta_1 - \theta_2|^2}{v_1}\right) \geq$$

$$1 - F\left(c_{1-\alpha}(1,0), \frac{|\theta_1 - \theta_2|^2}{v_1 + v_2}\right) \geq 1 - F\left(c_{1-\alpha}(1,\lambda_{02}), \frac{|\theta_1 - \theta_2|^2}{v_1 + v_2}\right) \quad (14.25)$$

The power of the test for replication will likely be much smaller than the power of study 1, since:

(a) We may wish to detect differences between effects that are themselves smaller than θ_1.
(b) Study 2 will not have an infinite sample size, and so $v_2 > 0$.
(c) We may have to adjust T_1 for publication bias, which will decrease the power of the test for replication.

Further, Hedges and Schauer (2019a) show that similar inequalities hold for tests when the burden of proof is on replication, for parameter estimation, and for Bayesian parameter estimation.

In practice, unless both studies have very high power (>99% power to detect effects θ_1 and θ_2, respectively), the power of the test for replication will be low. This result holds even if we conduct multiple replication studies and aggregate their results via a meta-analysis. Hence, Hedges and Schauer (2019a) argue that analyses framed in terms of falsifiability are likely to be underpowered, and that analyses framed in terms of consistency require more than one replication to ensure high power.

In the absence of conclusive designs and analyses about replication based on $k = 2$ studies, a series of methods have been proposed to better make sense of the evidentiary value of $k = 2$ studies regarding replicability. Maxwell, Lau, and Howard (2015) propose using an equivalence test to analyze a replication study when the original study finds a statistically significant effect. The idea behind this is if the original effect estimate is statistically different from zero, then one way to falsify that is if the effect estimate in the replication study is conclusively close to zero. Held's (2020) skeptical p-value is based on a Bayesian approach to prior skepticism

about the original effect estimate and evaluates whether that skepticism is consistent with the replication study effect estimate. Simonsohn's (2015) small telescopes approach involves estimating the statistical power of the original study relative to the effect found in the replication study. All three approaches allow for prospective sample size calculations to ensure that their conclusions are reasonably precise. However, none of these methods concerns definitions of replication focused on similarity of effect size: the equivalence test focuses on how negligibly small the effect is in the replication study, the skeptical *p*-value concerns the prior beliefs required to doubt the original study results, and small telescopes largely focuses on the sensitivity of the original and replication studies. Because of this, none support inferences about the agreement of effects explicitly, but rather give insight into the evidentiary value of $k = 2$ studies regarding the existence of effect. Though they support conclusions about more diffuse notions of replication, either approach may prove useful when replication research designs cannot include more than two studies (e.g., due to practical or resource constraints).

Random Effects Analyses for Replication ($k > 2$)

Estimation

If the framing of analysis is on consistency of effects, and studies are assumed to be random, then the relevant parameter to estimate is τ^2. There are a variety of possible estimators of τ^2 (see Veroniki et al., 2016, for a review). A common estimator due to DerSimonian and Laird (1986) is given in Eq. (14.9).

The standard error of this estimator is:

$$SE\left[\hat{\tau}^2_{DL}\right] = \sqrt{\frac{2(k-1)}{S^2} + \frac{4\tau^2}{S} + \frac{2\tau^4}{S^2}\left[S_2 - 2\frac{S_3}{S_1} + \frac{S_2^2}{S_1^2}\right]} \tag{14.26}$$

where S_j is defined in Eq. (14.7) and S in Eq. (14.8). To estimate this standard error, we can substitute the estimated variance component $\hat{\tau}^2_{DL}$ for τ^2 in the equation above. Note that the standard error will decrease as the v_i decrease (i.e., with large sample sizes within studies) and as k increases, but will increase as the amount of variation between studies τ^2 increases.

Statistical methodologists have argued that an alternative estimator due to Paule and Mandel (1982) tends to have slightly better properties in certain scenarios (see Veroniki et al., 2016; van Aert & Jackson, 2018). The Paule–Mandel estimator is based on the statistic $Q^*(\tau^2)$:

$$Q^*\left(\tau^2\right) = \sum_{i=1}^{k} \frac{\left(T_i - T_\cdot^*\right)^2}{v_i + \tau^2} \tag{14.27}$$

where

$$T_{\cdot}^{*} = \frac{\sum_{i=1}^{k} \frac{T_i}{v_i + \tau^2}}{\sum_{i=1}^{k} \frac{1}{v_i + \tau^2}} \tag{14.28}$$

The statistic $Q^{*}(\tau^2)$ is written this way because it is a function of τ^2. Note that $Q^{*}(\tau^2)$ and $T_{\cdot}^{*}$ differ from Q and $\bar{T}_{\cdot}$ in that they involve sums weighted by $1/(v_i + \tau^2)$ as opposed to $1/v_i$. Moreover, $T_{\cdot}^{*}$ differs from $\bar{T}_{\cdot}^{*}$ in Eq. (14.10) in that $T_{\cdot}^{*}$ uses weights that depend on the true value of τ^2, while $\bar{T}_{\cdot}^{*}$ uses weights that depend on an estimate $\hat{\tau}_{DL}^2$.

It can be shown that the expected value of $Q^{*}(\tau^2)$ is $k - 1$. The Paule–Mandel estimator is thus obtained by using an iterative program to solve the equation $Q^{*}(\tau^2) = k - 1$ for τ^2:

$$\hat{\tau}_{PM}^2 = \tau^2 : \sum_{i=1}^{k} \frac{\left(T_i - T_{\cdot}^{*}\right)^2}{v_i + \tau^2} = k - 1 \tag{14.29}$$

The Paule–Mandel estimator can be used in conjunction with a method for constructing confidence intervals for τ^2 called the Q-profile method (Viechtbauer, 2007). If the θ_i are normally distributed, then $Q^{*}(\tau^2)$ follows a chi-square distribution with $k - 1$ degrees of freedom. A $1 - \alpha$ confidence interval for τ^2 can be obtained by using an iterative program to solve two equations for τ^2: one equation is used to obtain the lower bound ($L_{1-\alpha}$) and one to obtain the upper bound ($U_{1-\alpha}$) of the confidence interval. These equations set $Q^{*}(\tau^2)$ equal to $c_{1-\alpha/2}(k - 1, 0)$, the $1 - \alpha/2$ percentile of the chi-square distribution with $k - 1$ degrees of freedom, and $c_{\alpha/2}(k - 1, 0)$, the $\alpha/2$ percentile, respectively:

$$L_{1-\alpha} = \tau^2 : \sum_{i=1}^{k} \frac{\left(T_i - \bar{T}_{\cdot}^{*}\right)^2}{v_i + \tau^2} = c_{1-\frac{\alpha}{2}}\left(k - 1, 0\right) \tag{14.30}$$

$$U_{1-\alpha} = \tau^2 : \sum_{i=1}^{k} \frac{\left(T_i - \bar{T}_{\cdot}^{*}\right)^2}{v_i + \tau^2} = c_{\frac{\alpha}{2}}\left(k - 1, 0\right) \tag{14.31}$$

In addition to reporting point estimates and uncertainty on the scale of τ^2, researchers may also report statistics such as the H^2 or I^2 values. Both of these statistics can be seen as depending on the ratio of τ^2/v. The statistic H^2 is an estimate of $1 + \tau^2/v$, while I^2 is an estimate of $\tau^2/(v + \tau^2)$ (Higgins & Thompson, 2002). Note that the precise value of H^2 and I^2 depends on an estimated variance τ^2, and hence will possibly differ between the DerSimonian–Laird and Paule–Mandel estimators.

Example. Consider the reverse gambler's fallacy experiment replicated by the Many Labs project described in previous sections. The effect sizes (on the scale of Cohen's *d*) and their variances from these replications are reported in Table 14.3. Based on the $k = 36$ replication studies, the DerSimonian and Laird estimator is $\hat{\tau}^2_{DL} = 0.013$, which has standard error 0.010 ($H^2 = 1.46$, $I^2 = 31.73\%$). The Paule–Mandel estimator is $\hat{\tau}^2_{PM} = 0.018$ ($H^2 = 1.66$, $I^2 = 39.91\%$), with 95% confidence interval [0.000, 0.060]. Thus, there is some evidence of variation between studies that could be considered modest to moderate (τ^2/v ranging from 0.46 to 0.66, depending on the estimator). The uncertainty in this estimate is such that the variation between studies could possibly be very near zero or as large as 0.06 ($\tau^2/v = 2.28$).

NHT: Burden of Proof on Nonreplication

A test of consistency when the burden of proof is on nonreplication would form a null hypothesis that the studies replicate successfully. Since the variance component τ^2 is the parameter that characterizes "replication," a relevant null hypothesis is

$$\mathrm{H}_0 : \tau^2 \leq \tau_0^{\,2} \tag{14.32}$$

where τ_0^2 constitutes the smallest amount of variation between studies considered non-negligible that would characterize replication failure. Note that if $\tau_0^2 = 0$, this is a test of exact replication, but when $\tau_0^2 > 0$, this is a test of approximate replication.

To test H_0, we can compute the Q statistic in Eq. (14.6), which follows a somewhat complex distribution that can be expressed as a linear combination of chi-square random variables. A reasonable approximation for that distribution is as follows. Denote the following moments of Q that are functions of τ^2

$$\mu_Q\left(\tau^2\right) = k - 1 + S\tau^2 \tag{14.33}$$

is the mean of Q where S is as in Eq. (14.8) and

$$\sigma_Q^{\,2}\left(\tau^2\right) = S^2\left(SE\left[\tau_{DL}^{\,2}\right]\right)^2 \tag{14.34}$$

is the variance of Q where $\mathrm{SE}[\tau_{\mathrm{DL}}^2]$ is as in Eq. (14.26) and S is in Eq. (14.9) (for further details, see Hedges & Pigott, 2001).

Given these functions, $2\mu_Q(\tau^2)\ Q/\sigma_Q^2(\tau^2)$ follows a chi-square distribution with $2\mu_Q^2(\tau^2)/\sigma_Q^2(\tau^2)$ degrees of freedom. Thus, under H_0, we can use the approximation that $2\mu_Q(\tau_0^2)\ Q/\sigma_Q^2(\tau_0^2)$ follows a chi-square distribution with $2\mu_Q^2(\tau_0^2)/\sigma_Q^2(\tau_0^2)$ degrees of freedom. When all studies have the same estimation error variance $v_i = v$ (i.e., all studies have the same sample size), then this approximation reduces to a

much simpler expression that can be written as a constant $(1 + \tau^2/v)$ times a chi-square distribution with $k - 1$ degrees of freedom:

$$Q \sim \left(1+\tau 2 / v\right) \chi^2_{k-1} \tag{14.35}$$

Moreover, when $\tau^2 = 0$, as in a null hypothesis of exact replication, then $\mu_Q(0) = k - 1$ and $\sigma_Q{}^2(0) = 2(k - 1)$ and Q follows a central chi-square distribution with $k - 1$ degrees of freedom.

An α-level test involves rejecting H_0 if $2\mu_Q(\tau_0{}^2)\, Q/\sigma_Q{}^2(\tau_0{}^2)$ exceeds $c_{1-\alpha}(2\mu_Q{}^2(\tau_0{}^2)/\sigma_Q{}^2(\tau_0{}^2), 0)$, the $1 - \alpha$ percentile of the chi-square distribution with $2\mu_Q{}^2(\tau_0{}^2)/\sigma_Q{}^2(\tau_0{}^2)$ degrees of freedom. For brevity, we will write

$$C\left(1-\alpha,\tau_0^2\right)=\frac{c_{1-\alpha}\left(\dfrac{2\mu_Q^2\left(\tau_0^2\right)}{\sigma_Q^2\left(\tau_0^2\right)},0\right)}{\dfrac{2\mu_Q\left(\tau_0^2\right)}{\sigma_Q^2\left(\tau_0^2\right)}} \tag{14.36}$$

to refer to this percentile/critical value, such that we reject H_0 when Q exceeds $C(1 - \alpha, \tau_0{}^2)$. The notation $C(1 - \alpha, \tau_0{}^2)$ denotes that it is a function of α and $\tau_0{}^2$. When all of the v_i are equal, the test reduces to rejecting H_0 when Q exceeds $C(1 - \alpha, \tau_0{}^2) = c_{1-\alpha}(k - 1, 0)(1 + \tau_0{}^2/v)$; that is, when Q exceeds a critical value from the central chi-squared distribution multiplied by $1 + \tau_0{}^2/v$. In tests of exact replication, $\tau_0{}^2 = 0$, so the critical value is simply the $1 - \alpha$ percentile of the chi-squared distribution with $k - 1$ degrees of freedom (akin to a traditional Q-test in meta-analysis).

The power of this test to detect some value $\tau^2 > \tau_0{}^2$ is given by

$$1-F\left(\frac{2\mu_Q\left(\tau^2\right)C\left(1-\alpha,\tau_0^2\right)}{\sigma_Q^2\left(\tau^2\right)}\Big|\frac{2\left[\mu_Q\left(\tau^2\right)\right]^2}{\sigma_Q^2\left(\tau^2\right)},0\right) \tag{14.37}$$

where F is the chi-square distribution function in Eq. (14.21); since the noncentrality parameter in this function is set to 0, this is a central chi-square distribution function. When all of the $v_i = v$, so that each study has the same estimation error variance, then the power reduces to:

$$1-F\left(\frac{C\left(1-\alpha,\tau_0^2\right)}{\left(1+\tau^2/v\right)}\Big|k-1,0\right) \tag{14.38}$$

The power of the test for replication is increasing as a function of the number of studies k, as well as in τ^2 in the metric of τ^2/v. Because of this, the power is higher

when τ^2/v is larger, which occurs when τ^2 is larger or when v is smaller (i.e., sample sizes within studies are larger). In addition, power is a decreasing function of τ_0^2, which means that tests of approximate replication have lower power than tests of exact replication. Power is discussed further at the end of this section.

Example. The reverse gambler's fallacy example comprises the effect sizes of the $k = 36$ replication studies. Based on the effect estimates and variances in Table 14.3, the Q statistic is 51.27. For an $\alpha = 0.05$ level test of exact replication (H_0: $\tau^2 = 0$), the relevant critical value is $c_{1-\alpha}(35, 0) = 49.8$. Because $Q > 49.8$, we reject H_0 and conclude the studies fail to replicate exactly ($p = 0.04$). This test had 85% power to detect variation on the order of $\tau^2 = 0.027$, or $\tau^2/v = 1$.

Consider a test of approximate replication such that we consider $\tau_0^2 = 0.01$ to be the largest amount of variation considered negligible. Note this would be consistent with roughly $\tau_0^2/v \approx 1/3$, which is roughly the convention specified by Hunter and Schmidt. The relevant critical value is $C(0.95, 0.01) = 69.17$. Since $Q < 61.17$, we do not reject H_0 and so do not conclude the studies failed to replicate approximately ($p = 0.35$). This test had 48% power to detect variation on the order of $\tau^2 = 0.027$, or $\tau^2/v = 1$.

NHT: Burden of Proof on Replication

When the burden of proof is on replication, then, as in the $k = 2$ case, we must form a null hypothesis that the studies fail to replicate. With our focus on between-study variation, our test will involve the following null hypothesis:

$$H_0 : \tau^2 \geq {\tau_0}^2 \tag{14.39}$$

This can be tested with the Q statistic in Eq. (14.6). Under H_0, we can use the approximation that $2\mu_Q(\tau_0^2)\ Q/\sigma_Q^2(\tau_0^2)$ follows a chi-square distribution with $2\mu_Q^2(\tau_0^2)/\sigma_Q^2(\tau_0^2)$ degrees of freedom, as derived above. An α-level test involves rejecting H_0 if $2\mu_Q(\tau_0^2)\ Q/\sigma_Q^2(\tau_0^2)$ is less than the α-percentile of that distribution $C(\alpha, \tau_0^2)$, where $C(\alpha, \tau_0^2)$ is described in Eq. (14.36). In other words, the test when the burden of proof is on replication proceeds in a similar manner as when the burden of proof is on nonreplication, except that: (1) the critical value now involves the α-percentile of the chi-square approximation, and (2) we reject H_0 if $2\mu_Q(\tau_0^2)\ Q/\sigma_Q^2(\tau_0^2)$ is less than that critical value.

The power of this test to detect some value $\tau^2 < \tau_0^2$ is given by

$$F\left(\frac{2\mu_Q\left(\tau^2\right)C\left(\alpha,\tau_0^2\right)}{\sigma_Q^2\left(\tau^2\right)}\middle|\frac{2\left[\mu_Q\left(\tau^2\right)\right]^2}{\sigma_Q^2\left(\tau^2\right)},0\right) \tag{14.40}$$

Table 14.3 Data from the Many Labs Replication Project replications of the reverse gambler's fallacy experiment

Site	Effect size	Variance
abington	0.590	0.051
brasilia	0.355	0.036
charles	0.886	0.063
conncoll	0.622	0.050
csun	0.517	0.046
help	0.516	0.043
ithaca	0.782	0.053
jmu	0.715	0.026
ku	0.527	0.039
laurier	0.961	0.042
lse	0.645	0.016
luc	0.528	0.029
mcdaniel	0.510	0.046
msvu	0.340	0.054
mturk	0.620	0.004
osu	0.111	0.038
oxy	1.188	0.048
pi	0.724	0.004
psu	0.605	0.048
qccuny	0.419	0.044
qccuny2	0.338	0.050
sdsu	0.616	0.026
swps	0.114	0.050
swpson	0.593	0.027
tamu	0.747	0.024
tamuc	0.749	0.054
tamuon	0.592	0.020
tilburg	0.687	0.059
ufl	0.378	0.034
unipd	0.765	0.035
uva	1.108	0.059
vcu	0.712	0.040
wisc	0.785	0.045
wku	0.441	0.044
wl	0.072	0.046
wpi	0.978	0.053

Source: Open Science Framework

where F is the chi-square distribution function in Eq. (14.21). When all of the $v_i = v$, so that each study has the same estimation error variance, then the power reduces to:

$$F\left(\frac{C\left(\alpha,\tau_0^2\right)}{1+\tau^2/v}\,|\,k-1,0\right) \tag{14.41}$$

The power of this test will increase as we test looser notions of replication failure; that is, when τ_0^2 is larger. It will increase when τ^2/v is smaller, and will be greatest when $\tau^2 = 0$, so that the studies replicate exactly. As discussed in the following section, the power of the test when the burden of proof is on replication will typically be lower than when the test puts the burden of proof on nonreplication.

Example. Suppose we wish to test a null hypothesis that the gambler's fallacy replications failed such that $\tau^2 = 0.027$, which would characterize $\tau_0^2/v = 1$. Recall that the Q statistic is 51.27. The relevant critical value is $C(0.95, 0.027) = 41.13$. Because $Q > 41.13$, we fail to reject H_0 and do not conclude the studies replicated successfully ($p = 0.17$). The power of this test will be greatest when $\tau^2 = 0$, and the studies actually replicate exactly, in which case the power would be 78%.

Power and Precision of Analyses

Note that the conclusions of the test for replication can depend on how the null hypothesis is formed. In the gambler's fallacy example, we rejected a null hypothesis of exact replication, but failed to reject a null hypothesis of approximate replication, nor did we reject a null hypothesis of replication failure. Thus, the ultimate conclusions reached about replication will be sensitive to the framing of the null hypothesis.

The sensitivity of analytic methods is key for both planning and analyzing replication studies. Proper interpretation of these tests, however, does not mean we conclude H_0 is true when we fail to reject it. In general, a failure to reject a null hypothesis is ambiguous, and must be interpreted in light of statistical power (or the Type II error rate). So too must interpretations of estimated variance be considered in light of their uncertainty. Moreover, because the power of these tests are known, as are the standard errors of estimators, researchers can plan replication studies prospectively to make sure that relevant analyses are sufficiently sensitive (i.e., have high power or precision). Note that the previous sections demonstrated that the sensitivity analysis will depend on v_i (and hence the within-study sample size n_i), as well as the total number of studies k. Thus, prospective planning of replication studies to ensure sensitivity analyses can be seen as choosing k and n_i or some minimum n for each study.

Understanding the sensitivity of analyses requires some idea of values of τ^2 considered large and what values might be considered negligible. For instance, the

power of tests for replication is a function of a value of variation τ_0^2 that distinguishes between negligible and non-negligible variations, as well as the variation one wishes to detect τ^2. The standard error of variance component estimates is a function of τ^2.

In a previous section, I argued that if the scale of the θ_i is understood, it is best practice to derive meaningful values of τ^2 in a manner that is consistent with both the scale of the θ_i, and knowledge regarding the finding under consideration and its scientific and clinical implications. Alternatively, a scale-free approach might conceive of these analyses and their properties as depending on τ^2/v. We noted that when effects are on the scale of standardized mean differences, the negligible values of τ^2 may range from 0 to as large as 0.035, and the values worth studying might range from 0.005 to 0.1.

Figure 14.1 shows the power of replication tests assuming $v_1 = v_2 = 4/100$, analogous to each study having a total sample size of 100 (on the scale of Cohen's d) and balanced two-arm designs. The first panel shows the power of a test (y-axis) for exact replication (H_0: $\tau^2 = 0$) to detect a given value of τ^2 (x-axis) for different numbers of studies $k = 2$, 5, 10, 20, and 40. The second panel shows the power of the test for nonreplication (H_0: $\tau^2 \geq \tau_0^2$) to detect exact replication ($\tau^2 = 0$) as a function of τ_0^2 (x-axis) for different numbers of k. In addition, the third panel shows the relative standard error (RSE) SE[$\hat{\tau}^2_{DL}$]/τ^2 for the DerSimonian–Laird estimator as a function of τ^2 (x-axis) for different numbers of k. The figure shows that unless there are a large number of studies, a design with $v = 4/100$ will likely be underpowered when analyzing moderate levels of heterogeneity.

To get a better sense of designs with a given number of studies k and sample size per study n (assuming balanced designs within studies) that can give sufficiently sensitive analyses, consider Table 14.4. Table 14.4 shows the required sample size per study necessary to obtain 80% power or a relative standard error less than 50%. The sample sizes presented assume that each study has at least n participants, each study is a balanced two-armed study, results are the scale of standardized mean differences, and the large sample approximation for $v_i = v \approx 4/n$ in Eq. (14.2) is valid.

The first panel in Table 14.4 gives the requisite sample size per study to ensure a test of exact replication (H_0: $\tau^2 = 0$) has 80% power for an $\alpha = 0.05$ level test.

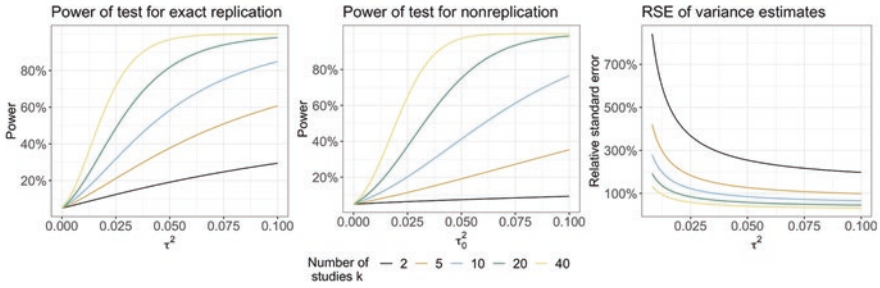

Fig. 14.1 Sensitivity of analyses for replication assuming $v_1 = v_2 = 4/100$

Table 14.4 Required within-study sample sizes to ensure 80% power in tests of exact replication, nonreplication, and estimates with a relative standard error (RSE) less than 50%

	Test of exact replication H_0: $\tau^2 = 0$ (80% power)				Test of nonreplication H_0: $\tau^2 \geq \tau_0^2$ (with $\tau^2 = 0$, 80% power)				Estimation of τ^2 (RSE $\leq$ 50%)			
	$\tau^2 = 0.03$	$\tau^2 = 0.05$	$\tau^2 = 0.07$	$\tau^2 = 0.1$	$\tau_0^2 = 0.03$	$\tau_0^2 = 0.05$	$\tau_0^2 = 0.07$	$\tau_0^2 = 0.1$	$\tau^2 = 0.03$	$\tau^2 = 0.05$	$\tau^2 = 0.07$	$\tau^2 = 0.1$
$k = 5$	635	381	272	191	991	595	425	298	—	—	—	—
$k = 10$	287	172	123	86	359	216	154	108	2199	1319	943	660
$k = 20$	161	97	69	49	183	110	79	55	247	148	106	74
$k = 40$	99	59	43	30	109	65	47	33	111	67	48	34

Note: — means no sample size is possible

Within-study sample sizes are reported for detecting various values of $\tau^2 > 0$ assuming a given number of studies k. The second panel in Table 14.4 gives requisite sample sizes for tests of nonreplication (H_0: $\tau^2 \geq \tau_0^2$) to detect exact replications ($\tau^2 = 0$) with 80% power ($\alpha = 0.05$) for various values of τ_0^2. The third panel gives the requisite sample size to ensure a relative standard error (RSE) $SE[\hat{\tau}^2_{DL}]/\tau^2 \leq 50\%$ for a given number of studies k and values of τ^2. Note that cells with "—" indicate that no sample size large enough could generate an RSE below 50%.

For reference, a test of exact replication with $k = 10$ studies would need a sample size of $n \geq 172$ per study to detect $\tau^2 = 0.05$ with 80% power. If $\tau^2 = 0.05$, and $k = 10$ studies are conducted, then we would need a sample size of $n \geq 1319$ to ensure a relative standard error less than 50% for an estimate of τ^2. Large sample sizes will be required for small RSE when τ^2 itself is small because of the relationship between τ^2 and the RSE. Relaxing the RSE slightly in such cases may not result in vastly larger standard errors. For reference, a relative standard error of 50% when $\tau^2 = 0.05$ would ensure a standard error $SE[\hat{\tau}^2_{DL}] \leq 0.025$. If instead we desire a standard error $SE[\hat{\tau}^2_{DL}] \leq 0.03$, then for $k = 10$ studies and $\tau^2 = 0.05$, we would require just $n = 294$ subjects per study, less than a quarter of the sample size indicated in the table.

Discussion

Questions about how to define replication as an estimand, analyze replication studies to make inferences on that estimand, and how to design replication studies to support sensitive analyses are intrinsically linked. In this chapter, I have showed that there are myriad approaches to defining replication that are functions of effect *parameters*, and hence there are a variety of analysis methods (functions of effect *estimates*) that are relevant for replication.

Different definitions of replication can and will be preferred in different settings and fields, and for different types of findings. Determining which definition is most relevant is subject to scientific and clinical judgment. Here, I have focused on definitions of replication that involve the consistency of effects (i.e., effects in replication studies are about the same size). This is particularly relevant to approaches for enhancing evidence-based practices, which support inferences about the stability of an effect.

Analyses for replication can support qualitative conclusions about the replicability of scientific findings in a manner consistent with NHT. Indeed, this chapter has presented a series of hypothesis tests for replication. However, recent moves by the American Statistical Association have urged researchers to focus on reporting point estimates and relevant uncertainty over p-values (see Wasserstein & Lazar, 2016). In keeping with those developments, I would suggest researchers studying replication to focus on estimating relevant parameters, though qualitative conclusions about replication may still be desired or warranted.

Designing replication studies requires some determination of the requisite number of studies k and sample size per study n to ensure sufficiently sensitive analyses. If one of the studies to be included in analyses is already published, a design with $k = 2$ studies (i.e., conducting only a single replication study) is unlikely to be sufficiently powered. If such $k = 2$ designs are unavoidable due, for instance, to budget or resource constraints, this chapter outlined some relevant methods (small telescopes, skeptical p-value) to make sense of the evidentiary value of $k = 2$ studies.

That a design involving a single replication of a published study is unlikely to be well powered suggests a few considerations for both primary research and replication research. The most obvious is that research seeking to replicate existing findings should (in most cases) have $k \geq 3$ studies. However, there are two possible ways around the statistical limitations of the $k = 2$ design that do not involve increasing the number of studies conducted. First, the research community could prioritize conducting primary studies with larger sample sizes. Sound statistical practice dictates that experiments be devised so that they are well powered. This typically means a power of at least 80%. However, even studies with 80% power will limit the power of analyses for replication. Thus, seeking designs of primary studies with higher power (e.g., at least 90%) may reduce the resources required to replicate them in the future.

Second, improving transparency of primary studies and their publication can improve statistical analyses for replication (see Collins & Tabak, 2014; Bollen et al., 2015; Schauer & Hedges, 2020). Recall that adjustments for publication bias will only reduce the sensitivity of replication analyses. Therefore, pre-registering studies and analysis plans, reporting all relevant effects, and reporting regardless of statistical significance may reduce the impact of selection and hence reduce bias in extant findings.

An alternative approach is to conduct replications prior to publishing any single study (e.g., Schweinsberg et al., 2016). Rather than designing a single study, the focus can be on designing an ensemble of replication studies. The results of these studies (including their consistency) can be reported as part of a single article or series of articles. This is analogous to the type of work done by the PPIR and to efforts possible under the Psychological Science Accelerator. Embedding replication into the process of primary inquiry can help improve our understanding about a phenomenon and the conditions under which it is studied.

Further Reading

For discussion regarding meta-analytic approaches to studying replication, Valentine et al. (2011) describe a general framework that was later refined by Hedges and Schauer (2019b). Finer points about fixed effects analyses were identified and discussed by Schauer (2018), including various ways for intuiting the scale of between-study variation. A more detailed demonstration of the methods discussed in this chapter was done by Schauer and Hedges (2020). Hedges and Schauer (2021)

proposed methods for identifying cost-effective or otherwise optimal designs of replication studies that support powerful analyses. In addition, Schauer (2018) derives some corrections to many of the analyses discussed in this chapter to account for small sample sizes in studies that could lead to violations of the assumption that the v_i are known and not estimated.

Though the focus of this chapter was on the replicability of results, particularly as operationalized as variation between effect parameters, replication research programs can provide insight into other parameters. For instance, meta-analytic methods can support estimation of mean effects across studies, as well as prediction intervals of effects (see Cooper, Hedges, & Valentine, 2019). Estimates of relevant parameters, including variance components, are possible with most meta-analytic software, including with the metafor library in the R computing language (Viechtbauer, 2010). Similarly, the Replicate library in R can conduct inference on the $P_{>q}$ and P_{orig} statistics (Mathur & VanderWeele, 2020).

The analyses presented here are primarily for direct replications, where studies are devised to be as similar as possible. Contrast that with conceptual replications, which may systematically vary aspects of a study to examine the potential impact of those variations on study results. This can be conceived of in a meta-analytic analysis of variance (ANOVA) framework, where studies can be grouped according to how they were conducted; if we denote a variable X that is systematically varied across studies, then we can group studies according to their value of X. Relevant analyses are discussed by Schauer (2018) and Schauer and Hedges (2020), which also includes empirical demonstrations.

As an alternative to the frequentist analysis methods that this chapter focused on, there are several possible Bayesian analyses of replications. Schauer (2018) describes Bayesian approaches to estimating λ and τ_F^2, and outlines various considerations for Bayesian estimation of τ^2. These latter discussions have been considered thoroughly in the statistical literature (for a good discussion, see Gelman et al., 2014). Alternatively, there have been approaches devised for $k = 2$ studies including those by Etz and Vandekerckhove (2016), van Aert and van Assen (2017), or Held (2020).

This chapter relied on meta-analytic notation as a matter of simplicity. This approach can be used even if data on individual participant are not available to the analyst. In programs of research regarding replication, this may not be the case; analysts may have access to individual participant data in all or a portion of relevant studies. In such cases, analyses can use multilevel models (also referred to individual participant data meta-analysis), which are analogous to the models presented in this chapter (see Raudenbush & Bryk, 1992; Riley et al., 2010; Tierney et al., 2015).

Finally, this chapter demonstrated that there are configurations of replication research programs wherein enough studies are conducted each with large enough sample sizes such that analyses are sufficiently powered. These are key design choices that should be made prior to collecting data. It is worth noting that there may be multiple configurations of study sample sizes n and number of studies k that provide sufficient power. Choosing between these can be as simple as what configurations can be implemented. Alternatively, Schauer (2018) and Hedges and Schauer (2021) provide an approach for optimally allocating sample sizes and numbers of studies.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*(5), 556–578.

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. In *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*. National Science Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., … Wollbrant, C. E. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science, 12*(3), 527–542.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., et al. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

Camerer, C. F., et al. (2016). Evaluating the reproducibility of laboratory experiments in economics. *Science, 351*, 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behavior, 2*, 637–644.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., … Yong, J. C. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science, 11*(5), 750–764.

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

Collins, F. S., & Tabak, L. A. (2014). NIH plans to enhance reproducibility. *Nature, 505*, 612–613.

Cooper, H. M., Hedges, L. V., & Valentine, J. (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). The Russell Sage Foundation.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177–188.

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). Wiley.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82.

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., … Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*(1), 158–171.

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

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*(6), 975–991.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). CRC Press.

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

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546–573.

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. https://doi.org/10.1371/journal.pbio.1002106

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

Hedges, W.L.V. (1982). Estimating effect size from a series of independent experiments. *Psychological Bulletin*, 92, 490–499.

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

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

Hedges, L. V., & Schauer, J. M. (2019a). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics, 44*(5), 543–570.

Hedges, L. V., & Schauer, J. M. (2019b). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods, 24*(5), 557–570.

Hedges, L. V., & Schauer, J. M. (2021). The design of replication studies. *Journal of the Royal Statistical Society, Series A, 184*, 868–886.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299–332. https://doi.org/10.3102/10769986021004299

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

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). Wiley.

Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society, Series A, 183*, 431–448. https://doi.org/10.1111/rssa.12493

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

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine, 21*, 1539–1558.

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

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

Klein, R. A., 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*(3), 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490.

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., … Ratliff, K. A. (2019). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement*. Retrieved from: https://psyarxiv.com/vef2c

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6*(1), 5–30.

Lawrance, R., Degtyarev, E., Griffiths, P., Trask, P., Lau, H., D'Alessio, D., Griebsch, I., Wallenstein, G., Cocks, K., & Rufibach, K. (2020). What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *Journal of Patient-Reported Outcomes, 4*(1), 68. https://doi.org/10.1186/s41687-020-00218-5

Mathur, M., & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society, Series A, 183*, 1145–1166.

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*(6), 487–498.
McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*(5), 730–749.
Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1*(4), 501–515. https://doi.org/10.1177/2515245918797607
Olive, K. A., et al. (2014). Review of particle properties. *Chinese Physics Journal C, 38*, 090001. http://iopscience.iop.org/issue/1674-1137/38/9
Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 943–951.
Oppenheimer, D. M., & Monin, B. (2009). Investigations in spontaneous discounting. *Memory & Cognition, 37*(5), 608–614. https://doi.org/10.3758/MC.37.5.608
Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep preferentially enhances memory for emotional components of scenes. *Psychological Science, 19*(8), 781–788. https://doi.org/10.1111/j.1467-9280.2008.02157.x
Perrin, S. (2014). Make mouse studies work. *Nature, 507*, 423–425.
Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Psychological Science, 7*, 531–536.
Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*(4), 539–544.
Paule, R., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards, 87*(5), 377–385. https://doi.org/10.6028/jres.087.022
Pigott, T. (2012). *Advances in meta-analysis*. Springer.
Raudenbush, S. W., & Bryk, A. S. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications.
Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ, 340*, c221. https://doi.org/10.1136/bmj.c221
Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Wiley.
Schauer, J. M. (2018). *Statistical methods for assessing replication: A meta-analytic framework*. (Doctoral thesis). Retrieved from https://search.proquest.com/docview/2164811196?accountid=12861
Schauer, J. M., Fitzgerald, K. G., Peko-Spicer, S., Whalen, M. C. R., Zejnullahi, R., & Hedges, L. V. (2021). An evaluation of statistical methods for aggregate patterns of replication failure. *Annals of Applied Statistics, 15*(1), 208–229. https://doi.org/10.1214/20-AOAS1387
Schauer, J. M., & Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin, 146*(8), 701–719.
Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods, 26*(1), 127–139. https://doi.org/10.1037/met0000302
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100.
Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55–67.
Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9*(5), 552–555.
Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569.

Tierney, J. F., Vale, C., Riley, R., Smith, C. T., Stewart, L., Clarke, M., & Rovers, M. (2015). Individual Participant Data (IPD) meta-analyses of randomised controlled trials: Guidance on their use. *PLoS Medicine, 12*(7), e1001855. https://doi.org/10.1371/journal.pmed.1001855

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

van Aert, R., & Jackson, D. (2018). Multistep estimators of the between-study variance: The relationship with the Paule-Mandel estimator. *Statistics in Medicine, 37*(17), 2616–2629. https://doi.org/10.1002/sim.7665

van Aert, R. C., & Van Assen, M. A. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS One, 12*(4), e0175302.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*(1), 55–79. https://doi.org/10.1002/jrsm.1164

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*, 428–443.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26*(1), 37–52. https://doi.org/10.1002/sim.2514

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

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

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., ... Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11, 917–928.

Wellak, S. (2002). *Testing statistical hypotheses of equivalence*. CRC Press.