Received: 8 January 2019 | Accepted: 22 November 2020

DOI: 10.1111/rssa.12688

ROYAL STATISTICAL SOCIETY
Series A
Statistics in Society

ORIGINAL ARTICLE

The design of replication studies

Larry V. Hedges[1] | **Jacob M. Schauer**[2]

[1]Department of Statistics, Northwestern University, Evanston, IL, USA

[2]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Evanston, IL, USA

Correspondence
Jacob M. Schauer, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Evanston, IL, USA.
Email: jms@u.northwestern.edu

Funding information
Directorate for Social, Behavioral and Economic Sciences, Grant/Award Number: DRL-1841075; Institute of Education Sciences, Grant/Award Number: R305B140042 and R305D140045

Abstract

Empirical evaluations of replication have become increasingly common, but there has been no unified approach to doing so. Some evaluations conduct only a single replication study while others run several, usually across multiple laboratories. Designing such programs has largely contended with difficult issues about which experimental components are necessary for a set of studies to be considered replications. However, another important consideration is that replication studies be designed to support sufficiently sensitive analyses. For instance, if hypothesis tests are to be conducted about replication, studies should be designed to ensure these tests are well-powered; if not, it can be difficult to determine conclusively if replication attempts succeeded or failed. This paper describes methods for designing ensembles of replication studies to ensure that they are both adequately sensitive and cost-efficient. It describes two potential analyses of replication studies—hypothesis tests and variance component estimation—and approaches to obtaining optimal designs for them. Using these results, it assesses the statistical power, precision of point estimators and optimality of the design used by the Many Labs Project and finds that while it may have been sufficiently powered to detect some larger differences between studies, other designs would have been less costly and/or produced more precise estimates or higher-powered hypothesis tests.

KEYWORDS

experimental design, meta-analysis, power, replication

1 | INTRODUCTION

There has recently been a great deal of empirical investigation into the replicability of research results. Concern about replication began to emerge in medicine when researchers compared highly cited clinical studies to subsequent replications of those studies (see, e.g. Ioannidis, 2005). Other researchers examined preclinical work and attempted to replicate those findings (see, e.g. Perrin, 2014). In all cases, the replicability of studies appeared to be poor. This became widely publicized by articles in the popular press including *Newsweek*, the *New Yorker* and the *Economist* suggesting that there were serious problems in medical science (Firger, 2015; Let's, 2016; Marcus, 2013). The controversy led to a policy response by the National Institutes of Health to promote replicability of medical science (Collins & Tabak, 2014).

However, the issue has not been limited to medicine. One of the areas most concerned by the issue is psychology (Pashler & Harris, 2012), which has galvanized efforts to organize pre-planned ensembles of studies to examine the replicability of findings. In perhaps the most prominent such project, the Open Science Collaboration (OSC) (2015) attempted to replicate 100 psychological findings each a single time and determined that about 61% of these attempts failed. A related project in behavioural economics tried to replicate 18 experiments each a single time and concluded that nearly a third of those attempts failed (Camerer et al., 2016). However, it has been more common for replication research programs to conduct multiple replications of the same experiment simultaneously. For instance, the Many Labs Project conducted 36 direct replications of each of 16 different experiments in order to investigate variation in experimental results (Klein, et al., 2014). Multi-laboratory replications have since been emulated by a variety of research programs, including the Psychological Science Accelerator, which has recruited over 500 laboratories across the world to facilitate large-scale inquiry into the replicability and generalizability of psychology experiments (Moshontz, et al., 2018).

An important aspect about such investigations is how they approach designing ensembles of studies to investigate replication. Researchers largely discuss design in terms of the steps they take to standardize and validate experimental protocols, which is difficult and a crucial element of studying replication (Klein et al., 2014; Open Science Collaboration, 2012). Yet none of these investigations offered a principled statistical argument that the design they chose was adequate to yield unambiguous conclusions. Such arguments, which typically involve examining the power or precision afforded by a design (e.g. a power analysis), might address how many subjects their replication studies should recruit, or how many studies should be conducted. The lack of statistical justification for designs has contributed to large differences in how such research is done: some programs run only a single replication study (per finding), while others have conducted up to 36. Moreover, while several scholars have commented on how to analyse replication studies, less attention has been paid to how to design them (see, e.g. Etz & Vandekerckhove, 2016; Gilbert et al., 2016; Hartgerink, Wicherts, & van Assen, 2017). Other researchers have pointed out limitations in the designs of past replication research programs, but have not made concrete proposals on how to formally approach the question of design *a priori* (see van Aert & van Assen, 2017; Hedges & Schauer, 2019a).

This paper addresses the problem of designing ensembles of studies intended to investigate replication. By *design* of replication studies, we mean intentionally planning how many studies m and how many subjects per study n are required to ensure hypothesis tests for replication have high power or point estimates of relevant quantities have sufficiently small standard errors. However, any consideration of design depends on the analysis that is to be conducted. Previous work by Hedges and Schauer (2019b) discussed possible quantities of interest for analyses of replication, proposed a series of hypothesis tests for replication, and examined their power. However, their work stopped short of explicitly addressing the purposive design of ensembles of replication studies, nor did they address

issues of optimal allocation of studies m and subjects per study n. This article expands on their findings by proposing methods for determining optimal designs of replication studies in order to ensure tests for replication are sufficiently powered or point estimates of relevant quantities have sufficiently small standard errors.

Because the analysis of replication involves all of the attempted replications (including the original study, if there is one), the proper conception of the design of replication studies must consider the entire ensemble of $m \geq 2$ studies. In this article, we assume that design choices affect all studies to be included in an analysis, which we refer to as *unconstrained* designs. In unconstrained designs, it is assumed that design choices pertain to the original or initial study, as in the scenario faced by research programs such as the Psychological Science Accelerator, or that the original study is omitted from analysis (see Schauer & Hedges, 2020). However, when the analysis involves extant published findings, then design choices only pertain to a subset of replication studies (because existing studies will be part of the analysis). In that case, the considerations are slightly different and more complicated, in part because the design will be *constrained* by features of the existing studies.

In the following section we review prior empirical work on replication and highlight how the Many Labs Project approached the types of design choices considered in this paper. Then, we describe a statistical model in order to shed light on the considerations of design and analysis of replication studies. The subsequent sections then detail hypothesis tests and point estimates for heterogeneity among replication study results and derive optimal designs that ensure such analyses are sufficiently sensitive (i.e. have high power or precision). Finally, we conclude with discussion regarding the implications of these findings for future replication research.

2 | STUDYING REPLICATION: PAST RESEARCH AND POTENTIAL DIFFICULTIES

As discussed in the introduction, the OSC (2015) may be the most well-known replication research program. For each of 100 findings, the OSC attempted a single replication intended to be identical to the original experiment, in some cases even obtaining materials from the original investigators. Replications intended to be the same are often referred to as *direct* replications (see Schmidt, 2009), and these have been prominent in replication research in the social sciences. Since a finding was only replicated once, the design of the OSC can be seen as having $m = 2$ studies: the original published finding, and the replication. A similar design involving direct replications used by Camerer et al. (2016) and Camerer et al. (2018).

However, it has been more common in the social sciences for systematic investigations of replication to involve multiple direct replication studies conducted simultaneously. The Many Labs Project, discussed below, appears to be the first such effort (Klein et al., 2014). Since then, there have been 10 Registered Replication Reports (RRR) completed in conjunction with the Association for Psychological Science (APS), including a subsequent Many Labs Project called Many Labs 2 (Klein et al., 2018). Like Many Labs, each RRR has involved multiple direct replications of a set of experiments. The pre-publication independent replication (PPIR) pipeline followed a similar multi-laboratory setup for a set of experiments that were still in development (Schweinsberg et al., 2016). Meanwhile, the Psychological Science Accelerator, which comprises some 500 laboratories, is just starting to produce results (Moshontz, et al., 2018).

Across all of these efforts, there are both constrained and unconstrained design problems. When designs are constrained in empirical research, analyses are typically framed as a comparison between the original published finding and subsequent replications. The goal of such designs and analyses

appears to be to confirm or falsify the original finding. The OSC, for instance, faced a constrained design problem since an original study for each finding had already been published. Even multi-site replication research programs like Many Labs assess whether the results of an original finding differ from those of replication studies.

When the goal of constrained designs is to compare the original finding to subsequent replications, the analysis inherently privileges the original finding and the sensitivity of analyses (i.e. the power of hypothesis tests or the precision of estimators) will depend in some way on the sample size of the original study. Hedges and Schauer (2019a) show that unless the original study has very high power (i.e. over 95%) it will be impossible to design replications in order for comparisons between the original and replication studies to be sufficiently sensitive, no matter how many replications are conducted or how large they are. Rather than privileging a single study (i.e. the original study), Hedges and Schauer (2019b) suggest that analyses can investigate heterogeneity among *all* studies (even including the original study). This has been part of the goal of multi-site replication research so far (see Klein et al., 2014; Schweinsberg et al., 2016).

When analyses of replication are framed in terms of variation across all studies, designs will be constrained if the original study is included in the analysis, and they will be unconstrained if the original study is excluded. In practice, the original study is often excluded from analyses. For instance, a primary goal of Many Labs was to investigate variation in experimental results across a series of direct replications, and analyses were conducted *excluding* the original findings. In that sense, their design could be seen as unconstrained. Similar assessments could be made of much of the multi-laboratory replication research.

Another important consideration involves how accurate the original finding is. There is a prevalence of empirical and theoretical research that suggests statistically significant results are more likely to be published. This type of selection can lead to substantial bias (e.g. Dickersin, 2005; Head et al., 2015). Thus, any constrained design problem will need to contend with issues of and potential correction for publication bias (e.g. Hedges & Vevea, 1996; Hedges, 1984, Vevea & Woods, 2005). However, since all studies are to be conducted simultaneously for unconstrained designs, publication bias would only be a factor if laboratories selectively shared results, which to date does not appear to be an issue in replication research.

This article addresses the unconstrained design of ensembles of replication studies. Consistent with Hedges and Schauer (2019b) and Schauer and Hedges (2020) analyses presented in this article concern the variation among all study results, rather than privileging a single study (e.g. the original study). Framing design in this way is appropriate for scenarios where all studies may be designed in concert, as with the Psychological Science Accelerator, or the original study will be excluded from analyses (e.g. because of publication bias), which has been common in replication research programs. In sum, the unconstrained design problem is a key issue for most modern replication research efforts, particularly in the social sciences.

Finally, to motivate our discussion of unconstrained design, we will focus on the Many Labs Project. The organizers of Many Labs recruited $m = 36$ laboratories to test each finding, requiring a sample size of at least $n = 80$ in each laboratory (Klein et al., 2014). Thus, for any one finding, they determined that $m = 36$ and $n = 80$ would be a sufficient baseline design, although it is worth noting that most laboratories recruited more than 80 subjects (median $n = 103$). Throughout this article, we will refer to the Many Labs design, its sensitivity, and its cost-effectiveness. We do this not to denigrate their design, nor to assume that it is necessarily suboptimal, but rather to illustrate how the principles discussed in this article might lead to different design choices under a variety of conditions.

While Many Labs replicated several different findings, to make this example more concrete, consider their replications of the ‘retrospective gambler's fallacy’, an experiment originally studied by

Oppenheimer and Monin (2009). In this experiment (and hence the Many Labs replications), participants were randomly assigned to one of two conditions and asked to imagine a man rolling dice at a casino. In one condition, participants imagined seeing the man roll three sixes. In the other condition, participants imagined the man 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. 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 carried out $m = 36$ replications of this experiment, and determined that $n = 80$ was a reasonable sample size per study, although most laboratories recruited more than 80 subjects.

3 | MODEL AND NOTATION

In this article, we approach the analysis of replication studies using common tools in meta-analysis, such as the Q-test or variance component estimation. Previous research on evaluating replication in psychology (e.g. Camerer et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015) and the sensitivity of tests for replication (Hedges & Schauer, 2019b) have used a meta-analytic approach in which the study results were represented by common effect sizes in meta-analysis (e.g. standardized mean differences). We follow that approach here, although we note that if all of the replication studies use exactly the same measurement protocol to assess the outcome, then analyses need not standardize effect sizes (e.g. they may use unstandardized mean differences). Analyses of replication can leverage large sample inference for multilevel models or methods related to analysis of variance. Either of these alternatives introduces further complexities (e.g. are within-study variances taken to be the same or different across studies) and neither of these methods is likely to yield substantially different results if within-study sample sizes are reasonably large (which this work suggests that optimal designs are likely to require).

Suppose that there are m experiments. Let θ_i denote the treatment effect size parameter, T_i denote the effect size estimate, n_i denote the sample size and v_i denote the estimation error variance of the effect size estimate in the ith study.

Investigations about replication focus on inference about the similarity of the θ_i between studies. Decompose T_i into

$$T_i = \mu + \gamma_i + \varepsilon_i, \tag{1}$$

where μ is the grand mean effect size, γ_i is the deviation of each study's mean from μ and ε_i is the estimation error ($\varepsilon_i \equiv T_i - \theta_i$). The two-level random effects model in meta-analysis would consider the γ_i as distributed independently with mean zero and variance τ^2, and the ε_i distributed independent of all other quantities on the right-hand side of Equation (1) with mean zero and variance v_i. Often, the γ_i are modelled as normally distributed.

Hedges and Schauer (2019b) studied analyses for replication whose properties depend on the v_i but did not present results that explored the relationship between v_i and the sample size within studies n_i. The methods described in this paper are appropriate for scenarios when the v_i can be written as $v_i = \omega/n_i$ where ω does not depend on θ or n_i. This is (approximately) true of a number of common effect sizes used in meta-analysis. For example, with standardized mean differences, $\omega = 4(1 + \theta^2/8)$. If θ is relatively small, then $\theta^2/8 \approx 0$, and hence a reasonable approximation is that $\omega \approx 4$. The formulation $v = \omega/n$ will be a better approximation for standardized mean differences transformed via a variance stabilizing transformation, such as

$$f(T) = \sqrt{2\sinh^{-1}}\left(T/\sqrt{8}\right),$$

in which case $\omega = 1$ (see Hedges & Olkin, 1985). Similarly, the z-transformed correlation coefficient has a variance of $v_i = 1/(n_i - 3)$ so that when n_i is large, $v_i \approx 1/n_i$ and $\omega \approx 1$.

4 | THE MAGNITUDE OF HETEROGENEITY

This article frames analyses of replication in terms of between-study variation τ^2. This is not the only way to assess replication, but it is one that is consistent with the stated goals of various replication research programs, many of which seek to examine variation across several direct replications (see, e.g. Klein et al., 2014).

There is empirical evidence in the social sciences (Hedges, 1987; Klein et al., 2014; Klein et al., 2018; Schauer & Hedges, 2020) and physical sciences (Olive et al., 2014; Rosenfeld, 1975) that study results (i.e. the θ_i) may vary even among direct replications. Replications can produce different results for a whole host of reasons. There may be minor differences in experimental procedures that may be unknown to researchers (for historical examples, see Collins, 1992). In the social and medical sciences, there is an understanding that context and sample composition can moderate the effects of interventions (see Keiding & Louis, 2018; Tipton and Hedges, 2017). While it may be of interest whether experimental results are correlated with specific components of an experimental procedure or the populations from which subjects are sampled (or an interaction between the two), it will be difficult to confirm these relationships if these factors are not systematically varied across studies.

Most modern replication research does not systematically vary relevant factors in an experiment. Important differences between studies may not be known or evident to researchers. Moreover, multisite replication research programs largely conduct one study per site, which inherently confounds variation in procedure and setting. Even when the goal of the program is to study sources of variation, these are often confounded. For example, the PPIR was interested examining the effects sample composition (i.e. were subjects primarily university students or not) and setting (i.e. is the study conducted in a laboratory vs. an online system like MTURK) (Schweinsberg et al., 2016). However, because most of the laboratories involved in PPIR were on university campuses, they were also more likely to involve samples of entirely university students, whereas studies conducted online were more likely to contain subjects who were not university students. We note that the analyses presented in this article do not require that factors that might affect study results are systematically varied, although they can still be conducted even if factors are systematically varied.

Understanding the analyses presented in this article (and the resulting designs) requires some idea of how much variation is considered meaningful or worth studying. For instance, determining the power of hypothesis tests regarding τ^2 involves specifying which values of τ^2 one wishes to detect. Different researchers from different fields will likely have different answers to this question. Thus, this section describes possible notions of between-study variation that may be considered meaningful across various fields.

We start by noting that researchers have tended to make judgements about heterogeneity in the metric of τ^2/v. For instance in meta-analysis, the H^2 statistic is an estimate of $1 + \tau^2/v$, and I^2 can be seen as an estimate of $(\tau^2/v)/(1 + \tau^2/v)$ (Higgins & Thompson, 2002). Various fields have established conventions about what constitutes a negligible amount of heterogeneity on this scale (for a larger discussion, see Hedges, 1987; Hedges & Pigott, 2001, 2004; Hedges & Schauer, 2019a,b). These

conventions range from $\tau^2/v = 1/4$ in high energy physics (see Olive, 2014), to $\tau^2/v = 1/3$ in personnel psychology (Hunter & Schmidt, 1990), to $\tau^2/v = 2/3$ in medicine (Higgins & Green, 2008). Others have offered alternative justifications for small or 'realistic' levels of heterogeneity in a meta-analysis (see van Erp et al., 2017; Pigott, 2012) that typically involve $\tau^2/v < 1$.

The scale on which we assess heterogeneity in this article is actually τ^2/ω. The following sections show that the power of hypothesis tests regarding τ^2 are a function of τ^2/ω, and the precision of estimators of τ^2 will depend on τ^2/ω. Reasoning about values of τ^2/ω is a little more complicated because researchers have tended to make judgments about $\tau^2/v = n\tau^2/\omega$ not τ^2/ω itself. Moreover, the scale of τ^2/v inherently depends on n, which means that while τ^2/v may be better understood, its scale will be affected by design choices (e.g. if designs require large n). One approach to determining negligible values of τ^2/ω would be to rely on the conventions of τ^2/v discussed above and use notions of a typical sample size n in an experiment to back out a value of $\tau^2/\omega = (\tau^2/v)/n$.

Suppose that a typical laboratory experiment has 50 to 100 total participants. The conventional values of τ^2/v used to define negligible heterogeneity described above range from 1/4 to 1, which correspond to values of I^2 of 0.2 to 0.5. Presumably, an ensemble of replication studies would not be designed to detect negligible heterogeneity, but somewhat larger values of heterogeneity, perhaps up to three times the negligible value ($I^2 = 0.42$ to 0.75). This implies that values of τ^2/v one might encounter in replications could range from 1/4 to 3 ($I^2 = 0.2$ and 0.75), which corresponds to values of τ^2/ω from 0.0025 to 0.06 (assuming n is around 50–100). While these are not the only values of τ^2/ω one may wish to study, they represent a range of values that could be considered meaningful across various fields.

The model presented in this section and the tests described in subsequent sections were presented by Hedges and Schauer (2019b), who also studied the power of those tests. This article extends those results in several ways. First, it augments the results on power found by Hedges and Schauer (2019b) by studying the explicit contribution of within-study sample sizes. Second, this article discusses point estimation as an alternative to null hypothesis tests for replication. Third, we describe a proscriptive approach to the purposeful planning of ensembles of replication studies to ensure analyses are sufficiently sensitive—hypothesis tests have high power or point estimates have high precision—an issue not addressed in previous work. Finally, this article addresses optimal allocations of the number of studies m and subjects per study n that minimize the costs of replication research programs.

5 | HYPOTHESIS TEST FOR VARIATION OF EFFECTS

A test of the hypothesis that there is no between-study variation in effects, that is, a test of

$$H_0: \tau^2 = 0$$

uses the test statistic

$$Q = \sum_{i=1}^{m} \left(T_i - T_\bullet\right)^2 / v_i = \sum_{i=1}^{m} n_i \left(T_i - \overline{T}\right)^2 / \omega, \tag{2}$$

where $T_\bullet$ is the grand weighted mean given by

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

and $\overline{T}$ is the unweighted grand mean (see Hedges, 1982). This test rejects the null hypothesis of no variation in effects at significance level α if the obtained value of Q exceeds the $100(1-\alpha)$ per cent point $c_{(1-\alpha)}$ of the chi-squared distribution with $m-1$ degrees of freedom.

Ensembles of planned replication studies may have a common sample size (for greatest efficiency), so that $n_1 = \ldots = n_m = \mathrm{n}$. This is not always true in practice, but even when it is not programs like Many Labs tend to require minimum sample sizes. When sample sizes are equal, the non-null sampling distribution of Q is equal to that of a constant times a central chi-squared random variable, so that

$$\left(\frac{v}{v+\tau^2}\right)Q = \left(\frac{1}{1+\tau^2/v}\right)Q \sim \chi^2_{m-1}, \tag{3}$$

(see Hedges & Pigott, 2001). It is worth noting that even if sample sizes are modest, this is a reasonable approximation, and the test retains levels that are nearly nominal (see Kulinskaya et al., 2011a,b). Thus, the statistical power of the test for between-study variation is

$$1 - F\left(\frac{c_{(1-\alpha)}}{1+\tau^2/v}\middle| m-1\right) = 1 - F\left(\frac{c_{(1-\alpha)}}{1+n\tau^2/\omega}\middle| m-1\right) \tag{4}$$

where $F(x|\ m-1)$ is the distribution function of the (central) chi-squared distribution with $(m-1)$ degrees of freedom. Note that the sampling distribution of Q is determined entirely by m and $v/(v+\tau^2) = 1/(1+n\tau^2/\omega)$. When the sample sizes are unequal, the non-null distribution of Q is more complex, but an approximation that is reasonably accurate when sample sizes are approximately equal was given by Hedges and Pigott (2001).

It is clear from Equation (4) that the power tends to 1 as $n \to \infty$. Noting that a chi-squared distribution with m degrees of freedom converges to a normal distribution with mean m and variance $2m$ as $m \to \infty$ and applying L'Hopital's rule to the implied normal deviate in the limiting distribution, it is clear that the power also tends to 1 as $m \to \infty$. Figure 1a shows the statistical power as a function of m for $n = 100$ and three values of τ^2/ω. Figure 1b shows the statistical power as a function of n for $m = 30$ and three values of τ^2/ω. Figure 1c shows the statistical power as a function of τ^2/ω for $m = 30$ and three values of n. Note that different configurations of m and n can yield approximately the same power. Because of this, there may be many designs that can detect a given amount of heterogeneity with a pre-specified power. A later section details how we might determine and optimal design from among those that provide the same amount of sensitivity.

We can use Equation (4) to determine the sensitivity of a particular design to test replication in at least two ways. We can compute the statistical power to detect a specified value of heterogeneity given the ensemble of studies being considered. Or, we can determine the smallest amount of heterogeneity that could be detected with some pre-specified power by an ensemble of studies. As an example, consider the Many Labs design ($m = 36$ and $n = 80$). The smallest amount of heterogeneity that this design could detect with 80% power is $\tau^2/v = 0.75$ ($I^2 = 43\%$). If the effect parameters θ are standardized mean differences and are reasonably small so that $\omega \approx 4$, this corresponds to $\tau^2/\omega = 0.009$ or $\tau^2 = 0.037$. These results are consistent with Hedges and Schauer's (2019b) assessment of this design.

6 | ESTIMATION OF VARIANCE OF EFFECTS

While there are several different methods for estimating τ^2, a review by Veroniki et al. (2016) found that when the effect parameters are normally distributed, the restricted maximum likelihood (REML)

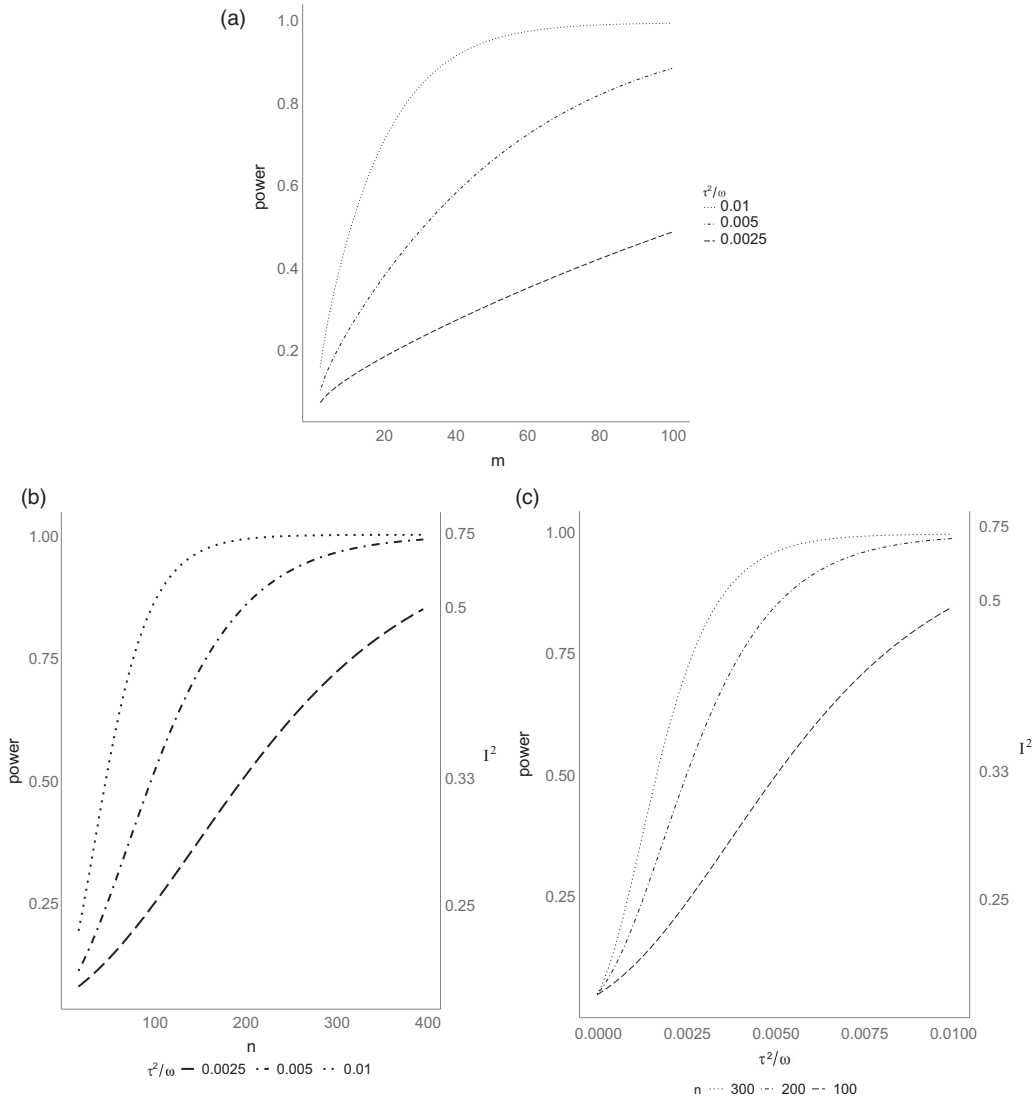

FIGURE 1 (a) Statistical power of the test for heterogeneity as a function of the number of studies m each of which have a sample size of $n = 100$. The three line types correspond to three values of variation of among study results on the scale of τ^2/ω, which correspond to $I^2 = 0.5$, 0.33 and 0.2. (b) Statistical power of the test for heterogeneity as a function of the within-study sample size n for $m = 30$ replication studies. The three line types correspond to three values of variation of among study results on the scale of τ^2/ω. The left y-axis shows the power of the test, and the right y-axis shows the values of I^2 for the curves, which is a bijective function of power. (c) Statistical power of the test for heterogeneity as a function of the between-study variance τ^2/ω for $m = 30$ studies and three different within-study sample sizes n. The left y-axis shows the power of the test, and the right y-axis shows the values of I^2 for the curves, which is a bijective function of power.

estimator tends to perform fairly well. The REML estimator of τ^2 cannot be expressed in closed form but can be computed via a simple iterative procedure starting with the method of moments estimator (see, e.g. Raudenbush, 2009; Rukhin, 2000; Veroniki et al., 2016). The variance of the maximum likelihood estimator of $\tau^2 2$ under the assumption of equal sample sizes and normally distributed random effects is

$$V\left\{\tau^2\right\} = \frac{2\left(v+\tau^2\right)^2}{m-1} = \frac{2\left(\omega+n\tau^2\right)^2}{(m-1)n^2} \tag{5}$$

Note that Equation (5) implies that the variance of $\widehat{\tau}^2$ tends to zero as $m \to \infty$, but the variance of $\widehat{\tau}^2$ tends to $2\tau^4/(m-1)$ as $n \to \infty$. This might seem odd because the power of the Q tends to 1 as $n \to \infty$, but the variance of the estimate of τ^2 does not tend to zero. The reason this happens is that Equation (3) implies that the distribution of Q is $(1+\tau^2/v)$ times a chi-squared with $m-1$ degrees of freedom. Thus, as $n \to \infty$ so that $v \to 0$, $Q \to \infty$. In other words, Q is in the metric of $\widehat{\tau}^2/v$, so that even if the variance of $\widehat{\tau}^2$ remains bounded, Q can tend to infinity as $n \to \infty$ (and therefore $v \to 0$). As a result of this difference, the relation between n and the distribution of Q (including properties of the distribution such as the power of significance tests) and the precision of estimates of τ^2, respectively, are somewhat different.

7 | OPTIMAL DESIGN FOR ESTIMATING VARIANCE COMPONENTS

It is clear from Equations (4) and (5) above that both the statistical power of the test for replication and the precision of estimates of τ^2 depend on the number of studies m and the sample size within each study n and thus on the total sample size mn. Yet, like many multilevel modelling situations, different configurations of m and n yield the same statistical power and estimates of variance components that have the same precision.

One way to approach the optimal design problem is to posit a cost function and an objective function (such as statistical power or precision of an estimate) and then maximize that objective function under the constraint of a fixed cost. For example, we might obtain the design that yields the estimate of τ^2 that has greatest precision for a fixed cost. An analogous problem has been studied in the context of average treatment effects in multi-site trials, which use a linear cost function that specifies an incremental cost c_2 for each additional experiment and c_1 of each additional individual within an experiment (see, e.g. Moerbeek et al., 2000; Raudenbush, 1997).

While the general setup may appear to have much in common with prior work on optimal designs for multi-site trials, there are some key differences. First, replication involves a different estimand than prior work; with replication we are concerned with precision in estimates of the variation of treatment effects τ^2 rather than the average treatment effect μ or some study-level moderator. Thus, we would expect that optimal designs for estimating μ or for detecting a study-level moderator may not be optimal for estimating τ^2. Indeed, in the results presented below (particularly in Tables 1 and 2), the optimal designs are quite different than those proposed by Raudenbush (1997) or by Raudenbush and Liu (2000).

We can specify the total cost as

$$C = mc_2 + mnc_1. \tag{6}$$

Obtaining an optimal design then, involves minimizing the variance given in Equation (5) with respect to m and n subject to the cost constraint given in Equation (6). Using the Lagrangian multiplier, it can be shown that the optimal within-study sample size n_O that solves this problem does not depend on C, and can be expressed as

$$n_O = \frac{1+\sqrt{1+8\tau^2 c_2/c_1}}{2\tau^2/\omega} \tag{7}$$

TABLE 1 Optimal within study sample size for estimating the between study variance component with one study per laboratory. In the table, c_2/c_1 is the ratio of the per-laboratory cost to the per-subject cost, τ^2/ω is the between-study variation, n_O is the optimal within-study sample size, and V_{ML}/V_O is the ratio of the variance of the estimate of τ^2 for the Many Labs design relative the variance of the optimal design

c_2/c_1	τ^2/ω	n_O	V_{ML}/V_O	c_2/c_1	τ^2/ω	n_O	V_{ML}/V_O
5	0.0025	410	1.68	5	0.02	59	1.03
10	0.0025	419	1.76	10	0.02	65	1.02
50	0.0025	483	2.39	50	0.02	100	1.01
100	0.0025	546	3.09	100	0.02	128	1.07
1000	0.0025	1116	9.68	1000	0.02	342	1.61
5	0.005	210	1.22	5	0.05	27	1.39
10	0.005	218	1.26	10	0.05	32	1.29
50	0.005	273	1.57	50	0.05	56	1.05
100	0.005	324	1.91	100	0.05	74	1.00
1000	0.005	740	4.63	1000	0.05	210	1.16
5	0.01	109	1.02	5	0.075	20	1.70
10	0.01	117	1.03	10	0.075	24	1.52
50	0.01	162	1.17	50	0.075	44	1.12
100	0.01	200	1.32	100	0.075	59	1.03
1000	0.01	500	2.50	1000	0.075	170	1.08

It is clear from Equation (7) that n_O depends on the ratio c_2/c_1 and τ^2/ω; n_O is increasing in c_2/c_1, which means that as setting up a new laboratory becomes more expensive, the optimal allocation involves recruiting more subjects instead. Also note that n_O is decreasing as τ^2/ω increases, which means that fewer subjects per laboratory are required for precise estimation of larger variance components.

Obtaining an optimal design in this framework can take one of two approaches. The first involves obtaining n_O, and then solving Equation (6) for m to get the optimal number of laboratories. However, while formula (7) provides a solution to the relevant optimization problem, it can (and often will) return a non-integer value. Moreover, depending on the cost constraint, finding the *true* optimal sample size may not necessarily be obtained simply by rounding n_O to the nearest integer. However, it does provide a good approximation to the optimal within-study sample size. As a practical matter, one may determine the optimal design either by brute force, or by testing values near n_O.

To interpret n_O values, some concept of reasonable values of c_2/c_1 and τ^2/ω are needed. Meaningful values of τ^2/ω (discussed in a previous section) might range from 0.0025 to 0.06. For costs, c_2 may be only a few times the magnitude of c_1 if data are collected via the internet (e.g. via a mechanism like Mechanical Turk). If experiments are performed in a laboratory, then a subject might be paid $10 for participating, but setup and administration of the experiment could cost hundreds or thousands of dollars. Thus c_2/c_1 could feasibly range from the single digits to a few thousand.

Table 1 shows the optimal within study sample size as a function of c_2/c_1, and τ^2/ω. These values could be used in planning an ensemble of replication studies by choosing an optimal within-study sample size n, and then choosing a number of studies m to obtain whatever precision is required of the variance component estimate. The values in Table 1 show that there is a strong dependence of n_O on the true heterogeneity that the experiments are intended to detect with values ranging over an order of magnitude as τ^2/ω varies. Optimal experiments to detect small degrees of heterogeneity will need

TABLE 2 Optimal within study sample size n_O and number of studies m_O for testing the hypothesis H_0: $\tau^2 = 0$ at 80% and 90% power. In the table, c_2/c_1 is the ratio of the per- laboratory cost to the per-subject cost, τ^2/ω is the between-study variation, n_O is the optimal within-study sample size, and m_O is the optimal number of studies

c_2/c_1	τ^2/ω	n_O	m_O	c_2/c_1	τ^2/ω	n_O	m_O
Optimal for power of 80%							
5	0.0025	951	9	5	0.02	119	9
10	0.0025	951	9	10	0.02	135	8
50	0.0025	1073	8	50	0.02	187	6
100	0.0025	1073	8	100	0.02	238	5
1000	0.0025	1902	5	1000	0.02	622	3
5	0.005	476	9	5	0.05	54	8
10	0.005	476	9	10	0.05	75	6
50	0.005	537	8	50	0.05	96	5
100	0.005	621	7	100	0.05	136	4
1000	0.005	1355	4	1000	0.05	249	3
5	0.01	238	9	5	0.075	36	8
10	0.01	269	8	10	0.075	50	6
50	0.01	311	7	50	0.075	91	4
100	0.01	373	6	100	0.075	91	4
1000	0.01	678	4	1000	0.075	166	3
Optimal for power of 90%							
5	0.0025	1011	12	5	0.02	153	10
10	0.0025	1011	12	10	0.02	153	1
50	0.0025	1106	11	50	0.02	199	8
100	0.0025	1224	10	100	0.02	236	7
1000	0.0025	2350	6	1000	0.02	619	4
5	0.005	506	12	5	0.05	69	9
10	0.005	506	12	10	0.05	69	9
50	0.005	612	10	50	0.05	118	6
100	0.005	689	9	100	0.05	159	5
1000	0.005	1585	5	1000	0.05	549	3
5	0.01	253	12	5	0.075	46	9
10	0.01	253	12	10	0.075	53	8
50	0.01	345	9	50	0.075	79	6
100	0.01	397	8	100	0.075	106	5
1000	0.01	1238	4	1000	0.075	366	3

to be very large, with sample sizes of several hundred when the relative costs of experiments versus subjects are high.

The designs in Table 1 differ from the original requirements of the Many Labs design of $m = 36$, $n = 80$. We can get a sense of how much this matters by comparing the variances attained by the optimal designs relative the Many Labs design. However, in order to do this, we would need to determine the number of laboratories m_O that could be used given the total budget C, which is not reported in the

table. To do this, note that for each cost ratio c_2/c_1, we can compute the total cost of the Many Labs design (up to a multiplicative constant)

$$C_{\mathrm{ML}} = \left(36c_2/c_1 + 36 * 80\right) c_1$$

Assuming C_{ML} is the total budget, we can then determine how many experiments m_{O} could be conducted given the optimal within-study sample size n_{O}, such that the cost of the optimal design is the same as the Many Labs cost C_{ML}:

$$C_{\mathrm{ML}} = \left(m_{\mathrm{O}}c_2/c_1 + m_{\mathrm{O}}n_{\mathrm{O}}\right) c_1 = \left(36c_2/c_1 + 36 * 80\right) c_1$$

Then, we can compute the variance of the REML estimator of both the optimal designs V_{O} and the Many Labs design V_{ML} for each value of τ^2/ω using Equation (5).

Two columns in Table 1 show the ratio of $V_{\mathrm{ML}}/V_{\mathrm{O}}$. What we see in those columns is that often the optimal design will lead to estimates of τ^2 that are 1.5–2 times as precise as the estimate from the Many Labs design, but when c_2/c_1 increases, optimal designs can offer a four- or nearly 10-fold increase in precision. However, we should note that the baseline within-study sample size for Many Labs of $n = 80$ is quite similar to some values of n_{O} in the table, implying that it may have been optimal to estimate *some* values of τ^2 (and assuming a given cost function).

8 | OPTIMAL DESIGN FOR STATISTICAL POWER

The previous section considered optimal designs for estimating variance the component τ^2. Another goal might be to find optimal designs for hypothesis tests. One conception of this would consider a design to be optimal if it has the highest statistical power for a given cost. Unlike with variance component estimates, optimal sample sizes that maximize power will depend on the total cost C. Conversely, because the goal is often to obtain high statistical power, it may be useful to examine instead optimal allocations that yield statistical power of 80% to 90%; that is, determine designs that attain a given power but minimize the total cost C. Analytic results for designs that maximize power (for a given cost) or minimize the cost (for a given power) do not seem feasible, but can be obtained numerically.

Optimal designs that minimize the cost for a desired power depend on the ratio of costs c_2/c_1 and the amount heterogeneity τ^2/ω that a researcher wishes to detect. The top panel of Table 2 gives the number of studies m_O and the total sample size per study n_O that yield power of 80% for the lowest total cost as a function of c_2/c_1 and τ^2/ω. The bottom panel of Table 2 gives the values of m_O and n_O that give power of 90% for the lowest total cost. Comparing the upper and lower panels of the table shows that the optimal allocation depends on the statistical power desired and that n_O is larger when the desired level of statistical power is 90% than when it is 80%. As in the case of optimal ensembles for estimating τ^2, the n_O depends strongly on τ^2/ω, varying by more than an order of magnitude from the largest to smallest value of τ^2/ω given in the table. It may be somewhat surprising that the optimal number of studies m_O is typically not large: less than $m = 10$ for 80% power and less than $m = 15$ for 90% power.

The results in the table show that it is possible to design ensembles that produce high power tests for heterogeneity with as few as a few hundred (to detect the largest values of heterogeneity) or several thousand total subjects (to detect smaller values of heterogeneity). The Many Labs Project, for example involved a few thousand total subjects (36 studies with an average sample size of over 100), so the optimally designed ensembles suggested here would seem to be feasible in social psychology.

However, the Many Labs design of $m = 36$ and $n = 80$ do not appear to be optimal for the cost functions considered in this table. For reference, the Many Labs design could detect heterogeneity of $\tau^2/\omega = 0.009$ with 80% power at a cost $C_{ML} = (c_2/c_1\ 36 + 36 * 80)c_1$ (under the linear function). We computed the optimal designs for detecting those same values of τ^2/ω with the same power as the Many Labs design for each value of c_2/c_1 in the table, and determined the cost of those optimal designs $C_O = (c_2/c_1\ m_O + m_O\ n_O)c_1$. The ratio of C_{ML}/C_O for those designs and cost ratios ranged from 1.31 (for $c_2/c_1 = 5$) up to 5.63 (for $c_2/c_1 = 1000$), which means that the same sensitivity could be attained at a fraction of the cost C_{ML}.

Comparing the corresponding values in Tables 1 and 2, we see that the n_O values in Table 2 are generally larger than those in Table 1 (typically almost twice as large). That is, within-study sample sizes that are optimal for minimizing variance of the estimate of τ^2 are smaller than optimal within-study sample sizes for obtaining high statistical power in testing whether $\tau^2 = 0$. This may seem surprising, but comparisons can be a bit misleading because (a) the goals of the two analyses are different, and (b) the relations between n and the statistics used in the two analyses are different, as described in previous sections. Thus, a small number of high precision (large n) studies may be optimal for testing whether $\tau^2 = 0$, but lower precision (smaller n) studies are optimal for estimating moderate values of τ^2.

9 | IMPACT ON OPTIMAL DESIGNS TO MISSPECIFICATION OF τ^2

The hypothesized value of τ^2 plays the same role in power analysis here as does the effect size in power analyses of, for example, cluster randomized experiments. It is clear from Tables 1 and 2 that power will depend strongly on τ^2, and one might expect optimal designs to be especially sensitive to misspecification of τ^2. Some indication of the impact of misspecification of τ^2 on optimal designs for power can be obtained by the optimal designs when τ^2 is under- or overestimated.

Suppose a researcher designs an ensemble to detect some τ^2, but the true value of heterogeneity is ${\tau_0}^2$. If ${\tau_0}^2 < \tau^2$ (i.e. the researcher overestimated ${\tau_0}^2$ in their planning), then the true power of the design will be lower than desired. Alternatively, if ${\tau_0}^2 > \tau^2$ (i.e. the researcher underestimated ${\tau_0}^2$ in their planning), then the power will be higher than necessary, but so will the cost of the ensemble.

Table 3 shows the sensitivity of optimal designs to misspecification of τ^2. Each vertical panel of Table 3 provides information about a configuration of small and large c_2/c_1 values (5 or 100) and small or large τ_0^2 values (0.005 or 0.05). Within each panel, the table shows the impact of over- or underestimating the true heterogeneity τ_0^2 by 50%. For example, the top panel concerns the situation when $c_2/c_1 = 5$ and $\tau_0^2 = 0.005$, where the optimal design has $m_O = 9$ and $n_O = 476$. This panel shows that if one had assumed that the true value of τ^2 was 50% less than it actually was (${\tau_0}^2$), then the optimal design computed would have had also had $m_O = 9$ but $n_O = 951$, a 100% overestimation of n_O, leading to an approximate doubling of cost and a resulting power of 95%. On the other hand, if one had assumed that the true value of τ^2 was 50% greater than it actually was (τ_0^2), then the optimal design computed would have also had $m_O = 9$ but $n_O = 317$, a 33% underestimation of n_O, leading to 33% reduction of cost but a resulting power of only 65%. The table illustrates that, as expected, underestimation τ^2 is conservative in terms of leading to larger power but results in higher costs.

10 | CONCLUSIONS

If empirical evaluations of replication are to continue, the planning of such research programs could be improved by designing ensembles of studies that support sufficiently sensitive analyses while

TABLE 3 Impact on optimal designs for achieving 80% power of misspecification of τ^2. Each horizontal panel shows the optimal designs for a given ratio of per-laboratory to per-subject costs c_2/c_1 and design parameter for the between-study variance τ_0^2. Columns in the left panel show the optimal designs for studying values of τ^2 that are multiples of τ_0^2, including the optimal number of studies m_O and subjects per study n_O. The right panel compares those designs to the design when $\tau^2 = \tau_0^2$.

Values	Comparison[b] to designs with τ_0^2							
τ^2	m_O	n_O	Cost	Power[a]	m_O	n_O	Cost	Power
$c_2/c_1 = 5$, $\tau_0^2 = 0.005$								
$0.5 \times \tau_0^2$	9	951	8604	0.95	100%	200%	199%	119%
τ_0^2	9	476	4329	0.80	—	—	—	—
$1.5 \times \tau_0^2$	9	317	2898	0.65	100%	67%	67%	81%
$c_2/c_1 = 5$, $\tau_0^2 = 0.05$								
$0.5 \times \tau_0^2$	8	108	904	0.95	100%	200%	192%	119%
τ_0^2	8	54	472	0.80	—	—	—	—
$1.5 \times \tau_0^2$	8	36	328	0.66	100%	67%	69%	83%
$c_2/c_1 = 100$, $\tau_0^2 = 0.005$								
$0.5 \times \tau_0^2$	8	1073	9384	0.95	114%	173%	186%	118%
τ_0^2	7	621	5047	0.80	—	—	—	—
$1.5 \times \tau_0^2$	6	497	3582	0.67	86%	80%	71%	84%
$c_2/c_1 = 100$, $\tau_0^2 = 0.05$								
$0.5 \times \tau_0^2$	5	191	1455	0.92	125%	140%	154%	116%
τ_0^2	4	136	944	0.80	—	—	—	—
$1.5 \times \tau_0^2$	4	91	764	0.70	100%	67%	81%	88%

Note: Here τ_0^2 is the actual value.

[a]Power is the power computed with $\tau^2 = \tau_0^2$, but m_O and n_O computed using the value of τ^2 defined for that row (50% smaller, equal to, or 50% larger than τ_0^2).; [b]Values here are computed as a percentage of the values for the designs when $\tau^2 = \tau_0^2$.

efficiently allocating resources. This article provided methods to do this when the mode of analysis is hypothesis testing or variance component estimation. It also demonstrated these methods on the Many Labs design, finding that while it could be argued that the design used was sensitive, it may not have been optimal. It is worth pointing out that the optimal designs depend on prior judgements about what values of τ^2 might be deemed plausible or worth detecting. Thus τ^2 is a design parameter necessary to compute the optimal design. This is much like the situation in which the effect size is a design parameter necessary to compute statistical power in randomized experiments. While design parameters cannot be known exactly in advance of carrying out the experiment, the results of this paper demonstrate that at least rough estimates of their value are important in the design of efficient experiments.

Previous discussion of the power of tests discussed in this article suggested that somewhat larger values of m would be necessary to obtain adequate statistical power (Hedges & Schauer, 2019b). However, those results and the ones presented here are not in conflict. The power of the tests discussed here depends on the metric of τ^2/v. Hedges and Schauer focused on statistical power in relation to conventions for heterogeneity used in empirical sciences, using various conventions for τ^2/v. This approach can be seen as assuming that v is the estimation error variance of 'typical' studies in the field. That is, the conventions are defined in terms of *normative* study designs. Conversely, in his article the sample size within studies, and hence v is affected by design choices. Hence, in the purposeful design

of studies, one may not be constrained by what the *typical* sample size might be, but rather one seeks to determine what the sample size *should be* in order to study meaningful heterogeneity among effects across studies. Consequently, some of our results suggest that adequate designs for testing heterogeneity are possible with far fewer studies provided that those studies have larger sample sizes than is typically required for to detect a non-null effect. For example, if θ is a standardized mean difference and we take $\omega \approx 4$, then a total sample size in a single study of $n = 126$ is adequate to achieve power of 80% to detect an effect of $\theta = 0.5$ (Cohen's medium sized effect). Yet, except for those associated with the largest values of heterogeneity, the within-study sample sizes n in Table 2 are all larger (and sometimes very much larger) than $n = 126$.

The findings presented here can be extended to different analyses, and potentially different designs. This article dealt exclusively with testing hypotheses about exact replication, that is testing the null hypothesis that $\tau^2 = 0$. We have argued previously that, while exact replication is logically appealing, it may be too strict a criterion to be scientifically useful (Hedges, 1987; Hedges & Schauer, 2019b). An alternative is to test for approximate replication, in which a specified small value of τ^2, call it $\tau_N{}^2$, is defined to constitute scientifically negligible heterogeneity, so that the null hypothesis tested is H_0: $\tau^2 \leq \tau_N{}^2$, not $\tau^2 = 0$. Because the test of the hypothesis of approximate replication can be accomplished by using the Q statistic with a different (larger) critical value (which will depend on the value of $\tau_N{}^2$ chosen), the methods used in this paper can be used to determine optimal designs for attaining the desired statistical power to test for approximate replication. Optimal designs to detect the same target value of τ^2, will however, tend to be larger in the case of approximate replication than in the case of exact replication because the corresponding tests are less powerful (see Hedges & Schauer, 2019b).

It is simpler to consider estimation and testing as separate goals for an analysis, but both are often of interest. We showed that the most efficient designs for estimation and testing are somewhat different. One can imagine a hybrid model in which the joint goals of statistical significance testing and precision of estimation are both taken into account. Such a hybrid model would make sense in situations where both estimation and testing are important, and the criterion for optimality would involve a loss function that combined error rates of the test and the variance of the variance component estimate.

The analyses that we considered in this paper involved conventional effect sizes having variances that could be treated as known for the purposes of the analysis. The use of conventional effect size measures is essential if somewhat different instruments measure the outcome in different studies. If the ensemble of studies is planned (and registered in advance), it might be possible to use the same instrument to measure the outcome in every study. If this were the case, the heterogeneity analysis could be conducted in the metric of the raw outcome scores. For example, the effect of interest might be the (untransformed) treatment-control mean difference. If the within-treatment-group-within-study variances were the same across studies, then analysis of variance style methods and a more conventional approach to design could be used. If the within-treatment-group-within-study variances were not the same across studies, different methods would be required (e.g. generalizations of methods like those of James, 1951 or Welch, 1951). It is an open question whether such methods lead to substantially more efficient designs or more powerful significance tests in this situation.

This article dealt with the design of ensembles of studies to assess the replicability of research findings when each laboratory conducts a single study. An alternate design is one in which each laboratory conducts $p > 1$ studies. In this model there are two components of variance: A between-studies-within-laboratories component τ_1^2 and a between laboratory (average) component τ_2^2. These two components of variance are confounded when only one study is conducted per laboratory. In the alternative design there is more than one possible target for estimation or hypothesis testing. One might be interested in τ_1^2, τ_2^2 or some combination like $\tau_1^2 + \tau_2^2$ (the quantity that we call τ^2 in this

article). If $\tau_1^2 > 0$, then efficiency of the design for estimating or testing hypotheses about τ_2^2 can often be improved by conducting more than one experiment in each laboratory ($p > 1$).

ORCID

Jacob M. Schauer https://orcid.org/0000-0002-9041-7082

REFERENCES

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M. et al. (2016) Evaluating replicability of laboratory experiments in economics. *Science*, 351, 1433–1436. https://doi.org/10.1126/science.aaf0918.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M. et al. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z.

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

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

Dickersin, K. (2005) Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In: H.R. Rothstein, A.J. Sutton and M. Borenstein (Eds.) *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: Wiley, pp. 11–33.

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.

Firger, J. (2015) Science's reproducibility problem: 100 psych studies were tested and only half held up. *Newsweek*. Retrieved from http://www.newsweek.com/reproducibility-science-psychology-studies-366744

Gilbert, D.T., King, S.G., Pettigrew, S. & Wilson, T.D. (2016) Comment on "estimating the reproducibility of psychological science". *Science*, 351, 1037–1038.

Hartgerink, C.H.J., Wicherts, J.M. & van Assen, M.A.L.M. (2017) Too good to be false: Nonsignificant results revisited. *Collabra. Psychology*, 3(1), 9. https://doi.org/10.1525/collabra.71.

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. (1982) Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92(2), 490–499. https://doi.org/10.1037/0033-2909.92.2.490.

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

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

Hedges, L.V. & Olkin, I. (1985) *Statistical methods for meta-analysis*. New York: 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. & Pigott, T.D. (2004) The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9, 426–445.

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 methods for studying replication: Meta-analytic perspectives. *Psychological Methods*, 24(5), 557–570.

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.

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

Higgins, J.P. & Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558. https://doi.org/10.1002/sim.1186.

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

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

James, G.S. (1951) The comparison of several groups of observations when the ratios of the variances are unknown. *Biometrika*, 38, 324–329.

Keiding, N. & Louis, T. (2018) Web-based enrollment and other types of self-selection in surveys and studies: Consequences for generalizability. *Annual Review of Statistics and Its Application*, 5, 25–47. https://doi.org/10.1146/annurev-statistics-031017-100127.

Klein, R.A., Ratliff, K.A., Vianello, M., Adams, R.B., Bahník, Š., Bernstein, M.J. et al. (2014) Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), 142–152. https://doi.org/10.1027/1864-9335/a000178.

Klein, R.A., Vianello, M., Hasselman, F., Adams, B.G., Adams, R.B., Alper, S. et al. (2018) Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490. https://doi.org/10.1177/2515245918810225.

Kulinskaya, E., Dollinger, M.B. & Bjørkestøl, K. (2011) Testing for homogeneity in meta-analysis I. The one-parameter case: standardized mean difference. *Biometrics*, 67(1), 203–212. https://doi.org/10.1111/j.1541-0420.2010.01442.x.

Kulinskaya, E., Dollinger, M.B. & Bjørkestøl, K. (2011) On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research Synthesis Methods*, 2(4), 254–270. https://doi.org/10.1002/jrsm.1446.

Let's just try that again. (2016) *The Economist*. Retreived from https://www.economist.com/science-and-technology/2016/02/06/lets-just-try-that-again

Marcus, G. (2013) The crisis in social psychology that isn't. *The New Yorker*. Retrieved from https://www.newyorker.com/tech/elements/the-crisis-in-social-psychology-that-isnt.

Moerbeek, M., van Breukelen, G.J.P. & Berger, M.P.F. (2000) Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, 25(3), 271–284.

Moshontz, H., Campbell, L., Ebersole, C.R., IJzerman, H., Urry, H.L., Forscher, P.S. et al. (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., Agashe, K., Amsler, C., Antonelli, M., Arguin, J.-F., Asner, D.M. et al. (2014) Review of particle properties. *Chinese Physics C*, 38, 090001. https://doi.org/10.1088/1674-1137/38/9/090001.

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) The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, 4(5), 326–334.

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

Perrin, S. (2014) Make mouse studies work. *Nature*, 507, 423–425.

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

Raudenbush, S.W. (1997) Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185.

Raudenbush, S.W. (2009). Analyzing effect sizes: Random-effects models. In Cooper, H., Hedges, L.V. & Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation, pp. 295–315.

Raudenbush, S.W. & Liu, X. (2000) Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199.

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

Rukhin, A. (2000) Approximate entropy for testing randomness. *Journal of Applied Probability*, 37(1), 88–100. https://doi.org/10.1239/jap/1014842270.

Schauer, J.M. & Hedges, L.V. (2020) Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*, 146(8), 701–719.

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(2), 90–100. https://doi.org/10.1037/a0015108.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S.A., Jordan, J., Tierney, W. et al. (2016) The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, 66, 55–67. https://doi.org/10.1016/j.jesp.2015.10.001.

Tipton, E. & Hedges, L.V. (2017). The role of the sample in estimating and explaining treatment effect variation: A commentary on three papers. *Journal of Research on Educational Effectiveness*, 10(4), 903–906.

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

van Erp, S., Verhagen, J., Grasman, R.P.P.P. & Wagenmakers, E.-J. (2017) Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013. *Journal of Open Psychology Data*, 5(1), 4. https://doi.org/10.5334/jopd.33.

Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., et al. (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(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428.

Welch, B.L. (1951) On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330–336.

How to cite this article: Hedges LV, Schauer JM. The design of replication studies. *J R Stat Soc Series A*. 2021;184:868–886. https://doi.org/10.1111/rssa.12688