Assessing Heterogeneity in Replication Research

Jacob M. Schauer and Larry V. Hedges

Department of Statistics
Northwestern University

Abstract

Recent empirical research has questioned the replicability of published findings in psychology. However, analyses of these studies have proceeded with conflicting definitions of what it means for a finding to replicate. In this paper, we use a meta-analytic approach to highlight different ways to define "replication failure," and argue that analyses can focus on exploring variation among replication studies or assess whether their results contradict the findings of the original study. We then apply this framework to experiments that have been subject to systematic replications in psychology. Among these experiments, we find that fewer studies conclusively failed to replicate than previous reporting would suggest. However, this finding must be interpreted with an important caveat: even the most powerful tests for replication failure tend to have low power for these data, which means that for the majority of experiments, these analyses are inconclusive. Further, while common interpretations of the replication crisis in psychology involve underpowered initial studies overestimating an effect, in half of the findings in this data, this is reversed: the original study understates the magnitude of the effect relative to the replications and would have been well powered to detect the effects estimated by the replication studies. We conclude by suggesting that efforts to assess replication would benefit from further methodological work on designing replication studies to ensure analyses are sufficiently sensitive.

Public Significance Statement

This article clarifies subjective notions about "replication" and applies these principles to analyses of existing replication research data. It finds that fewer experiments failed to replicate than previous reporting would suggest. It also finds that most replication research programs are underpowered to detect replication failures.

Assessing Heterogeneity in Replication Research

The idea that experiments can be replicated is fundamental to the logic and rhetoric of science (McNutt, 2014). However, recent empirical evaluations have cast doubt on the replicability of findings in several fields, giving rise to a "replication crisis" in science (Lindsay, 2015; Pashler & Wagenmakers, 2012). This has been particularly acute in psychology, where programs of research investigating the replicability of psychological experiments have suggested startlingly high replication failure rates (e.g., Open Science Collaboration, 2015).

The most appropriate analysis for these types of meta-research programs is not a settled matter. The Open Science Collaboration (2015) notes that, "No single indicator sufficiently describes replication success." Thus, a variety of metrics have been used to determine whether a finding replicates (see Open Science Collaboration, 2015; Camerer, et al., 2016). These have involved methods that rely on *p*-values, compare confidence intervals, and attempt to assess the sensitivity of various experiments (see Schweinsberg et al., 2016). Most of these cannot adequately incorporate more than two studies: an original study and a single replication. In addition, in the face of multiple analyses, which can support contradictory conclusions about replication, it is unclear which has or should be given priority over the others.

The focus on analysis methods is important for several reasons. One is that each analysis depends on some operational definition of what it means for a replication to be successful; what do we mean when we say that a finding replicates? On its face, defining "replication" seems trivial: simply check that studies get the same results. However, based on the analyses conducted by replication research programs, a study's "results" could mean several different things: its effect size, a *p*-value, or statistical significance (see Schauer, 2018). Moreover, each method can lead to erroneous inferences about replication. Conclusions about replication must be interpreted in light of the relevant operational definition, and how likely a method is to make incorrect determinations about replication.

In general, it would seem that there are multiple ways to frame what we mean by replication. For instance, it has been argued that replication is fundamental to the idea that science is self-correcting, since replication failures will help identify spurious findings (see McNutt, 2014). Viewed this way, "replication" means that the results of an original finding are consistent with subsequent replications, though what "consistent" means is similarly ambiguous. If effects in the original and replication studies are both positive or the same size, that may be seen as consistent. In psychology, it has been common for replications to align their protocols and materials with original studies, even obtaining the original authors' input and validation (see Open Science Collaboration, 2015). In that case, we might expect the size of the effects in the original and replication studies to be similar.

Replication is also an important way to identify sources of variation in experimental procedures and results. In some sense, the idea that an experiment is replicable means that if we repeat a procedure, we can expect a certain result (Bollen, et al., 2015). This is a concept that may be further explored when multiple replications are conducted, which has become the modal approach to replication research in psychology (see Simons, Holcombe, & Spellman, 2014). Viewed this way, we might define "replication" as *all* of the studies producing the same effect or roughly the same effect (see Hedges & Schauer, 2018).

The type of replication research in psychology seems relevant to both of these notions of replication. Since replications typically proceed from an original (often published) finding, it would seem natural to determine whether the original study is consistent with the replication(s), which we will call *question 1*. However, since many replication research programs in psychology involve multiple independent, pre-registered studies conducted simultaneously, they offer a way to study the sources and magnitude of variation between studies. Thus, one may also be interested in *question 2*, which concerns whether the results from a series of replications are themselves consistent.

Analysis methods so far have tended to focus on question 1. As a result, nearly all of the methods used in empirical research involve some sort of comparison between the original study and the replication(s). Most of these methods are only appropriate for only two studies. Effects from multi-site replications are often aggregated into a single effect estimate via a meta-analysis in order to make direct comparisons with the original study (e.g., Schweinsberg et al., 2016). Methods that do so often ignore information about replication effects, including their potential heterogeneity, which impacts how we might interpret such analyses.

The ambiguity regarding “replication” and how it is defined in analyses of individual experiments has important implications. First, signals of a replication crisis are built on individual determinations of replication failure or success, such as the Open Science Collaboration’s (2015) determination that 61% of their attempted replications failed. If it is unclear what is meant by “replication” or “failure to replicate,” such statistics are impossible to interpret. Moreover, the emphasis on question 1 in empirical research means that there has been less attention paid to question 2, despite programs such as Many Labs that were designed to investigate heterogeneity among replications (Klein, et al., 2014).

In this paper, we re-analyze the results of several sets of replication studies in psychology from a meta-analytic perspective. Our goal in doing so is not to promote or falsify individual findings, but rather to provide a broader picture of replication and replicability in psychology, and to demonstrate that analyses of replication, and hence signals of a crisis, can depend on how "replication" is operationalized. Thus, this article examines two main questions about an experiment: (1) if the original study in some way differs from the (distribution of) effects found by replication studies, and (2) whether the body of evidence about a finding is consistent (i.e., effects are relatively similar). We discuss how one might formulate and test hypotheses about whether experimental results are similar, and show that these tests can be sensitive to the precise definition of replication. We also provide some idea of the power for these analyses to detect meaningful differences between study results.

Replications in Psychology

The replication crisis, as it is referred to today, gained greater attention in psychology throughout the 2010s. Controversies surrounding failed replication attempts of high profile findings gave way to contentious debate, such as in the cases of Doyen et al.'s (2012) failed attempt to replicate Bargh et al.'s (1996) work on age priming, or Rahehill et al.'s (2015) failed replication of Carney et al.'s (2010) power pose experiments. At the same time, articles addressing potential issues of replicability in psychology emerged highlighting factors such as small sample sizes, publication selection, and suspect research practices (e.g., Francis, 2012; Schmidt & Oh, 2016). These threats to building reproducible scientific knowledge have been at the core of a growing reform movement in the field (Lindsay, 2015; Bollen, et al., 2015; Nosek, et al., 2015).

But perhaps the most important evidence of a crisis has come from programs of meta-research that systematically attempt to replicate scientific findings. Though not the first such programs, the Replication Project: Psychology (RPP) (Open Science Collaboration, 2015) and the Replication Project: Economics (RPE) (Camerer et al., 2016) have been among the most prominent in this discussion. Both of these took a series of experiments and attempted a single replication of each: the RPE involved 18 different experiments in behavioral economics, while the RPP attempted to replicate 100 social and behavioral psychology experiments, 73 of which they identified as a "meta-analytic subset" for which meta-analysis methods would be appropriate. Claims that only 39% of findings in social and behavioral psychology replicated in the RPP have been seemingly ubiquitous, popping up everywhere from *The Atlantic*, to the *Wall Street Journal* (Yong, 2016; Wood & Randall, 2018). However, the criteria used to determine if a finding had been replicated were quickly challenged and alternative methods have been proposed (e.g., Etz & Vandekerckhove, 2016; Hartgerink et al., 2017; van Aert & van Assen, 2016). Indeed, work by Hedges and Schauer (in press) suggests that statistical analyses about replication may be subject to severe limitations when only a single replication study is conducted.

While the RPP and RPE were particularly high profile, it has been more common that replication research in psychology involves conducting multiple independent replication studies. The Many Labs Replication Project recruited 36 labs to conduct the same 16 experiments (Klein, et al., 2014). Somewhat paradoxically, they concluded that despite evidence of heterogeneity among replications for eight experiments, ultimately 14 findings were successfully replicated.

It appears that the Many Labs approach has become something of a norm for replication research in psychology. The same year they published their results, the Association for Psychological Science (APS) announced a program of Registered Replication Reports for replicating published findings (Simons, Holcombe, & Spellman, 2014). Since then, six such reports have published results, and four others are in progress (completed efforts have been published as: Alogna, et al., 2014; Bouwmeester, et al., 2017; Cheung, et al., 2016; Eerland, et al., 2016; Hagger, et al., 2016; Wagenmakers, et al., 2016). Published efforts have each attempted to replicate between one and four experiments (for a total of 13 experiments) and have involved 13 to 33 independent laboratories. These reports have suggested that the replication results have often contradicted published findings (eight out of 13), and only one has reported significant heterogeneity between replication study results. The recently published results of Many Labs 2 (itself part of the Registered Replication Reports) suggested that of the 28 experiments they attempted to replicate, fourteen to fifteen replications found significant effects in the same direction as the original study (Klein, et al., 2018).

A related program was the Pre-Publication Independent Replication (PPIR) project that sought to replicate findings that had yet to be published (Schweinsberg, et al., 2016). They recruited 25 independent research groups to conduct subsets of 11 different experiments, so that each experiment was independently replicated between 12 and 18 times. Like with the Many Labs analyses, PPIR concluded that despite finding evidence of heterogeneous effects among some of the replication studies, all but two findings were successfully replicated.

Finally, forthcoming results are expected of the Psychological Science Accelerator, an international collaboration of psychology laboratories designed to conduct the type of multi-lab replication research that has become more common in psychology. So far, the Accelerator has recruited over 500 laboratories across the world to facilitate large-scale inquiry into the replicability and generalizability of psychology experiments (see Moshontz, et al., 2018).

Just as designs have become somewhat normative, so too (to some extent) have analysis methods. The RPP posited that there is no one way to determine if a replication attempt is successful, and analyzed their data in several, often conflicting ways. The methods used by the RPP included procedures that determined replications failed if:

- they did not correspond in sign and statistical significance with the original study (e.g., the original study was positive and statistically significant and the replication was not),
- the original effect estimate was not contained in a 95% confidence interval of the replication study effect estimate,
- the weighted average of effect estimates from the original and replication studies was not statistically significant.

The RPE followed a similar analysis plan, including an additional analysis called the "prediction interval" that is equivalent to the meta-analytic *Q* test when there are only two studies (see Patil, Peng, & Leek, 2016; Hedges & Schauer, 2018). However, this was not limited to programs involving one-off replications. PPIR and Many Labs used some of the same analyses, aggregating the results of multiple replications into a single estimate, often using models from meta-analysis. These programs included determinations about whether the original and aggregate replication findings were both statistically significant, as have many of the Registered Replication Reports (e.g., Klein et al., 2018).

At the same time, additional analysis methods have been proposed and applied to replication research data. For instance, Etz and Vandekerckhove (2016) propose a Bayes factor analysis comparing the original and replication studies that corrects for publication bias. van Aert and van Assen (2017, 2018) describe Bayesian approaches for examining agreement between the original and replication results, and Hartgerink et al. (2017) apply Fisher's method to non-significant *p*-values in replications to detect false negatives.

It is worth pointing out two ways in which interpretations of replication research programs' reported results can be difficult. The first is that any analysis method inherently relies on some operational definition of replication, and that definition should frame the interpretation of analytic results. However, researchers rarely define replication in a statistical way, but rather apply analysis methods and present their own interpretations, leaving the reader to infer the relevant operational definition. Moreover, the different analyses that have been used (even in the same paper) can imply different definitions of replication. Consider the most common approach to assessing replication, which is to check whether the original study and the replication (or the average of several replications) correspond in sign and statistical significance (e.g., both are positive and significant). This relies on a definition of replication that requires effects to be in the same direction, for example both effects are positive. Yet, this definition puts no bound on how different effects can be. For instance, an original study that finds an effect of $d = 20$ and a replication that finds an effect of $d = 0.2$ would be considered successful so long as they are both statistically significant. Conversely, the prediction interval approach (Patil et al., 2017) used by the PPIR implies that successful replication requires effects to be the same size; that is, the original and replication effects are the same, rather than merely both being positive.

Second, it is possible to that determinations about replication based on these methods may arise in error. This is particularly true of the statistical significance criterion, for which a failed replication might involve an initial finding that is statistically significant (and positive), but a replication that is not statistically significant. Implicitly, this is taken to mean that an effect is positive in one study but is not in another. Yet, this conclusion rests on interpreting the nonsignificant result as proof of the null hypothesis, which is a logical fallacy. Put another way, concluding that a replication failed can involve misinterpreting a failure to reject the null hypothesis.

More generally, erroneous conclusions about replication may arise simply because of random variation (e.g., in effect estimates). The probability of making such errors can be analogous to the Type I and Type II error rates in null hypothesis tests. Interpretation of published results about replication, then, must take into account the error rates of the procedures used. In other words, while a statistic like the 61% replication failure rate in psychology seems to have grabbed headlines, this number is without context: we do not know precisely how conflicts among operational definitions were resolved, nor is there a full appraisal of the sensitivity of the methods that produced that number.

Research Questions and Types of Replications

In this paper we are concerned with replicability rather than reproducibility. Similar to the guidance of Bollen et al. (2015), we refer to replication studies as independent runs of the same experiment. Scientists have long noted that there are different types of replication studies that support different types of research questions. Several researchers have proposed taxonomies of types of replication studies (Anderson & Maxwell, 2016; Bahr et al., 1983; Lykken, 1968; Valentine et al., 2011). Schmidt (2009) argues that these taxonomies largely make a distinction between *direct* and *conceptual* replications. In direct replications, studies are designed to be identical: from the experimental protocol, to materials and instrumentation, to (if possible) the experimental units themselves. The goal of direct replications is often to obtain results that are, in some sense, the same.

Conceptual replications, meanwhile, involve studies that differ in some way, such as their protocol or sample composition. Such differences can be deliberate in order to evaluate potential sources of experimental variation (e.g., White et al., 2014). While the distinction between direct and conceptual replications is clear-cut in theory, it may be more difficult to make in practice. Important differences between studies may be unknown to researchers; tacit knowledge about seemingly innocuous (and so undocumented) details have more than once marked the difference between successful and unsuccessful replication attempts in various fields of science (Collins, 1992).

When multiple replications are conducted across different labs, the question of direct versus conceptual is twofold. First, it could refer to only the replication studies (excluding the original study): they can be designed to be identical to each other or to vary in known ways. Second, it could refer to whether or not replications are designed to be identical to the original experiment.

With a few exceptions, the studies used in this article can be thought of as direct (or at least attempts at direct) replications. One of the hallmarks of replication research in psychology so far is that great care is taken to ensure replications are as similar as possible (see Open Science Collaboration 2012, 2015). All of the multi-site replication programs used standardized materials, protocols, and measurements. Further, most replication efforts sought to synchronize (to the extent possible) their protocol with that of the original experiment. Most of these efforts required consulting with and obtaining the approval of the original authors, which in some cases allowed those carrying out the replications to use the original experiment’s materials (see Alogna et al., 2015; Bouwmeester, et al., 2017; Cheung et al., 2016; Eerland et al., 2016; Hagger et al., 2016; Schweinsberg et al., 2016; Wagenmakers et al., 2016). This was also true for the RPP and RPE, which made such consultations and approvals a required step prior to pre-registration.

Operational Definitions of Replication

On its face, assessing replication seems simple: just repeat an experiment and check that you get the same results. However, there are at least two aspects of this that have proven difficult to nail down in practice. First, analyses of past replication efforts would appear to imply different possible notions of "results" being "the same." As discussed above, this could involve effects that are the same size, or merely in the same direction. Second, it is not necessarily clear what it means to say an experiment "replicated" or "is replicable," particularly when it comes to multi-site replications. It could mean that across all studies, the results are relatively consistent. Alternatively, it could refer to whether the results of the original study agree with those of the replication studies. These two conceptions are focused on slightly different questions, and both can shed light on the replicability of a finding. What would it mean, for instance, if an original study was consistent with the replications, but the replications varied so much that many warranted divergent scientific interpretations?

In this section, we attempt to clarify both of these issues. To address the first, we use a common model in meta-analysis. Suppose there are *k* studies, and denote the parameter θ_i as the effect for study *i*. This is what we would observe if experiments had perfect precision. Instead, we observe an estimate T_i, which has variance v_i. A common assumption in meta-analysis is that T_i is normally distributed with mean θ_i and known variance v_i.

$$T_i \sim \mathrm{N}(\theta_i, v_i), \quad i = 1, \ldots, k$$

This is a very accurate approximation for some effect sizes, such as *z*-transformed correlations, and a very good approximation for others, such as standardized mean differences (see Cooper, Hedges, & Valentine, 2009).

While research programs of been equivocal about what exactly a study's result is (i.e., *p*-value, effect size, statistical significance, etc.), in meta-analysis a study's result is its effect parameter θ_1. It is what would be observed in the absence of any estimation error. Thus, one way to frame the definition of replication is that the θ_i are the same size. Differences between them arise from differences between experiments, rather than variation due to sampling. For instance, deviations in experimental contexts or the populations studied may lead to variation among the θ_i. Sources of variation across experimental results can be documented, or there may be hidden moderators of effects.

This article defines replication as when the effect parameters θ_i are similar in value. This is not the only way to define replication; for instance, it can be argued that a successful replication would mean effect parameters agree qualitatively in that they are in the same direction (e.g., $\theta_i > 0$). However, we focus on the similarity of effect sizes for a few reasons. First, experiments conducted in the same way, with the same materials, on individuals from the same population should involve the same effect parameter (see Steiner and Wong, 2018 for a causal inference perspective). Second, just because effects are in the same direction does not necessarily mean they have the same interpretation, as our example in the previous section suggests ($d = 0.2$ versus $d = 20$). Finally, guidance from various scientific bodies, including the APA's *Publication Manual* and Journal Article Reporting Standards and the American Statistical Association (Wasserstein & Lazar, 2016), emphasize that scientific interpretations of studies should focus on effect sizes, rather than just statistical significance (which implies just interpreting the direction of the effect). We would argue that defining replication in terms of the similarity of effect sizes is consistent with this guidance.

Within this model, there may be more than one focus of analysis. One approach would be to determine if replications of an experiment produce consistent results (question 2). In this case, analyses are about how similar all of the θ_i are to each other. Researchers have also been interested in whether the original finding is corroborated or contradicted by what is found by the replications. Here, the original effect θ_{orig} would be compared to the distribution of replication effects (question 1). Multi-site replication designs can provide insight into both of these questions. They can assess whether the replication studies are consistent the initial finding (i.e., the original finding was replicated), and if replication studies produce similar results (i.e., a finding replicates).

Regardless of the focus of analyses, defining replication in the meta-analytic context depends on two additional important theoretical considerations, particularly when it comes to assessing the consistency of results across replications. The first is whether the studies are treated as fixed or random. If the studies are treated as fixed, then the θ_i are treated as fixed, but unknown constants (Hedges & Olkin, 1985; Laird & Mosteller, 1990). In this case, inferences about replication are inferences about how similar the results of the observed studies are. An alternative is to treat the θ_i as draws from the same distribution or population, which is equivalent to a random effects meta-analysis model (Hedges & Vevea, 1998). This means that inferences about replication pertain to the distribution of study results, and not just to those of the observed studies. A more extensive discussion of this distinction is available in Hedges and Schauer (2018) and Schauer (2018).

In this paper, we use both models. To test hypotheses about whether a completed ensemble of studies produced consistent results (i.e., they replicated successfully), we use the fixed studies approach. Empirical evaluations of replication have focused to a certain extent on the results of observed studies (see Schweinsberg et al., 2016). Moreover, this framework provides more powerful tests; they will be more sensitive to smaller differences among the observed studies. This paper also assesses the magnitude and potential sources of variation in effects across replications, as well as whether original findings are consistent with the distribution of effects found by the replications. In these analyses, we rely on standard random effects models.

The other consideration is whether replication is considered *exact* or *approximate*. A logically appealing definition of replication requires all of the effect parameters to be equal ($\theta_1 = \cdots = \theta_k$), which we might call *exact replication*. However, this may too stringent in practice for a few reasons. First, Wong and Steiner (2018) spell out the conditions necessary for exact replication, which are a high bar. In hard sciences, like physics, there is an understanding that even in the case of strong theory and sound scientific practice that there might be some slight variation among experiments deemed to have replicated (Hedges, 1987; Henrion & Fischhoff, 1986). Second, effects that differ slightly in size may still have the same scientific or clinical interpretation. Thus, a more practical definition might take into account approximate replication, where the θ_i are "almost the same." The following sections describe ways to operationalize this more precisely.

Testing for Replication: Are Results Consistent?

In this section, we describe tests for whether the effects from replications are roughly the same. Here, when we refer to "replication" we mean that effects found in direct replications are consistent with each other. To test hypotheses about replication, Hedges and Schauer (2018) propose an adaptation to the *Q* test, a standard meta-analytic tool to test for differences between studies. The properties of the standard meta-analytic *Q* test, including its power, were first studied by Hedges and Pigott (2001, 2004), and Jackson (2005) later derived similar equations. These tests involve computing the *Q* statistic:

$$Q = \sum_{i=1}^{k} \frac{(T_i - \overline{T}.)^2}{v_i} \qquad (1)$$

where $\overline{T}. = \left(\sum_{i=1}^{k} T_i / v_i\right) / \left(\sum_{i=1}^{k} 1 / v_i\right)$. Under the model, Q has a noncentral chi-squared distribution with $k - 1$ degrees of freedom and noncentrality parameter λ:

$$\lambda = \sum_{i=1}^{k} \frac{(\theta_i - \overline{\theta}.)^2}{v_i} \tag{2}$$

where $\overline{\theta}. = \left(\sum_{i=1}^{k} \theta_i / v_i\right) / \left(\sum_{i=1}^{k} 1 / v_i\right)$ (see Hedges and Pigott, 2001). Note that when $\theta_1 = \cdots = \theta_k$, so that the studies replicate exactly, then $\lambda = 0$, and Q has a central chi-squared distribution with $k - 1$ degrees of freedom.

Hedges and Schauer (2018) argue that potential definitions of replication can be operationalized explicitly in terms of λ. Larger values of λ are associated with greater differences across studies, and the following section shows that $\lambda/(k - 1)$ can be interpreted as the ratio of between-study differences to within-study variance, and hence the quantity $1 + \lambda/(k - 1)$ is similar in scale to the H^2 statistic (Higgins & Thompson, 2002). Thus, while the parameter of this test is more intuitive in terms of $\lambda/(k - 1)$, for simplicity of notation we describe it here in terms of λ. Let λ_0 denote a value of λ that corresponds to a specific definition of replication. If $\lambda_0 = 0$, then this would refer to exact replication, and $\lambda_0 > 0$ would refer to approximate replication. Then a null hypothesis that the studies replicate can be written as:

$$\mathrm{H}_0\colon \lambda \leq \lambda_0 \tag{3}$$

An α-level test proceeds by computing Q as in equation (1), and comparing it to the critical value c_α:

$$c_\alpha(\lambda_0) = F^{-1}(1 - \alpha | k - 1, \lambda_0) \tag{4}$$

where $F(x \mid a, b)$ is the noncentral chi-squared distribution function with a degrees of freedom and noncentrality parameter b. Note that we write $c_\alpha(\lambda_0)$ since the critical value depends on the value of λ_0 defining the null hypothesis.

When $\lambda_0 = 0$, this corresponds to a test of exact replication, which is equivalent to the standard Q test in meta-analysis, and c_α is the $1 - \alpha$ quantile of the central chi-squared distribution with $k - 1$ degrees of freedom. When $\lambda_0 > 0$, this is a test of approximate replication, and c_α is the $1 - \alpha$ quantile of the noncentral chi-squared distribution with $k - 1$ degrees of freedom and noncentrality parameter λ_0.

The power of this test is given by

$$\pi(\lambda) = 1 - F(c_\alpha(\lambda_0)|k - 1, \lambda) \tag{5}$$

where $1 - \alpha$ and F are described in equation (4). For a given α, c_α is an increasing function of λ_0. Therefore, testing looser notions of approximate replication (i.e., larger λ_0) requires larger values of Q to reject the null hypothesis. Thus, the test for approximate replication will be less powerful for larger values of λ_0 than smaller ones.

Note that failure to reject the null hypothesis for the tests described in this section does not mean that the studies successfully replicate. A failure to reject the null hypothesis might arise when the studies successfully replicate, or it could happen if when the studies fail to replicate but the test has low power. Using equations (4) and (5), we can assess the sensitivity of a given meta-research program in the context of these tests. One way is to specify some value of λ that would be worth detecting and computing the power of the program to detect it. Alternatively, we can consider the smallest amount of heterogeneity the tests above might be suitably powered to detect, called the minimally detectable heterogeneity (MDH). Computing the MDH involves setting the desired power π, level α, and null hypothesis λ_0 and solving equation (5) for λ.

When multiple replication studies have been conducted, the analyses described above can be seen as focusing on whether all studies obtained the same results (question 2). Such analyses may include or exclude the original published finding. There are various reasons why one may exclude the original finding from these analyses, including concerns over publication selection. There is considerable empirical evidence that statistically significant results are more likely to be published, and such selection can violate the assumptions of the tests above (see Dickersin, 2005; Rothstein et al., 2005). There have been various proposed corrections for publication bias, including maximum likelihood estimates (Hedges, 1984; McShane et al., 2016). Often, these corrections require the data from the original study. Thus, if one suspects that the original study was subjected to publication selection they can either attempt to correct that bias (if possible), or they may opt to exclude it.

Magnitude and Scale of Heterogeneity

Conducting hypothesis tests for replication and computing their power will depend on the noncentrality parameter through the values of λ_0 and λ. One way to gain insight into the scale of λ (and hence λ_0) is that when all of the estimation error variances are the same, so that $v_1 = \ldots = v_k = v$, then λ can be expressed as:

$$\lambda = \frac{k-1}{v} \sum_{i=1}^{k} \frac{(\theta_i - \bar{\theta}.)^2}{k-1} = (k-1)\frac{\tau^2}{v} \tag{6}$$

where τ^2 is a descriptive statistic akin to the variance of the θ_i. In other words, $\lambda/(k-1)$ is roughly the ratio of between- to within-study variation. Note that this holds even if the v_i are unequal, so long as they are not too different. In that case, heterogeneity is categorized in terms of τ^2/v where v is the typical estimation error variance. In this article, we follow the guidance of Higgins and Thompson (2002, Eq. 9) for defining the "typical" sampling variance v.

The scale of $\lambda/(k-1)$ is a natural scale in meta-analysis, as various metrics can be seen as depending on the ratio of τ^2/v (Higgins & Thompson, 2002). For instance, $H^2 = Q/(k-1)$ is an estimate of $1 + \tau^2/v$, and $(1/I^2 - 1)^{-1}$ can be interpreted on the same scale as τ^2/v. Thus, the tests described in the previous section can be conducted and interpreted on this common scale. However, precisely which values of on this scale might be considered negligible or worth detecting will depend on scientific judgement.

In this article, we compare the heterogeneity in replication studies to benchmarks in meta-analyses from different scientific disciplines. Three fields have expressed ideas of negligible heterogeneity that can be interpreted on the scale of τ^2/v. In high energy physics, the Particle Data Group (PDG), which has been conducting systematic reviews on physical constants for the past 50 years, has suggested that a Birge ratio of $H^2 = Q/(k-1) \leq 1.25$ could be considered negligible (Olive, 2014). This means that $\tau^2/v \leq 1/4$ would be negligible. In personnel psychology, Hunter & Schmidt (2004) propose a rule wherein if v is 75% of the total variation $\tau^2 + v$, then the between-study variation could be considered negligible. This corresponds to negligible heterogeneity of $\tau^2/v \leq 1/3$. Finally, in medicine, an $I^2 \leq 0.4$ is considered "not important" (Higgins & Green, 2008). This would imply that $\tau^2/v \leq 2/3$ would characterize negligible heterogeneity. Thus, in this paper, we conduct analyses to assess exact replication and approximate replication as operationalized by these three conventions of negligible heterogeneity, so that $\lambda_0 = 0$, $(k-1)/4$, $(k-1)/3$, and $2(k-1)/3$.

Finally, it is worth noting that λ, like the H^2 and I^2 statistics, quantifies heterogeneity τ^2 relative to the within-study estimation error variance v. Thus, it can be sensitive to the value of v, which itself tends to decrease as the sample size within studies n increases (i.e., where n is the number of participants in *each* study). Thus, the scale of λ depends on the within-study sample sizes, which means that comparing values of λ across replications of different experiments is not the same as comparing values of τ^2 (see Borenstein, et al., 2017).

Data

This paper re-analyzes data from and assesses the sensitivity of several research programs: RPP's "meta-analytic subset" of 73 findings, RPE, Many Labs, PPIR, and six Registered Replication Reports. Data from these programs are available online at the Open Science Framework (https://osf.io), and effect sizes were computed from this data, often using the researchers' own code. All effect sizes are on the scale of (bias-corrected) standardized mean differences (Cohen's *d*). Our analytic code and data are included as an online supplement to this article. Summary information about these programs are available in Table 1, which shows how many experiments each program attempted to replicate, how many times those experiments were conducted, and how many of those replication attempts were deemed to have failed by the authors, as well as summary results of analyses presented in the following sections.[1]

Results

In each of the sections that follow, we present the results of analyses of replication. The first section focuses solely on the RPP and RPE, which both involve only $k = 2$ studies, and so the test for replication reduces to a test of a difference in normal means. For larger ($k > 2$) ensembles, we conduct tests for exact and approximate replication, and examine the effect of a few study-level moderators. Then, we explore potential discrepancies between the initial published finding and the subsequent (pre-registered) replications. Finally, we assess the magnitude of heterogeneity among only the pre-registered replication studies in order to gain some insight into the amount of variation that might be expected in replications in psychology.

[1] Note that the failed replication rate for the RPP in Table 1 is computed from the 73 experiments in the meta-analytic subset. This differs from the 61% failure rate that is widely attributed to that program, which was computed on the entire 100 studies that the RPP attempted to replicate.

RPP and RPE

The RPP and RPE replication efforts involve pairs of $k = 2$ studies: an initial finding and a single replication study. When only $k = 2$ studies are involved, the focus of the replication study, and hence the analysis, would seem to be on falsifying the original finding (study 1). For $k = 2$ studies, the Q test reduces to a test for differences between the two effect parameters, akin to a test of the difference of normal means. Tests for exact replication ($\lambda_0 = 0$) in this case are identical to the prediction interval analysis method that has been used in some replication research programs (see Patil, Peng, & Leek, 2017). Such tests will only be conclusive when they determine that the replication failed. Part of this is due to the structure of the null hypothesis test, but as Hedges and Schauer (in press) point out, these tests are also bound to have poor statistical power. Thus, while this section reports when these tests concluded that a replication failed and whether that differed from the determination made by the original authors, the primary purpose of this section is to assess just how insensitive these tests can be.

For the RPP, the Q test concluded that that 22 of 73 (30%) studies failed to replicate exactly ($\lambda_0 = 0$), and that between 11 and 17 (15%–23%) did not replicate approximately ($\lambda_0 =$ 1/4, 1/3, and 2/3). Two of the failed replications according to the Q test were actually determined to be successes by the RPP (Larsen & McKibban, 2008; and Halevy, Bornstein, & Sagiv, 2008). Both study pairs exhibited effects that differed by about $d = 0.7$, or on the order of (Cohen's) large effect, however both were significant. While they failed the RPP's confidence interval criterion, the RPP concluded that they successfully replicated according to two other criteria. For the RPE, which actually conducted an equivalent analysis, three of 18 (17%) studies were determined to have not replicated exactly or approximately. These determinations were in line with the RPE conclusions about these studies.

While these results seem more optimistic, we must reiterate that for the lion's share of experiments, the Q test was inconclusive. Just because the test does not conclude that a replication failed does not mean that it necessarily succeeded. Failure to reject the null hypotheses is inherently ambiguous and must be interpreted in light of the sensitivity of the test. Recall that we can compute the MDH, the smallest value of λ that a set of studies was well powered to detect in a given the null hypothesis test. For $k = 2$ studies, λ can be expressed as $|\theta_1 - \theta_2| = \sqrt{\lambda(v_1 + v_2)}$, which means we can compute the smallest difference between effects that the tests would have had 80% power to detect.

Figure 1 shows the distribution of MDH (computed for level $\alpha = 0.05$ and power $\pi = 0.8$) on the scale of $|\theta_1 - \theta_2|$ for both the RPP and RPE. For these programs, it shows the MDH for the four hypothesis tests ($\lambda_0 = 0$, 1/4, 1/3, and 2/3). Note that the bulk of the RPP studies were well powered only to detect effect differences greater than $d = 1.0$. Although the RPE had somewhat smaller median MDH values, they were all greater than $d = 0.8$, which on this scale would seemingly be a large difference. Put another way, these studies were only well powered for scenarios where one study had a very large effect ($d > 1$) and the other effect was zero; or where both effects were moderate ($|d| = 0.5$) but in different directions. Moreover, most studies had less than 50% power to detect a difference of 0.5.

The figure also shows that as we incorporate less stringent definitions of replication (larger λ_0), the sensitivity of the test gets worse. The median MDH increases by about 0.3 in Cohen's d units moving from exact to approximate replication. Thus, the test for exact replication will be the most sensitive to smaller differences between studies.

[INSERT FIGURE 1 HERE]

This low power is not necessarily the result of an inappropriate analysis method. The *Q* test is the uniformly most powerful test of this null hypothesis, which means that no other test would be more powerful. Rather, this has more to do with the design of conducting only a single additional study to assess replication. The power of *Q* test for $k = 2$ studies will be limited by the power of the original study to detect an effect (Hedges & Schauer, in press). Unless the original study has high power, then it may be impossible to design a single replication to detect meaningful differences in their effects. The few studies in the data for which the analyses were more sensitive had initial published findings with a large sample size. For instance, the replication based on Ranganath and Nosek (2008), was the only finding that had 80% power to detect a difference as small as 0.5 in a test of exact replication. The original experiment had a sample size of 564 and the replication involved 3,597 participants.

One might be tempted to think that programs such as Many Labs or PPIR, which aggregated replicates into a single estimate, might be better powered to detect meaningful differences. The idea is that since the replication effect estimate pools information across experiments, its sampling variance will be small, and thus the design can detect smaller differences between the original and replication effects. However, the power of the *Q* test for $k = 2$ studies (even if one study is a synthetic effect derived by combing many effects) is determined by the uncertainty of the most uncertain of the two effects (typically the original study). Most of these larger ensembles were only well powered to detect a difference between the original and replicate effects on the order of $d = 0.5$.

Multi-lab Replication Programs

A more common design in replication research in psychology involves multiple ($k > 2$) labs independently conducting the same experiment, which was the design used by the Many Labs project, PPIR, and the APS's Registered Replication Reports. For these programs, we can test null hypotheses of exact and approximate replication as operationalized by four values of λ_0: 0, $(k - 1)/4$, $(k - 1)/3$, and $2(k - 1) /3$. At the insistence of reviewers, we have excluded the original published findings from these analyses in order to ensure they are unaffected by publication selection. Summary results from these tests are shown in Table 1, and Table 2 shows results for individual findings, including values of Q, p-values for each test, and the MDH for each test, as computed for level $\alpha = 0.05$ and power $\pi = 0.8$. Findings that reject the null hypothesis and conclude that the studies do not replicate are highlighted in gray. The MDH values in Table 2 are reported on the scale of τ^2/v.

Making precise statements about which findings do or do not replicate based on Table 2 will depend on which studies are considered, and how we account for multiple comparisons. However, reading the test results panel more heuristically, it highlights two important aspects about tests for replication. First, it shows that determinations about replication can be sensitive to what values of λ are considered negligible. Larger values of λ_0 correspond to definitions of replication wherein study results may exhibit greater heterogeneity (and still be considered successful replications), and tests may be less likely to rule out successful replication. Thus, for these tests, it is important to specify *a priori* how large a difference between studies can still be considered negligible.

The results panel also shows that psychology studies designed to be direct replications of each other can obtain different results. Indeed, the Q test indicates that between 11 and 15 ensembles of replications (27.5%–37.5%) produced heterogeneous effects, and depending on how much heterogeneity one considers negligible, these could be seen as having inconsistent results. In particular, the four Many Labs experiments on anchoring tend to exhibit substantial heterogeneity. These experiments involved tasks where participants estimated certain quantities, such as the number of babies born in a single day in the US, after being given "anchor" values of these quantities that were clearly too large or too small (see Klein et al., 2014). For each of these experiments, the effect of the anchor values appears to vary substantially across direct replications.

Likewise, most of the PPIR studies seem to give rise to variable results. In fact, for 11 studies, we can conclude that the heterogeneity is at least as large as two-thirds the sampling variance ($H^2 \geq 1.67$, $I^2 \geq 40\%$). These studies involve predictions of participants' moral judgements, including how notions of morality may affect perceptions of economic processes (see Schweinsberg, et al., 2016). The variation among results of these studies may be due in part to the nature of the PPIR, which attempted to replicate experiments that had not been published, and whose procedures were still "in the pipeline." Thus, the greater amount of heterogeneity exhibited by the PPIR studies may speak to the fact that when experimental procedures are still under development it can be difficult to control sources of variation between laboratories.

These analyses do not include the original findings over concerns of potential publication bias. Thus, one might expect that the full ensemble of studies (including the original study) may be more heterogeneous than merely the pre-registered replications (excluding the original study). While we investigate this more fully in a later section, we would note that including the original study only changes the results of Q tests for two experiments: the ‘Math/Art/Gender’ experiment from Many Labs, and the ‘Intentionality’ experiment replicated by Eerland et al. For the former, including the original study would lead us to reject the null hypothesis that the studies replicated exactly ($p = 0.04$), however the amount of heterogeneity is roughly the same whether we include the original finding ($H^2 = 1.47$) or not ($H^2 = 1.42$). For the latter, there is a substantial drop in heterogeneity ($H^2 = 2.34$ with the original study, and $H^2 = 1.69$ without it), and the test that includes the original study rejects null hypotheses that the studies replicate exactly ($p = 0.01$) or approximately: $p = 0.03$ for $\lambda = (k - 1)/4$ and $p = 0.04$ for $\lambda = (k - 1)/3$.

[INSERT TABLE 2 HERE]

For most findings, we do not reject the null hypothesis of the Q tests. However, this does not mean we can conclude that the studies successfully replicate. As in the case with $k = 2$ studies, failure to reject the null hypothesis of replication (either exact or approximate) may happen because the studies do actually replicate, or because studies failed to replicate, but the test did not detect that failure because it had low power. The “Sensitivity” panel of Table 2 gives some idea of the power of the tests performed. It shows the MDH that could be detected with 80% power at level $\alpha = 0.05$ for each ensemble of studies and null hypothesis; values are reported on the metric of τ^2/v. Various conventions for negligible values of τ^2/v in meta-analysis range from 1/4 to 2/3 ($H^2 = 1.25$ to 1.67; $I^2 = 20\%$ to 40%), and Hedges and Schauer (2018) argue that ratios larger than 0.75 or 1.0 would likely be worth detecting. However, fewer than half of the ensembles were well powered to detect this in a test of exact replication ($\lambda_0 = 0$). None of the ensembles could detect this level of heterogeneity in tests of approximate replication. In fact, for tests of the most stringent definition of approximate replication, $\lambda_0 = (k -$

1)/4, most ensembles could detect heterogeneity on the order of $\tau^2/v = 1$ with only about 50% power.

Potential Moderators

Table 2 suggests that replication results may exhibit some heterogeneity. However, this may not be entirely surprising for at least two reasons. The first is that the original studies may differ in some way from the replications, either in protocol or due to selection bias. The second is that for some of the replication ensembles, experimental contexts varied across studies in known ways. With the Many Labs Project, while all experiments were computer-based, some were conducted in a lab, and some were conducted online. Moreover, some labs were located in the US and some were not, and differences in cultures may have led to differences in study results. PPIR made note of the same factors, as well as if study samples were comprised primarily of university students.

That experiments were not entirely identical means that the heterogeneity found in Table 2 could be due to these known differences between studies. To assess this, we can group Many Labs and PPIR studies according to these observed covariates and conduct tests for residual heterogeneity as described by Hedges (1982) and Hedges and Pigott (2004). These tests assume there are p groups and m_i studies in group i, and denote θ_{ij}, T_{ij}, and v_{ij} as the parameter, estimate, and variance of the jth study in the ith group. Note that the total number of studies is $k = \sum m_i$. To test the null hypothesis that for each group i, that $\theta_{i1} = \cdots = \theta_{im_i}$ compute the statistic Q_E:

$$Q_E = \sum_{i=1}^{p} \sum_{j=1}^{m_i} \frac{(T_{ij} - \overline{T}_{i \cdot})^2}{v_{ij}} \tag{8}$$

where $\overline{T}_{i\cdot} = \left(\sum_{j=1}^{m_i} T_{ij} / v_{ij}\right) / \left(\sum_{j=1}^{m_i} 1 / v_{ij}\right)$ is the weighted mean effect within group *i*. If the studies replicate exactly within groups, then Q_E follows a chi-squared distribution with $k - p$ degrees of freedom. Note that these analyses exclude the original study, as in the previous section.

In the data, study results for ten experiments appeared to depend on whether they were conducted online or in a lab, in the US or abroad, or if the sample was primarily university students. Table 3 shows which findings were moderated by which covariate, their moderated and unadjusted statistics Q_E and Q, and *p*-values for the for tests of exact replication. It also reports the difference in the average effect across subgroups; for example, in PPIR's 'Bad Tipper' replications, studies conducted in the US found effects that were on average about 0.42 larger than those that were not. What we see is that while residual heterogeneity appears to decrease (i.e., $Q_E < Q$), actual determinations about the existence of heterogeneity did not. In other words, studies that exhibited heterogeneity in the Q test still exhibited heterogeneity even after controlling for the study-level moderators we considered.

[INSERT TABLE 3 HERE]

For some of the studies in Table 3, the country in which experiments were conducted appeared to matter. For instance, PPIR's 'Bad Tipper' experiment in which participants compare a person who leaves a full tip at a restaurant in pennies versus one who leaves a smaller tip in bills was moderated by whether the study was conducted in the US. Similarly, for PPIR's 'Bigot-Misanthrope' experiment, where participants compare a manager who mistreats minority employees to a manager who mistreats *all* of their employees. While the results in Table 3 for both of these experiments seemingly point to cultural context of the morality judgements involved, they may also be an artifact of how difficult it is to translate the materials of such experiments into a different language.

Initial Published Findings

While analyses so far have focused on whether replications of an experiment obtain similar results (i.e., does this experiment replicate?), another goal of conducting replications is to assess whether those replications corroborate an initial finding (i.e., was this finding replicated?). There are at least two reasons why the initial finding may differ from the replication results. The first is that the standardized protocols used among replication studies may differ in potentially important ways from the procedures used in the initial experiment. Some of these differences may be known to researchers, as with the adaptations made by Wagenmakers et al. or Hagger et al. in their replication reports, but that may not always be the case. These research programs demonstrate that replication attempts in psychology involve a sort of translation; they must take the original study and determine which components in that experiment are required to reproduce it, and how those components can be standardized across labs (Open Science Collaboration, 2012; Klein et al., 2014). This is a difficult process that, even with strong theory of methodology and causal mechanisms, can result in a replication protocol that differs from the original experiment. Indeed, past efforts to do this in different scientific fields have often run into bits of tacit (and undocumented) knowledge that were at one point considered innocuous, but that turned out to be the difference between a successful and failed replication (for examples, see Collins, 1992).

Another reason to focus on the initial findings is that most were not pre-registered, and most were published. Thus, they could be subject to some sort of publication selection, or to potentially suspect research practices (e.g., *p*-hacking). Both can induce bias in the initial effect size estimate (see Hedges, 1984). Since the replication studies in the data were pre-registered, these factors are not likely to affect them. Thus, the initial result may disagree with those of the replication studies both because of differences in experimental procedures, or because of the vagaries of research and publication without pre-registration.

Determining whether the initial study is consistent with the replications must contend with the fact that there may be variation among the replication results themselves. To illustrate this, suppose the effects found in the replications (excluding the original study) are not identical. The effect in the original study may be different from the average of those replication effects, but it could still be in line with their distribution. Thus, in this section, we assume a random effects model among the replication studies, and denote the variance of the random effects as τ^2.

To assess the extent to which initial findings may be incongruous with the replications, we examine their externally standardized residuals (Hedges & Olkin, 1985; Viechtbauer & Cheung, 2010):

$$r_i = \frac{T_i - \bar{T}_{\cdot(i)}}{\sqrt{v_i + \hat{\tau}^2 + v_{\cdot(i)}}}$$

where $\bar{T}_{\cdot(i)} = \left(\sum_{j \neq i} T_j/(v_j + \hat{\tau}^2)\right) / \left(\sum_{j \neq i} 1/(v_j + \hat{\tau}^2)\right)$ is the weighted mean effect size excluding study i, $v_{\cdot(i)} = \left(\sum_{j \neq i} 1/(v_j + \hat{\tau}^2)\right)^{-1}$ is its variance, and $\hat{\tau}^2$is the estimated between-studies variance. For the eight results that were moderated by study-level covariates, residuals were computed within groups as delineated by those covariates. The externally standardized residual can be seen as a comparison of the *i*th effect parameter θ_i and the distribution of the effect parameters from the other studies. Assuming that all of the studies in an ensemble involve effects from the same distribution (i.e., θ_i is drawn from the same distribution as the other θ's), the variance of these residuals should be about one. It should be noted that, just with null hypothesis tests, these residuals can be indicative and perhaps even conclusive about a failure to replicate, but will likely be ambiguous about successful replications.

Figure 2 shows the distribution of original study residuals relative to the distribution of replication study residuals. The vertical dashed lines correspond to positive and negative 2.0. The residuals for the replication studies exhibit the behavior one would expect: their variance is 1.00, and 95% of them are between -1.96 and 1.96. However, the residuals for the original studies tend to be more variable than those for the replications; their variance is 3.84 and only 67% of them lie between -1.96 and 1.96. In fact, 12 of 39 (31%) of original experiments had standardized residuals greater than 2.0 in magnitude.[1] These constituted either the largest or smallest (most negative) standardized residual among their respective ensembles.

[INSERT FIGURE 2 HERE]

That several original studies appear to differ from the results of replications seems to align with a common narrative about the replication crisis in psychology. Authors have proposed several reasons for the crisis that tend to center on the impact of small sample sizes and publication selection, which is also part of the logic underpinning some proposed analyses methods for replications (see Schmidt & Oh, 2016; Simonsohn, 2015). This reasoning typically involves a small initial study overstating the magnitude of the effect, while a subsequent (larger) replication finds a smaller or null effect, a phenomenon that has been well documented in the medical sciences (Ioannidis, 2005; Dumas-Mallet et al., 2017). Likewise, both the RPP and RPE found that often the initial study had a larger effect estimate than the replication, though often

[1] Note that one original finding did not have a design-comparable effect size and variance (the Quote Attribution experiment from Many Labs), and thus there are 39 externally standardized residuals for original studies in the data, but 40 experiments were replicated.

this difference was not statistically significant (see previous sections). Indeed, this has been a part of some of the more contentious debates about findings replicating in psychology. Thus, it may be of interest if this dynamic plays out in empirical evaluations of replication.

One approach to evaluating this is to determine *ex ante* if the initial study would have been well powered to detect the average effect of the replications. To do so, we compute the average effect among the replications (excluding the original study); where findings were moderated by study-level covariates the average effect considered was computed using the covariates of the initial study. Then, using the initial study's sampling variance, we can determine its power to detect that effect. The results of this are displayed in Table 4, which for each finding shows estimated heterogeneity among the replications excluding the original study (H^2), the initial effect estimate and standard error (T_{orig}), the average replication effect and standard error (T_{rep}), the externally standardized residual for the initial finding (r_{orig}), and the power of the original study to detect the average replication effect. Rows in which the residual is greater than two in magnitude ($|r_{\text{orig}}| > 2$) are highlighted in gray.

[INSERT TABLE 4 HERE]

What can be seen in Table 4 is that about half of the initial findings follow the standard narrative about failed replications. The bottom half of the table is comprised of findings where the initial study obtains a larger (in magnitude) effect than the replications, and that initial study would have very low power to detect the average replication effect. Many of these include scenarios where the replications find effects near zero. However, the top half of Table 4 comprises findings where the opposite is true. For these findings, the initial study tends to have a smaller (in magnitude) effect than the replications, and that initial study would have high power to detect the effects found by replications. This pattern, where half of the initial studies appear to understate the size of the effect, is evident even if we focus on just the initial studies with large residuals, which are in shaded rows.

Aside from this issue, however, the results in Table 4 largely corroborate the analyses and conclusions of the replication research programs with a few exceptions. First, PPIR found that four of their original findings were inconsistent with their replications according to a confidence overlap criterion. However, the residuals in Table 4 suggest that most of the original studies obtained effects that were not notably different the replications. One reason for this difference is that the criterion used by PPIR did not account for the fact that their replications obtained fairly heterogeneous results, which can be seen both in Table 2 and Table 4.

Second, the Many Labs Project concluded that their replications contradicted the results of only two original experiments (flag and money priming). Yet in Table 4 we see that there is evidence that the effect in the original study differed from effects in the replications for seven different experiments. Some of these differences are interpreted as under- or over-estimates by the Many Labs project, however some of the largest standardized residuals in this data are associated with original studies for which Many Labs concluded the replications succeeded. For instance, consider the original 'Gain/Loss' study, which examined participants' willingness to take risks when consequences of their actions are framed in terms of expected gains versus expected losses (see Klein, et al., 2014). The original Gain/Loss study found an effect nearly double that of the replications. The discrepancy between Table 4 and the Many Labs conclusions is largely due to the fact that Many Labs typically determined replication success by a correspondence in sign and statistical significance. Thus, since both the original and average replication effect estimates T_{orig} and T_{rep} were statistically significant, Many Labs deemed that a successful replication.

Evidence about Heterogeneity in Replication Research: Benchmarks and Guidance

This article has argued that there are a few reasons why we may wish to define replication as approximate rather than exact. First, effect sizes that are not identical but are similar in value may have the same clinical or scientific interpretations. For instance, there may not be much of a difference in how an effect of $\theta = 0.48$ (in Cohen's *d* units) is inteterpreted versus how an effect of $\theta = 0.52$ is interpreted. Further, it has proven difficult historically in various fields, such as physics, to obtain identical results. Based on the previous few sections of this article, it would seem that we might expect some heterogeneity in results even among direct replications in psychology.

Defining replication and interpreting analyses is a matter of scientific and clinical judgement about how much heterogeneity would be considered negligible. Analyses in this article have based this judgement on conventions from meta-analyses in various fields. But, there is no set convention in psychology. In this section, we explore how much heterogeneity there is in existing direct replications in psychology, and how that compares to other conventions described in this article.

First, Table 4 shows H^2 values for each set of pre-registered replication studies (excluding the original study). Recall that each ensemble of studies in this table used standardized protocols and identical materials, and H^2 values reported in the table control for known differences between studies (see the section on moderators). From the H^2 values reported in the table, the median amount of heterogeneity among replications appears to be about $\tau^2/v = .24$ (mean $\tau^2/v = 1.36$). However, there are several sets of studies that exhibit zero heterogeneity, as well as some that exhibit considerable heterogeneity. For reference, H^2 has a mean of 1.0

when the studies replicate exactly. Moreover, we would expect about 21 of the H^2 values to be greater than one assuming all of the studies replicated exactly for each experiment, and $H^2 > 1.0$ for 24 experiments.

Comparing the heterogeneity reported in Table 4 to conventions in other fields requires consideration of which data are included and how heterogeneity is quantified. As an example, the idea of negligible heterogeneity in physics depends on how outliers are handled. Stigler (1977) suggests consensus among findings involves trimming 10% of outliers, meaning that one would remove the highest and lowest 10% of observations. In its systematic reviews, the Particle Data Group note that results are excluded for a variety of reasons, including that they are inconsistent with other results, which has led to the deletion of nearly 40% of data in some instances (Rosenfeld, 1975). These practices are based on the idea that excluded results are, in some way, wrong, and those studies were not estimating the same quantity as the ones that are included.

Applying similar rules to the replications in this study will naturally result in less heterogeneity. To get a sense of how much less, we can delete replication studies with the largest and smallest (most negative) standardized residuals for their given ensemble and re-compute H^2. Under this procedure, H^2 for the ensembles of studies in the article is largely in the range from one (implying $\tau^2/v = 0$) to 2.68 (so that $\tau^2/v = 1.68$), with a median of about 1.00 and a mean of 1.25. Note that this mean value of H^2 (1.25) corresponds with the convention from physics ($\tau^2/v = ¼$).

Finally, while this data suggest that we might expect some heterogeneity in direct replications, some ensembles of replications exhibited almost zero heterogeneity. This appears to be more common among replications that have effects closer to zero. The bottom 12 rows of table 4 all involve point estimates of $H^2 = 1.0$, which corresponds to no heterogeneity. All 12 of these ensembles involve average effects that were smaller than 0.1 in magnitude. This pattern does not hold for all findings, such as the Many Labs Allow/Forbid experiment, where replications did not vary much around a larger effect ($H^2 = 1.0$, $T_{\text{rep}} = 0.74$). The correlation between the magnitude of the average replication effect $|T_{\text{rep}}|$ and H^2 is $r = 0.45$. Thus, it would seem that some of the most reliably re-created results are those in which the manipulation is very weakly correlated with the outcome (or not correlated at all).

Discussion

This article argued that the question of replicability can be framed in at least two different ways. One way concerns whether a set of replications consistently get the same result, either exactly or approximately. The other involves determining whether replication studies contradict the findings of an original study. Both of these align with the logic of science, and when considered jointly they can provide a more complete picture about the replicability of a finding. Our approach has been to use analyses that are conclusive about replication failures. While this has limitations (discussed below), it provides an alternative view of the replication crisis in psychology using methods with known and measurable error rates.

Among the 40 ensembles of multi-site direct replications analyzed here, the Q test concluded that between 11 and 17 of them produced heterogeneous effects. For several of the these ensembles, some of their heterogeneity could be explained by documented differences between studies. Tests of whether the original and replication studies produced different effects concluded that only 25 of 91 (27%) RPP and RPE studies failed to replicate exactly, and only 14 (15%) failed to replicate approximately. However, as we note below, this does not mean that the remaining 66-77 experiments successfully replicated, and our analysis showed that most of these tests were severely underpowered.

Comparisons between original studies and ensembles of multiple replications found that many (just under half) of the experiments involved initial effects that appear inconsistent with the results of the pre-registered replications. For 11 other experiments (eight from PPIR and three from Many Labs), we could not conclusively say that the original study differed from the replications, but the replications themselves had variable results. The apparent divergence between initial results and replications is something that would seem predictable given current understandings about the replication crisis in psychology. The common narrative is that initial findings, which are often subject to some publication selection, tend to overestimate an effect, which does occur in the data. Yet, somewhat surprisingly, this only occurs for about half of the findings, while the other half involved initial studies that understated the magnitude of the effect as implied by the replications. Moreover, the experiments for which heterogeneity was often small tended to have effect estimates near zero.

That the effect estimates from original studies were frequently smaller than the average effects found in multi-site replications would seem to run counter to prior research on replication. Empirically, in various fields, it is more common for original studies to have larger effects than subsequent replications (Ioannidis, 2005; Open Science Collaboration, 2015; Camerer et al., 2016; Dumas-Mallet et al., 2017). There are a variety of explanations for this finding, including how experiments are chosen for study by replication research programs, how the protocols were standardized, and even whether experimental protocols were still in development.

An important caveat to these findings, though, is that the inferential structure of the analyses means that they will only be conclusive about failures to replicate. The null hypothesis of the Q test is that the studies replicate, and hypothesis tests cannot prove that the null hypothesis is true. Thus, these tests will never be conclusive about replication success (unless they have very high power). Test that *are* necessarily conclusive about successful replication could flip how analyses are framed, so that the null hypothesis is that the studies failed to replicate (e.g., H_0: $\lambda > \lambda_0$). Details on this are discussed in Hedges and Schauer (2018).

Given this limitation, we felt it appropriate to examine the statistical power of the analyses presented here. Prior research suggests that the power of analyses that involve only k = 2 studies will be bounded, even if the second study is infinitely large (see Hedges & Schauer, in press). This was demonstrated in reporting on the MDH of the RPP and RPE, which found that most of these tests were severely underpowered. A similar limitation pertains to the use of externally residuals to compare the original effect size to the distribution of effects produced by an ensemble of replication studies. Thus, while values of $|r_{\text{orig}}| > 2$ can be seen as indicative that an original study's results are inconsistent with those of the replications, $|r_{\text{orig}}| < 2$ does not imply successful replication.

Similarly, low power was also found in tests of heterogeneity involving multi-site replications. However, the Q test is the likelihood ratio test under the model, and so is the most powerful test of heterogeneity. The low power is not necessarily a fault of the method, but rather, there less information in a set of replications about effect heterogeneity than we might think. Hedges and Schauer (2018) show that it is possible for ensembles of replication studies to support a well-powered Q test, but the studies analyzed in this article do not appear to have been designed to ensure that. Thus, future work on improving the design of replication studies is still needed, including how to design ensembles of replications to ensure sufficiently sensitive analyses (e.g., high power for hypothesis tests). Moreover, it would seem useful to explore designs that systematically vary experimental conditions and contexts to more explicitly examine how such variation affects study results.

So too is work on examining notions for negligible heterogeneity in psychology. Analyses of heterogeneity can be sensitive to what one might consider to be a negligible difference between effects. As has been pointed out by reviewers, the idea that we can use potentially different values of λ_0 in tests of replication would seem to provide researchers an opportunity to misuse these tests by choosing values of λ_0 that support a desired conclusion. However, we stress that λ_0 must be specified prior to analysis. Ideally, it should be specified prior to conducting replications, so that they can be designed to ensure well-powered tests. Pre-registration of analyses and greater transparency, which have played a large role in replication research, can help prevent such misuse of the method.

Alternatively, analyses could rely on relevant scientific conventions about negligible heterogeneity. The analyses in this article demonstrated how this could work by borrowing from conventions in meta-analyses from other fields. Empirical results from this article suggest that heterogeneity among designed replications in psychology were not inconsistent with these conventions, and that we might expect variation among replications that ranges from $\tau^2/v = 0$ to $\tau^2/v = 1$. However, precisely operationalizing this conception of replication would seem difficult, particularly given the information of only a single published finding.

Finally, conceptions of heterogeneity presented in this paper are on the relative scale of between study variance to within-study variance (τ^2/v). This is a common scale of heterogeneity in meta-analysis, however it is not the only scale we can use. Instead, we can specify tests and interpret results on the raw scale of τ^2. Since the magnitude of τ^2 will depend on the individual study effects, interpretation of heterogeneity will depend on the type of effect size used in the analysis. Results of these analyses may differ if they are carried out and interpreted on this raw scale as opposed the relative scale, and further work is required to understand which should be preferred when.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101 (2), 470–92.

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.

Bargh J. A., Chen M., & Burrows L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype-activation on action. *Journal of Personality and Social Psychology*, 71, 230–244.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Reproducibility, replicability, and generalization in the social, behavioral, and economic sciences.* Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. Arlington, VA: National Science Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester: 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.

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

Carney, D., Cuddy, A. J., & Yap, A. (2010). Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. *Psychological Science*. 21. 1363-8.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

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

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

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PLoS ONE*, *7*(1), e29081.

Dumas-Mallet, E., Smith, A., Boraud, T., & Gonon, F. (2017). Poor replication validity of biomed- ical association studies reported by newspapers. *PLoS ONE*, 12(2), e0172650.

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.

Ericson, K. M. M. & Fuster A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *The Quarterly Journal of Economics*, 126(4), 1879–1907.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PloS one*, *11*(2), e0149794.

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

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on Estimating the reproducibility of psychological science. *Science*, 351(6277), 1037–1037.

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.

Halevy, N., Bornstein, G., & Sagiv, L. (2008). "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science*, 19(4), 405–411.

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

Hedges, L. V. (1982). Fitting Categorical Models to Effect Sizes from a Series of Experiments. *Journal of Educational Statistics,* 7(2), 119-137.

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(5), 443–455.

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

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

Hedges, L. V., & Pigott, T. D. (2004). The Power of Statistical Tests for Moderators in Meta-Analysis. *Psychological Methods,* 9(4), 426-445.

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

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

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

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions*. Hoboken, NJ: Wiley-Blackwell.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed). Thousand Oaks, Calif: Sage.

Henrion, M., & Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics*, 54(9), 791–798.

Ifcher, J., and Zarghamee, H. (2011). Happiness and time preference: The effect of positive affect in a random-assignment experiment. *American Economic Review*, 101(7), 3109–29.

Ioannidis J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2), 218–228.

Jackson, D. (2006), The power of the standard test for the presence of heterogeneity in meta-analysis. Statistics in Medicine, *25*, 2688–2699.

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.

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.

Larsen, J. T. & McKibban, A. R. (2008). Is happiness having what you want, wanting what you have, or both? *Psychological Science* 19(4), 371–377.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, 26(12), 1827–1832.

McNutt, M. (2014). Reproducibility. *Science*, 343(6168), 229–229.

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.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science (New York, N.Y.)*, 348(6242), 1422–1425.

Olive, K. A. (2014). Review of particle physics. Chinese Physics C, 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), pp. 657–660.

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

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530.

Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science, 11*(4), 539-44.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S. & Weber, R. (2015). Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5), 653-656.

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science*, 19(3), 249–254.

Rosenfeld, A. H. (1975). The particle data group: Growth and operations-eighteen years of particle physics. *Annual Review of Nuclear Science*, 25(1), 555–598.

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis : Prevention, assessment and adjustments*. Chichester, England ; Hoboken, NJ: 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

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology,* 4(1), 32-37.

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.

Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, 5(6), 1055–1098.

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

van Aert, R. C., & van Assen, M. A. (2018). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior research methods*, *50*(4), 1515-1539.

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

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., ... Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928.

Wong, V. C., & Steiner, P. M. (2018). Replication designs for casual inference. EdPolicy Works. Working Paper No. 62.

Wood, P., & Randall, D. (2018). How bad is the government's science? *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/how-bad-is-the-governments-science-1523915765.

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

Table 1: Failed Replication Rates in Meta-Research Programs on Replication in Psychology. This table summarizes determinations about replication failures for empirical evaluations of replication. For each paper, the table shows the number of experiments they attempted to replicate *m*, the number of times each experiment was replicated *k*, and proportion of those experiments that the initial papers determined failed to replicate ("Published Analyses"). The "Meta-analysis Results" panel shows the proportion of experiments that the *Q* test determined failed to replicate for tests of exact replication ($\tau^2/v = 0$) and approximate replication ($\tau^2/v = 1/4$ to $2/3$).

					Meta-analysis Results			
Paper	Year	*m*	*k*	Published Analyses	$\tau^2/v = 0$	$\tau^2/v = 1/4$	$\tau^2/v = 1/3$	$\tau^2/v = 2/3$
Alogna	2014	2	24-33	0%	0%	0%	0%	0%
Bouwmeester	2016	1	22	100%	0%	0%	0%	0%
Cheung	2016	4	17	50%	0%	0%	0%	0%
Eerland	2015	3	13	100%	0%	0%	0%	0%
Hagger	2016	2	24	50%	0%	0%	0%	0%
Many Labs	2014	16	36-37	19%	38%	25%	25%	19%
PPIR	2016	11	12-18	18%	82%	82%	82%	73%
RPE	2016	18	2	39%	22%	22%	22%	17%
RPP	2015	73	2	63%	30%	23%	22%	15%
Wagenmakers	2016	1	18	100%	0%	0%	0%	0%

Figure 1: MDH of RPP & RPE. This figure shows the smallest difference detectable for RPP and RPE studies with 80% power and level $\alpha = 0.05$. The y-axis displays the magnitude of the detectable difference in Cohen's d units, and the dashed line corresponds to $d = 0.5$, or a medium sized effect. Each pair of boxplots corresponds to a different test of replication ranging from exact ($\lambda_0 = 0$) to various definitions of approximate (λ_0 = ¼, 1/3, 2/3).

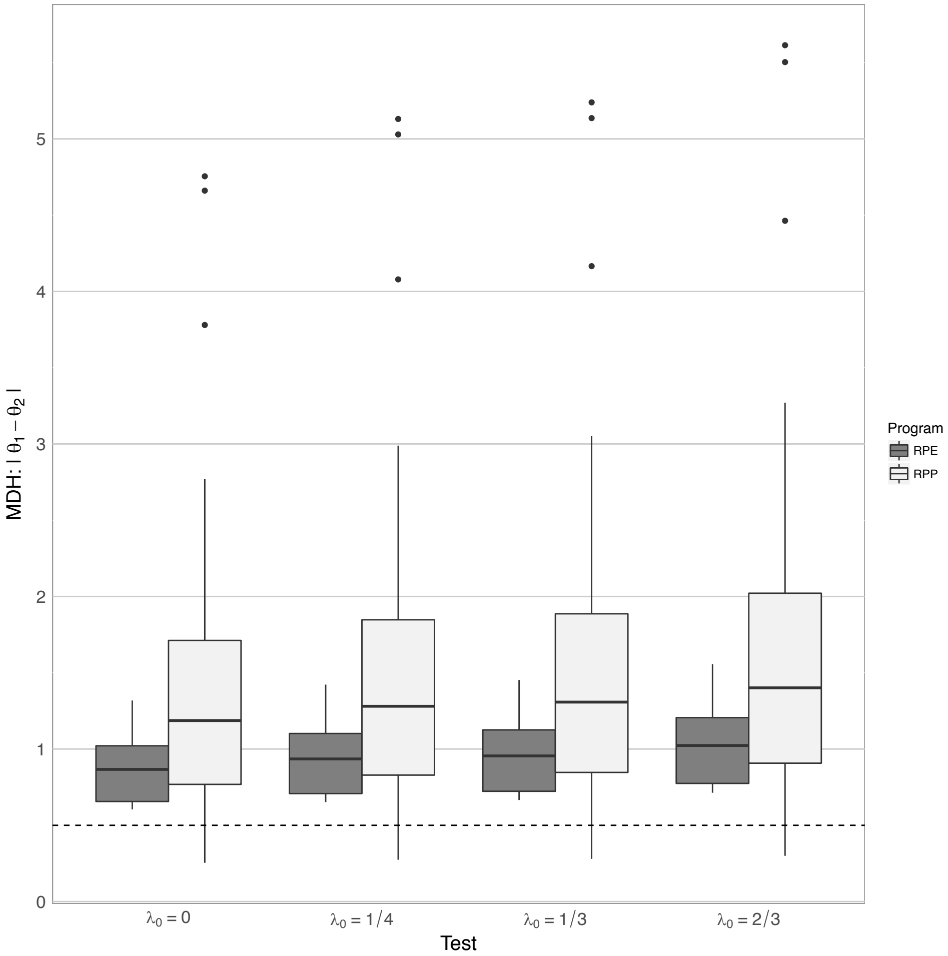

Table 2: Hypothesis Tests for Ensembles of Replication Studies. The “Q-Test Results” panel shows the p-value of tests of replication for different values of λ_0 ranging from 0 to $2(k-1)/3$; cells with $p < 0.05$ are shaded in gray. The “Sensitivity” panel shows the minimal amount of heterogeneity (referred to as MDH) on the scale of τ^2/v those tests could detect with 80% power.

				Q-Test Results				Sensitivity			
				p-values				MDH (τ^2/v)			
Paper	Experiment	k	Q	0	(k-1)/4	(k-1)/3	2(k-1)/3	0	(k-1)/4	(k-1)/3	2(k-1)/3
Many Labs	Allowed/Forbidden	36	27.66	0.81	0.93	0.95	0.98	0.75	1.14	1.26	1.74
	Anchoring1	36	61.57	0.00	0.06	0.11	0.38	0.75	1.14	1.26	1.74
	Anchoring2	36	156.73	0.00	0.00	0.00	0.00	0.75	1.14	1.26	1.74
	Anchoring3	36	317.14	0.00	0.00	0.00	0.00	0.75	1.14	1.26	1.74
	Anchoring4	36	90.60	0.00	0.00	0.00	0.03	0.75	1.14	1.26	1.74
	Flag Priming	36	30.71	0.68	0.90	0.93	0.98	0.75	1.14	1.26	1.74
	Gain/Loss	36	37.01	0.38	0.71	0.78	0.93	0.75	1.14	1.26	1.74
	Gambler’s Fallacy	36	51.36	0.04	0.23	0.32	0.65	0.75	1.14	1.26	1.74
	IAT	35	46.82	0.07	0.31	0.41	0.71	0.76	1.15	1.28	1.76
	Imagined Contact	36	46.44	0.09	0.38	0.48	0.78	0.75	1.14	1.26	1.74
	Math/Art/Gender	35	48.17	0.06	0.28	0.38	0.69	0.76	1.15	1.28	1.76
	Money Priming	36	28.80	0.76	0.94	0.96	0.99	0.75	1.14	1.26	1.74
	Quote Attribution	36	68.49	0.00	0.02	0.04	0.24	0.75	1.14	1.26	1.74
	Reciprocity	36	38.89	0.30	0.59	0.66	0.86	0.75	1.14	1.26	1.74
	Scales	36	33.08	0.56	0.82	0.86	0.93	0.75	1.14	1.26	1.74
	Sunk Costs	36	36.07	0.42	0.75	0.81	0.94	0.75	1.14	1.26	1.74
PPIR	Bad Tipper	16	173.44	0.00	0.00	0.00	0.00	1.25	1.73	1.87	2.43
	Belief-Act Inconsistency	13	83.85	0.00	0.00	0.00	0.00	1.45	1.94	2.10	2.68
	Bigot-Misanthrope	12	50.86	0.00	0.00	0.00	0.00	1.53	2.04	2.20	2.80
	Burn in Hell	15	37.32	0.00	0.01	0.02	0.08	1.31	1.79	1.94	2.50
	Cold-Hearted Prosociality	12	53.01	0.00	0.00	0.00	0.00	1.53	2.04	2.20	2.80
	HS - Chairty	11	92.46	0.00	0.00	0.00	0.00	1.62	2.15	2.31	2.92
	HS - Company	11	96.17	0.00	0.00	0.00	0.00	1.62	2.15	2.31	2.92
	Intuitive Economics	15	47.38	0.00	0.00	0.00	0.02	1.31	1.79	1.94	2.50
	Moral Cliff	15	9.13	0.82	0.92	0.94	0.97	1.31	1.79	1.94	2.50
	Moral Inversion	14	61.22	0.00	0.00	0.00	0.00	1.37	1.86	2.01	2.59
	Presumption of Guilt	17	25.34	0.06	0.22	0.28	0.52	1.20	1.67	1.81	2.36
Alogna	RR1	32	32.85	0.38	0.71	0.79	0.94	0.80	1.20	1.33	1.82
	RR2	23	16.28	0.80	0.93	0.95	0.99	0.99	1.42	1.55	2.07
Cheung	Exit	16	13.95	0.53	0.74	0.79	0.91	1.25	1.73	1.87	2.43
	Neglect	16	17.12	0.31	0.55	0.62	0.80	1.25	1.73	1.87	2.43
	Voice	16	11.04	0.75	0.89	0.91	0.97	1.25	1.73	1.87	2.43
	Loyalty	16	8.20	0.92	0.97	0.98	0.99	1.25	1.73	1.87	2.43
Eerland	Imagery	12	7.93	0.72	0.85	0.88	0.94	1.53	2.04	2.20	2.80
	Intention Attribution	12	10.45	0.49	0.69	0.73	0.86	1.53	2.04	2.20	2.80
	Intentionality	12	18.54	0.07	0.20	0.25	0.45	1.53	2.04	2.20	2.80
Hagger	RTV	23	20.12	0.58	0.82	0.86	0.96	0.99	1.42	1.55	2.07
	RT	23	23.17	0.39	0.68	0.75	0.91	0.99	1.42	1.55	2.07
Bouwmeester	Time/Delay	21	16.50	0.69	0.87	0.90	0.97	1.05	1.49	1.63	2.15
Wagenmakers	Facial Feedback Hyp.	18	17	9.01	0.91	0.97	0.98	0.99	1.20	1.67	1.81
	Failed Replications:			37.5%	32.5%	32.5%	27.5%				

Table 3: Tests for Homogeneity with Moderators. This table presents the effects of controlling for moderators on tests for heterogeneous effects. For each experiment, the table shows the raw (unadjusted) Q statistic and p-value for the test for exact replication. It also shows the test for exact replication after accounting for moderators with the relevant statistic Q_E and p-value p_E, as well as the effect of the moderator and standard error.

Paper	Experiment	Q	p_Q	Q_E	p_E	Factor	Difference
PPIR	Bad Tipper	173.44	0.00	126.86	0.00	us	0.42 (0.22)
PPIR	Belief-Act Incon.	83.85	0.00	42.55	0.00	online	-0.64 (0.17)
PPIR	Bigot-Misanthrope	50.86	0.00	20.73	0.00	us	-0.53 (0.15)
PPIR	Moral Cliff	9.13	0.82	6.63	0.76	students	-0.11 (0.07)
Many Labs	Anchoring1	61.57	0.00	50.26	0.00	online	-0.19 (0.08)
Many Labs	Anchoring2	156.73	0.00	105.68	0.00	online	-0.33 (0.15)
Many Labs	Anchoring3	317.14	0.00	173.38	0.00	online	-0.80 (0.22)
						us	0.57 (0.22)
Many Labs	Flag Priming	30.71	0.68	28.5	0.63	online	0.08 (0.05)
Many Labs	Quote Attribution	68.49	0.00	60.04	0.00	us	0.20 (0.08)
Many Labs	Scales	33.08	0.56	25.77	0.51	online	-0.28 (0.10)

Figure 2: Externally Standardized Residuals of Replications. This plot shows the distributions of externally standardized residuals across findings in meta-research programs, which is the standardized difference between a given effect T_i and the meta-analytic average of all studies excluding study *i*. The dark density plot corresponds to residuals of the original findings in the ensemble, and the lighter plot corresponds to the pre-registered replications. Dashed lines correspond to positive and negative two. The variance of the residuals for the original studies is 3.84, and the variance of the residuals for the replication studies is 1.0.

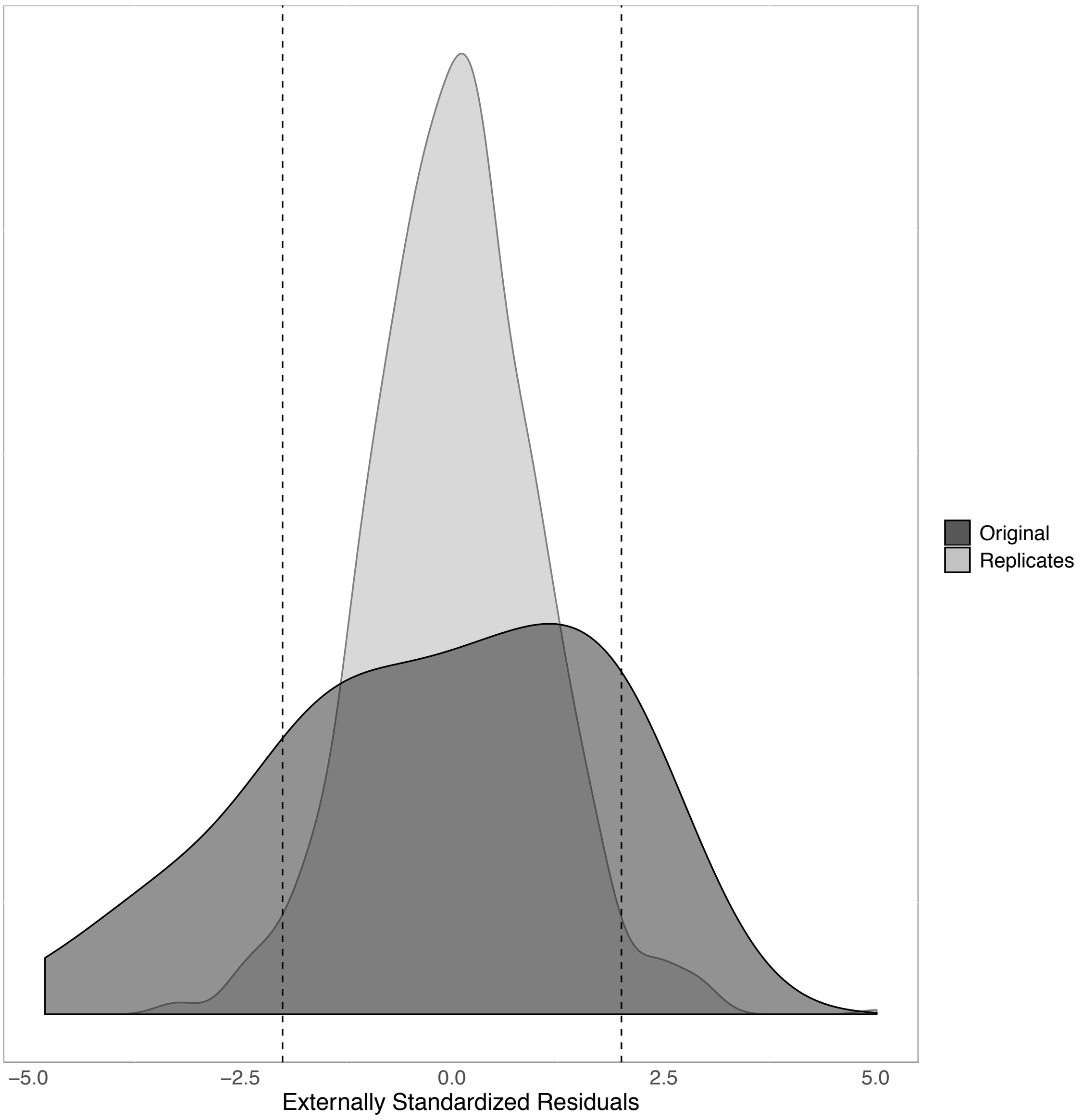

Table 4: Power of Original Studies. For each finding, the table shows the heterogeneity among the replications H^2, the original and replication effect estimates T_{orig} and T_{rep} with standard errors, the externally standardized residual of the original study r_{orig},[2] and the power of the original study to detect an effect as large as T_{rep}. Rows with $|r_{\text{orig}}| > 2$ are highlighted in gray.

Paper	Experiment	H^2	T_{rep}	T_{orig}	r_{orig}	Power
Many Labs	Anchoring 3	5.3	2.80 (0.12)	0.93 (0.18)	-3.70	1.00
PPIR	Cold-Hearted Pros.	4.8	2.04 (0.1)	2.26 (0.29)	0.52	1.00
Many Labs	Anchoring 2	4.5	2.02 (0.08)	0.93 (0.18)	-2.53	1.00
PPIR	Belief-Act Incon.	3.9	0.88 (0.14)	0.37 (0.18)	-1.59	1.00
Many Labs	Anchoring 4	2.6	2.54 (0.06)	0.93 (0.18)	-4.80	1.00
PPIR	Bigot-Misanthrope	2.1	1.16 (0.07)	0.9 (0.15)	-1.14	1.00
Many Labs	Anchoring 1	1.5	1.28 (0.05)	0.93 (0.18)	-1.53	1.00
Many Labs	IAT	1.4	0.82 (0.04)	0.93 (0.14)	0.56	1.00
Many Labs	Gain/Loss	1.1	-0.66 (0.03)	-1.21 (0.15)	-3.53	1.00
Many Labs	Reciprocity	1.1	0.37 (0.04)	0.16 (0.05)	-2.23	1.00
Many Labs	Allow/Forbid	<1.0	0.74 (0.04)	0.51 (0.05)	-3.43	1.00
PPIR	HS - Company	9.6	0.92 (0.13)	0.34 (0.21)	-1.26	0.99
Many Labs	Scales	<1.0	0.95 (0.09)	0.61 (0.23)	-1.39	0.99
PPIR	Moral Cliff	<1.0	0.79 (0.06)	0.71 (0.2)	-0.38	0.98
PPIR	HS - Charity	9.2	0.90 (0.12)	0.92 (0.24)	0.04	0.96
PPIR	Intuitive Economics	3.4	0.51 (0.07)	0.85 (0.15)	1.27	0.94
PPIR	Bad Tipper	9.1	0.73 (0.14)	0.64 (0.23)	-0.18	0.87
Many Labs	Math/Art/Gender	1.4	0.58 (0.04)	1.01 (0.24)	1.61	0.67
Many Labs	Gamblers' Fallacy	1.5	0.61 (0.04)	0.69 (0.27)	0.26	0.63
Many Labs	Sunk Costs	<1.0	0.29 (0.03)	0.23 (0.14)	-0.43	0.56
PPIR	Moral Inversion	4.7	0.47 (0.08)	0.81 (0.27)	0.86	0.41
Alogna	RR2	<1.0	-0.15 (0.02)	-0.25 (0.1)	-0.91	0.32
PPIR	Burn in Hell	2.7	0.21 (0.06)	0.27 (0.16)	0.23	0.26
PPIR	Pres. of Guilt	1.6	0.19 (0.04)	0.03 (0.23)	-0.63	0.13
Eerland	Intentionality	1.7	-0.17 (0.08)	0.77 (0.3)	2.65	0.09
Many Labs	Imagined Contact	1.3	0.12 (0.03)	0.86 (0.4)	1.79	0.06
Alogna	RR1	1.1	-0.03 (0.02)	-0.22 (0.11)	-1.70	0.06
Cheung	Neglect	1.1	-0.05 (0.05)	-0.45 (0.21)	-1.71	0.06
Hagger	RT	1.1	0.08 (0.04)	0.29 (0.29)	0.70	0.06
Cheung	Exit	<1.0	-0.05 (0.05)	-0.60 (0.22)	-2.47	0.06
Eerland	Imagery	<1.0	-0.08 (0.06)	0.73 (0.3)	2.68	0.06
Many Labs	Flag Priming	<1.0	0.02 (0.03)	0.50 (0.25)	1.92	0.05
Many Labs	Money Priming	<1.0	-0.02 (0.03)	0.80 (0.38)	2.15	0.05
Bouwmeester	Time/Delay	<1.0	-0.02 (0.03)	0.27 (0.17)	1.67	0.05
Cheung	Voice	<1.0	0.02 (0.05)	0.34 (0.21)	1.45	0.05
Cheung	Loyalty	<1.0	0.01 (0.05)	0.21 (0.21)	0.94	0.05
Eerland	Intent.-Attrib.	<1.0	0.01 (0.06)	0.67 (0.3)	2.20	0.05
Hagger	RTV	<1.0	0.00 (0.04)	0.68 (0.3)	2.24	0.05
Wagenmakers	Facial Feedback Hyp.	<1.0	0.02 (0.05)	0.47 (0.26)	1.74	0.05

[2] Externally standardized residuals are the standardized difference between the original effect estimate T_{orig} and T_{rep}, the weighted average of the replication studies (excluding T_{orig}).