Statistical Methods for Replicability Assessment

Kenneth Hung* William Fithian†

March 14, 2019

Abstract

Large-scale replication studies like the Reproducibility Project: Psychology (RP:P) provide invaluable systematic data on scientific replicability, but most analyses and interpretations of the data fail to agree on the definition of "replicability" and disentangle the inexorable consequences of known selection bias from competing explanations. We discuss three concrete definitions of replicability based on (1) whether published findings about the signs of effects are mostly correct, (2) how effective replication studies are in reproducing whatever true effect size was present in the original experiment, and (3) whether true effect sizes tend to diminish in replication. We apply techniques from multiple testing and post-selection inference to develop new methods that answer these questions while explicitly accounting for selection bias. Re-analyzing the RP:P data, we estimate that 22 out of 68 (32%) original directional claims were false (upper confidence bound 47%); by comparison, we estimate that among claims significant at the stricter significance threshold 0.005, only 2.2 out of 33 (7%) were directionally false (upper confidence bound 18%). In addition, we compute selection-adjusted confidence intervals for the difference in effect size between original and replication studies and, after adjusting for multiplicity, identify five (11%) which exclude zero (exact replication). We estimate that the effect size declined by at least 20% in the replication study relative to the original study in 16 of the 46 (35%) study pairs (lower confidence bound 11%). Our methods make no distributional assumptions about the true effect sizes.

1 Introduction

Growing concerns about selection bias, p-hacking, and other questionable research practices (QRPs) have raised urgent questions about the reliability of scientific findings. While concerns about replicability cut across scientific disciplines, psychologists have led large-scale efforts to assess the replicability of their own field. The largest and most systematic of these efforts has been the Reproducibility Project: Psychology (RP:P),[1] a major collaboration by several hundred psychologists to replicate a representative sample of 100 studies published in 2008 in three top psychology journals, *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition.*[2]

*Department of Mathematics, University of California, Berkeley

†Department of Statistics, University of California, Berkeley

[1] In some parts of the literature, "reproducibility" has taken on a computational connotation, meaning only that other scientists can repeat the analysis using the original study's data; we will lean toward the more unambiguous term "replicability."

[2] The test statistics, effect sizes and most pertinent information are all publicly available on at the Open Science Foundation website at `https://osf.io/ezcuj/`.

While the RP:P dataset is an invaluable resource, scientists disagree on how to quantify or measure replicability (Goodman et al., 2016; Amrhein et al., 2017). Open Science Collaboration (OSC; 2015) reported three main metrics: it found that 64% ($= 1 - 36\%$) of the replication studies did not find statistically significant results in the same direction as the original studies, that 53% ($= 1 - 47\%$) of 95% confidence intervals for the replication studies do not contain the point estimates for their corresponding original studies, and that 83% of the effect size estimates declined from original studies to replications. All three summary statistics were widely reported as indicating a dire crisis for the credibility of experimental psychology research. For example, the *Washington Post* reported that RP:P "affirms that the skepticism [of published results] was warranted" (Achenbach, 2015); the *Economist* noted that OSC "managed to replicate satisfactorily the results of only 39% of the studies investigated" (Ano, 2016); and the *New York Times* reported that "more than half of the findings did not hold up when retested" (Carey, 2015).

This negative gloss was challenged in a comment by Gilbert et al. (2016a), who criticized both the fidelity of some of the replications' experimental designs and the aptness of the metrics reported by Open Science Collaboration (2015). In particular, Gilbert et al. pointed out that, because there is sampling error in the replication point estimates, we should not expect 95% of the estimates to fall into the replication confidence intervals even under ideal conditions. Moreover, any small or large variations in the true effect sizes between the original and replication studies could further deflate the expected fraction of "successful replications," as measured in this way. Gilbert et al. concluded that "OSC seriously underestimated the reproducibility of psychological science," sparking further debate between defenders of OSC's conclusions (Anderson et al., 2016; Srivastava, 2016; Nosek and Gilbert, 2016) and the critics (Gilbert et al., 2016c,b).[3]

To determine whether OSC truly underestimated replicability, we must first pin down the rather slippery question of what "replicability" actually is. Although the three metrics used by OSC are simply descriptive statistics that do not purport to estimate any explicitly defined underlying quantity, we can loosely characterize the 64%, 53% and 83% numbers respectively as qualitative answers to three questions:

False directional claims. *What fraction of the original studies were erroneous* in claiming that the true effect was nonzero, in the claimed direction (positive or negative)? Gelman and Tuerlinckx (2000) called such mistakes *type S* errors.

Effect shift. *How much do the effect sizes shift* from the original study to the replication study? We call the discrepancy between the original and replication effect *effect shift*.

Effect decline. *What fraction of the effect sizes decline?* More precisely, what fraction of the true effect sizes shift in a direction opposite to the original claims when the studies were replicated, and by how much?

The first question concerns a type of *false discovery rate* (FDR) of the statistical hypotheses, viewing the field of social psychology as a collective enterprise in large-scale multiple testing: it quantifies the fraction of findings that would be confirmed if the exact same studies could be carried out again with much larger samples from the same populations. The second question concerns a basic form of repeatability: whether scientists are typically successful in closely

[3]While much of the ensuing discussion focused on the question of whether the confidence interval metric 53% is too pessimistic, analogous criticisms apply to the "significant replications" metric of 64% as well: the replication studies could be underpowered even when a true effect is present.

replicating each others' experimental conditions, so that the true effect being measured is stable across different experiments. The third question builds upon the second question: whether true effect sizes tend systematically to attenuate in replications. An overall trend of declining true effects could suggest various interpretations, including systematic biases in the original experiments or failures by the replication teams to reproduce key experimental conditions that produced the original effects.

As we will see, however, none of the three reported metrics can be taken at face value as *estimates* of the answers to the corresponding questions, due to the confounding factor of pervasive selection bias. By using techniques from multiple testing and post-selection inference, we will develop methods to rigorously address these questions without assuming a model for the prior distribution of effect sizes. For the RP:P data we estimate the rate of false directional claims at roughly 32% among studies with $p < 0.05$, which would be considered unacceptably high in most multiple testing applications. By contrast, among studies with $p < 0.005$, a lower threshold proposed by Benjamin et al. (2018), our estimate drops to 7%, with an upper confidence bound of 18%. We also compute confidence intervals for the effect shift in each individual study pair and find that, after adjusting for multiplicity, about 11% of the intervals exclude zero, an idealized null hypothesis of perfect replication. For effect decline, we find in aggregate that 35% of the true effects declined, and 35% declined by at least 20%.

In addressing each question, we define our estimands in terms of the true effects present in the statistical populations actually sampled in each study. Because some studies may be biased or lack external validity — for example, because of flaws in the study design, or because survey participants are unrepresentative of the broader population of scientific interest — these effect sizes may not reflect the latent scientific quantities the experiments purport to measure. Uncovering such discrepancies is beyond the reach of data analysis alone, but we should keep them in mind as we interpret the results.

1.1 The role of selection bias

The RP:P data shows unmistakable signs of selection for statistically significant findings in the original experiments: 91 of the 100 results replicated by OSC were statistically significant at the 0.05 level in the original study and four of the others had "marginally significant" p-values between 0.05 and 0.06. This is due partly to publication bias (that the studies might not have been published, or the results discussed, if the p-values had not been significant), but also partly to OSC's method for choosing which results to replicate. Each OSC replication team selected a "key result" from the last experiment presented in the original paper, and evidently most teams chose a significant finding as the key result (justifiably so, since positive results usually draw the most attention from journal readers and the outside world). Figure 1 shows the empirical distribution of p-values from the original and replication studies.

The resulting selection bias in the original studies leads to many well-known and predictable pathologies, such as systematically inflated effect size estimates, undercoverage of (unadjusted) confidence intervals, and misleading answers from unadjusted meta-analyses. Indeed, most of the phenomena reported by OSC, including the three metrics discussed above, could easily be produced by selection bias alone. This would be true *even if there are few false directional claims, all replications are exact, and true effects do not decline*, as illustrated in the following simulation study.

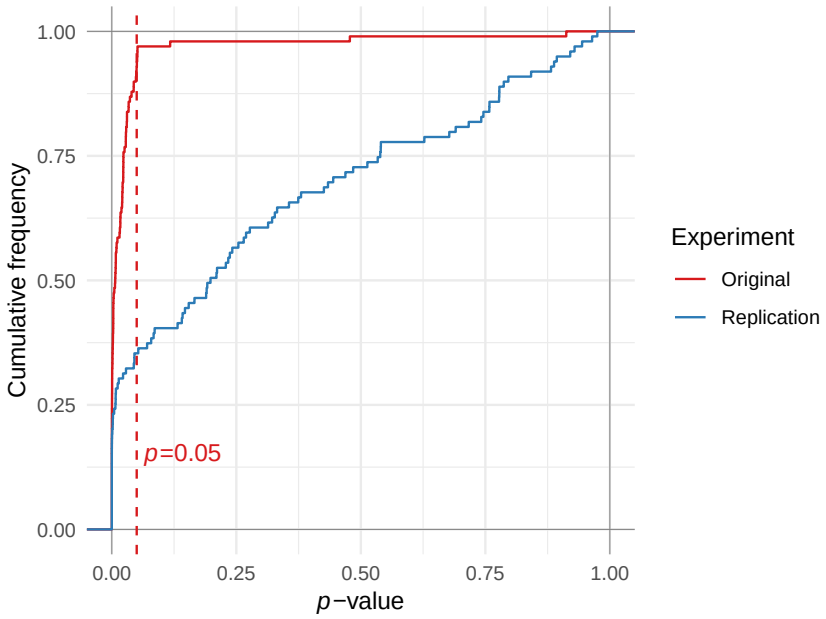


Figure 1: The empirical distribution of the original and replication p-values. Nearly all of the original p-values (in red) are smaller than 0.05.

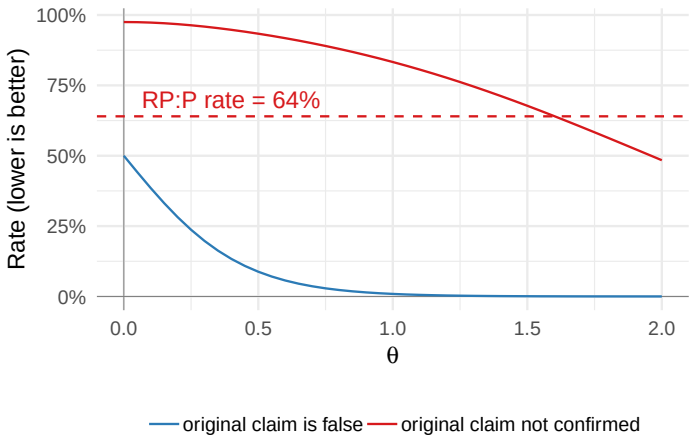
Example 1. Consider a stylized setting where all experiments (both original and replication) have an identical effect size θ, producing an unbiased Gaussian estimate with standard error 1. Assume, however, that we observe only study pairs for which the original study is significant at level 0.05.

Figure 2a shows the expected fraction of replication studies which are not statistically significant in the same direction as the corresponding original studies, as a function of effect size θ, along with the true proportion of false directional claims; or type S errors. Even when the true error rate is low, e.g. at $\theta = 1$ as shown in Figure 2b, the proportion of replications reporting the same directional findings as the original studies can remain low.

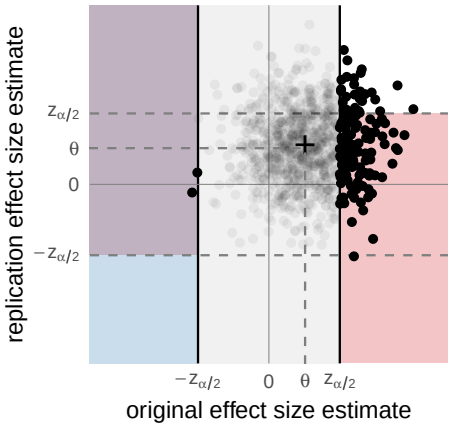
Likewise, we simulate the expected fraction of 95% replication confidence intervals that fail to cover their original point estimates in Figure 3 and the expected fraction of effect sizes that decline in Figure 4. In both cases, we see that selection bias is more than sufficient to produce the metrics in RP:P, even in our idealized simulation with exact replications and relatively few type S errors.

Because selection bias could, in principle, provide a sufficient explanation for the metrics reported in RP:P, those metrics do not, in and of themselves, provide any evidence of any other problems. In particular, they shed no light on whether the FDR is actually high, or how much the effect sizes shifted, or whether effect sizes tend to decline. Nor do they provide evidence for any competing accounts of the replication crisis, such as QRPs like p-hacking, high between-study variability in effect sizes, or systematic biases in the original studies. To discern anything about other explanations, we must adjust for the pervasive effects of selection bias.

Another good reason to disentangle selection bias from other sources of error is that the former is, in some sense, the most innocuous explanation for the phenomena observed by OSC while the others present much deeper scientific issues. The technical issues of selection bias

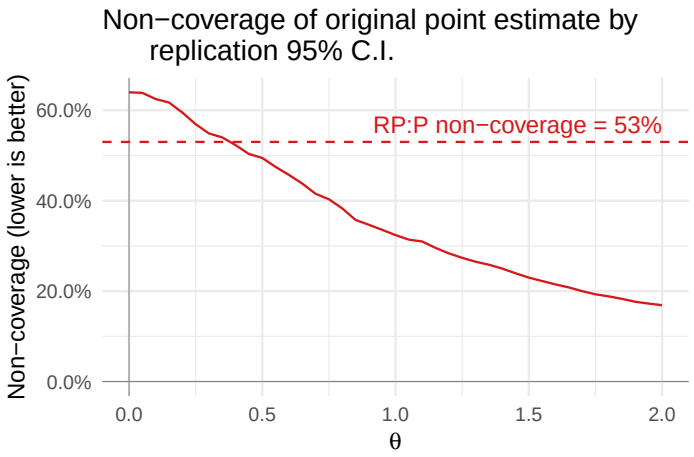


(a) The expected fraction of replications that do not confirm (at level 0.05) the original directional claim (red), and the proportion of false directional claims in the original studies (blue), as a function of effect size θ. For small θ, the fraction of replications that do not confirm the claims in the original studies may dramatically overestimate the fraction of false original claims.

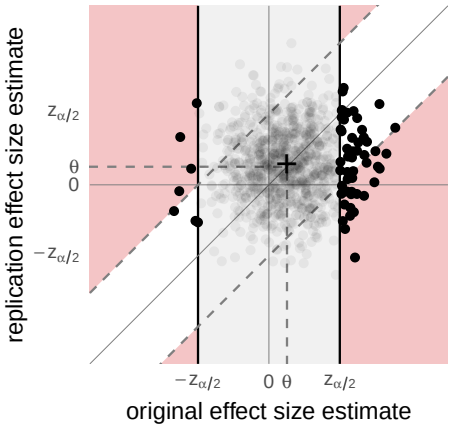


(b) $\theta = 1$. The gray region is unobserved. For points in the red region, the replication does not confirm the original directional claim, and for points in the blue region, the original claim is directionally false. The red and blue regions overlap in the purple region.

Figure 2

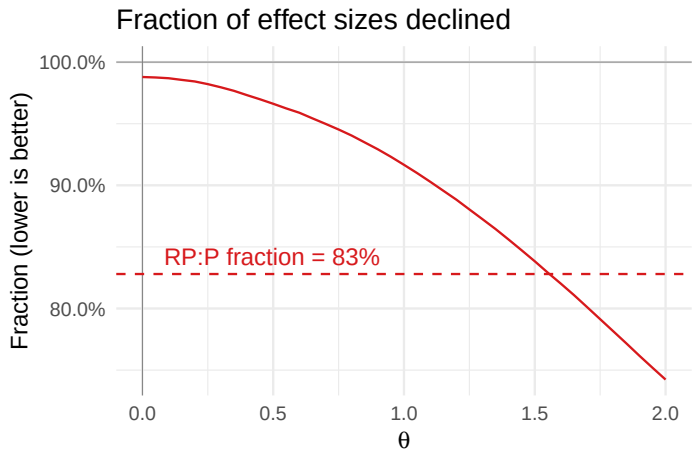


(a) Expected fraction of original point estimates falling outside the replication confidence interval, as a function of effect size θ. For small θ, the fraction of original point estimates falling outside the replication 95% confidence intervals can easily exceed the RP:P reported metric of 53%, even when all replications are perfectly exact.

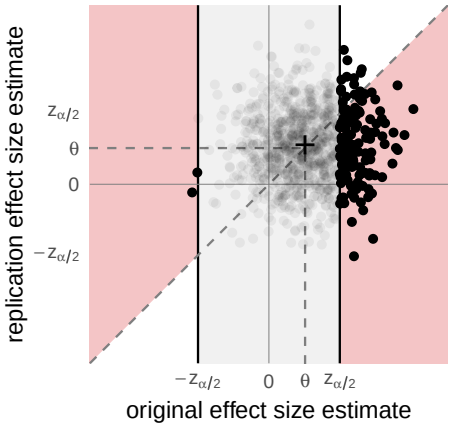


(b) $\theta = 0.5$. The gray region is unobserved. For points in the red region, the original point estimate differs from the replication estimate by more than $z_{\alpha/2}$ and hence the original point estimate falls outside in the replication 95% confidence interval.

Figure 3



(a) Expected fraction of effect size point estimates that declined toward zero in replication, as a function of effect size of θ. For small θ, the fraction of effect size estimates declining from original to replication studies can easily exceed the RP:P reported metric of 83%, even when there is no decline in the true effect sizes.



(b) $\theta = 0.5$. The gray region is unobserved. Points in the red region represent declining point estimates in replications. When the original point estimate is positive, a decline is marked by a smaller replication estimate; on the other hand, if the original estimate is negative, a decline is indicated by a larger replication estimate.

Figure 4

can be addressed either retrospectively by statistical adjustments (e.g. Duval and Tweedie, 2000; Hedges, 1992; Simonsohn et al., 2014a; Fithian et al., 2014; Andrews and Kasy, 2018), or prospectively with more preregistration or larger sample sizes. By contrast, it would be deeply worrying if psychologists were systematically unable to repeat their colleagues' experiments, or if most published claims about effect sizes were directionally incorrect.

1.2 Formalizing replicability

We now introduce a simple formal model for replication studies with selection bias. For study $i = 1, \ldots, m$, let $\theta_{i,O}$ and $\theta_{i,R}$ denote the true effect sizes in the original and the replication studies, respectively. Abstracting away experimental design details, assume that each study pair produces two normally distributed effect size estimators $\hat{\theta}_{i,O}$ and $\hat{\theta}_{i,R}$. Assume additionally that for the study pair to appear in our replication data, $\hat{\theta}_{i,O}$ must be statistically significant at level $\alpha = 0.05$;[4] then for some significance threshold $c > 0$ we have

$$\hat{\theta}_{i,O} \sim N\left(\theta_{i,O}, \sigma^2_{i,O}\right) 1_{\{|\hat{\theta}_{i,O}|>c\}} \qquad \text{and} \qquad \hat{\theta}_{i,R} \sim N\left(\theta_{i,R}, \sigma^2_{i,R}\right), \tag{1}$$

with all estimates assumed to be independent of each other. The indicator $1_{\{|\hat{\theta}_{i,O}|>c\}}$ beside the normal distribution in (1) means that the distribution of $\hat{\theta}_{i,O}$ has been truncated to the event where $|\hat{\theta}_{i,O}| > c$ and renormalized so that it integrates to 1. For the moment, we assume

[4] We relax this assumption in Section 2.

that the variances $\sigma^2_{i,O}$ and $\sigma^2_{i,R}$ are known; in that case $c = z_{0.05/2}\,\sigma_{i,O}$. We will relax this assumption in Section 2.

False directional claims To formalize false directional claims in terms of the parameters of model (1), we note that a type S error occurs when a statistically significant finding gets the sign of the parameter wrong:

$$H_i^{S,O}:\ \text{sign}(\theta_{i,O}) \neq \text{sign}(\hat{\theta}_{i,O}), \qquad \text{where } \text{sign}(x) = \begin{cases} +1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0 \end{cases}.$$

Note that $|\hat{\theta}_{i,O}|$ is always larger than c, so $\text{sign}(\hat{\theta}_{i,O}) \in \{-1, +1\}$. Letting $S_i = \text{sign}(\hat{\theta}_{i,O})$, we can rewrite the hypothesis as

$$H_i^{S,O}:\ S_i \cdot \theta_{i,O} \leq 0.$$

Here $H_i^{S,O}$ is fundamentally data-dependent as it is determined by S_i. Nonetheless it is a meaningful hypothesis: when $S_i = +1$, we want to test the null that $\theta_{i,O} \leq 0$; otherwise we want to test the null that $\theta_{i,O} \geq 0$. Our strategy is to condition on the value of S_i, since the null hypothesis is fixed again once we know S_i. We defer the discussion of valid testing of data-dependent hypotheses for now.

The question of false directional claims, then, boils down to asking how many $H_i^{S,O}$ are true: a multiple testing problem. Our estimand, the proportion of type S errors that occurred, is V/R, where V is the number of type S errors and R is the number of "discoveries," i.e. rejections. If we classify the hypotheses by whether $H_i^{S,O}$ is true and whether the test for $H_i^{S,O}$ is significant, then V and R correspond to the cell counts in Table 1.

Original p-value	$H_i^{S,O}$ is true	$H_i^{S,O}$ is false	Total
Significant	V	$*$	R
Not-significant	$*$	$*$	$*$
Total	$*$	$*$	$*$

Table 1: Classification of the hypotheses, in the style of Benjamini and Hochberg (1995). Only R is observed and we wish to infer on V.

In the multiple testing literature, V/R is called the *directional false discovery proportion* (directional FDP, or FDP_{dir}), the type S error analog of false discovery proportion (FDP; Benjamini and Hochberg, 2000). In addition to an estimate, we also provide an upper confidence bound for the directional FDP in Section 2. Both the estimator and the confidence bound are based on a "p-curve" analysis, i.e. an analysis of the distribution of significant p-values (Simonsohn et al., 2014b). We further modify these methods to evaluate the proposal to lower the statistical significance threshold by Benjamin et al. (2018).

Although $\hat{\theta}_{i,R}$ is irrelevant to testing $H_i^{S,O}$, it is informative for the closely related question of whether $\hat{\theta}_{i,O}$ incorrectly predicts the direction of the effect in a replication study, i.e.

$$H_i^{S,R}:\ S_i \cdot \theta_{i,R} \leq 0.$$

Note that S_i is computed from the original study, so this hypothesis is a measure of external validity as to the (claimed) directions of effects. If an experimental result has external validity, then any directional claim about the true effect should apply not only to the original study, but also to direct replications thereof. We provide analogous methods for multiple testing of the hypotheses $H_i^{S,R}$.

Effect shift To assess the effect shift in a specific replication attempt, we can test the hypothesis $H_i^E : \theta_{i,O} = \theta_{i,R}$ (an exact replication). As Anderson et al. (2016) noted, "there is no such thing as exact replication"; nevertheless, exactness serves usefully as an idealized null hypothesis. By inverting a test for H_i^E we can obtain a predictive interval for $\hat{\theta}_{i,R}$. Furthermore, by inverting tests for a related hypothesis $H_i^{E,\delta} : \theta_{i,O} - \theta_{i,R} = \delta$, we obtain a confidence interval for $\theta_{i,O} - \theta_{i,R}$, the effect shift in study i. Our methods explicitly take into account the truncation of $\hat{\theta}_{i,O}$.

Effect size decline The null hypothesis for effect size decline is closely related to effect shift, and can be formalized as the null hypothesis where the true effect size has declined by no more than a fraction $\rho \in [0, 1]$:

$$H_i^{D,\rho} : S_i \cdot \theta_{i,R} \geq S_i \cdot (1 - \rho)\theta_{i,O}.$$

If $S_i = +1$ and $\rho = 0.2$, for example, rejecting $H_i^{D,\rho}$ amounts to an assertion that $\theta_{i,R} < 0.8\,\theta_{i,O}$, i.e. the true effect declined by more than 20%, or is negative.

In particular, if $\rho = 0$ then $H_i^{D,0}$ is a one-sided version of H_i^E, and when $\rho = 1$, $H_i^{D,1}$ is equivalent to $H_i^{S,R}$. We can subsequently ask how many of $H_i^{D,\rho}$ are false: another multiple testing problem. We provide two estimators (one overestimate and one underestimate) and confidence interval for the proportion of false $H_i^{D,\rho}$.

1.3 Data-dependent hypotheses and conditional inference

Our hypotheses above, $H_i^{S,O}$, $H_i^{S,R}$ and $H_i^{D,\rho}$, are all innately data-dependent. While data-dependent hypotheses may at first sound unusual, they are commonplace in practice, for example when pilot studies are performed to generate hypotheses that are tested later on with fresh data. There is no inherent conceptual problem with testing these data-dependent hypotheses: intuitively, we understand that the test remains valid because the type I error rate is controlled for whatever hypothesis is selected, conditional on that hypothesis having been selected.

Conditional inference is well-established in the statistical literature as a means of constructing valid confidence intervals for parameters that were selected in a data-dependent way (e.g. Sampson and Sill, 2005; Zöllner and Pritchard, 2007; Weinstein et al., 2013; Yekutieli, 2012). Fithian et al. (2014) generalized the intuition about pilot studies to argue that a test of a data-dependent hypothesis is valid, so long as the type I error rate is controlled conditioned on the portion of the data that generated the hypothesis. For our hypotheses here, S_i is the part of the data that determines the hypothesis: in effect, we can imagine ourselves in the position of having observed the signs of all the original estimators, but knowing nothing else about the data. At that stage, it is valid to formulate a hypothesis that depends on S_i, and plan to test it using the still-unobserved data: namely, $|\hat{\theta}_{i,O}|$ and $\hat{\theta}_{i,R}$.

After conditioning on S_i, each hypothesis discussed above amounts to testing a fixed linear hypotheses about $(\theta_{i,R}, \theta_{i,O})$, the natural parameter of the truncated bivariate normal model (1);

as a result, they are all amenable to post-selection inference using the selective z-test built on the work of Lee et al. (2016). Section 2 discusses the methodology in detail.

1.4 Related work

There has been much commentary on how to define replicability for scientific experiments. Valentine et al. (2011) pointed out that the definition should depend on the scientific context. For example, sometimes one may wish to test the robustness of conclusions to subpopulation differences, but in other times, to changes in experimental conditions. Goodman et al. (2016) expanded on this, and gave a few useful definitions for what replicability is, such as *methods reproducibility*, *results reproducibility*, *inferential reproducibility*, etc., but stopped short of an operational statistical criterion for replicability. False directional claims and effect shift can be loosely interpreted as inferential and results reproducibility, respectively.

Operationally, Valentine et al. (2011) and Nosek and Errington (2017) proposed the metrics used in RP:P and Camerer et al. (2018), a similar replication effort in experimental economics. These metrics however suffer the shortcomings discussed earlier, in that they do not answer a concrete statistical question and cannot disentangle selection bias from other explanations.

In this article, our definitions of replicability are inspired primarily by the statistical literature on multiple testing and meta-analysis, such as the estimator in Storey (2002), the FDP and directional FDP from Benjamini and Hochberg (2000); Benjamini and Yekutieli (2005), and the partial conjunction testing framework of Heller et al. (2007); Benjamini and Heller (2008). Related error rates have also been estimated before: Jager and Leek (2013) have modeled the p-value distributions under alternatives and the selection for statistical significance to estimate the FDR in the medical literature, accompanied by useful discussions from Gelman and O'Rourke (2013); Goodman (2013); Ioannidis (2013); in addition, Camerer et al. (2018) used Bayesian methods to estimate the false positive rate, instead of the FDR, for published social science results in *Nature* and *Science*.

Furthermore, there are many past efforts to model and quantify selection bias, specifically using the RP:P dataset. For instance, Johnson et al. (2017) considered a publication bias model where the probability of publication is a step function of the p-value, which is generalized nonparametrically in Andrews and Kasy (2018). The two analyses estimated that a statistically significant result was 200 (Johnson et al., 2017) or 30 (Andrews and Kasy, 2018) times as likely to be published as a statistically insignificant one.

Adjusting for selection, van Aert and van Assen (2017, 2018) have combined the evidences from both the original and replication experiments to provide estimates for the effect sizes. Specifically with a truncated Gaussian model, Etz and Vandekerckhove (2016) have also analyzed the RP:P dataset from a Bayesian perspective, and investigated the discrepancies between the original and replication studies. Our analysis provides a complementary point of view with frequentist hypothesis testing without any prior on the effect sizes, with the help of recent advances in post-selection inference, including primarily the selective z-test framework of Lee et al. (2016).

1.5 Outline

Section 2 details the methodology and assumptions used in this analysis, and is somewhat technical. Section 3 applies the developed methodology to the RP:P dataset, summarizes and interprets the results. Section 4 concludes.

2 Methodology

In this section we will construct an estimator for directional FDP, a test for the effect shift in replication i and an estimator for the proportion of effect sizes that declined. We also use $X \geq_{\text{st}} Y$ to denote that X is stochastically larger than Y. The index i is suppressed when there is no risk of ambiguity.

Since we need a well-defined notion of direction to consider the proportion of false directional claims, we restrict our attention to univariate tests, namely z-, t-, $F(1, \cdot)$-tests or correlations. Thus, studies that are not univariate or have p-values greater than $\alpha_0 = 0.05$ are discarded: our estimates and analyses below consider only the $m = 68$ remaining studies with univariate structure and conventionally significant original p-values.

2.1 Selection bias model

Model (1) assumes that results are only published if they achieved statistical significance at some conventional threshold level α_0, which is 0.05 in our data. While this assumption is not literally true in the case of RP:P since some original p-values are above 0.05, we note that the model can be relaxed to the following milder assumption:

Assumption 1. $p_O < \alpha_0$ is "significant enough": that is, not all results with significant p-values are necessarily published, but a result with $p_O < \alpha_0$ would have been equally likely to be published (or selected for replication), had the p-value taken on some other statistically significant value.

If Assumption 1 holds, then we can model the original test statistics as following their theoretical distribution, truncated to the event where the corresponding p-values are below α_0, as in Model 1.

Note that Assumption 1 contemplates a fairly straightforward mechanism for selection on statistical significance, which may not be adequate to describe the effects of more complex and difficult-to-model QRPs. In particular, p-hacking — the iterative tweaking of an analysis until the p-value drops below the researcher's desired significance level α_0 — is commonly suspected to produce a pileup of p-values just below the significance threshold (see e.g. Simonsohn et al., 2014b). Because p-hacking is such a vaguely defined practice, it is unclear how we might incorporate it into our model, but in any case there is no evidence of a pileup just below 0.05 in the original RP:P studies (see Figure 6a).

2.2 False directional claims

We will adapt the method in Storey (2002) to estimate the directional FDP while accounting for selection bias. Furthermore, if we believe the chosen studies are representative of the publications in the journal or discipline (e.g. Stroebe, 2016), then this estimator can also be regarded as an estimator for the journal-wide or discipline-wide directional false discovery rate (FDR_{dir}), the expectation of the directional FDP (Benjamini and Yekutieli, 2005).

Adjusting for selection bias While dividing a post-selection p-value by α_0 intuitively adjusts for selection, it is not immediately valid when the null is one-sided with a true effect not on the boundary. We demonstrate below that this adjustment typically remains valid even in this case.

Recall that a valid p-value is a random variable that is stochastically larger than Uniform$[0,1]$ (i.e. superuniform) under the null hypothesis. If we only observe the original p-value when it is significant, it is not superuniform after selection under $H^{S,O}$, and it is therefore not valid for testing the hypothesis of a false directional claim. To adjust these p-values for selection, we follow the principle in Fithian et al. (2014) by conditioning on the event that the p-values are selected, and also on the variable $S = \text{sign}(\hat{\theta}_O)$ which determines the hypothesis $H^{S,O}$ that we test. We consider two cases: when the original study is a one-sided test and when it is a two-sided test. As we will see, the adjustment in either case is to divide by α_0.

First we consider the case where the original study was a one-sided test. Assume p_O is a p-value for a test of the hypothesis $H_0 : \theta_O \leq 0$, in which case $S = +1$ deterministically (the opposite case with $H_0 : \theta_O \geq 0$, and $S = -1$ deterministically, is directly analogous). Suppose p_O is the original p-value, which we observe only when it is significant at the conventional threshold, i.e. when $p_O < \alpha_0$. Under mild assumptions satisfied by both z-tests and t-tests,[5] $p_O \geq_{\text{st}} \text{Uniform}[0, \alpha_0]$ under $H^{S,O}$, in which case $p_O/\alpha_0 \geq_{\text{st}} \text{Uniform}[0,1]$.

Next we consider the case where p_O is a p-value for a two-sided test of $H_0 : \theta_O = 0$, and where $S = +1$ (the case with $S = -1$ is analogous). If p_O^+ was the original one-sided p-value for $H_0 : \theta_O \leq 0$, then $p_O = 2p_O^+$ when $S = +1$ ($p_O = 2 - 2p_O^+$ if $S = -1$). In our truncated model, under the same assumptions as above and conditional on $S = +1$, $p_O^+ \geq_{\text{st}} \text{Uniform}[0, \alpha_0/2]$ and therefore $p_O/\alpha_0 = 2p_O^+/\alpha_0 \geq_{\text{st}} \text{Uniform}[0,1]$ under $H^{S,O}$. We write $p'_O = p_O/\alpha_0$ for the adjusted p-value.

Inference on FDP: estimate and upper confidence bound Using the adjusted original p-values, we can estimate the directional FDP in the original studies. Recall from Table 1 that

$$R = \#\{p_{i,O} \leq \alpha_0\} = m,$$
$$V = \#\{p_{i,O} \leq \alpha_0 \text{ and } H_i^{S,O} \text{ is true}\}.$$

Since all of the studies were deemed discoveries, $R = m$ is the total number of studies here. Table 2 classifies the m conventionally significant studies according to whether $H_i^{S,O}$ is true and whether the adjusted p-value is larger than some fixed value λ in $(0,1)$, e.g. $\lambda = 0.5$.

Adjusted p-value	$H_i^{S,O}$ is true	$H_i^{S,O}$ is false	Total
$p'_{i,O} < \lambda$	$*$	$*$	$*$
$p'_{i,O} \geq \lambda$	U	$*$	B
Total	V	$*$	$R = m$

Table 2: Classification of the $R = m$ significant original studies. Here only R and B are observed, and we wish to infer on V.

Note that $B = \#\{\lambda\alpha_0 \leq p_{i,O} < \alpha_0\}$ from Table 2 is observable, while V and U are not. Under the one-sided null, the p-value is superuniform, and so

$$B \geq_{\text{st}} U \geq_{\text{st}} \text{Binomial}(V, 1-\lambda). \tag{2}$$

[5] namely, that the test statistic has monotone likelihood ratio in the parameter

As a result, $\mathbb{E}[B] \geq (1-\lambda)V$ and a conservative (upwardly biased) estimator of the directional FDP is

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{B}{(1-\lambda)R}.$$

This estimate is conservative in the sense that it overestimates the type I error, and is equivalent to the estimator $\hat{\pi}_0$ of the true null proportion in Storey (2002). Using $\lambda = 0.5$ and $\alpha_0 = 0.05$, the estimate boils down to

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{2}{m} \cdot \#\{0.025 \leq p_{i,O} < 0.05\}.$$

While the above is formally an estimator for the number of directional errors, it can be interpreted practically as an estimate of the fraction of directional claims where *either* the direction is wrong *or* the effect has a negligible magnitude, cf. type M error from Gelman and Carlin (2014). This is because p-values whose effect sizes are very close to zero are nearly uniform and contribute to our estimator similarly as if the true effect were exactly zero.

Additionally, we can exploit (2) to obtain an upper confidence bound for the directional FDP, by testing the hypothesis $H_0 : V \geq v_0$, a partial conjunction hypothesis investigated in Heller et al. (2007). Here we combine only the coarse information of whether each p-value is greater than λ,[6] and reject for small values of B. We can compute the largest v_0 such that the test still accepts, which gives an upper confidence bound of V. Dividing this bound by R gives an upper confidence bound for the directional FDP.

Directional FDP at smaller thresholds One proposal to address the replicability crisis is to lower the conventional significance threshold from $\alpha_0 = 0.05$ to some smaller value α, such as 0.005 (Benjamin et al., 2018). As suggested by Goodman (2013), an empirical method to evaluate the hypothetical scenario with a smaller threshold can be helpful. We now discuss methods for inference on the directional FDP for those studies with $p_O < \alpha < \alpha_0$, based on comparing the number of adjusted p-values below α with the number above $\lambda\alpha_0$, for some $\lambda > \alpha/\alpha_0$. We call this method the *external comparison method* in contrast to the earlier *internal comparison method*. This method will be less conservative as we are not constrained to only using the p-values in $[0, \alpha)$.

Let $N \leq m$ denote the total number of original p-values in $[0, \alpha) \cup [\lambda\alpha_0, \alpha_0)$ (or equivalently, the number of adjusted p-values in $[0, \alpha') \cup [\lambda, 1)$ for $\alpha' = \alpha/\alpha_0$). Table 3 classifies these N studies according to whether $H_i^{S,O}$ is true and whether the adjusted p-value is larger than λ or smaller than α'. The numbers of false directional claims and all directional claims under the hypothetical threshold are V_α and R_α, respectively. Auxiliary counts, T_α and W, are defined according to Table 3 as well. The directional FDP, V_α/R_α, remains as our quantity of interest.

Our method is inspired by the following stochastic inequality.

Lemma 1. *Conditional on N, T_α and W, we have*

$$B \mid N, T_\alpha, W \geq_{st} \text{Binomial}(N - T_\alpha, \beta). \tag{3}$$

Proof. All adjusted p-values are independent, and are either small ($p \leq \alpha'$) or big ($p \geq \lambda$). The adjusted p-values corresponding to a true null are big with probability at least $\beta = \frac{1-\lambda}{1-\lambda+\alpha'}$.

[6] More precisely, we count number of p-values that are greater than λ and consider its distribution under the partial conjunction null hypothesis

Adjusted p-value	$H_i^{S,O}$ is true	$H_i^{S,O}$ is false	Total
Small ($p'_{i,O} < \alpha'$)	V_α	T_α	R_α
Big ($p'_{i,O} \geq \lambda$)	U	W	B
Total	N_0	$*$	N

Table 3: Classification of the $N \leq m$ original studies with adjusted p-values in $[0, \alpha'] \cup [\lambda, 1]$. Only R_α, B and N are observed. Auxiliary unobserved quantities, N_0, T_α and R_α, are defined accordingly. Our goal is to infer on V_α.

We proceed to condition on T_α and W, so they are now considered deterministic. So the total number of big adjusted p-values, B, satisfies

$$B = U + W \geq_{\text{st}} \text{Binomial}(N - N_0, \beta) + W \geq_{\text{st}} \text{Binomial}(N - T_\alpha, \beta).$$

□

With (3), we can estimate $N - T_\alpha$ conservatively with B/β. Since $V_\alpha = N - T_\alpha - B$, a reasonable estimator for the directional FDP is

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{1-\beta}{\beta} \cdot \frac{B}{R_\alpha}.$$

Furthermore (3) gives us a 95% upper confidence bound for the directional FDP:

$$\text{FDP}^*_{\text{dir}} = \frac{Q - B}{R_\alpha}, \qquad \text{where } Q = \max\{q : \mathbb{P}[\text{Binomial}(q, \beta) \geq B] \geq 0.95\}.$$

Proposition 2. *The expectation of $\widehat{FDP}_{dir}$ is at least the expectation of the true directional FDP, and FDP^*_{dir} is greater than the true directional FDP, with probability at least* 95%.

Proof. For the estimator, we start by taking the expectation of $\widehat{\text{FDP}}_{\text{dir}} - \text{FDP}_{\text{dir}}$, conditional on N, T_α and W:

$$\begin{aligned}
\mathbb{E}[\widehat{\text{FDP}}_{\text{dir}} - \text{FDP}_{\text{dir}} \mid N, T_\alpha, W] &= \mathbb{E}\left[\frac{\frac{1-\beta}{\beta}B - V_\alpha}{R_\alpha} \,\middle|\, N, T_\alpha, W\right] \\
&\geq \mathbb{E}\left[\frac{\frac{1-\beta}{\beta}(N_0 - V_\alpha) - V_\alpha}{V_\alpha + T_\alpha} \,\middle|\, N, T_\alpha, W\right] \\
&= \mathbb{E}\left[\frac{(1-\beta)N_0 - V_\alpha}{\beta(V_\alpha + T_\alpha)} \,\middle|\, N, T_\alpha, W\right] \\
&\geq \frac{(1-\beta)N_0 - \mathbb{E}[V_\alpha \mid N, T_\alpha, W]}{\beta(\mathbb{E}[V_\alpha \mid N, T_\alpha, W] + T_\alpha)} \qquad (4) \\
&\geq 0, \qquad (5)
\end{aligned}$$

where (4) follows from applying Jensen's inequality to the convex function $f(x) = \frac{(1-\beta)N_0 - x}{\beta(x+T_\alpha)}$, and (5) follows from $V_\alpha \mid N, T_\alpha, W \leq_{\text{st}} \text{Binomial}(N_0, 1-\beta)$. Taking expectation on both sides completes the proof.

For $\text{FDP}^*_{\text{dir}}$, we can directly compute the probability that it is greater than FDP_{dir}, conditional on N, T_α and W:

$$\begin{aligned}\mathbb{P}[\text{FDP}^*_{\text{dir}} \geq \text{FDP}_{\text{dir}} \mid N, T_\alpha, W] &= \mathbb{P}\left[\frac{Q-B}{R_\alpha} \geq \frac{V_\alpha}{R_\alpha} \,\middle|\, N, T_\alpha, W\right] \\ &= \mathbb{P}[Q \geq B + V_\alpha \mid N, T_\alpha, W] \\ &= \mathbb{P}[Q \geq N - T_\alpha \mid N, T_\alpha, W] \\ &\geq 0.95,\end{aligned}$$

from the construction of Q. Taking expectation on both sides hence yields the desired marginal coverage. □

Remark. This proof of conservativeness actually shows something stronger than marginal guarantees: the estimator and confidence upper bound are both conservative conditionally, even when we condition on the signs S_i.

Methods using replication p-values As mentioned in Section 1, we can use the replication p-values in lieu of the adjusted original p-values above, providing an estimate and confidence bound for the frequency of when the $\hat{\theta}_O$ incorrectly predicts the replication effect direction. While this approach requires potentially costly replications in future applications, it provides valuable additional information. In particular, the replication p-values are more likely to be free of QRPs or p-hacking that may violate our assumption that adjusted p-values are superuniform under the null, providing more robust evidence regarding replicability. The corresponding estimator for unadjusted replication p-values with $\lambda = 0.5$ is

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{2}{m} \cdot \#\{p_{i,R} \geq 0.5\}.$$

2.3 Effect shift

We will derive a test for the hypothesis $H^E : \theta_O = \theta_R$ at level 0.05. Our test is based on a normal distribution, so we start by demonstrating that the effect size estimates of the univariate studies can be reasonably modeled by our truncated bivariate normal distribution in model (1). We classify these studies into two categories and provide a rough rationale in our definition of effect size in each category: (1) t-tests and $F(1, \cdot)$ ANOVAs, where all independent variables are categorical; and, (2) correlations and regressions, where one or more independent variables are continuous.

For a t-test or $F(1, \cdot)$ ANOVA, we can define the effect size as the noncentrality parameter, scaled for cell sizes. In other words, the t-statistic is distributed as $T \sim t_{df}(k\theta)$, for some real constant k chosen based on the study design. For example, $k = \sqrt{n}$ for a one-sample t-test. When df is sufficiently large, the t-statistic is approximated well by a z-statistic, and distributed approximately as

$$T \sim N(k\theta, 1).$$

For our analysis, we consider studies where the original and replication degrees of freedom are at least 30.[7]

[7] The choice of 30 complies with the analysis in Andrews and Kasy (2018). Further discussion on the approximation in available in the supplement.

For a (partial) correlation coefficient estimate, R, we can apply Fisher transformation (1921; 1924) to convert it into a z-statistic, which approximately follows

$$\sqrt{n-3-p}\tanh^{-1}(R) \sim N(\sqrt{n-3-p}\,\theta, 1),$$

where p is the number of controlled covariates and θ is a quantity that can be taken as the effect size.

In either case, the test statistic in 46 studies can be transformed to an approximate z-score $Z \sim N(k\theta, 1)$ for some real constant k. Additional considerations in certain studies are detailed in the supplement.

Adjusting for selection bias We turn next to address the issue of post-selection inference. Again, we condition on the event where the z-scores are observed, but we do not need to condition on S as the hypothesis H^E is no longer random. Since the statistic is only observed if it is statistically significant, the original and replication z-statistics follow a truncated bivariate normal joint distribution:

$$\begin{bmatrix} Z_O \\ Z_R \end{bmatrix} \sim N\left(\begin{bmatrix} k_O\theta_O \\ k_R\theta_R \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) 1_{\{Z_O \in A\}}.$$

Here A is the selection event, which contains the statistically significant values of Z_O. We are interested in testing $H^E : \theta_O = \theta_R$ and more generally the null hypothesis $H^{E,\delta} : \theta_O - \theta_R = \delta$, which can be inverted to yield a confidence interval.

We cast this as a more general testing problem here to benefit later derivations on effect decline. Suppose we have a truncated bivariate distribution

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\mu, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) 1_{\{Z_1 \in A\}}, \quad \text{where } \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

and we want to test $\eta'\mu = \delta$ for some constant vector $\eta = (\eta_1, \eta_2)$ with $\eta_1 > 0$. Test for H^E and $H^{E,\delta}$ are special cases where $\eta = (1/k_O, -1/k_R)$.

We can perform this general testing problem with the *selective z-test*, based on the framework in Lee et al. (2016).

Definition 1 (Selective z-test)**.** *Let $\eta_\perp = (\eta_2, -\eta_1)$, $D = \eta' Z$ and $M = \eta'_\perp Z$. We now consider M as a constant and test $\eta'\mu = \delta$ using the test statistic D against the null distribution*

$$N(\delta, \|\eta\|^2) 1_{\left\{D \in \frac{\|\eta\|^2 A - \eta_2 M}{\eta_1}\right\}}.$$

Specifically, we reject $\eta'\mu = \delta$ when D is below the $\frac{0.05}{2}$-quantile or over the $(1 - \frac{0.05}{2})$-quantile of this null distribution.

We proceed to show that this is a valid test by construction.

Proposition 3. *The selective z-test defined in Definition 1 has level* 0.05*.*

Proof. Leveraging the fact that $\eta'\eta_\perp = 0$, we reparametrize the joint distribution of (Z_1, Z_2) under the null such that δ is a parameter, i.e.

$$\begin{bmatrix} D \\ M \end{bmatrix} = \begin{bmatrix} \eta'\mu \\ \eta'_\perp\mu \end{bmatrix} \sim N\left(\begin{bmatrix} \delta \\ \eta'_\perp\mu \end{bmatrix}, \begin{bmatrix} \|\eta\|^2 & 0 \\ 0 & \|\eta\|^2 \end{bmatrix}\right) 1_{\{Z_1 \in A\}}.$$

In particular, the event $Z_1 \in A$ can be rewritten as

$$D \in \frac{\|\eta\|^2 A - \eta_2 M}{\eta_1}.$$

And so the distribution of D conditional on M under H_0^δ is a truncated Gaussian distribution,

$$[D \mid M] \sim N\left(\delta, \|\eta\|^2\right) 1_{\left\{D \in \frac{\|\eta\|^2 A - \eta_2 M}{\eta_1}\right\}}$$

and we obtain a valid test by rejecting when D is smaller than the $\frac{0.05}{2}$-quantile or larger than the $\left(1 - \frac{0.05}{2}\right)$-quantile. □

The construction above is represented graphically in Figure 5, in the style of Lee et al. (2016). We can represent the observation (Z_1, Z_2) as a point in $\mathbb{R}^2$. Conditioning on M is equivalent to conditioning on $M/\|\eta_\perp\|$, which means we are now considering the conditional distribution on the truncated line ℓ. The test statistic D, or equivalently $D/\|\eta\|$, indicates the position on ℓ. Under the null that $\eta'\mu = \delta$, the conditional distribution on ℓ is known and a valid p-value can be obtained, yielding the selective z-test.

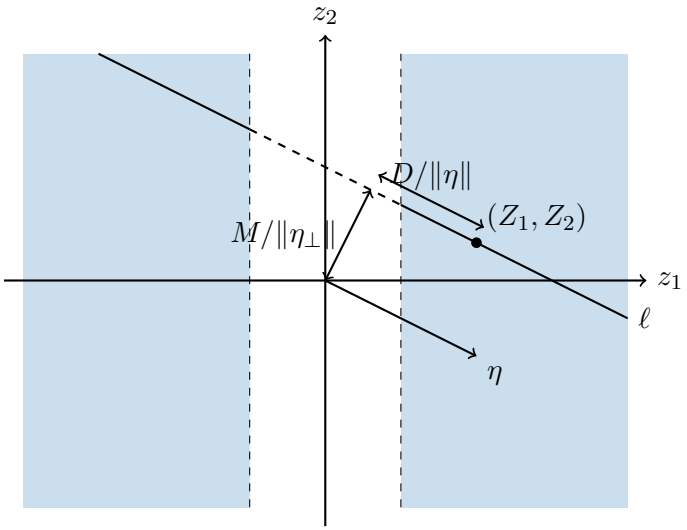


Figure 5: Graphical representation of the selective z-test. The observation (Z_1, Z_2) is a point and the truncation on Z_1 means that the shaded area is the support of the joint distribution (Z_1, Z_2). Conditioning on M is the same as conditioning on $M/\|\eta_\perp\|$, so we now consider the conditional distribution on the truncated line ℓ. The test statistic D indicates the position on ℓ. Under the null $H^{E,\delta} : \theta_1 - \theta_2 = \delta$, the conditional distribution on ℓ is known and a valid p-value can be obtained, yielding the selective z-test.

Remark. It is not necessary to use $\frac{0.05}{2}$- and $(1 - \frac{0.05}{2})$-quantiles of the null distribution, as long as the desired significance level is achieved under the null distribution. For example, a uniformly most powerful unbiased test can be used in lieu of a test with equal tail cutoffs. Furthermore, if we are interested in a one-sided hypothesis, e.g. $\eta'\mu \leq 0$, we can reject on one tail only. This will be particularly useful for derivations about effect decline later.

Interval estimation Given a valid test $\phi(Z_O, Z_R)$ for testing $H^{E,\delta} : \theta_O - \theta_R = \delta$, we can obtain two intervals: a predictive interval for the replication effect size estimate, and a confidence interval for effect shifts.

Under the null hypothesis $H^E : \theta_O = \theta_R$, $\mathbb{P}[\phi(Z_O, Z_R) \text{ rejects}] = 0.05$, or equivalently,

$$\mathbb{P}[\{z_R : \phi(Z_O, z_R) \text{ accepts}\} \ni Z_R] = 0.95.$$

Hence $\{z_R : \phi(Z_O, z_R) \text{ accepts}\}$ is a predictive interval for Z_R, which translates to a predictive interval for the point estimate $\hat{\theta}_R$ of the replication effect size.

By the duality of hypothesis testing and confidence set, the set

$$\{\delta : H^{E,\delta} \text{ is rejected}\}$$

covers the difference of the original and replication effect sizes with probability 95%.

2.4 Effect decline

We will estimate the proportion of effect sizes that declined by at least a fraction of ρ. Our procedure consists of two parts: (1) for each study i, test and produce a p-value for the hypothesis $H_i^{D,\rho}$, and (2) adapt the method for the directional FDP to estimate the proportion of $H_i^{D,\rho}$ that are false.

Adjusting for selection bias As with the exactness test, we condition not only on the event where the z-scores are observed, but also on $S = \text{sign}(\hat{\theta}_O)$ as our hypothesis $H^{D,\rho}$ is determined by this random variable. In other words, we consider the z-statistic Z_O to be drawn from the set A_+, where A is the selection event from our test for effect shift and

$$A_+ = A \cap \mathbb{R}_+ = \{z_O : z_O \text{ is statistically significant}\} \cap \mathbb{R}_+.$$

Putting Z_O and Z_R together, they follow a truncated bivariate normal joint distribution:

$$\begin{bmatrix} Z_O \\ Z_R \end{bmatrix} \sim N\left(\begin{bmatrix} k_O\theta_O \\ k_R\theta_R \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) 1_{\{Z_O \in A_+\}}.$$

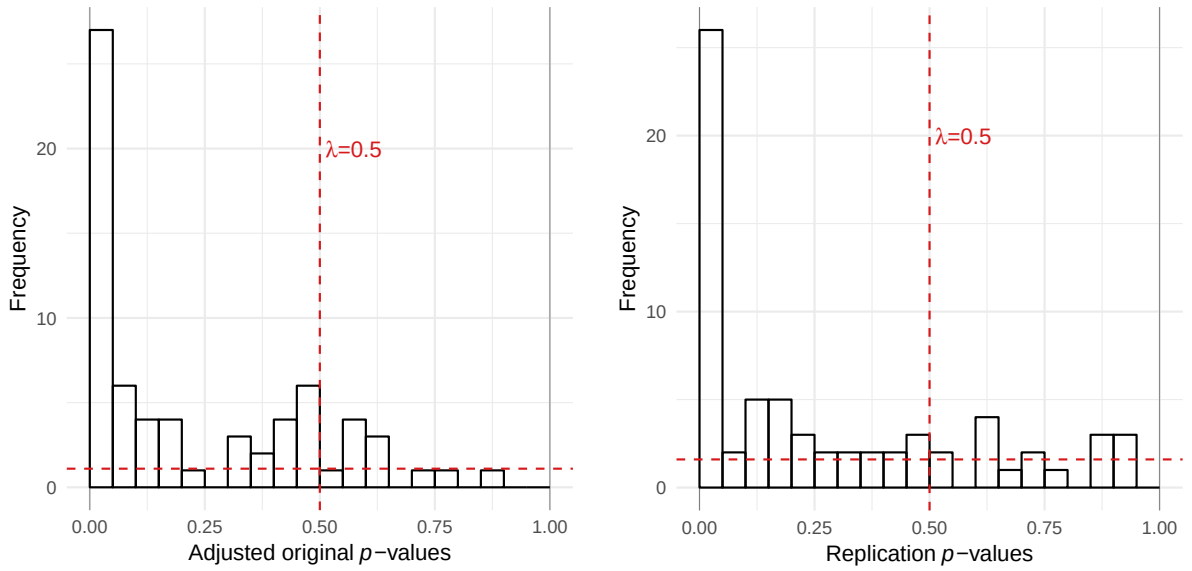
By convention RP:P chose $\hat{\theta}_O > 0$ so the hypothesis $H^{D,\rho}$ reduces to $\theta_{i,R} \geq (1-\rho)\theta_{i,O}$, or equivalently $\theta_{i,R} - (1-\rho)\theta_{i,O} \geq 0$. This can be tested using the selective z-test with $\eta = (1/k_O, -1/(1-\rho)k_R)$ and rejecting on one tail only.

Inference on effect decline: estimates and confidence bounds With the resulting p-values, our earlier methods on directional FDP can provide an overestimate and a upper confidence bound for the proportion of true $H^{D,\rho}$. Subtracting these from 1 yields an underestimate and a lower confidence bound for the proportion of false $H^{D,\rho}$. On the other hand, by considering the complement of the hypothesis $H^{D,\rho}$, we can also provide an overestimate and an upper confidence bound for the proportion of false $H^{D,\rho}$. These estimators and bounds together provide an overestimate, an underestimate and a 90% confidence interval for the proportion of effect sizes that at least declined by a fraction of ρ.

3 Re-analysis of RP:P

3.1 False directional claims

We implemented our method with $\lambda = 0.5$ to estimate the number of one-sided nulls and the directional FDP.[8] The adjusted original p-values and replication p-values are given in Figures 6a and 6b respectively. Using the original p-values, we estimate that 22 of the 68 (32%) original directional claims are false, with a 95% upper confidence bound of 47%. Using the replication p-values, we estimate that 32 of the 68 (47%) original directional claims incorrectly predict the direction of the replication effect, with a 95% upper confidence bound of 63%. In particular both of our FDP estimates are much lower than the 64% which could be suggested by a naive reading of RP:P (e.g. Baker, 2015). These numbers are summarized again in Table 4 later. Furthermore, while we can compute a lower confidence bound, it will always be 0% as the data is obviously consistent with many null hypotheses being slightly false.



(a) Histogram of the adjusted original p-values.

(b) Histogram of the replication p-values.

Figure 6: Histograms of p-values. We estimate the expected number of true nulls in each bin by the method from Storey (2002), shown by the horizontal red line. A net excess of p-values above this line means false directional claims.

We proceeded to evaluate the proposal to reduce the statistical significance threshold (Benjamin et al., 2018). We considered three candidates for the new threshold, 0.001, 0.005 and 0.01, using the external comparison method. The directional FDP estimates and upper confidence bounds are given in Table 4.

These estimates corroborate Benjamin et al. (2018)'s suggestion that reducing the statistical significance threshold may improve replicability, at least regarding the directional FDP of the original statistical hypotheses (of course, there is no way to account for potential change in

[8] Choosing $\lambda = 0.5$ follows the convention in the multiple testing literature for a bias-variance trade off: if λ is too small, many true discoveries are counted as false; if λ is too big, the estimator can have large variance.

α	Adjusted original		Replication	
	Est.	U.C.B.	Est.	U.C.B.
0.001	0.4/22 = 2%†	2/22 = 9%†	6/22 = 27%	12/22 = 55%
0.005	2.2/33 = 7%†	6/33 = 18%†	12/33 = 36%	20/33 = 61%
0.01	4.4/41 = 11%†	9/41 = 22%†	16/41 = 39%	25/41 = 61%
0.05	22/68 = 32%	32/68 = 47%	32/68 = 47%	43/68 = 63%

Table 4: The directional FDP estimates and 95% upper confidence bounds, using the adjusted original and replication p-values. The statistical significance level is α. The external comparison method was used for computing the directional FDP estimates and the upper confidence bounds marked with daggers(†) above, as information of p-values between α and 0.05 can improve the precision. The estimates and upper confidence bounds in the "Replication" column are relatively noisy due to the small number of p-values below the stricter rejection thresholds, and give little basis for any conclusions.

researcher's behavior in response to the lowered threshold). Shall this be of interest, this method provides an empirical way to determine a better significance threshold, as no replications are needed. Nonetheless, potential effect heterogeneity is often a bigger concern. In this case, we are more concerned about the directional FDP for replications, which remains unacceptably high and requires replication experiments. Note, however, that a replication with low power could contribute to our estimates, even if there were no type S error.

3.2 Effect shift

We performed the selective z-test for the hypothesis $H^E : \theta_O = \theta_R$ while adjusting for selection, where seven (15%) studies are rejected. In contrast, without adjusting for selection, 18 (39%) studies are rejected at 0.05 significance. If we wish to correct for multiplicity, we can apply Benjamini–Hochberg procedure (1995), which rules five (11%) replication studies as inconsistent with the original studies at false discovery rate 0.10.[9] Applying the more stringent Holm's method (1979) to control the familywise error rate rules only the replication of Farris et al. (2008) as inconsistent at familywise error rate 0.05.

We inverted the test for the hypothesis H^E, to yield a predictive interval for Z_R and hence a predictive interval for the replication effect size estimate $\hat{\theta}_R$, shown in Figure 7. By definition H^E is rejected when $\hat{\theta}_R$ is not included in the predictive interval. Adjusting for selection generally stretches the predictive intervals, resulting in fewer rejections.

We also inverted the test for $H^{E,\delta}$ and obtained a confidence interval for the effect shifts, $\theta_O - \theta_R$, given in Figure 8. By construction the null hypothesis $H^E : \theta_O = \theta_R$ is rejected when the confidence interval does not include 0. Adjusting for selection also generally lengthens the confidence intervals, resulting in fewer rejections.

If all procedures are replicated perfectly, we should expect to reject 5% of the tests on average, rather than the observed 15%, and after the Benjamini–Hochberg correction, there would be no rejection with 90% probability. In other words, while selection bias can partly

[9]The five rejected studies are Dodson et al. (2008); van Dijk et al. (2008); Purdie-Vaughns et al. (2008); Farris et al. (2008); Larsen and McKibban (2008).

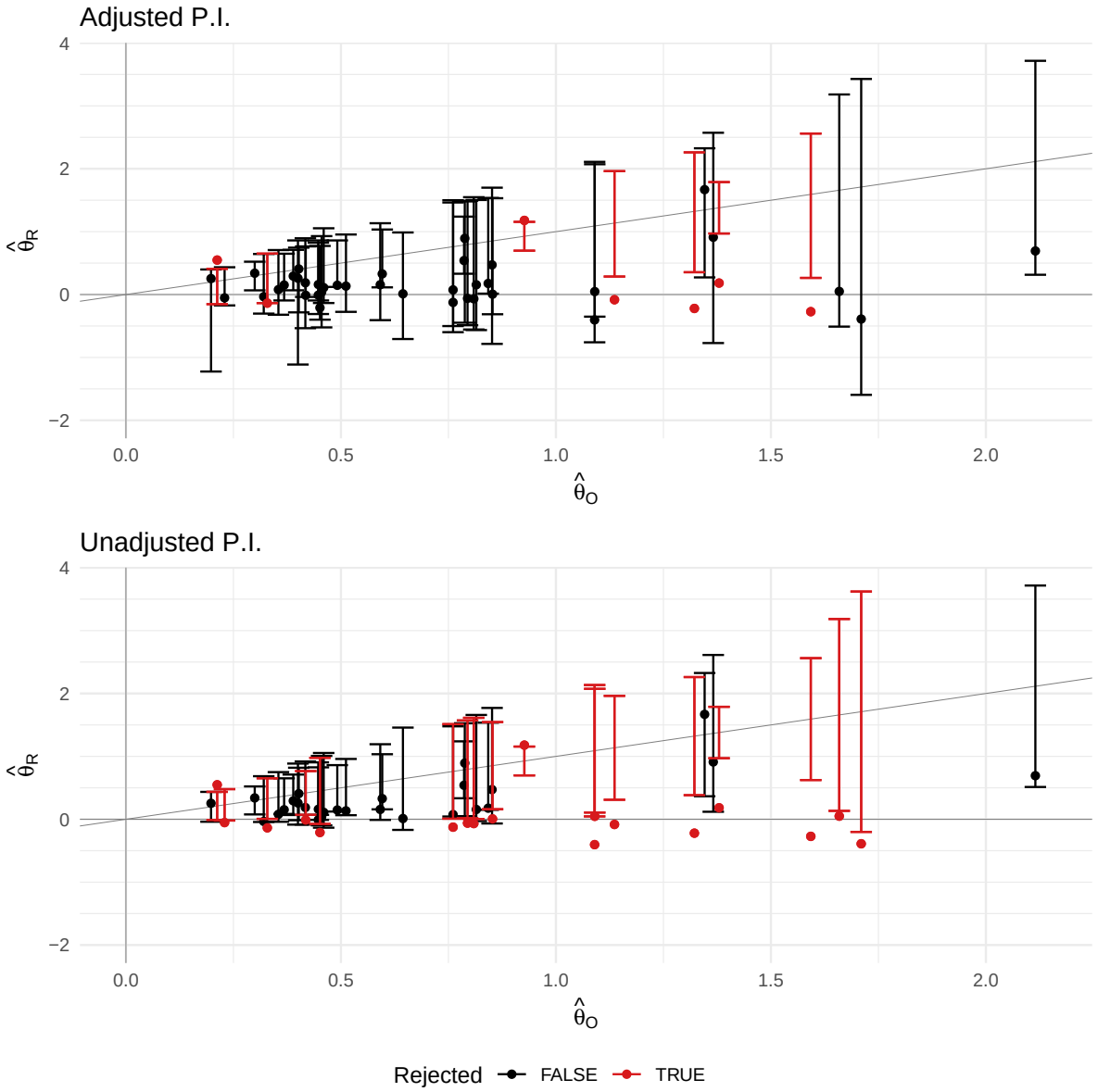


Figure 7: Predictive intervals for $\hat{\theta}_R$, both adjusted and unadjusted for selection, overlay with a plot of $\hat{\theta}_R$ against $\hat{\theta}_O$. Studies 36 and 145 are not shown here. By definition we reject $H_0 : \theta_O = \theta_R$ whenever the replication effect size estimate lies outside of the predictive interval. The intervals are generally longer after adjusting for selection.

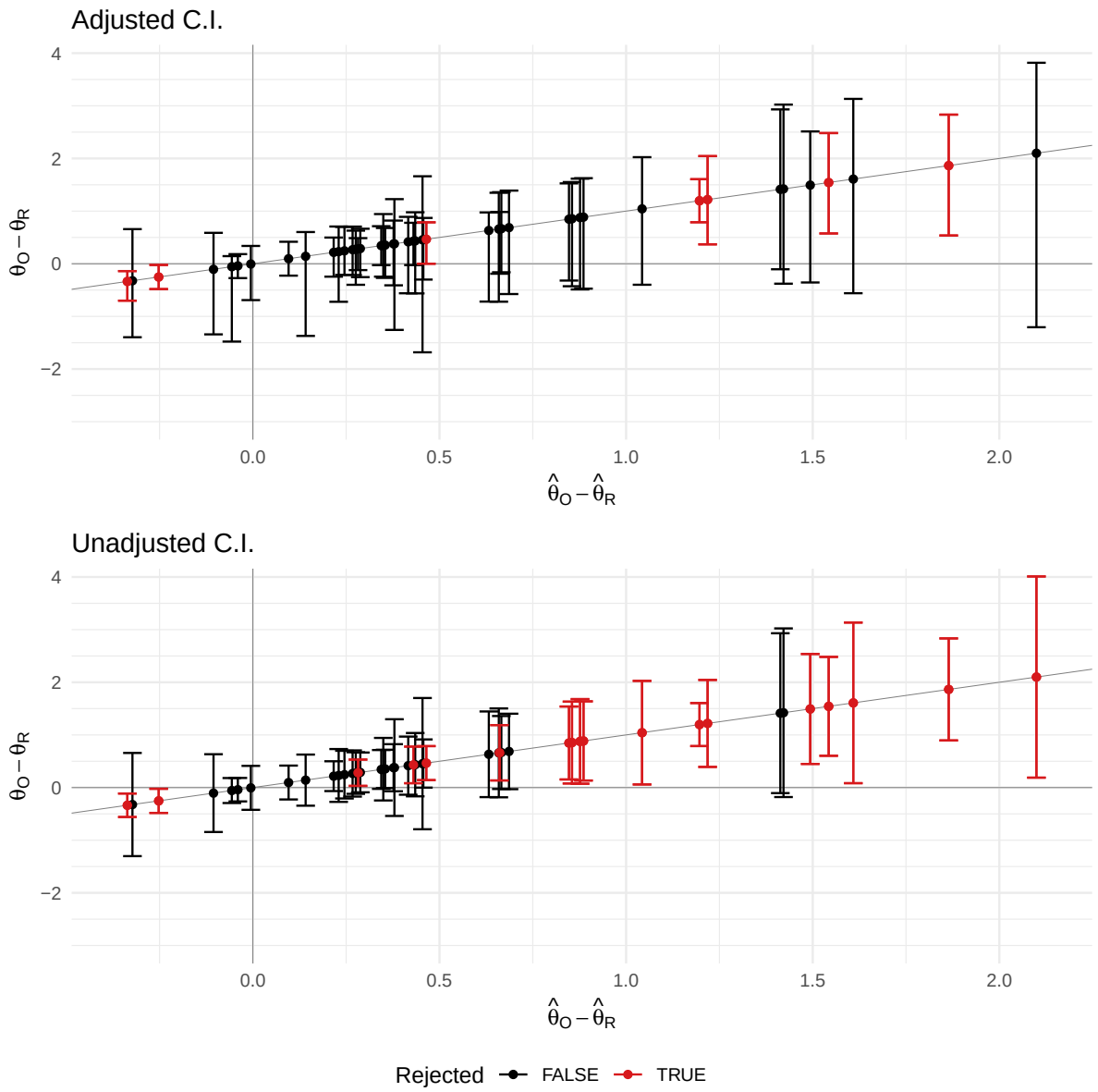


Figure 8: Confidence intervals for $\theta_O - \theta_R$, both adjusted and unadjusted for selection. By construction the null hypothesis $H_0 : \theta_O = \theta_R$ is rejected when the confidence interval does not include 0. Many of the adjusted intervals are fairly long as either the replication studies suffer low power or the original effect size estimate is near the rejection threshold. The intervals are generally longer after adjusting for selection.

explain the discrepancies between the original and replication studies, it does not explain all of it. Nevertheless, the RP:P data cannot be taken as strong evidence of widespread failure by replication teams to satisfactorily repeat the same experiment performed in the original study. The lack of strong evidence is hardly surprising: if the original study lacks power (Morey and Lakens, 2017) or $\hat{\theta}_O$ is closed to the rejection boundary, little can be said about θ_O and hence $\theta_O - \theta_R$. Furthermore, the replication sample sizes were determined based on the original effect size to achieve at least 80% in power. Selection bias inflated the original effect size, leading to lower test power and statistically insignificant replications (Etz and Vandekerckhove, 2016; Camerer et al., 2018). The lack of information about $\theta_O - \theta_R$ is evident in generally wider confidence intervals after adjustment in Figure 8.

3.3 Effect decline

Finally, we considered the proportion of effect sizes that declined. Using the selective z-test, we tested the hypothesis H^D, conditioning on the event where the z-scores are observed and the variable S. The resulting p-values are given in Figure 9. Our underestimate and overestimate are 35% (= 16/46) and 100% respectively, with a 90% confidence interval of (11%, 100%).

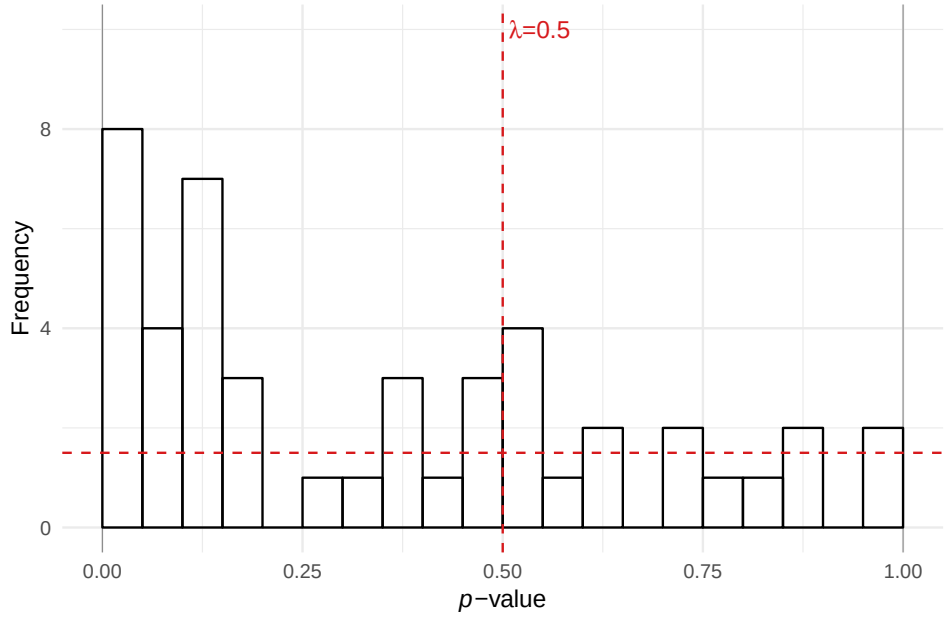


Figure 9: Histogram of the p-values for the null hypothesis $\theta_R \geq \theta_O$. p-values to the left gives more evidence for $\theta_R < \theta_O$ whereas p-values to the right gives more evidence for $\theta_R \geq \theta_O$. The estimate of the expected number of null p-values within each bin is given by the horizontal red line.

More generally, we used the hypothesis $H^{D,\rho}$ to estimate the proportion of effect sizes that declined by at least a fraction of ρ. The underestimate, overestimate and the 90% confidence interval are given in Figure 10. For example, we estimate that 16 of the 46 effect sizes (35% (with a 95% lower confidence bound 11%) decreased by at least 20%, even after adjusting for selection on measurement noise. Note that this does not exclude explanations by other forms of selection, e.g. selecting a large effect when there is a random effect.

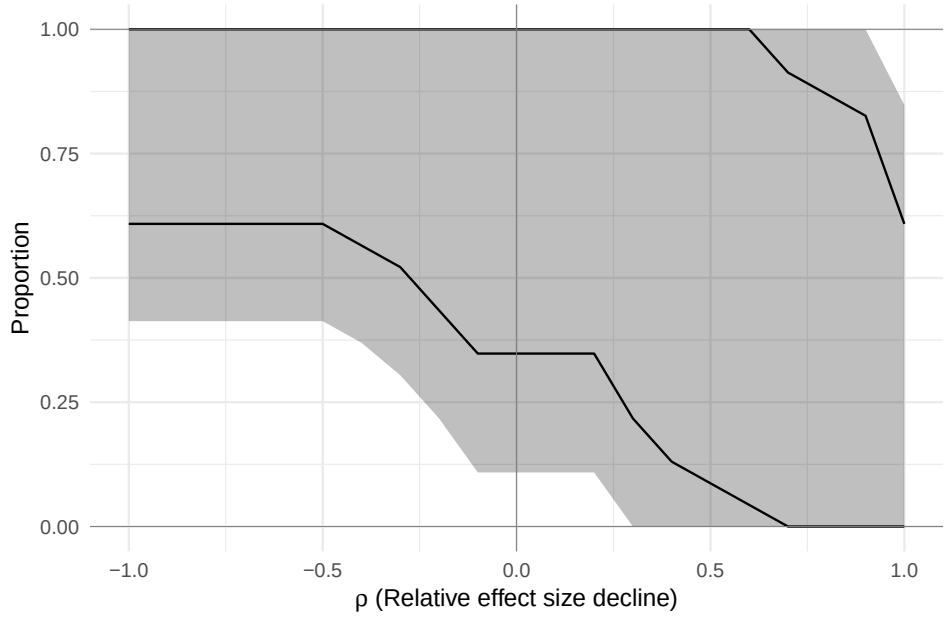


Figure 10: The underestimate, overestimate and the 90% confidence interval. The lower black line is the underestimate, the high black line is the overestimate and the gray band is the 90% confidence interval.

4 Discussion

4.1 Importance of adjusting for selection bias

As we have seen, selection bias plays a powerful and pervasive role in shaping the data we observe in large-scale replication studies (and, by extension, the data we observe in published studies that have not yet been replicated!). It leads to many predictable pathologies and should be viewed as a proverbial "elephant in the room" whenever we discuss descriptive statistics computed from such studies. In particular, we should avoid leaping to any conclusions about how many false claims there were in the original studies, whether effect sizes declined or by how much, or which replication studies suffered from infidelities, until we have carefully ruled out the possibility that publication bias alone is to blame for whatever descriptive statistic we have computed.

Fortunately, the truncated Gaussian model, properly combined with modern multiple testing and post-selection inference methods, opens many avenues for analyses that directly answer questions about true effect sizes with appropriate uncertainty quantification. We have explored several such avenues here (see also Andrews and Kasy, 2018) but many others are possible.

4.2 Importance of statistical formality

In addition, we hope this article serves to advocate for the benefits of careful formal statistical modeling in analyzing replication studies, in place of (or in addition to) descriptive statistics. In particular, using vaguely specified models or eschewing models altogether can lead to analyses from which it is difficult to draw firm conclusions. For example, in Open Science Collaboration

(2015), McNemar's test was applied to a 2×2 contingency table of whether the original and replication studies are equally likely to be statistically significant. The very small p-value reported for this test establishes nothing more than that the original studies were selected to be statistically significant, a fact which is likely already known by most in the field. In fact, the test does not quite establish even that, because it is unclear whether this hypothesis would be true even without the effect of selection bias: The proportion of statistically significant p-values is a measure of the average power, which depends on the sample sizes, and the sample sizes often differed substantially between the original and replication studies.

Another example is RP:P's use of sample correlation coefficients between independent and dependent variables as a standardized measure of effect size for comparison between the original and replication studies. This comparison implicitly assumes that the distribution of the independent variable is the same in the original and replication studies, an assumption that was violated by many of the replications. In an extreme case, an ANOVA in Purdie-Vaughns et al. (2008) with race as one of the factors used 40 African Americans and 37 Whites, but was replicated with 120 African Americans and 1370 Whites. With such a dramatic change in the distribution of an independent variable, there is no reason why the correlation coefficients should remain the same, as illustrated in the following example.

Example 2. A study with a two-sample t-test for some treatment condition is replicated. Suppose the treatment and control group are drawn from $N(1, 1)$ and $N(0, 1)$, respectively. If the ratio of the two group sizes changes from one study to another, the correlation coefficients may differ as well, even without any infidelities or hidden moderators. Borrowing the numbers from Purdie-Vaughns et al. (2008) for instance, if the original study contains 40 treatment and 37 control units, the true correlation coefficient is 0.45, whereas in a replication with 120 control and 1370 treatment units the true coefficient is 0.26 instead.

Replication projects similar to RP:P have since materialized, but few stated an explicit statistical hypothesis. For example, in economics, Camerer et al. (2016) used the same flawed metric of proportion of statistically significant results in the original direction. A statistical analysis with explicitly stated models and hypotheses will give us more meaningful estimates, particularly valuable given how costly these large scale replication efforts are.

4.3 Interpretation of effect shifts

While we have proposed several methods for quantifying discrepancies between the effect sizes in the original and replication studies, the data alone cannot tell us why they might differ. Several potential explanations include:

1. design failures, systematic biases or calculation errors in either the original or the replication study;
2. major differences in experimental conditions between the original and replication studies, which most researchers would recognize *a priori* as likely to affect the results; which Gilbert et al. (2016a) call *infidelities*; and
3. minor differences in experimental conditions between the studies — such as lighting, weather, or the passage of time — which cannot all be controlled but whose effects may nevertheless alter the true effect size in unforeseeable ways, often referred to as *hidden moderators* (e.g. Srivastava, 2015).

While there may be no sharp distinction in principle between infidelities and hidden moderators, there is a scientifically crucial difference between moderating factors that can be anticipated by experimenters and those that cannot. If we can anticipate in advance when replications are likely to fail by carefully evaluating their designs, we might hope to solve the problem simply by being more careful in setting up experiments. By contrast, if hidden moderators confound most attempts to replicate most psychological studies, it would raise profound questions about the entire enterprise of experimental psychology. In the extreme case, if even trivial changes to those conditions have large and unpredictable effects on most phenomena of interest, we might begin to despair of gaining generalizable knowledge about psychology through laboratory experimentation.

Our analyses point to several conclusions regarding effect shifts: First, that there are a few studies where we can be confident the effect in the replication study was significantly different than in the original study; second, that in aggregate, when effects do shift, they tend to decline (shift toward zero) in replications rather than increase; and third, that there is insufficient evidence to conclude that the vast majority of experimental effects simply evaporated upon replication. In particular, 83% should not be treated as a reasonable estimator of the fraction of *true* effect sizes that declined; rather, it likely reflects that the estimates in the original studies overestimated their corresponding true effects due to selection bias.

One possible explanation for systematically declining effect involves a subtler form of selection bias, where every experiment's effect size is random, buffeted by hidden moderators, and those experiments whose moderators primarily magnify the effect size are more likely to be published. That is, in the same way that experimenters select studies whose sampling error is large, they also selects for studies whose true effect size is larger than usual. Further systematic replication studies may help to shed light on which factors are most often the culprits in moderating true effect sizes, possibly improving the reliability of experiments and leading to new scientific insights (Barrett, 2015; Klein et al., 2018).

4.4 Future work

As large-scale replicability studies are becoming more common in assessing the "well-being" of a scientific domain, this paper serves as a stepping stone for improving methodologies in future replicability studies.

First, selection for significance is an inevitable consequence of the current scientific process. Our adjustments for selection is admittedly crude, but necessitated by the limitations in the given data. With more available information, a better model for selection can be used. For example, with the advancement of preregistration, we can use the external comparison method to produce less conservative estimates of the directional FDP at level $\alpha = 0.05$ if we have more information about statistically nonsignificant studies. With more replications carried out, we can estimate the publication bias model in Andrews and Kasy (2018) more precisely; together with higher powered design in replications (e.g. Camerer et al., 2018), we can enhance the precision of our estimators and power of our tests.

Second, we emphasized the importance of statistical formality. Our proposed criteria are based on clearly defined parameters. While these criteria may not suit all needs in future replicability studies, additional formal hypotheses can also be analyzed under the post-selection inference framework similarly.

With our proposed criteria and procedures, researchers can perform more informative infer-

ences than the current practice, and provide a clearer picture of the replicability crisis.

Reproducibility

A git repository containing with the code generating the images in this article is available at `https://github.com/kenhungkk/assessing-replicability.git`.

Supplement

The supplement is available in the git repository, or directly on `https://github.com/kenhungkk/assessing-replicability/raw/public/supplement.pdf`.

Acknowledgment

We thank Marcel A L M van Assen, Yoav Benjamini, Dean Eckles, Philip B Stark, Jacob Steinhardt, Jonathan Taylor, Alexa Tulett, Stefan Wager, Daniel Yekutieli, and Bin Yu for helpful comments and discussions.

References

The scientific method. *The Economist*, February 2016.

Joel Achenbach. Many scientific studies can't be replicated. That's a problem. *The Washington Post*, August 2015.

Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ*, 5(2):e3544, 2017.

Christopher J Anderson, Štěpán Bahník, Michael Barnett-Cowan, Frank A Bosco, Jesse Chandler, Christopher R Chartier, Felix Cheung, Cody D Christopherson, Andreas Cordes, Edward J Cremata, Nicholas Della Penna, Vivien Estel, Anna Fedor, Stanka A Fitneva, Michael C Frank, James A Grange, Joshua K Hartshorne, Fred Hasselman, Felix Henninger, Marije van der Hulst, Kai J Jonas, Calvin K Lai, Carmel A Levitan, Jeremy K Miller, Katherine S Moore, Johannes M Meixner, Marcus R Munafò, Koen I Neijenhuijs, Gustav Nilsonne, Brian A Nosek, Franziska Plessow, Jason M Prenoveau, Ashley A Ricker, Kathleen Schmidt, Jeffrey R Spies, Stefan Stieger, Nina Strohminger, Gavin B Sullivan, Robbie C M van Aert, Marcel A L M van Assen, Wolf Vanpaemel, Michelangelo Vianello, Martin Voracek, and Kellylynn Zuni. Response to Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037c, March 2016.

Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. *GitHub*, pages 1–85, May 2018.

Monya Baker. Over half of psychology studies fail reproducibility test, August 2015. URL `http://www.nature.com/doifinder/10.1038/nature.2015.17433`.

Lisa Feldman Barrett. Psychology Is Not in Crisis. *The New York Times*, page A23, September 2015.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, Eric-Jan Wagenmakers, Richard A Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin F Camerer, David Cesarini, Christopher D Chambers, Merlise Clyde, Thomas D Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P Field, Malcolm Forster, Edward I George, Richard Gonzalez, Steven N Goodman, Edwin Green, Donald P Green, Anthony G Greenwald, Jarrod D Hadfield, Larry V Hedges, Leonhard Held, Teck-Hua Ho, Herbert Hoijtink, Daniel J Hruschka, Kosuke Imai, Guido W Imbens, John P A Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E Maxwell, Michael McCarthy, Don A Moore, Stephen L Morgan, Marcus R Munafò, Shinichi Nakagawa, Brendan Nyhan, Timothy H Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J Watts, Christopher Winship, Robert L Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2:6–10, January 2018.

Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64 (4):1215–1222, December 2008.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 57(1):289–300, 1995.

Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000.

Yoav Benjamini and Daniel Yekutieli. False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters. *Journal of the American Statistical Association*, 100(469): 71–81, March 2005.

Colin F Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351 (6280):1433–1436, March 2016.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, 343:229–268, August 2018.

Benedict Carey. Many psychology findings not as strong as claimed, study says. *The New York Times*, page A1, August 2015.

Chad S Dodson, James Darragh, and Allison Williams. Stereotypes and retrieval-provoked illusory source recollections. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):460–477, 2008.

Sue Duval and Richard Tweedie. Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*, 56(2):455–463, June 2000.

Alexander Etz and Joachim Vandekerckhove. A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE*, 11(2):e0149794–12, February 2016.

Coreen Farris, Teresa A Treat, Richard J Viken, and Richard M McFall. Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. *Psychological Science*, 19(4):348–354, April 2008.

Ronald Aylmer Fisher. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

Ronald Aylmer Fisher. The distribution of the partial correlation coefficient. *Metron*, 3:329–332, 1924.

William Fithian, Dennis L Sun, and Jonathan E Taylor. Optimal Inference After Model Selection. *arXiv*, October 2014.

Andrew Gelman and John Carlin. Beyond Power Calculations. *Perspectives on Psychological Science*, 9(6):641–651, November 2014.

Andrew Gelman and Keith O'Rourke. Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics*, 15(1):18–23, December 2013.

Andrew Gelman and Francis Tuerlinckx. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.

Daniel T Gilbert, Gary King, Stephen Pettigrew, and Timothy D Wilson. Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037a, 2016a.

Daniel T Gilbert, Gary King, Stephen Pettigrew, and Timothy D Wilson. More on "Estimating the Reproducibility of Psychological Science", March 2016b. URL `https://projects.iq.harvard.edu/files/psychology-replications/files/gkpw_post_publication_response.pdf`.

Daniel T Gilbert, Gary King, Stephen Pettigrew, and Timothy D Wilson. A Response to the Reply to Our Technical Comment on "Estimating the Reproducibility of Psychological Science" , March 2016c. URL `https://projects.iq.harvard.edu/files/psychology-replications/files/gkpw_response_to_osc_rebutal.pdf`.

Steven N Goodman. Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):23–27, December 2013.

Steven N Goodman, Daniele Fanelli, and John P A Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12, June 2016.

Larry V Hedges. Modeling publication selection effects in meta-analysis. *Statistical Science*, 7 (2):246–255, 1992.

Ruth Heller, Yulia Golland, Rafael Malach, and Yoav Benjamini. Conjunction group analysis: an alternative to mixed/random effect analysis. *Neuroimage*, 37(4):1178–1185, 2007.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

John P A Ioannidis. Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics*, 15(1):28–36, December 2013.

Leah R Jager and Jeffrey T Leek. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):1–12, December 2013.

Valen E Johnson, Richard D Payne, Tianying Wang, Alex Asher, and Soutrik Mandal. On the Reproducibility of Psychological Science. *Journal of the American Statistical Association*, 112(517):1–10, March 2017.

Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams, Sinan Alper, Mark Aveyard, Jordan R Axt, Štěpán Bahník, Rishtee Batra, Mihály Berkics, Michael J Bernstein, Daniel Berry, Olga Bialobrzeska, Evans Binan, Konrad Bocian, Mark J Brandt, Robert Busching, Anna Cabak Rédei, Huajian Cai, Fanny Cambier, Katarzyna Cantarero, Cheryl L Carmichael, Francisco Ceric, David C Cicero, Jesse Chandler, Armand Chatard, Eva E Chen, Jen-Ho Chang, Winnee Cheong, Sharon Coen, Jennifer A Coleman, Brian Collisson, Morgan A Conway, Katherine S Corker, Paul G Curran, Fiery Cushman, Zubairu K Dagona, Ilker Dalgar, Anna Dalla Rosa, William E David, Maaike de Brujin, Leander De Schutter, Thierry Devos, Canay Doğulu, Nerisa Dozo, Kristin Nicole Dukes, Yarrow Dunham, Kevin Durrheim, Charles R Ebersole, John E Edlund, Alexander Scott English, Anja Eller, Carolyn Finck, Natalia Frankowska, Miguel-Ángel Freyre, Mike Friedman, Elisa Maria Galliani, Joshua C Gandi, Tanuka Ghoshal, Steffen R Giessner, Tripat Gill, Timo Gnambs, Ángel Gómez, Roberto González, Jesse Graham, Jon E Grahe, Ivan Grahek, Eva G T Green, Kakul Hai, Matthew Haigh, Elizabeth L Haines, Michael P Hall, Marie E Heffernan, Joshua A Hicks, Petr Houdek, Jeffrey R Huntsinger, Ho Phi Huynh, Hans IJzerman, Yoel Inbar, Åse H Innes-Ker, William Jiménez-Leal, Melissa-Sue John, Jennifer A Joy-Gaba, Anna Kende, Roza G Kamiloğlu, Heather Barry Kappes, Serdar Karabati, Haruna Karick, Victor N Keller, Nicolas Kervyn, Goran Knežević, Carrie Kovacs, Lacy E Krueger, German Kurapov, Jamie Kurtz, Daniël Lakens, Ljiljana B Lazarević, Carmel A Levitan, Jr Neil A Lewis, Samuel Lins, Nikolette P Lipsey, Joy Losee, Esther Maassen, Angela T Maitner, Winfrida Malingumu, Robyn K Mallett, Saita A Marotta, Janko Međedović, Fernando Mena Pacheco, Taciano L Milfont, Wendy L Morris, Sean Murphy, Andriy Myachykov, Nick Neave, Koen Neijenhuijs, Anthony J Nelson, Félix Neto, Austin Lee Nichols, Aaron Ocampo, Susan L ODonnell, Elsie Ong, Malgorzata Osowiecka, Gábor Orosz, Grant Packard, Rolando Pérez-Sánchez, Boban Petrović, Ronaldo Pilati, Brad Pinter, Lysandra Podesta, Gabrielle Pogge, Monique M H Pollmann, Abraham M Rutchick, Alexander Saeri, Patricio Saavedra, Erika Salomon, Kathleen Schmidt, Felix D Schönbrodt, Maciej B Sekerdej, David Sirlopú, Jeannie L M Skorinko, Michael A Smith, Vanessa Smith-Castro, Karin Smolders, Agata Sobkow, Walter Sowden, Manini Srivastava, Oskar K Sundfelt, Philipp Spachtholz, Troy G Steiner,

Jeroen Stouten, Chris N H Street, Stephanie Szeto, Ewa Szumowska, Andrew Tang, Norbert Tanzer, Morgan Tear, Manuela Thomae, Jakub Traczyk, David Torres, Jordan Theriault, Joshua M Tybur, Adrienn Ujhelyi, Robbie C M van Aert, Marcel A L M van Assen, Paul A M van Lange, Marije van der Hulst, Anna Elisabeth van 't Veer, Alejandro Vásquez Echeverría, Leigh Ann Vaughn, Alexandra Vásquez, Luis Diego Vega, Catherine Verniers, Mark Verschoor, Ingrid Voermans, Marek A Vranka, Marieke de Vries, Cheryl Welch, Aaron Wichman, Lisa A Williams, Michael Wood, Julie A Woodzicka, Marta K Wronska, Liane Young, John M Zelenski, Zhijia Zeng, and Brian A Nosek. Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. oct 2018. URL `https://psyarxiv.com/9654g/`.

Jeff T Larsen and Amie R McKibban. Is Happiness Having What You Want, Wanting What You Have, or Both? *Psychological Science*, 19(4):371–377, 2008.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Richard D Morey and Daniël Lakens. Why most of psychology is statistically unfalsifiable. 2017. URL `https://github.com/richarddmorey/psychology_resolution`.

Brian A Nosek and Timothy M Errington. Reproducibility in Cancer Biology: Making sense of replications. *eLife*, 6:e23383, January 2017.

Brian A Nosek and Elizabeth Gilbert. Let's not mischaracterize replication studies: authors, March 2016. URL `https://retractionwatch.com/2016/03/07/lets-not-mischaracterize-replication-studies-authors/`.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):943, 2015.

Valerie Purdie-Vaughns, Claude M Steele, Paul G Davies, Ruth Ditlmann, and Jennifer Randall Crosby. Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, 94(4): 615–630, 2008.

Allan R Sampson and Michael W Sill. Drop-the-losers design: Normal case. *Biometrical Journal*, 47(3):257–268, 2005.

Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*, 9 (6):666–681, November 2014a.

Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. P-Curve: a Key to the File-Drawer. *Journal of Experimental Psychology: General*, 143(2):534–547, 2014b.

Sanjay Srivastava. Moderator interpretations of the reproducibility project, September 2015. URL `https://thehardestscience.com/2015/09/02/moderator-interpretations-of-the-reproducibility-project/`.

Sanjay Srivastava. Evaluating a new critique of the reproducibility project, March 2016. URL `https://thehardestscience.com/2016/03/03/evaluating-a-new-critique-of-the-reproducibility-project/`.

John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 64(3):479–498, July 2002.

Wolfgang Stroebe. Are most published social psychological findings false? *Journal of Experimental Social Psychology*, 66(C):134–144, September 2016.

Jeffrey C Valentine, Anthony Biglan, Robert F Boruch, Felipe González Castro, Linda M Collins, Brian R Flay, Sheppard Kellam, Eve K Mościcki, and Steven P Schinke. Replication in Prevention Science. *Prevention Science*, 12(2):103–117, May 2011.

Robbie C M van Aert and Marcel A L M van Assen. Bayesian evaluation of effect size after replicating an original study. *PLoS ONE*, 12(4):e0175302–23, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0175302. URL `http://dx.plos.org/10.1371/journal.pone.0175302`.

Robbie C M van Aert and Marcel A L M van Assen. Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods*, 50(4):1515–1539, 2018. ISSN 15543528. doi: 10.3758/s13428-017-0967-6.

Eric van Dijk, Gerben A van Kleef, Wolfgang Steinel, and Ilja van Beest. A social functional approach to emotions in bargaining: When communicating anger pays and when it backfires. *Journal of Personality and Social Psychology*, 94(4):600–614, 2008.

Asaf Weinstein, William Fithian, and Yoav Benjamini. Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, 108 (501):165–176, 2013.

Daniel Yekutieli. Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):515–541, 2012.

Sebastian Zöllner and Jonathan K Pritchard. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4): 605–615, 2007.