Using Unbiased Many-Teams Replication Data to Evaluate Meta-Analytic Estimators

Amanda Kvarven

January 2, 2022

Abstract

Meta-analysis is often considered to be at the top of the scientific hierarchy of evidence. However, there is a growing worry among researchers that meta-analysis suffers from publication bias. To combat this problem, researchers have proposed multiple methods of adjusting meta-analyses. In this paper, I test five of these adjustment methods by applying them to unbiased meta-analytic data to see if the bias adjustment methods can detect the true level of bias and thereby yield the same result as the unadjusted analysis. This data comes from large pre-registered replication projects where experiments are redone in Many Labs, thus creating the meta-analytic data structure. I find that all regression-based adjustment methods correct the results of unbiased replication and might therefore yield misguiding results.

1 Introduction

Meta-analysis is often placed at the top of the hierarchy of scientific evidence (Ioannidis, 2016; Siddaway et al., 2019). Typically, it is difficult for individual studies to obtain a sufficient amount of statistical power, making meta-analysis very useful. Therefore, meta-analysis is used increasingly in several fields of social sciences such as psychology (Batz-Barbarich et al., 2018) and economics (Eckel and Füllbrunn, 2015; Havránek and Kokes, 2015; Smith and Huang, 1995). A meta-analysis can combine several already-existing studies, hence achieving high statistical power at a low cost. This property also makes it possible to include population heterogeneity in the analysis, possibly making meta-analysis superior to individual studies. This is because adding heterogeneity may help generalizability across populations. Furthermore, meta-analysis can even out mistakes and errors across studies, as long as these errors are not systematic.

There is, however, a growing concern that meta-analysis might suffer from publication bias (Gurevitch et al., 2018; Sterling et al., 1995). In the last few years, there have been several attempts at trying to find and quantify the level of publication bias in the literature (Camerer et al., 2018; Klein et al., 2014, 2018; Ebersole et al., 2016). Since meta-analyses often are based on published studies, such a bias might taint the meta-analysis. To combat this issue, several methods have been developed to correct for bias in meta-analyses (Andrews and Kasy, 2019; Duval and Tweedie, 2000; Stanley and Doucouliagos, 2014; Vevea and Hedges, 1995).

Kvarven et al. (2020) makes an attempt to quantify the level of bias in meta-analyses by comparing pre-registered replications with meta-analyses and tries to test the function of several methods for correcting publication bias in meta-analyses. The replications used come from different Many Labs projects. These projects try to replicate findings by establishing a pre-analysis plan and having different labs across the world execute the experiment. As all details of the analysis are pre-registered and the project has been guaranteed publication before they start data collection, these data should be unbiased. Kvarven et al. (2020) finds that, on average, meta-analytic effect sizes are almost three times as large as effect sizes from pre-registered replication on the same topic. This study also shows that applying adjustment methods for publication bias has little or no effect on the difference between the estimates.

In a recent paper, Lewis et al. (2020) raises the following question: Should one expect correction methods to eliminate the difference between meta-analysis and replication? For publication bias correction methods to eliminate the discrepancy, one must assume that the entire difference between meta-analysis and replication is due to publication bias. However, there are other explanations than publication bias for the difference, such as heterogeneity. The replications in Kvarven et al. (2020) include some level of population heterogeneity, as the experiment is performed in different populations. However, these replications do not include other types of heterogeneity such as design heterogeneity, since all the different experiments included in the replication follow the exact same pre-analysis plan. Hence, the difference between meta-analysis and replications might be caused by several things, and publication bias correction methods should only be expected to remove the part of the difference caused by publication bias.

This paper offers a cleaner way to investigate publication bias adjustment methods, as I only use data from pre-registered replications, not meta-analyses. I investigate the possibility that some correction methods might correct the results even when no bias is present in the meta-analysis. This means that, while Kvarven et al. (2020) examines how the adjustment methods adjust in the presence of bias, I check whether the methods correct the result, even if there is no bias present. Thus, my paper looks at the cost of using adjustment methods when they are not needed. Hence, if a method passes the tests in this paper, it will not harm the meta-analysis if the bias correction method is applied. This paper cannot, however, determine whether the method provides bias correction when needed. Therefore, methods that perform well in this paper will only have the potential of improving the meta-analysis, without any potentially harmful side effects. I accomplish this test by applying correction methods to unbiased meta-analytic replication data and then comparing the estimates to the unadjusted results calculated from the same dataset.

My paper differs from that of Kvarven et al. (2020) in its use of meta-analysis and replication pairs, as I only use data from pre-registered Many Labs replication reports. Therefore, this paper does not need to assume that the difference between meta-analysis and replication is due to publication bias. In addition, there is also no need to assume that meta-analyses and replications are comparable, only that the replications are unbiased. Furthermore, as I use only replication data and do not require matching of meta-analyses, I can draw data from all Registered Replication Reports (Alogna et al., 2014; Eerland et al., 2016; Hagger et al., 2016; Cheung et al., 2016; Wagenmakers et al., 2016; Bouwmeester et al., 2017; O'Donnell et al., 2018; McCarthy et al., 2018; Verschuere et al., 2018), Many Labs 1 (Klein et al., 2014), and Many Labs 3 (Ebersole et al., 2016) replications, so my study uses a larger and possibly more varied sample than Kvarven et al. (2020) does, in combination with providing a cleaner test. I therefore end up with a sample size of 24 meta-analytic replication datasets, compared to the 15 study pairs included in Kvarven et al. (2020).

2 Method

For this analysis, I will use data from Many Labs Replication Projects in psychology. These projects use a Many Labs format, meaning that multiple labs across the world have followed the same pre-analysis plan and conducted the same experiment. Hence, each experiment within a given replication is comparable, but the samples and sample populations will be different, allowing population heterogeneity. Therefore, I can use the individual estimates from each lab to create a meta-analytic dataset based on the replication data, with one effect size and one standard error from each lab.

A Many Labs replication is conducted by different research teams volunteering to replicate a certain finding, based on a set pre-analysis plan. This pre-analysis plan has been made in cooperation with the authors of the original study being replicated, to ensure that the design of the replication is a fair test of the hypothesis. When the pre-analysis plan is ready, it is sent to a journal. The goal of this process is for the journal to promise publication for the replication project when it is completed. This should then ensure that the replication is published regardless of the findings, thus eliminating the possibility for publication bias among such replication reports. When the pre-analysis plan is ready and the project has been promised to be published, the different labs will follow the pre-analysis plan and perform the replication separately. In the replications used in this paper, all the different experiments follow an identical pre-analysis plan and design, and therefore there is no design heterogeneity. When all the data is collected, an overall analysis of all the data is performed. Then the results from this analysis are presented in a paper and published in the aforementioned journal. These studies are therefore pre-registered and guaranteed to be published regardless of outcome. As the pre-registration of these studies is thorough, there should be no opportunity to p-hack. Hence, they should not be affected by publication bias or bias arising from p-hacking.

These data are collected by different research teams, thus creating a structure similar to a meta-analysis. Therefore, my data are based on the effect sizes of each individual lab and combined into a meta-analysis. Hence, I use the terms ”replications” and ”meta-analyses” interchangeably, as the data can be best described as replication estimated through meta-analysis.

I apply five different bias correction methods to the replication data and then compare the corrected estimate with the uncorrected estimate. This process is illustrated in Figure 1. With this method, I can investigate for overcorrection for publication bias of the meta-analytic estimate. The unadjusted effect size is calculated with the meta-analytic random effects method.

The replications I use in this study should already be unbiased, making bias correction redundant. Hence, an adjustment method that responds only to publication bias should not correct the estimates at all. No difference is expected between the uncorrected replication estimate and the corrected estimate. If the corrected estimate differs from the uncorrected estimate in either direction, this would indicate overcorrection.

In addition to estimating the level of overcorrection, I will measure how statistical power is affected by different adjustment methods. The results regarding statistical power will be directly comparable to those from (Carter et al., 2019) and Kvarven et al. (2020).

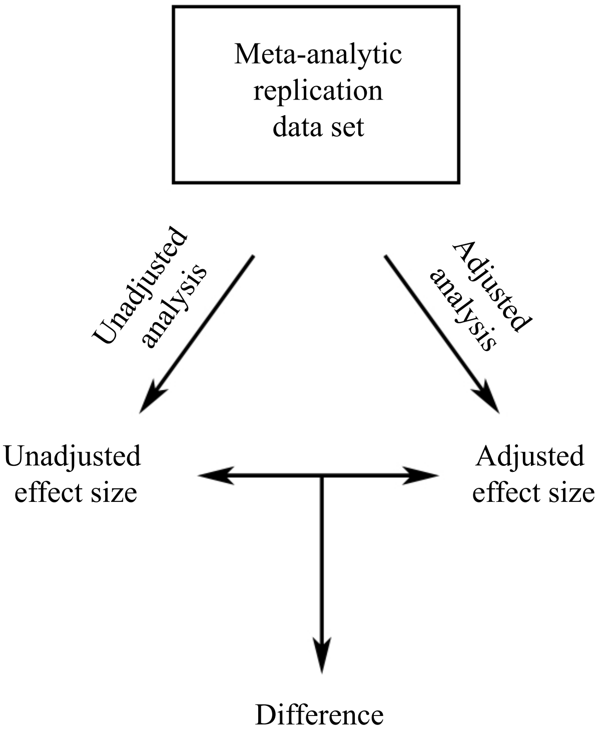

Figure 1: Illustration of method.

Table 1 compares my paper to those of Carter et al. (2019) and Kvarven et al. (2020). Carter et al. (2019) evaluated bias-adjustment methods by simulating different conditions believed to be present in the literature that could create problems for meta-analyses. Kvarven et al. (2020) investigates issues regarding bias in practice by comparing meta-analyses to registered Many Labs replications. The main difference between this paper and that of Carter et al. (2019) is that Carter et al. (2019) uses simulation studies and tests the methods in various conditions, such as degrees of publication bias, questionable research practices, and heterogeneity. The most important difference between my paper and Kvarven et al. (2020) is that I investigate performance when no publication bias is present. I also use a larger sample of studies, test more adjustment methods, and do not need a matching meta-analysis to my replication estimates.

Table 1: Overview of previous research approches

Study	Evaluation method	Publication bias condition	Adjustment methods tested
Carter et al. (2019)	Simulations	No, medium and high levels of bias	Trim & Fill, WAAP, p-curve, p-uniform, PET-PEESE and 3PSM
Kvarven et al. (2020)	Empirical	Unknown level of bias	Trim & Fill, PET-PEESE and 3PSM
Current study	Empirical	No bias	Trim & Fill, PET, PEESE, PET-PEESE and 3PSM

Note: PET=the precision-effect test, PEESE=the precision-effect estimate with standard error, 3PSM=the three-parameter selection model.

2.1 Estimation of meta-analysis bias-adjustment methods

Following Carter et al. (2019) and Kvarven et al. (2020), I limit my analysis to five bias-correction methods: Trim & Fill, the precision-effect test (PET), the precision-effect estimate with standard error (PEESE), PET-PEESE (a combination of PET and PEESE), and the three-parameter selection model (3PSM).

2.1.1 Trim & Fill

Duval and Tweedie (2000) introduced Trim & Fill, a bias-correction method that uses an algorithm to correct the estimate. This method is one of the most common correction models for publication bias (Simonsohn et al., 2014). First, the algorithm removes studies that create funnel plot asymmetry. When the new center is calculated, the algorithm fills in holes in the plot based on the trimmed studies. Filling is accomplished by adding one study for every study that was trimmed. This should increase the symmetry of the funnel plot. The method then runs a meta-analysis containing all the studies, including trimmed and imputed effect sizes.

This is then the corrected effect size.

Carter et al. (2019) finds that Trim & Fill does not adjust sufficiently for publication bias. Kvarven et al. (2020) also finds that Trim & Fill corrects for bias to a small degree, returning estimates that are close to those of the unadjusted analyses.

2.1.2 Regression-based adjustment methods

I also test out tree regression-based adjustment methods, the precision-effect test, the precision-effect estimate with standard error, and a conditional estimator that combines the two. These methods are being used frequently in several fields, such as psychology (Batz-Barbarich et al., 2018) and economics (Havránek and Kokes, 2015; Havránek and Irsova, 2011; Havránek, 2015).

PET

The first regression-based method is the precision-effect test (PET), a regression-based correction method developed by Stanley (2008). PET is estimated through a weighted-least-squares regression, weighted by the inverse of the variance. The corrected effect size is represented by the constant term.

PET is specified by the following regression equation:

$$Y_i = \gamma_0 + \alpha SE_i + \varepsilon_i$$

,

where γ_0 is the constant term, α is the regression coefficient, and SE_i is the standard error.

PEESE

The precision-effect estimate with standard error (PEESE) was created by Stanley and Doucouliagos (Stanley and Doucouliagos, 2007). PEESE is described by the following equation, the only difference from PET being the use of the squared standard error as the independent variable:

$$Y_i = \gamma_0 + \alpha SE_i^2 + \varepsilon_i$$

,

where γ_0 is the constant term, α is the regression coefficient, and SE_i is the standard error.

PET-PEESE

PET-PEESE is a conditional estimator that combines the estimates from PET and PEESE (Stanley and Doucouliagos, 2014). When the PET estimate is statistically insignificant, it is chosen as the PET-PEESE estimate. This is because PET has been found to outperform PEESE when there is no underlying effect. When there is a non-zero underlying effect, however, PEESE is found to outperform PET, so when the PET estimate is statistically significant, the PEESE estimate becomes the PET-PEESE estimate.

Carter et al. (2019) concludes that, regarding the correction of publication bias, PET-PEESE performs better than Trim & Fill.The results from Kvarven et al. (2020), however, indicate that PET-PEESE seems to change the effect size of the meta-analysis, strongly decreasing statistical power, but does not decrease the mean squared error. The results from Kvarven et al. (2020) regarding PET-PEESE indicate that this method eliminates systematic bias in the meta-analysis but that the high mean squared error indicates that the PET-PEESE corrected estimate differs from the replication to the same degree, on average, as estimates made using the other methods examined.

2.1.3 Three-parameter selection model

The three-parameter selection model (3PSM) was designed by Vevea and Hedges (1995). This method uses maximum likelihood to estimate the probability that an insignificant result will enter the literature, as well as the underlying effect and the heterogeneity based on the meta-analytic dataset. These parameters are then used to estimate an adjusted effect size and standard error. As this method is based on one-sided p-values, I use a cut-off of 0.025, as this is the equivalent of a two-sided p-value of 0.05.

Carter et al. (2019) finds that, although 3PSM cannot return an estimate every time, it outperforms PET-PEESE almost every time it does return an estimate. However, the results of Carter et al. (2019) will depend on the level and types of problems in the literature. Kvarven et al. (2020) observes that 3PSM performs slightly better than Trim & Fill regarding bias correction, but that substantial levels of bias remain after its use. Moreover, 3PSM comes with a loss of statistical power.

2.2 Data collection

The starting point of this paper is to collect data from pre-registered Many Labs replications in psychology. Two data sources were identified: 1) replications of the Registered Replication Report format published in the journals *Advances in Methods and Practices in Psychological Science* and *Perspectives on Psychological Science*, and 2) the Many Labs projects.

The data from Many Labs 2 (Klein et al., 2018) are not included, as errors have been discovered in this dataset, and some work is left to be done to correct the data. As Many Labs 2 is the largest of the replication projects, it will be important to add Many Labs 2 to the dataset when the data have been controlled and approved.

Furthermore, I also exclude replications reporting in Pearson's r or eta-squared. This excludes three studies. This is because these effects would have to be converted to Cohen's d, and this process could introduce inaccuracy into the dataset. Such inaccuracy could in turn affect the bias correction method's performance, as inaccuracy can mask patterns in the data that the methods use to identify bias[1]. This yields a total of 24 meta-analytic replication datasets, when excluding effects that would need converting.

Most of the replications included in this paper were originally reported in Cohen's d. However, seven of the 24 effects were reported in unstandardized beta coefficients. Cohen's d for these replications is constructed using the raw data available from the studies in question.

For the data from Many Labs 1 and Many Labs 3, a positive effect in the meta-analysis is defined as finding an effect in the same direction as the original study that was replicated. Therefore, I change the sign of the effect if the original study found a negative effect for the studies from the registered replication reports (RRR) as well, making the data comparable.

[1] When effect size and standard error must be converted, the converted estimates can become slightly inaccurate. This inaccuracy can then be amplified when the converted estimates are used to calculate z-values. This is not expected to systematically shift z-values, but to randomly create a small increase or decrease of the z-value. One important category of correction methods, selection models, typically uses a fraction of significant tests when identifying bias. Inaccurate z-values might therefore tamper with the adjustment method's ability to detect bias, if the inaccuracy causes some z-values to cross the significance threshold. Other methods, such as Trim & Fill, use patterns in the data to identify bias. If the data are inaccurate, these patterns might change, and the method will not be able to correct bias as intended. Hence, I exclude effects that need converting, as the sample size is sufficient without them.

In some cases, multiple effect sizes were replicated from the same original paper. In these cases, I use the most accurate estimate with the smallest standard error, as long as the replication does not state explicitly that one of the estimates is the main estimate.

One such case is the replication of Schooler and Engstler-Schooler (1990). An error in the replication protocol resulted in the need for new data collection (Alogna et al., 2014). Both instances of data collection are considered to be valid tests of the hypotheses presented in the original paper and are therefore valid replications. The two data collections present three options for this replication: 1) use the data collected in the first attempt, 2) use the data collected in the second attempt, or 3) combine the two datasets. Option three would have the highest level of statistical power, but combining the two attempts would introduce design heterogeneity into the meta-analysis. As there is no design heterogeneity in any of the other replications, I will use the dataset from one of the two first options. In the main analysis, I use the data collected in the first attempt, as it is the most accurate of the two attempts[2].

2.3 Indicators used to compare the performance of the bias-adjustment methods

To evaluate the performance of the adjustment methods, I measure the accuracy of the estimate, the level of statistical power, and the false-positive and false-negative rates. The methods of measurement I use in this paper follow Carter et al. (2019) and Kvarven et al. (2020).

To measure the change in the effect size, I use the mean difference between the adjusted and unadjusted estimates. This measurement detects systematic bias in the adjusted effect size. An estimate close to zero then indicates no systematic bias.

I also measure the average deviance from the unadjusted estimate with the root mean squared error. This measure focuses on the absolute distance between the two estimates, regardless of the direction, thus directing attention to the average size of the difference rather than to systematic differences. The root mean squared error is calculated by the following formula:

$$RMSE = \sqrt{mean((adjusted - unadjusted)^2)}$$

[2] For one of the 31 labs included in Alogna et al. (2014), the raw data was unavailable. For this lab I use converted data

.

Again, no difference between the adjusted and the unadjusted methods produces an estimate close to zero.

Statistical power is measured by the minimum detectable effect size (MDE) at 80 % power. In this paper, I use the 5 % significance level. Hence, the MDE is given by the following formula:

$$MDE_i = 2.8 * SE_i$$

.

To compare the methods, I look at the mean MDE for each method compared to the MDE for the unadjusted estimate. A low MDE entails high levels of statistical power, while a high MDE means low levels of statistical power.

In this paper, a "false positive" will be defined as a case where the unadjusted meta-analysis does not find a statistically significant effect, but the adjusted meta-analysis does. Consequently, the "false-negative rate" is defined as the percentage of cases for which the adjusted meta-analysis finds a null effect but for which the unadjusted meta-analysis finds a statistically significant effect. Thus, the true effect is defined based on the result of the unadjusted meta-analytic replication result. Hence, in a scenario where the bias adjustment methods do not adjust for bias and therefore reach the same conclusion as the unadjusted analysis in terms of statistical significance, the false-positive and false-negative rate will be exactly zero. This is likely to be the case for Trim & Fill when the method does not adjust for bias, as the method then would produce the exact same result as the random effects model used for the unadjusted analysis. If the methods change the level of statistical power, however, the method might yield a small false-positive or false-negative rate even if there is no adjustment for bias. It is therefore important to also take the RMSE into account when determining if the false-positive or negative rate is due to adjusting for non-existing bias. I calculate these rates at the 5 % level.

3 Results

Figure 2 depicts the different effect sizes with 95 % confidence intervals. This plot shows that 3PSM did not return an estimate in 4 cases. The plot also indicates that PET-PEESE has the widest confidence intervals, indicating a low degree of accuracy in the analysis.

Table 2 shows that the results for PET and PET-PEESE are very similar. Both PET and PET-PEESE adjust for bias that is not present. The root mean squared error is 0.26 and the mean difference is 0.05 for PET. PET-PEESE yields a root mean squared error of 0.23 and a mean difference of 0.04 , indicating that there is no over- or underestimation but that there is a substantial difference between the adjusted and unadjusted effect sizes. Furthermore, the MDE at the 5 % level is 0.47 for PET and 0.44 for PET-PEESE, indicating a considerable loss of statistical power compared to the unadjusted MDE of 0.11 . This is also indicated by the false-negative rate of 0.27 .

PEESE also has a false-negative rate of 0.27 , but a higher level of statistical power with a mean MDE of 0.27 . PEESE also has the lowest RMSE of the three regression-based adjustment methods, with an RMSE of 0.13 .

Moreover, 3PSM has a similar false-negative rate and a small false-positive rate. Since the unadjusted method finds a statistically insignificant result in 12 cases where 3PSM yields an estimate, and 3PSM finds a statistically significant effect in one of these cases, the false-positive rate for 3PSM is 0.08 . This can be observed from Figure 2. Regarding statistical significance, the method came to a different conclusion than the unadjusted analysis three times, as reported in Figure 2. The mean difference and root mean square error are both close to zero. Hence, they show no indications of bias adjustment, and the MDE is near the unadjusted estimate.

Trim & fill is the only method tested that yields neither false positives nor false negatives. Hence, the false-positive and false-negative rate for this method is zero. This is illustrated in Figure 2, as Trim & fill reaches the same conclusion in terms of statistical significance as the unadjusted method for all 24 cases. Furthermore, all indicators show that trim & fill produces a result very close to the unadjusted method, regarding both effect size and statistical power.

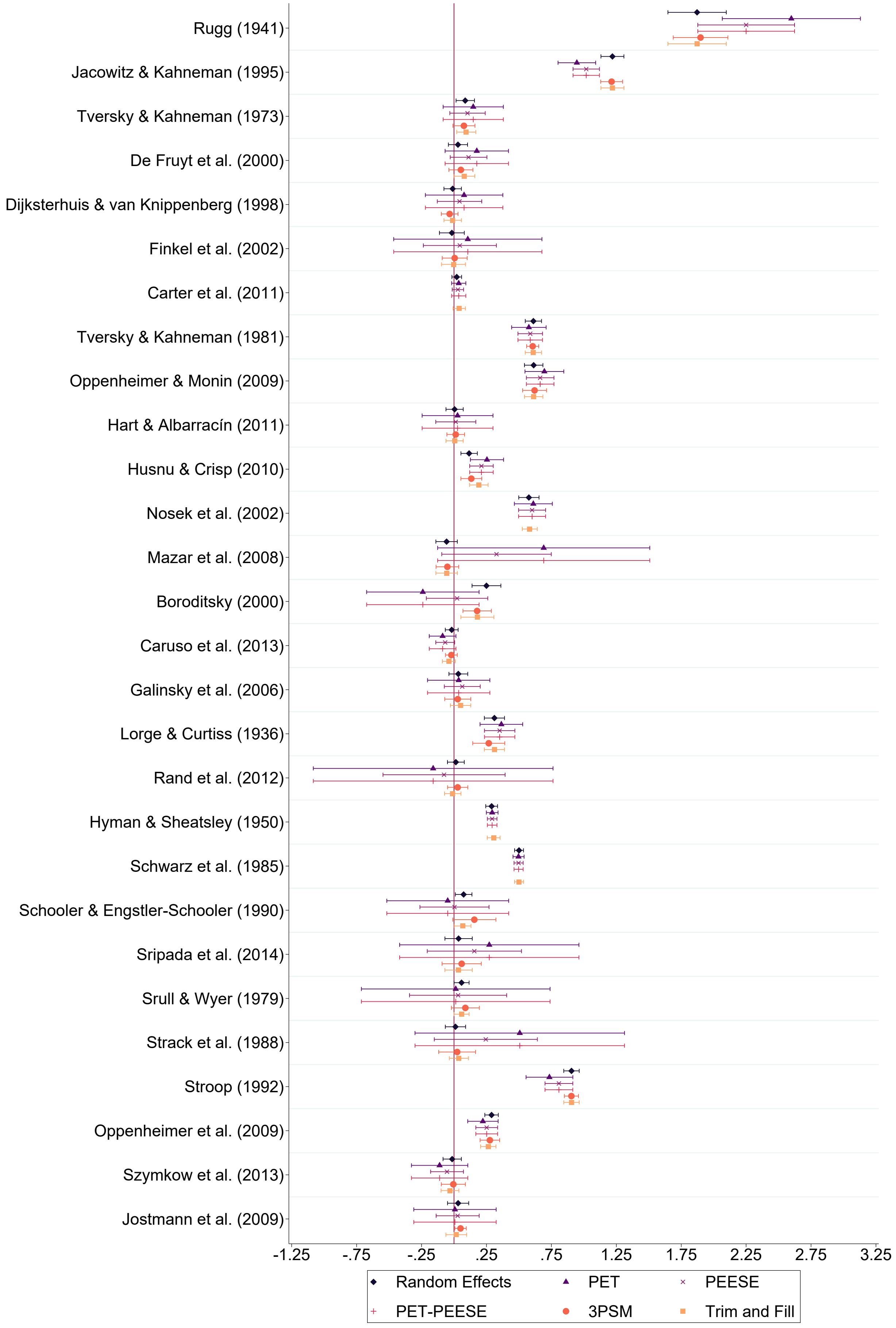

Figure 2: Effect sizes with 95 % confidence interval for all methods in Cohen's d.

Table 2 – Results for indicators used to measure performance

Method	False-positive rate	False-negative rate	Mean difference	Root mean squared error	Mean MDE
Trim & fill	0.00	0.00	0.00	0.02	0.11
Regression based methods					
PET	0.00	0.27	0.05	0.26	0.47
PEESE	0.00	0.27	0.03	0.13	0.27
PET-PEESE	0.00	0.27	0.04	0.23	0.44
3PSM	0.08	0.25	0.00	0.03	0.14

False-positive and -negative rates calculated at the 5 % level with the replication as the baseline. Mean difference measured in Cohen's d. MDE calculated at the 5 % level. Minimum detectable effect size (MDE) is for the 5 % level.

3.1 Robustness tests

I have also performed several robustness tests. The first test is to find out if the bias correcting methods perform equally across my dataset, or if the methods perform differently if the number of studies included in the meta-analysis is either high or low. The number of included studies could affect statistical power, as well as the method's ability to adjust for bias, creating the need for this robustness test.

I divide my sample into two equal parts based on the number of included studies in the meta-analysis. Hence, I create one group with the half of the meta-analyses that has the fewest included studies and another group with the highest amount of studies. The false-negative rate and the RMSE are slightly lower in the group with many studies. This indicates a higher level of statistical power in this group. The mean difference and RMSE are near the main estimate for both groups, indicating that the level of bias is close to the estimates for the full sample. The exception is for PET-PEESE, which is now producing a similar mean difference, and RMSE to PEESE. As PET-PEESE is a conditional estimator, it is equal to PEESE in 76.92 % of cases for the group with a high number of included studies, versus 6.67 % for the group with a low number of included studies.

Furthermore, I also limit the sample to cases where 3PSM returns an estimate to better compare the three methods and to rule out the possibility that the comparative results are due to the difference in the sample. This excludes 4 meta-analyses from my sample, leaving a sample size of 24 meta-analyses. The results from this test are almost identical to the main results.

Table 3 – Robustness results for indicators used to measure performance

Method	**False-positive rate**	**False-negative rate**	**Mean difference**	**Root mean squared error**	**Mean MDE**
For meta-analyses where the number of included studies are below the median					
Trim & fill	0.00	0.00	-0.00	0.03	0.11
Regression based methods					
PET	0.00	0.75	0.06	0.29	0.66
PEESE	0.00	0.75	0.03	0.14	0.35
PET-PEESE	0.00	0.75	0.07	0.28	0.66
3PSM	0.09	0.50	0.00	0.02	0.13
For meta-analyses where the number of included studies are above the median					
Trim & fill	0.00	0.00	0.00	0.02	0.11
Regression based methods					
PET	0.00	0.09	0.01	0.22	0.24
PEESE	0.00	0.09	0.02	0.13	0.17
PET-PEESE	0.00	0.09	0.01	0.13	0.19
3PSM	0.00	0.13	0.01	0.03	0.15
Only including meta-analyses where 3PSM returned an estimate					
Trim & fill	0.00	0.00	0.00	0.03	0.12
Regression based methods					
PET	0.00	0.33	0.06	0.28	0.50
PEESE	0.00	0.33	0.03	0.15	0.30
PET-PEESE	0.00	0.33	0.05	0.24	0.50
3PSM	0.08	0.25	0.00	0.03	0.14

False-positive and -negative rates calculated at the 5 % level with the replication as the baseline. Mean difference measured in Cohen’s d. MDE calculated at the 5 % level. Minimum detectable effect size (MDE) is for the 5 % level.

4 Discussion

This work found that 3PSM seems to make only small changes to the meta-analyses. However, these minor changes are in some cases enough to create false positives and negatives. It is therefore important to proceed with caution when applying this method.

When comparing the methods tested in this paper, it is important to note that 3PSM did not return an estimate in some cases. It is possible that the results for this method would be different if it had always returned an estimate. This also makes it difficult to evaluate the method, considering that the inability to return an estimate could be a limitation. In a robustness test, I limit the sample to include only cases where 3PSM yields an estimate. This test shows similar results as the main estimate.

PET & PET-PEESE is the method that creates the largest differences compared to the unadjusted effects. As PET-PEESE is a conditional estimator, and equal to PET in the 60.71 % of cases where PET is statistically insignificant, the difference between these two methods is small. The differences for PET and PET-PEESE range from -0.49 to 0.75 . For Cohen's d, 0.4 is considered a small effect, while 0.6 would indicate a medium effect. Hence, PET and PET-PEESE will in some cases over- or underestimate the effect by a small-to-medium amount. In addition, the method also decreases statistical power and creates wide confidence intervals, compared to the unadjusted analyses.

PEESE outperforms both PET and PET-PEESE when it comes to bias and statistical power, but has an identical false-negative rate as PET and PET-PEESE. The root mean squared error for PEESE is approximately half of the RMSE for PET and PET-PEESE, though still considerably larger than that for 3PSM and Trim & Fill.

Of the methods tested in this paper, Trim & Fill seems to perform the best, in that it does not correct for bias when none is present. However, the tests in this paper are a necessary but not sufficient condition for validity. It is also necessary for the method to adjust for bias when bias is present. If a method never adjusts for bias, it would perform good in the tests in this paper, though this does not make the method a good bias adjustment method. Thus, the results for Trim & Fill might simply indicate that the method does not adjust well for bias whether bias is present or not, as indicated by Carter et al. (2019) and Kvarven et al. (2020). It is therefore important to examine the results of this analysis in combination with those of other studies.

A central caveat for the results of this paper is the representativity of the meta-analytic datasets included in this work. All the included meta-analyses come from Many Labs projects. This means that there is no design heterogeneity of any sort, as would be expected in an average meta-analysis. In other words, the heterogeneity of these meta-analyses is artificially low. This is important first because these anti-biasing methods might have been developed for a dataset containing such heterogeneity and might therefore work differently on the datasets included in this study. Second, it is important to note that such heterogeneity can be incorrectly interpreted by the adjustment methods as bias. The methods are therefore likely to perform worse if design heterogeneity were present. Thus, the conclusions of this paper might change if design heterogeneity were included.

Furthermore, the selection of hypotheses tested in the included datasets might not be representative of the literature as a whole. There might be a replicator selection effect, meaning that researchers look for an effect that they do not believe to be replicable and hence end up with a high fraction of null findings. This is, however, tested for in Kvarven et al. (2020), where there was no evidence of such an effect.

It is also possible that there is no replicator selection effect but that the percentage of null findings might differ from the rest of the scientific literature because statistically significant findings are typically being replicated. The existence of statistically significant findings means there is previous support for the hypothesis, which can increase the chance that a significant effect will be produced, when the study is replicated. Thus, since the projects included in this study are replications, it is possible that the percent of true effects deviates from the rest of the literature, which could affect the mean difference and the root mean squared error if some methods overadjust only when there is a true effect and not otherwise (or vice versa). This would not affect internal validity but might slightly limit generalizability.

Another limitation of my analysis is that the replication datasets include between 12 and 37 studies each. The number of studies included in the dataset can affect important factors such as statistical power and the adjustment method's ability to adjust for bias. Some methods try to assess the level of bias based on differences in effect and standard error between the included studies. When there are few studies, it might become difficult for these adjustment methods to accurately estimate relevant parameters. A low number of included studies has also previously been raised as a limitation for the performance of PET-PEESE (Stanley, 2017). Hence, the conclusions of the paper may not be possible to extrapolate to bigger meta-analyses. To test for these kinds of differences, I have included a robustness test where I split my sample into two groups. The first group consists of the 50 % of meta-analyses that have the lowest number of included studies. The other group is the 50 % with the highest number of studies.

As expected, the false-negative rate and the level of statistical power are better among the meta-analyses with the largest number of included studies. This is not surprising, as more included studies typically means higher statistical power. The mean difference and RMSE are almost identical between the two groups, indicating that the performance of the bias adjustment methods does not improve with a larger sample. However, it is possible that this conclusion would be different with even larger numbers of included studies.

It is important to mention, however, that there is a difference in the mean difference and RMSE for PET-PEESE. When in the group with a high number of included studies, the RMSE and mean difference are very close to the results for PEESE, while they are close to the results for PET in the other group. This is probably due to the group with a high number of included studies having more statistical power, which in turn makes the meta-analytic effect size more likely to be statistically significant. As previously stated, PET-PEESE is a conditional estimator equal to PET when PET is not statistically significant, and equal to PEESE when PET is statistically significant. Thus, it stands to reason that PET-PEESE will be closer to PET when statistical power is low, and more likely to be close to PEESE when statistical power is high. This is supported by PET-PEESE being equal to PEESE in 6.67 % of cases in the group with a low number of the included studies, and 76.92 % in the group of studies with a high number of included studies.

One important point, however, is that most meta-analyses typically include few studies. Based on data from van Erp et al. (2017), which has gathered information regarding the included number of studies from 747 meta-analyses from 61 published papers in *Psychological Bulletin*, the median number of included studies in a meta-analysis is 12. Furthermore, 75 % of the meta-analyses in their sample includes 33 studies or less. In this context, my dataset, with a median of 23 included studies, includes fairly large meta-analyses. This implies that the robustness test with the smaller half of the sample might be more representative for how the adjustment methods would perform on actual unbiased data in the literature.

5 Conclusion

I find that the regression-based adjustment estimators come with a considerable loss of statistical power when applied to unbiased data. For these methods, the minimum detectable effect size is between 2.45 and 4.27 times larger than for the unadjusted analysis. This loss of power is further demonstrated by the 26.67 % false negatives. In addition, the regression-based methods also make changes to the meta-analytic estimate, when there is no bias and no change is required. This indicates that the regression-based methods come with a cost when applied to unbiased data.

The remaining methods seem to outperform the regression-based methods in this test. The three parameter selection model leads to a small loss of statistical power. The power loss is, however, minimal compared to the power loss of the regression-based methods. The Trim & Fill method does not adjust the unbiased estimate at all, and is thus the method that performs the best in this paper. However, not adjusting the effect in an unbiased dataset is only a necessary, but not sufficient, condition. This paper does not evaluate how well the methods actually adjust for publication bias. These results do, however, indicate that there is little or no cost of using these two methods in cases where there is no bias. Thus, one would expect, based on Carter et al. (2019) and Kvarven et al. (2020), that these methods might give some limited benefit when applied to biased data, but without being harmful to the meta-analysis if the dataset turns out to be unbiased. I therefore conclude that Trim & Fill and the three parameter selection model can come with potential benefits, but without the same costs as the regression-based adjustment methods.

References

V. Alogna, M. K. Attaya, P. Aucoin, Š. Bahník, S. Birch, A. R. Birt, B. H. Bornstein, S. Bouwmeester, M. A. Brandimonte, C. Brown, et al. Registered replication report: Schooler and engstler-schooler (1990). *Perspectives on Psychological Science*, 9(5):556–578, 2014.

I. Andrews and M. Kasy. Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94, 2019.

C. Batz-Barbarich, L. Tay, L. Kuykendall, and H. K. Cheung. A meta-analysis of gender differences in subjective well-being: estimating effect sizes and associations with gender inequality. *Psychological science*, 29(9):1491–1503, 2018.

L. Boroditsky. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28, 2000.

S. Bouwmeester, P. P. Verkoeijen, B. Aczel, F. Barbosa, L. Bègue, P. Brañas-Garza, T. G. Chmura, G. Cornelissen, F. S. Døssing, A. M. Espín, et al. Registered replication report: Rand, greene, and nowak (2012). *Perspectives on Psychological Science*, 12(3):527–542, 2017.

C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9): 637–644, 2018.

E. C. Carter, F. D. Schönbrodt, W. M. Gervais, and J. Hilgard. Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2):115–144, 2019.

T. J. Carter, M. J. Ferguson, and R. R. Hassin. A single exposure to the american flag shifts support toward republicanism up to 8 months later. *Psychological science*, 22(8):1011–1018, 2011.

E. M. Caruso, K. D. Vohs, B. Baxter, and A. Waytz. Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, 142(2):301, 2013.

I. Cheung, L. Campbell, E. P. LeBel, R. A. Ackerman, B. Aykutoğlu, Š. Bahník, J. D. Bowen, C. A. Bredow, C. Bromberg, P. A. Caprariello, et al. Registered replication report: Study 1 from finkel, rusbult, kumashiro, & hannon (2002). *Perspectives on Psychological Science*, 11 (5):750–764, 2016.

F. De Fruyt, L. Van de Wiele, and C. Van Heeringen. Cloninger's psychobiological model of temperament and character and the five-factor model of personality. *Personality and individual differences*, 29(3):441–452, 2000.

A. Dijksterhuis and A. Van Knippenberg. The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of personality and social psychology*, 74(4):865, 1998.

S. Duval and R. Tweedie. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the american statistical association*, 95(449):89–98, 2000.

C. R. Ebersole, O. E. Atherton, A. L. Belanger, H. M. Skulborstad, J. M. Allen, J. B. Banks, E. Baranski, M. J. Bernstein, D. B. Bonfiglio, L. Boucher, et al. Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82, 2016.

C. C. Eckel and S. C. Füllbrunn. Thar she blows? gender, competition, and bubbles in experimental asset markets. *American Economic Review*, 105(2):906–20, 2015.

A. Eerland, A. M. Sherrill, J. P. Magliano, R. A. Zwaan, J. D. Arnal, P. Aucoin, S. A. Berger, A. R. Birt, N. Capezza, M. Carlucci, et al. Registered replication report: Hart & albarracín (2011). *Perspectives on Psychological Science*, 11(1):158–171, 2016.

E. J. Finkel, C. E. Rusbult, M. Kumashiro, and P. A. Hannon. Dealing with betrayal in close relationships: Does commitment promote forgiveness? *Journal of personality and social psychology*, 82(6):956, 2002.

A. D. Galinsky, J. C. Magee, M. E. Inesi, and D. H. Gruenfeld. Power and perspectives not taken. *Psychological science*, 17(12):1068–1074, 2006.

J. Gurevitch, J. Koricheva, S. Nakagawa, and G. Stewart. Meta-analysis and the science of research synthesis. *Nature*, 555(7695):175–182, 2018.

M. S. Hagger, N. L. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, M. J. Brandt, G. Brewer, S. Bruyneel, et al. A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4):546–573, 2016.

W. Hart and D. Albarracín. Learning about what others were doing: Verb aspect and attributions of mundane and criminal intent for past actions. *Psychological Science*, 22(2):261–266, 2011.

T. Havránek. Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association*, 13(6):1180–1204, 2015.

T. Havránek and Z. Irsova. Estimating vertical spillovers from fdi: Why results vary and what the true effect is. *Journal of International Economics*, 85(2):234–244, 2011.

T. Havránek and O. Kokes. Income elasticity of gasoline demand: A meta-analysis. *Energy Economics*, 47:77–86, 2015.

S. Husnu and R. J. Crisp. Elaboration enhances the imagined contact effect. *Journal of Experimental Social Psychology*, 46(6):943–950, 2010.

H. H. Hyman and P. B. Sheatsley. The current status of american public opinion. pages 11–34. National Council of Social Studies, New York, 1950.

J. P. Ioannidis. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3):485–514, 2016.

K. E. Jacowitz and D. Kahneman. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11):1161–1166, 1995.

N. B. Jostmann, D. Lakens, and T. W. Schubert. Weight as an embodiment of importance. *Psychological science*, 20(9):1169–1174, 2009.

R. A. Klein, K. A. Ratliff, M. Vianello, R. B. Adams Jr, Š. Bahník, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh, et al. Investigating variation in replicability. *Social psychology*, 2014.

R. A. Klein, M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Š. Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

A. Kvarven, E. Strømland, and M. Johannesson. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4):423–434, 2020.

M. Lewis, M. Mathur, T. VanderWeele, and M. C. Frank. The puzzling relationship between multi-lab replications and meta-analyses of the rest of the literature. 2020.

I. Lorge and C. C. Curtiss. Prestige, suggestion, and attitudes. *The Journal of Social Psychology*, 7(4):386–402, 1936.

N. Mazar, O. Amir, and D. Ariely. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644, 2008.

R. J. McCarthy, J. J. Skowronski, B. Verschuere, E. H. Meijer, A. Jim, K. Hoogesteyn, R. Orthey, O. A. Acar, B. Aczel, B. E. Bakos, et al. Registered replication report on srull and wyer (1979). *Advances in Methods and Practices in Psychological Science*, 1(3):321–336, 2018.

B. A. Nosek, M. R. Banaji, and A. G. Greenwald. Math= male, me= female, therefore math≠ me. *Journal of personality and social psychology*, 83(1):44, 2002.

M. O'Donnell, L. D. Nelson, E. Ackermann, B. Aczel, A. Akhtar, S. Aldrovandi, N. Alshaif, R. Andringa, M. Aveyard, P. Babincak, et al. Registered replication report: Dijksterhuis and van knippenberg (1998). *Perspectives on Psychological Science*, 13(2):268–294, 2018.

D. M. Oppenheimer and B. Monin. The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, 4(5):326, 2009.

D. M. Oppenheimer, T. Meyvis, and N. Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45 (4):867–872, 2009.

D. G. Rand, J. D. Greene, and M. A. Nowak. Spontaneous giving and calculated greed. *Nature*, 489(7416):427–430, 2012.

D. Rugg. Experiments in wording questions: Ii. *Public opinion quarterly*, 1941.

J. W. Schooler and T. Y. Engstler-Schooler. Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive psychology*, 22(1):36–71, 1990.

N. Schwarz, H.-J. Hippler, B. Deutsch, and F. Strack. Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, 49(3): 388–395, 1985.

A. P. Siddaway, A. M. Wood, and L. V. Hedges. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology*, 70:747–770, 2019.

U. Simonsohn, L. D. Nelson, and J. P. Simmons. p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6): 666–681, 2014.

V. K. Smith and J.-C. Huang. Can markets value air quality? a meta-analysis of hedonic property value models. *Journal of political economy*, 103(1):209–227, 1995.

C. Sripada, D. Kessler, and J. Jonides. Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological science*, 25(6):1227–1234, 2014.

T. K. Srull and R. S. Wyer. The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social psychology*, 37(10):1660, 1979.

T. Stanley and H. Doucouliagos. Identifying and correcting publication selection bias in the efficiency-wage literature: Heckman meta-regression. *Economics Series*, 11:2007, 2007.

T. D. Stanley. Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and statistics*, 70(1):103–127, 2008.

T. D. Stanley. Limitations of pet-peese and other meta-analysis methods. *Social Psychological and Personality Science*, 8(5):581–591, 2017.

T. D. Stanley and H. Doucouliagos. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78, 2014.

T. D. Sterling, W. L. Rosenbaum, and J. J. Weinkam. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The american statistician*, 49(1):108–112, 1995.

F. Strack, L. L. Martin, and S. Stepper. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of personality and social psychology*, 54(5):768, 1988.

J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, 121(1):15, 1992.

A. Szymkow, J. Chandler, H. IJzerman, M. Parzuchowski, and B. Wojciszke. Warmer hearts, warmer rooms. *Social Psychology*, 2013.

A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

S. van Erp, J. Verhagen, R. P. Grasman, and E.-J. Wagenmakers. Estimates of between-study heterogeneity for 705 meta-analyses reported in psychological bulletin from 1990–2013. *Journal of Open Psychology Data*, 5(1), 2017.

B. Verschuere, E. H. Meijer, A. Jim, K. Hoogesteyn, R. Orthey, R. J. McCarthy, J. J. Skowronski, O. A. Acar, B. Aczel, B. E. Bakos, et al. Registered replication report on mazar, amir, and ariely (2008). *Advances in Methods and Practices in Psychological Science*, 1(3):299–317, 2018.

J. L. Vevea and L. V. Hedges. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3):419–435, 1995.

E.-J. Wagenmakers, T. Beek, L. Dijkhoff, Q. F. Gronau, A. Acosta, R. Adams Jr, D. Albohn, E. Allard, S. D. Benning, E.-M. Blouin-Hudon, et al. Registered replication report: strack, martin, & stepper (1988). *Perspectives on Psychological Science*, 11(6):917–928, 2016.