Meta-Analysis and Publication Bias: How Well Does The Fat-Pet-Peese Procedure Work?

Nazila Alinaghi and W. Robert Reed

WORKING PAPER 14/2017
October 2017

Working Papers in Public Finance

VICTORIA UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School

The Working Papers in Public Finance series is published by the Victoria Business School to disseminate initial research on public finance topics, from economists, accountants, finance, law and tax specialists, to a wider audience. Any opinions and views expressed in these papers are those of the author(s). They should not be attributed to Victoria University of Wellington or the sponsors of the Chair in Public Finance.

Further enquiries to:
The Administrator
Chair in Public Finance
Victoria University of Wellington
PO Box 600
Wellington 6041
New Zealand

Phone: +64-4-463-9656
Email: cpf-info@vuw.ac.nz

Papers in the series can be downloaded from the following website:
http://www.victoria.ac.nz/cpf/working-papers

Working Papers in
Public Finance

META-ANALYSIS AND PUBLICATION BIAS: HOW WELL DOES THE FAT-PET-PEESE PROCEDURE WORK?

by

Nazila Alinaghi and W. Robert Reed
Department of Economics and Finance
University of Canterbury
Christchurch, NEW ZEALAND

Abstract

This paper studies the performance of the FAT-PET-PEESE (FPP) procedure, a commonly employed approach for addressing publication bias in the economics and business meta-analysis literature. The FPP procedure is generally used for three purposes: (i) to test whether a sample of estimates suffers from publication bias, (ii) to test whether the estimates indicate that the effect of interest is statistically different from zero, and (iii) to obtain an estimate of the mean true effect. Our findings indicate that the FPP procedure performs well in the basic but unrealistic environment of "fixed effects," where all estimates are assumed to derive from a single population value and sampling error is the only reason for why studies produce different estimates. However, when we study its performance in more realistic data environments, where there is heterogeneity in the population effects across and within studies, the FPP procedure becomes unreliable for the first two purposes, and is less efficient than other estimators when estimating overall mean effect. Further, hypothesis tests about the mean true effect are frequently unreliable. We corroborate our findings by recreating the simulation framework of Stanley and Doucouliagos (2017) and repeat our tests using their framework.

KEYWORDS: Meta-analysis, publication bias, funnel asymmetry test (FAT), Precision Effect Estimate with Standard Error (PEESE), Monte Carlo, Simulations

JEL CLASSIFICATION: B41, C15, C18

Contact Information: W. Robert Reed, Private Bag 4800, Christchurch 8140, New Zealand. Phone: +64-3-364-2846; Email: bob.reed@canterbury.ac.nz .

Acknowledgments. We thank Tom Stanley and participants at the 2016 MAER-Net Colloquium for helpful comments. We are especially indebted to Sanghyun Hong for invaluable research assistance. Lastly, we gratefully acknowledge the comments of anonymous reviewers which resulted in a much improved manuscript.

I. INTRODUCTION

Meta-analysis is the statistical analysis of estimates from multiple studies that are all concerned with measuring a similar "effect." Two main goals of meta-analysis are (i) to reach a single conclusion about the size and significance of that effect, and (ii) to understand why studies differ in their estimates of that effect. Meta-analysis has become an increasingly popular research tool in economics and business. FIGURE 1 shows a time series bar chart that lists all Web of Science journal articles in economics and business that have the word "meta-analysis" in the title. The trend is clearly upwards.

It is widely recognized that publication selection bias distorts the distribution of estimated effects that appear in the literature, either because statistically insignificant estimates may not be considered sufficiently interesting to publish, or because they may be wrong-signed according to the established theory in the field, the researcher's personal beliefs, or other reasons. This is a problem. As the data for meta-analysis consist of estimated effects from the literature, if that distribution is distorted, so will be the conclusions that derive from them. Thus, a crucial component of a meta-analysis is to detect, and correct, publication selection bias.

A common procedure for doing this in the economics and business literature is the FAT-PET-PEESE procedure (Stanley and Doucouliagos, 2012; 2014). The starting point for the FAT-PET-PEESE procedure is the equation

$$\hat{\alpha}_j = \beta_0 + \beta_1 SE_j + e_j \, , j = 1,2, \dots, M, \tag{1}$$

where the $\hat{\alpha}_j$ is the j^{th} estimated effect in the meta-analyst's sample, β_0 represents the mean of the distribution of estimated effects, and SE_j is the estimated effect's standard error. The latter is included to control for publication bias.

Because the estimated effects have different precisions, the error term in Equation (1) is heteroskedastic. Accordingly, WLS is used to estimate this equation, with weights given by $\omega_j = \left(\frac{1}{SE_j}\right)$. Mutiplying through by the weights produces the specification,

(2) $$\frac{\hat{\alpha}_j}{SE_j} = \beta_1 + \beta_0\left(\frac{1}{SE_j}\right) + \frac{e_j}{SE_j},$$

which can now be estimated by OLS.

With this as the starting point, FIGURE 2 depicts the different steps associated with the FAT-PET-PEESE procedure. First is the Funnel Asymmetry Test (FAT). It tests whether $\beta_1 = 0$. If the estimate of β_1 is significant, that is taken to indicate that the estimates suffer from publication bias. Next is the Precision Effect Test (PET). It tests whether $\beta_0 = 0$. This test is designed to determine whether the mean of the distribution of estimated effects is zero; i.e., whether "an effect" exists.

If the PET fails to reject the null hypothesis of no effect, then $\hat{\beta}_0$ is taken as the estimate of overall effect with the understanding that it is statistically insignificant from zero. If the PET rejects the null, then a new specification is estimated, and the associated estimate of β_0 represents the best estimate of overall effect. This is known as the PEESE, or Precision Effect Estimate with Standard Error.

Examples of recent studies in the economics and business literature that use the FAT-PET-PEESE procedure are Costa-Font, Gemmill, and Rubert (2011), Doucouliagos, Stanley, and Viscusi (2014), Doucouliagos and Paldam (2013), Efendic, Pugh, and Adnett (2011), Haelermans and Borghans (2012), Havránek (2010), Iwasaki and Tokunaga (2014), Laroche (2016), Lazzaroni and van Bergeijk (2014), Linde Leonard, Stanley, and Doucouliagos (2014), and Nelson (2013).

Despite the widespread use of this (and similar) procedures, a number of researchers have questioned the use of funnel plot-inspired procedures that rely on coefficient standard

errors to detect publication bias. An early contributor is Terrin et al. (2003). They caution that heterogeneity in true effects across studies can impair the ability of coefficient standard errors to both identify the presence of publication selection bias, and correct for it. Similar warnings can be found in Lau et al. (2006) and Sterne et al. (2011). The latter write: "Because it is impossible to know the precise mechanism(s) leading to funnel plot asymmetry, simulation studies (in which tests are evaluated on large numbers of computer generated datasets) are required to evaluate test characteristics" (page 598). Accordingly, this study uses simulation to investigate how well the FAT-PET-PEESE (FPP) procedure

- correctly detects the existence of publication bias,
- correctly tests whether a population effect exists, and
- compares with two common meta-analysis estimators that do not correct for publication bias.

A distinctive feature of our simulations is that we simulate meta-analyses under three "data environments". In the simplest data environment ("Fixed Effects"), each study reports a single regression equation and there is one true effect underlying all regressions. A second data environment generalizes this by allowing heterogeneity in true effects across studies, while each study still reports a single regression equation ("Random Effects"). The third data environment generalizes further by allowing studies to contain multiple regression equations, with the true effects underlying these regressions differing both within and across studies ("Panel Random Effects"). Our analysis focuses on this last case because it comes closest to matching the situation faced by most meta-analyses in economics and business. We also separately investigate two different types of publication selection bias: sample selection that is biased against insignificant estimates, and sample selection that is biased against "wrong signs", here taken to mean negative estimates.

We find that the FPP procedure works well in the "Fixed Effects" environment. The nominal sizes of the FAT and PET tests are close to their significance levels when there is no

publication selection bias and the true effect is zero, respectively. Further, the FPP procedure has good power. Rejection rates are either 100% or close to 100% when publication selection bias is nontrivial, and when the true effect is nonzero. However, the FPP procedure works progressively worse as the data environment is generalized to the "Random Effects" and "Panel Random Effects" environments.

We next analyse the performance of the FPP procedure within the Panel Random Effects environment, and compare it with two related WLS estimators that do not include a *SE* variable to correct for publication bias. While the FPP procedure has smallest bias, it consistently is less efficient, and hypothesis testing about the true effect is sufficiently distorted as to render it useless in many instances.

We are mindful that our results differ substantially from previous research supporting the use of the FPP procedure (e.g., Stanley and Doucouliagos, 2012; Stanley and Doucouliagos, 2014; Stanley and Doucouliagos, 2017). This raises concern that our results are due to idiosyncratic aspects of our simulation procedure. To address this concern, we replicate the simulation environment of Stanley and Doucouliagos (2017) – henceforth S&D – and repeat our analysis within their simulation framework. While there are differences, we find that the FPP procedure suffers from instances of poor performance even within the S&D framework: it performs poorly on the FAT when there is publication selection bias, does poorly on the PET when true effects have substantial heterogeneity, and is generally less efficient than the two WLS estimators that do not include a *SE* term. The main difference in results compared to our framework is that hypothesis testing about the mean true effect is more reliable in the S&D framework. This last result is not too surprising given that the error structure under Panel Random Effects is more complex.

We proceed as follows. Section II describes our simulation framework. Section III describes the simulated datasets used in our analysis. Section IV presents our results. Section V compares our results with those from previous studies. Section VI concludes.

II. DESCRIPTION OF THE SIMULATION FRAMEWORK

General framework. The general framework for our simulations is presented in TABLE 1.[1] We model a situation where a meta-analyst is interested in studying the effect that a variable *x* has on an outcome *y*. The simulation begins by generating individual observations for a primary study. It then collects these observations into a sample. The primary study estimates the effect of *x* on *y* using this sample. Other studies are then simulated until a large number of estimates (1000 estimates) are generated and stored in a conceptual "holding tank." Publication selection bias then filters out estimates that are "unpreferred." The meta-analyst applies the FPP procedure to this censored sample. The process is repeated 1000 times and the aggregated results from these simulated meta-analyses are analysed across several performance measures.

As noted above, we consider three types of "data environments." The "Fixed Effects" (FE) data environment models primary studies as having only one regression and assumes that there is one true effect of *x* on *y* underlying all primary studies. The "Random Effects" (RE) data environment also assumes only one regression per primary study, but allows true effects to differ across studies. This could arise because different studies sample different populations, with the effect of *x* differing across populations due to demographics, culture, institutions, etc. The "Panel Random Effects" (PRE) data environment assumes primary studies report multiple regression equations, with true effects differing both across and within studies. True effects could differ for regressions from the same primary study because even though they may work with similar samples, the effect of *x* in a given regression could be moderated by other

[1] This framework borrows heavily from Reed, Florax, and Poot (2015) and Reed (2015).

variables, and the different regressions could include different sets of control variables. We assume the ultimate goal of the meta-analyst is to estimate the mean true effect across studies, α. The subsequent discussion focuses on the PRE case, but is easily modified for the FE and RE cases.

Each regression r in a study i is assumed to be estimated from a sample of 100 observations generated from the data generating process (DGP):

(3) $y_{irt} = 1 + \alpha_{ir} x_{irt} + \varepsilon_{irt}$,

where α_{ir} is the true effect of x on y in regression r of study i. The true effect is the same for all observations belonging to the same regression, but differs across regressions both within the same study, and across different studies. Likewise, the error terms, ε_{irt}, have the same variance for observations belonging to the same regression, but have different variances for observations belonging to different regressions. A consequence of the latter is that estimated effects from different regressions will have different standard errors. Both true effects and error variances/coefficient standard errors are constructed so that they are more similar for regressions from the same study compared to regressions from different studies.

Each regression r from primary study i uses OLS to estimate the effect size, α_{ir}, from the following specification:

(4) $y_{irt} = \psi_{ir} + \alpha_{ir} x_{irt} + \epsilon_{irt}$.

producing regression results, $\{\hat{\alpha}_{ir}, s.e.(\hat{\alpha}_{ir})\}$. We note that OLS is appropriate in this setting because the error terms for the associated sample are homoskedastic.

In this manner, results are produced for a 1000 regressions, with 100 studies each producing results for 10 regressions. These 1000 estimates comprise a "Pre-Publication Selection Bias" sample. The meta-analyst never sees this sample. Instead publication selection bias filters out estimates that are deemed to be "unpreferred."

We separately consider two types of publication selection bias: (i) bias against estimates that are statistically insignificant, and (ii) bias against estimates that are "wrong signed", where our simulations assume that the "correct" sign is positive. This latter selection bias is intended to model scenarios where theory predicts that the sign of a coefficient should be positive, as in the case of value of life studies, or studies of supply curve price elasticities. Our framework models incomplete selection bias, as it is unrealistic to think that all insignificant or negative estimates are eliminated. Some of the unpreferred estimates are allowed to escape the publication selection bias filter, albeit at a substantially reduced rate.[2]

Publication selection bias reduces the sample to M observations, $(\{\hat{\alpha}_1, SE_1\}, \{\hat{\alpha}_2, SE_2\}, \ldots, \{\hat{\alpha}_M, SE_M\})$, where $M \leq 1000$ and is endogenously determined. This is the "Post-Publication Selection Bias" sample and is the sample that the meta-analyst has available to work with.

The meta-analyst proceeds by following the FPP procedure outlined in FIGURE 2 and described above. The resulting estimates of β_1 and β_0 in Equation (2) are used for the FAT and PET, respectively. Subsequent testing and estimation leads to a final estimate of β_0, which provides the meta-analyst with an estimate of the mean true effect, α.

In this manner, a meta-analysis is simulated starting from the generation of individual observations, to the estimation of regressions in primary studies, to creating a collection of estimated effects constituting "the literature" which the meta-analyst uses to (i) test for publication bias (FAT), (ii) test for the existence of an effect (PET), and (iii) obtain an estimate

[2] For the FE and RE data environments, significant or positive estimates were selected into the Post-Publication Selection Bias sample with probability 1.00. Insignificant or negative estimates were included with probability 0.10. The PRE case is more complicated because PRE has multiple estimates per study. This raises an issue that does not arise with FE/RE: What should be done when some estimates from a study are significant and some are insignificant? One option that we considered but rejected was to only use the estimates from a study that satisfied the selection criterion. However, we felt that was not realistic. Journals don't publish some of the estimates from a study. They either publish all or none. So we had to have a rule that incorporated this distinctive feature of multiple estimates per study. We settled on the following rule: In order to be accepted into the Post-Publication Selection Bias sample, a study must have most of its estimates (at least 7 out of 10) be statistically significant/positive. 7 out of 10 was chosen rather than 6 out of 10 or 8 out of 10 because this produced overall "removal rates" that were roughly consistent with those from the FE and RE data environments.

of the overall effect of *x* on *y*. We assume the meta-analyst follows good practice and uses cluster robust standard errors to account for correlations in estimates from the same study.

The last stage of our simulation repeats the above process 1000 times and analyses how well the FPP procedure performs. In addition to tracking the performance of the FPP with respect to the FAT and PET, we calculate the mean value of the estimated $\hat{\beta}_0's$, their associated mean squared error (MSE), and how well hypothesis tests about the mean true value perform ($H_0: \beta_0 = \alpha$).

Further, to get a better appreciation of the FPP procedure, we compare its performance with another procedure that does not include the *SE* variable to correct for publication selection bias. This procedure estimates the equation

(5) $\hat{\alpha}_j = \beta_0 + e_j\ , j = 1,2, \ldots, M,$

using one of two different WLS estimators. The first of these is the same WLS described above, with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. The second uses the weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2 + \tau^2}}\right)$, where τ^2 is the estimated variance of the true effect across studies.[3] This is the familiar random effects estimator. As these two WLS estimators match up conceptually with the Fixed Effects and Random Effects cases, and to underscore their similarity, we identify them below as "WLS-FE" and "WLS-RE."

The framework described above sets up six classes of experiments, arising from the different combinations of data environments and publication selection bias types (FE/bias against insignificance, FE/bias against wrong signs, RE/bias against insignificance, etc.). Within each class, we construct nine experiments corresponding to nine different values of the

[3] τ^2 is estimated using the *metareg* procedure in Stata. We used the method of moments (mm) option, which is a generalization of the commonly used DerSimonian and Laird (1986) method. We chose this option rather than the maximum likelihood option (reml) because we found that a non-trivial portion of the maximization routines failed to converge. The method of moments option avoids this problem.

mean true effect, $\alpha = 0.0, 0.5, 1.0, 1.5, \ldots, 4.0$. Each individual experiment creates 1,000 simulated meta-analysis studies. Thus, for each class of experiment, and for each value of α, (a total of 54 experiments) we have 1,000 tests for publication bias (FAT), 1,000 tests for the existence of a non-zero overall true effect (PET), and 1,000 estimates of α.

It is worth noting that there is an important interplay between the size of the mean true effect, α, and the incidence of publication selection bias. This is most clearly seen in the case of selection bias against negative estimates. When $\alpha = 0$, approximately half of the Pre-Publication Selection Bias sample will be negative and targeted for removal, causing the average of the remaining estimates to have a substantial positive bias. As α increases, more and more of the distribution moves into positive territory, so that fewer and fewer estimates are withheld from the Post-Publication Selection Bias sample. As a result, the positive bias becomes smaller. As α increases further, the percent of negative estimates becomes negligible, almost all estimates are selected into the Post-Publication Selection Bias sample, and publication bias is eliminated.

The situation for selection bias against insignificant estimates is more complicated. When $\alpha = 0$, no bias in the estimate of α is expected because the distribution of estimates is symmetrically distributed around 0. Removal of insignificant estimates eliminates negative and positive estimates with equal effect, so that the mean of the remaining estimates is still 0. As α increases, more negative than positive estimates are eliminated, and publication selection causes estimates of α to be positively biased. However, as the distribution moves further to the right, the percent of insignificant estimates becomes negligible, so that almost all estimates are selected into the Post-Publication Selection Bias sample, and publication bias is eliminated. Thus, when selection bias targets insignificant estimates, the effect of increases in α on publication bias is non-monotonic, with increases causing the bias to go from zero to positive and back to zero again. The relationship between effect size and the incidence of publication

selection bias has not been noted by previous studies. Further details about the simulation framework used for our analysis are given in APPENDIX 1.

III. CHARACTERISTICS OF THE SIMULATED META-ANALYSIS SAMPLES

In setting the specific parameter values underlying our simulation framework, we tried to simulate meta-analysis studies that met two general criteria. First, the simulated meta-analysis studies should appear "realistic", as measured by (i) the range of effect estimates and associated t-values, (ii) the percent of statistically significant estimates, and (iii) the degree of effect heterogeneity across estimates, commonly measured by I^2 (Higgins and Thompson, 2002). Second, we wanted the meta-analysis studies to display a wide range of publication selection incidences. In particular, we aimed to have selection bias eliminate a wide range of estimates between 0 and 100 percent of the Pre-Publication Selection Bias sample. On top of that, we wanted the pool of estimates in the Pre-Publication Selection Bias sample to be the same for the two types of selection bias. But this meant that the same parameter settings had to produce two Post-Publication Selection Bias samples, with each one satisfying our criteria.This was challenging.

TABLE 2 gives sample characteristics for a representative meta-analysis dataset in the Random Effects data environment with a mean true effect equal to 1. The top panel reports characteristics before selection bias. When $\alpha = 1$, estimated effects range, on average, from -7.47 to 9.46, with 90% of the estimated effects lying between -2.38 and 4.39. The median estimated effect is insignificant, with a t-value of 0.79. While t-values range widely within the meta-analysis sample, 90% lie between -1.47 and 5.90. Roughly a quarter of the estimated effects are statistically significant. This meta-analysis sample is characterized by substantial heterogeneity, with a median I^2 value of 0.86.

The middle panel reports what a representative meta-analysis sample looks like after selection bias filters out insignificant estimates. On average, only 33.0 percent of the original

estimates survive to the Post-Publication Selection Bias sample. While the range of estimates remains largely the same, selection bias has disproportionately eliminated negative estimates. This is evident by comparing the (P5%, P95%) range across the two samples. This has shifted from (-2.38, 4.39) to (-2.07, 5.69). The result is that the median estimated effect in the post-selection bias sample now has an average positive bias of 81%. As one would expect, t-values are substantially larger, with an average median value of 2.54. After selection against insignificance, over 90% of the estimated effects in the meta-analyst's sample are statistically significant. Measured heterogeneity has also increased. The median I^2 value across meta-analysis samples is now 0.94.

The bottom panel allows one to compare how the meta-analyst's sample would look differently if publication selection bias targeted negative estimates rather than insignificant ones. Our parameter settings result in a substantially larger number of effects surviving to the Post-Publication Selection Bias sample, so that the meta-analyst sees an average of 74.6% of the full set of estimated effects. This moderates the bias from eliminating negative estimates, though the median estimated effect still shows a substantial positive bias of 55%. The median t-statistic is 1.28. While this is higher than the value in the uncensored sample, it is lower than what would result if selection bias targeted insignificant estimates, as one would expect. Approximately 50% of the estimates are significant, and heterogeneity remains substantial with a median I^2 value of 0.81.

Reviewing the bottom two panels, we deem the simulated meta-analysis samples to have met our two general criteria. The simulated samples look "realistic." In particular, the I^2 values are consistent with Stanley and Doucouliagos' statement that "it is our experience that I^2 values of 80% to 90% are the norm" (Stanley and Doucouliagos, 2017, page 28). Further, our samples allow for a substantial degree of selection bias, with large differences in the incidence of selection bias evident across the different types of publication selection bias.

As discussed above, we conduct a total of 54 different experiments and thus are unable to show sample characteristics for all of them. However, it may be interesting to compare sample characteristics for the Panel Random Effects case with the same mean true effect, $\alpha = 1$. This is done in TABLE 3. Clearly, the data environment makes a difference. Publication selection bias produces larger positive biases in the PRE case, with larger t-statistics, and a higher percentage of significant effects. The (P5%, P95%) range is higher than we would like to see when sample selection targets insignificance, but the corresponding range when selection bias targets negative estimates seems "realistic." The I^2 values continue to lie in the range that Stanley and Doucouliagos (2017) identify as the norm. Finally, the samples display a wide range of publication selection incidences.

IV. RESULTS

TABLE 5 reports rejection rates associated with the Funnel Asymmetry Test (FAT) and Precision Effect Test (PET).[4] As noted above, there are six classes of experiments based on the pairing of (i) type of DGP (FE, RE, PRE), and (ii) type of publication bias (insignificance/wrong sign). The table is divided vertically into three panels according to type of DGP, from least realistic (FE) to most realistic (PRE). It is divided horizontally into left and right halves based on type of publication bias. The far left column reports the mean true effect, α. We also report the percent of estimates (from the original 1000) that survive selection bias to become included in the meta-analyst's Post-Publication Selection Bias sample ("Percent").

We start with the basic case of FE, where each study produces only one estimate and there is one population effect underlying all studies. When $\alpha = 0$ and publication bias is directed against insignificance (cf. left side of the table), the average meta-analysis sample

[4] Stata do files that allow the user to replicate all the results of TABLES 1 through 5 can be downloaded from Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F4IOLOP . R codes that allow the user to replicate all the results of TABLES 8A through 11B can be downloaded from Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Z9KYOV .

contains 144 estimates (14.4 percent). Each of the 1,000 simulated meta-analysis samples is tested for publication selection bias (FAT). As discussed above, under publication selection bias against insignificance, when $\alpha = 0$, insignificant estimates are as likely to be negative as positive. Thus eliminating insignificant estimates should not bias the estimate of the mean true effect. Accordingly, the null hypothesis ($H_0: \beta_1 = 0$) is true and the rejection rate should equal 0.05. The FPP procedure performs well in this case, producing a rejection rate of 7 percent. The PET tests whether the mean true effect is zero ($H_0: \beta_0 = 0$). Again, when $\alpha = 0$, and selection is biased against insignificant estimates, the null is true and the rejection rate should equal 0.05. The FPP procedure does not perform as well in this case. The actual rejection rate is 16 percent. To provide some structure for the large number of tests reported here and subsequently, we colour-code the cells to help identify bad performance. Henceforth, when the null hypothesis is true, and the rejection rate is greater than 15%, we indicate poor performance by shading the cell red.[5] (A full schedule of the colour-coding used here and in subsequent tables is given in TABLE 4.)

As α increases, fewer estimates lie in the insignificant range, and the percent of estimates that survive selection bias monotonically increases. For all values of $\alpha > 0$, publication selection bias disproportionately eliminates negative estimates. As a result, the null hypothesis ($H_0: \beta_1 = 0$) is false and the rejection rate should equal 100% for the FAT. The null hypothesis ($H_0: \beta_0 = 0$) is also false since there is a genuine non-zero true effect, so that rejection rate for the PET should likewise be 100%. Indeed, the FPP does very well on both tests for all values of $\alpha > 0$, displaying a power of 100%.

Continuing with publication bias against insignificance (left side of the table), we move down a panel to the more realistic case of RE. When $\alpha = 0$, the FPP produces rejection rates

[5] When printing in black and white, the greyscale equivalents are dark grey = red, medium grey = grey, and light grey = yellow.

of 9% for both the FAT and PET. These are relatively close to their significance levels (0.05). In contrast, the FAT under the FPP procedure has poor power when $\alpha > 0$. For example, when $\alpha = 0.5$, the FAT has a power of only 33%. Here and elsewhere, when rejection rates are less than 80% for false hypotheses, we shade the cells red to again indicate poor performance (cf. TABLE 4). Going down the rows of the panel, we see that that the FPP procedure generally performs poorly on the FAT when $\alpha > 0$. The FPP procedure does better on the PET. Other than the case where $\alpha = 0.5$, it rejects the false null hypothesis of no effect either 99% or 100% of the time.

The bottom panel reports results for the PRE data environment, where studies contain more than one estimate and there is heterogeneity in true effects both within and across studies. The FPP procedure performs substantially worse on both the FAT and the PET in this data environment.[6] When $\alpha = 0$ and selection bias targets statistical insignificance, the FAT rejects the (true) null of no publication bias over half the time (56%). The PET rejects the true null of no effect 31% of the time. When $\alpha > 0$, the FAT consistently displays poor power, with the PET showing poor power for values of $\alpha < 2.5$.

Moving to the right side of the table and beginning again with the top panel, we see that the FPP procedure generally performs well in the simplistic, FE data environment. When publication bias is directed against negative estimates and $\alpha = 0$, so that approximately half of all estimates are wrong signed, the FAT has a power of 100%. The FAT continues to show excellent power as α increases, up to the point where $\alpha = 3.5$. Here the power of the FAT falls to 52%, falling further to 29% when $\alpha = 4.0$. It should be noted, however, that for these values of α, the incidence of selection bias is very small, with 97% and 98% of the original estimates surviving to the Post-Publication Selection Bias sample. Whenever power is less than 80%, but

[6] Heteroskedasticity-robust standard errors were used when testing hypotheses in the FE and RE cases. Clustered robust standard errors were used in the PRE case.

less than 10% of the estimates have been eliminated due to selection bias, we shade the cells grey (cf. TABLE 4). The FPP procedure also performs well on the PET. When $\alpha = 0$, it rejects the null of no effect 9% of the time. When $\alpha > 0$, it rejects the null of no effect 100% of the time, as it should.

As it did in the case of selection bias against insignificant estimates, the performance of the FPP procedure declines as the simulated data environments become more realistic. In the RE data environment, when $\alpha = 0$ the FAT shows poor power, rejecting the null hypothesis of no selection bias 62% of the time (less than the expected 100%), despite the fact that 45% (=1-0.55) of the Pre-Publication Selection Bias sample is filtered out of the meta-analyst's sample. The power of the FAT drops further as α increases, though this is to be expected as selection bias eliminates fewer estimates. When $\alpha = 2.5$, power falls to 21%, but less than 10% of estimates are eliminated from selection bias. The poor performance of the FPP procedure on the FAT is indicated by red cells for α values 0 through 2.0, and by grey cells for α values 2.5 through 4.0 (since less than 10% of the estimates have been censored due to publication selection bias). For positive values of α in the RE data environment, the FPP procedure continues to perform well on the PET. However, it falsely rejects the true null of no effect 89% of the time, earning the respective cell a red colouring.

Moving to the PRE data environment causes FPP's performance on the FAT to decline yet further, with power consistently less than 50%. Its performance on the PET also shows a general deterioration, with rejection rates falling below 100%, though still above the 80% threshold for "poor performance."

In summary, we see a consistent pattern of declining performance of the FPP procedure on the FAT and PET as we move from the unrealistic, simplistic environment of FE, to the more realistic RE and PRE data environments. This is represented by the increasing prevalence of red cells as one moves from the top panel of TABLE 5 down through the bottom panel. It is

true for both selection bias against statistical insignificance, and selection bias against negative estimates.

For many if not most meta-analyses, the FAT and PET are preliminary to the main issue, which is an estimate of the size of the overall effect. While it is interesting to know whether a literature is affected by publication bias, and whether the estimate of the effect is statistically significant, a primary goal of meta-analysis is to aggregate the estimates in the literature and arrive at an estimate of the mean true effect. In the context of our experiments, that means using $\hat{\beta}_0$ to estimate α.

Accordingly, we continue our analysis of FPP performance by comparing the FPP procedure with two WLS estimators, WLS-FE and WLS-RE, which do not attempt to control for publication bias by including a *SE* variable. The three alternatives are compared with respect to their mean values of $\hat{\beta}_0$, their mean squared error (MSE), and the results of hypothesis tests which test the null that $\beta_0 = \alpha$. TABLE 6 reports the results of this analysis. We conserve space by only reporting results for the PRE case, however we include results for both types of publication selection bias.

When $\alpha = 0$ and selection is biased against insignificance, the FPP procedure's average estimate of $\hat{\beta}_0$ over the 1000 simulated meta-analysis studies is 0.06. This compares with average $\hat{\beta}_0$ values of 0.04 and -0.01 for WLS-FE and WLS-RE, respectively. As the WLS-RE estimate is closest to the mean true value of α, we colour this cell yellow (cf. TABLE 4). For other values of α, the FPP procedure is consistently less biased. While the FPP procedure often overestimates the mean true value of α, other than the one case ($\alpha = 0$, selection bias against insignificance), it always produces an estimate of α that is less biased than either WLS-FE or WLS-RE.

When selection bias targets negative estimates (the right hand side of TABLE 6), FPP's estimates of α are less biased than WLS-FE and WLS-RE across the full range of α values,

despite suffering itself from substantial bias. For example, when $\alpha = 0$, the FPP procedure calculates an average value of 1.69 for $\hat{\beta}_0$. While the bias is substantial, it is less than that associated with the WLS-FE and WLS-RE estimates (1.77 and 1.88, respectively).

Interestingly, superior performance on the first moment of the distribution does not translate into greater efficiency. When $\alpha = 0$ and publication bias is targeted against insignificance, the MSE for the FPP estimates is 1.688. This compares to 0.928 and 0.448 for the two WLS estimators. For both types of publication bias, and for every value of α, the FPP estimates are less efficient than the WLS-FE estimates. They are also less efficient than the WLS-RE estimates, with one lone exception ($\alpha = 2.5$). Nor is this situation unique. When selection is biased against estimates with wrong/negative signs, once again FPP is dominated on efficiency by both WLS-FE and WLS-RE. For some values of α (such as $\alpha = 0$), WLS-FE is most efficient. For other values of α (such as $\alpha = 1.5$), WLS-RE is most efficient. But for every value of α, WLS-FE and WLS-RE are each more efficient than the FPP procedure.

The reason for why FPP can be least biased but also least efficient is due to the substantial differences in their variances. This is illustrated in FIGURE 3, which produces a kernel plot of $\hat{\beta}_0$ estimates for each of the three estimating procedures for the experiment PRE, $\alpha = 1$, and selection bias against negative estimates. The FPP procedure produces the least biased estimates, but also the estimates with greatest variance. WLS-RE is most biased, but has smallest variance. In the middle is WLS-FE, whose combination of moderate bias and moderate variance makes it most efficient.

Finally, as was foreshadowed by the PET results in TABLE 5 when $\alpha = 0$, the FPP procedure does not do well when testing hypotheses about α. Sometimes it performs better than the WLS estimators, and sometimes worse. But as is clear from the large swathes of red in the bottom panel of TABLE 6, the rejection rates for the true null hypothesis, $H_0: \beta_0 = \alpha$,

are sufficiently large that hypothesis testing about α should generally not be relied upon for any of these approaches.

Before proceeding it is good to summarize the main results above. We find that the FPP procedure works well in the "Fixed Effects" environment. With one exception, its Type I and Type II error rates on the FAT and PET are acceptable to excellent. However, it performs progressively worse as the data environment is generalized to the "Random Effects" and "Panel Random Effects" environments. Type II error rates for the FAT are particularly poor in the RE data environment. And the poor performance of the FPP procedure on the FAT is accompanied by declining performance for the PET in the PRE data environment. When we compare the performance of FPP with WLS-FE and WLS-RE, we find that the FPP procedure generally produces the least biased estimates of the mean true effect. However, it is consistently less efficient than either WLS-FE or WLS-RE. Further, hypothesis testing about α is sufficiently poor across all values of α and for both types of selection bias that the results are rendered useless in most instances.

V. FURTHER ANALYSIS USING THE STANLEY AND DOUCOULIAGOS (2017) FRAMEWORK

One concern with the preceding analysis is that the simulation framework differs significantly from the simulation framework employed by Stanley and Doucouliagos in a series of papers that investigate the properties of alternative meta-analysis estimators and procedures (Stanley, 2008; Stanley and Doucouliagos, 2014; Stanley and Doucouliagos, 2015; Stanley and Doucouliagos, 2017). This is to some extent unavoidable because the S&D framework does not readily lend itself to scenarios where studies have multiple estimates. It does, however, raise concerns that our finding of poor FPP performance may be due to idiosyncrasies of our simulation framework. Accordingly, in this section, we adopt the framework used in Stanley and Doucouliagos (2017) and use it to apply the same tests that we performed above.

S&D are primarily interested in testing the performance of an estimator they call the "weighted least squares meta-regression analysis" estimator, or WLS-MRA. This estimator differs from the standard fixed effects estimator because its standard error is unrestricted, whereas the variance in the standard fixed effects estimator is standardized. In fact, S&D's WLS-MRA estimator is the identical to the WLS estimator with weights $\omega_j = \left(\frac{1}{SE_j}\right)$ that we used above, in Section IV, and called WLS-FE.

Stanley and Doucouliagos are interested in testing this estimator in a data environment where there is heterogeneity in true effects across studies. They model this heterogeneity three ways, which they call "Indirect" heterogeneity, "Direct" heterogeneity, and "Random Mean" heterogeneity. Similar to our FE and RE data environments, they assume that each study reports one regression.

To ensure that we are correctly recreating S&D's simulation framework, we proceed to replicate their findings. While we were unable to obtain the code that Stanley and Doucouliagos used in their simulations, their associated description is very clear. Based on that description, we reproduced the tables that they present in their paper.[7] As these are simulation exercises, our results will not be able to exactly reproduce the numbers reported in their tables. However, in every case we are able to closely replicate their results. Our replications are provided in APPENDIX 2 of our paper and easily compared with the original tables in S&D.

The analysis we present in this section focuses on their "Indirect" and "Direct" models of effect heterogeneity. A summary of the corresponding simulation frameworks is provided in TABLE 7 and we briefly describe them here: The meta-analyst is interested in summarizing the literature that estimates the effect a variable X_1 has on an outcome variable Z. In the "Indirect" model, the DGP for generating observations i for a given primary study is given by

[7] Specifically, we reproduced Tables 1-6, and 9-10. We were unable to reproduce their Tables 7 and 8 for "Random Mean" heterogeneity. As a result, our experiments focus on their "Indirect" and "Direct" heterogeneity models.

(5.a) $Z_i = 100 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \varepsilon_i$, $\varepsilon_i \sim N(0, 100^2)$,

where

(5.b) $X_{1i} \sim U(100, 200)$; $X_{2i} = X_{1i} + N(0, 50^2)$; $X_{3i} = X_{1i} + N(0, 50^2)$

and

(5.c) $\alpha_1 = \{0, 1\}$; $\alpha_2 = 0.5$; $\alpha_3 \sim N(0, \sigma_h^2)$; $\sigma_h = \{0, 0.125, 0.25, 0.5, 1, 2, 4\}$.

The meta-analyst is interesting in aggregating the estimates of α_1 across primary studies.

Note that the variables X_2 and X_3 are constructed to be correlated with X_1. Some of the primary studies include X_2, while others do not. In the latter case, this generates omitted variable bias. None of the primary studies control for X_3. As a result, if $\sigma_h \neq 0$, estimates of α_1 in individual studies will be biased, some positively biased, some negatively biased, with the bias depending on the particular value of α_3. S&D construct α_3 such that it is randomly distributed across studies, having a mean of 0 and a variance of σ_h^2, $\alpha_3 \sim N(0, \sigma_h^2)$. [8]As a result, the biases in individual primary studies should cancel out in the aggregate, so that the mean true effect of X_1 on Z will be α_1. To induce differences in coefficient standard errors, S&D allow sample sizes to random vary across studies, $n_j \in \{62, 125, 250, 500, 1000\}$.

S&D conduct a variety of experiments. The experiments differ first on MRA sample size, with meta-analysis studies assumed to consist of 20 primary studies in some experiments, and 80 primary studies in others. Next they differ on the degree of "excess heterogeneity," given by the parameter σ_h, $\sigma_h \in \{0, 0.125, 0.25, 0.5, 1, 2, 4\}$. The larger σ_h, the greater the degree of effect heterogeneity across studies. Experiments also differ on the value of the mean true effect, which takes the value 0 in some experiments, and 1 in others. Finally, some of the experiments build in publication selection bias where estimates that are positive and statistically significant are disproportionately represented in the meta-analyst's sample. In

[8] $\sigma_h = 0$ ($\sigma_h \neq 0$) is analagous to our FE (RE) data environment.

particular, only positive and significant estimates of α_1 are selected for half of the meta-analyst's study. The remaining estimates are unfiltered. S&D call this set-up "50% Publication Selection Bias."

The "Direct" model is very similar, with the essential difference being that effect heterogeneity is directly built into the slope coefficient for X_1, so that the DGP for a given primary study is given by

(5.a') $Z_i = 100 + (\alpha_1+\alpha_3)X_{1i} + \alpha_2 X_{2i} + \varepsilon_i,\ \varepsilon_i \sim N(0, 100^2)$.

As X_3 no longer serves as the channel by which effect heterogeneity enters the DGP, it is dropped from the model. The distributions of X_1 and X_2 are identical to what they were under the "Indirect" model:

(5.b') $X_{1i} \sim U(100, 200)$; $X_{2i} = X_{1i} + N(0, 50^2)$;

as are the other model parameters:

(5.c') $\alpha_1 = \{0, 1\}$; $\alpha_2 = 0.5$; $\alpha_3 \sim N(0, \sigma_h^2)$; $\sigma_h = \{0, 0.125, 0.25, 0.5, 1, 2, 4\}$.

Note that the mean true effect of X_1 on Z remains α_1 under the "Direct" model.

In the analysis that follows, we repeat the analysis of TABLES 5 and 6 using S&D's simulation framework as defined above. We follow S&D in simulating 10,000 meta-analysis studies for analysis. The experiments are ordered first by size of mean true effect size (0, then 1). Within each effect size, experiments are next ordered by MRA study size (first 20, then 80 studies); and within each MRA study size, in order of increasing effect heterogeneity, as measured by σ_h. To help interpret these, we also report the corresponding empirical I^2 measure. Note that σ_h values larger than 0.5 correspond to I^2 values greater than 80%; and σ_h values larger than 1.0 to I^2 values greater than 90%. As this study is concerned with the performance of the FPP procedure in the presence of publication selection bias, we focus on the experiments with "50% Publication Selection Bias."

TABLES 8A and 8B report the results of using the first step of the FPP procedure (as represented in FIGURE 2) to perform FAT and PET tests for the "Indirect" and "Direct" models, respectively.[9] Since every meta-analysis study in every experiment is characterized by publication selection bias, the rejection rates for the FAT in TABLES 8A and 8B should be 100%. Rarely are they even 50%. For example, in TABLE 8A, when the mean true effect size is 0, MRA sample size is 20, excess heterogeneity is 0 (I^2 = 7.2%) , we find that the average rejection rate for the FAT is 16.8%. When excess heterogeneity increases to 0.125 (I^2 = 27.0%), the rejection rate for the FAT falls to 11.7%. As all the rejection rates for the FAT in both tables are less than 0.80, we colour-code the cells red to indicate poor performance (cf. TABLE 4). The results for TABLE 8B are similar, with all cells being shaded red for poor performance.

The results for the PET are somewhat better, but still not good. When the mean true effect is 0, rejection rates for the PET should be 0.05. In the 14 corresponding experiments in TABLE 8A (cf. top two panels), 3 have rejection rates above 15%, earning a red shading. In TABLE 8B, 8 of the 14 have rejection rates above 15%. Turning to the bottom two panels of TABLES 8A and 8B, when the mean true effect is 1, rejection rates for the PET should be 1.00. Of the 14 corresponding experiments in TABLE 8A, 5 have rejection rates below 80%. In TABLE 8B, 6 out of 14 experiments have rejection rates below 80%. There also appears to be an inverse relationship between excess heterogeneity (I^2) and rejection rates when the mean true effect equals 1, with the PET losing power as heterogeneity increases. Summarizing the PET results, across both tables and for both values of the mean true effect (0 and 1), FPP tends to perform poorly, particularly for I^2 values above 90%, which falls in the range of

[9] S&D do not use robust standard errors for their hypothesis testing. Accordingly, neither do we when we use their simulation frameworks.

heterogeneity values that S&D identify as "the norm" for meta-analyses in economics and business.

Interestingly, these results are qualitatively similar to the results we report in TABLE 5 using our Panel Random Effects (PRE) framework (which is also characterized by high I^2 values; cf. TABLE 3). Under PRE, FPP performed very poorly on the FAT both when publication selection bias targeted statistical insignificance, and when it targeted estimates that were wrong-signed/negative. FPP's performance on the PET was better, but still bad given selection bias against insignificance (though better when selection was biased against negative estimates). Note that S&D combine the two types of publication selection bias, preferring estimates that are both significant and positive. Our results suggest that when FPP does poorly in the S&D framework, this may be due primarily to selection bias against insignificance.

TABLES 9A, 10A, and 11A continue our evaluation of the FPP procedure. All three tables assume "Indirect" heterogeneity, and correspond to the three panels of TABLE 6 – top, middle, and bottom, which respectively report (i) the mean value of β_0, (ii) MSE, and (iii) Type I error rates corresponding to testing $H_0: \beta_0 = \alpha$. TABLES 9B, 10B, and 11B do the same for "Direct" heterogeneity. As before, we yellow-code cells to indicate which of the three estimation procedures – FPP, WLS-FE, and WLS-RE – has smallest bias (TABLES 9A and 9B) or best efficiency (TABLES 10A and 10B), and red-code cells for which the Type I error rate is greater than 15%.

FPP almost always performs best in terms of smallest bias. For example, in TABLE 9A, for meta-analyses consisting of 20 primary studies, with 0.125 excess heterogeneity and a mean true effect of 0, the mean estimate of β_0 is -0.0417 for the FPP procedure, versus 0.0620 and 0.0633 for WLS-FE and WLS-RE. As excess heterogeneity increases to 0.25, the bias in the FPP procedure increases, but the bias associated with WLS-FE and WLS-RE increases even more. In fact, across all experiments with "Indirect" heterogeneity, FPP is least biased

with only two exceptions: the two experiments where the mean true effect is 0 and there is no excess heterogeneity.

The results for "Direct" heterogeneity are qualitatively identical (cf. TABLE 9B). As a result, we find that the FPP procedure generally produces estimates of the mean true effect that are least biased compared to WLS-FE and WLS-RE. Interestingly, these results are also very similar to the results we report in the top panel of TABLE 6 using our Panel Random Effects (PRE) framework.

When we evaluate FPP's performance on mean squared error (MSE), the results again generally confirm the findings from our earlier simulation framework (cf. the middle panel of TABLE 6). In 41 of the 56 experiments reported in TABLES 10A and 10B, either WLS-FE or WLS-RE displays better efficiency than FPP in estimating the mean true effect than FPP. In 37 experiments, the FPP procedure is least efficient.

The main difference that we observe in FPP's performance across the two simulation frameworks occurs when testing hypotheses about the mean true effect. A comparison of TABLES 11A and 11B with the bottom panel of TABLE 6 shows that FPP performs much better in the S&D simulation environment. In our framework, tests of the hypothesis $H_0: \beta_0 = \alpha$ always produced rejection rates greater than 15%, which is our threshold for demarcating good versus poor performance. In contrast, corresponding Type I error rates in S&D's simulation environment were larger than 15% in only 16 out of the 56 experiments. This difference, however, is not particularly surprising. The PRE simulation environment in our framework has a very different error structure than S&D, with mean true effects having compound heterogeneity, first across studies, and then again within studies. As a result, it would not be surprising if the FPP standard errors in our framework were more severely underestimated, causing higher rejection rates.

To summarize, when we repeat our analysis of the FPP procedure using the simulation framework employed by S&D, we identify a number of similar deficiencies in FPP's performance. FPP performs poorly on the FAT when there is publication selection bias, does poorly on the PET when true effects have substantial heterogeneity, and is generally less efficient than the two WLS estimators that do not include a *SE* term. The main difference in results compared to our framework is that hypothesis testing about the mean true effect is more reliable in the S&D framework, but even here the FPP procedure struggles at times.

VI. CONCLUSION

This paper studies the performance of the FAT-PET-PEESE (FPP) procedure, a commonly employed procedure for addressing publication bias in economics and business meta-analyses. The FPP procedure is generally used for three purposes: (i) to test whether a sample of estimates suffers from publication bias, (ii) to test whether the estimates indicate that the effect of interest is statistically different from zero, and (iii) to obtain an estimate of the overall, mean effect.

We investigate the performance of the FPP procedure using a simulation framework that progressively generalizes the data environment from the simplistic case where there is one underlying true effect across all studies, and each study reports only one regression ("Fixed Effects"); to the more general case of heterogeneous true effects across studies, where each study continues to report only one regression ("Random Effects"); to the most general case where studies report multiple regressions and true effects differ both across across and within studies ("Panel Random Effects").

Our findings indicate that the FPP procedure performs well in the basic but unrealistic environment of "Fixed Effects," where all estimates are assumed to derive from a single, population value and sampling error is the only reason for why studies produce different estimates. However, when we study its performance in more realistic data environments, where there is heterogeneity in population effects across and within studies, the FPP procedure

becomes unreliable for the first two purposes, and less efficient than some other estimators that do not correct for publication bias. Further, hypothesis tests about the overall mean effect often cannot be trusted.

We then attempt to corroborate our findings by replicating the simulation framework used in Stanley and Doucouliagos (2017). We first reproduce the findings reported by S&D to demonstrate that we can recreate their data environments. We then use their "Indirect" heterogeneity and "Direct" heterogeneity simulation frameworks to repeat our analysis of the FPP procedure. We generally confirm our main findings with one exception: In the S&D data environments, the FPP procedure performs better in testing hypotheses about the overall mean. However, this is not surprising given that our "Panel Random Effects" data environment has a very different error structure than S&D, making hypothesis testing more challenging.

There are two main conclusions we draw from our research. The first is that meta-analyses should routinely report measures of heterogeneity such as I^2. This is not standard practice in the economics and business literatures and should be because the FPP procedure does not perform well when there is substantial effect heterogeneity. The second conclusion is that we still do not have a reliable method for detecting and correcting publication selection bias in realistic research environments. Publication selection bias is perhaps the greatest obstacle hindering the ability of meta-analyses to uncover population effect values. This study demonstrates that much more remains to be done to address this problem.

REFERENCES

Costa-Font, J., Gemmill, M., and Rubert, G. (2011). Biases in the healthcare luxury good hypothesis?: A meta-regression analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(1), 95-107.

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

Doucouliagos, H., and Paldam, M. (2013). The robust result in meta-analysis of aid effectiveness: A response to Mekasha and Tarp. *The Journal of Development Studies*, *49*(4), 584-587.

Doucouliagos, H., Stanley, T. D., and Viscusi, W. K. (2014). Publication selection and the income elasticity of the value of a statistical life. *Journal of Health Economics*, *33*, 67-75.

Efendic, A., Pugh, G., and Adnett, N. (2011). Institutions and economic performance: A meta-regression analysis. *European Journal of Political Economy*, *27*(3), 586-599.

Haelermans, C., and Borghans, L. (2012). Wage effects of on-the-job training: A meta-analysis. *British Journal of Industrial Relations*, *50*(3), 502-528.

Havránek, T. (2010). Rose effect and the euro: is the magic gone? *Review of World Economics*, *146*, 241-261.

Higgins, J.P. and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-58.

Iwasaki, I., and Tokunaga, M. (2014). Macroeconomic impacts of FDI in transition economies: a meta-analysis. *World Development*, *61*, 53-69.

Laroche, P. (2016). A meta-analysis of the union–job satisfaction relationship. *British Journal of Industrial Relations*.

Lau, J., Ioannidis J.P.A., Terrin, N., Schmid, C.H., and Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal, 333*, 597-600.

Lazzaroni, S., and van Bergeijk, P. A. (2014). Natural disasters' impact, factors of resilience and development: A meta-analysis of the macroeconomic literature. *Ecological Economics*, *107*, 333-346.

Linde Leonard, M., Stanley, T. D., and Doucouliagos, H. (2014). Does the UK minimum wage reduce employment? A meta-regression analysis. *British Journal of Industrial Relations*, *52*(3), 499-520.

Nelson, J. P. (2013). Meta-analysis of alcohol price and income elasticities – with corrections for publication bias. *Health Economics Review*, *3:17.*

Reed, W. R. (2015). A Monte Carlo analysis of alternative meta-analysis estimators in the presence of publication bias. *Economics: The Open-Access, Open-Assessment E-Journal*, 9 (2015-30): 1—40. http://dx.doi.org/10.5018/economics-ejournal.ja.2015-30

Reed, W. R., Florax, R. J. G. M., and Poot, J. (2015). A Monte Carlo analysis of alternative meta-analysis estimators in the presence of publication bias. Economics Discussion Papers, No 2015-9, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2015-9

Stanley, T.D. (2008). 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.

Stanley, T.D., and Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. London: Routledge.

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

Stanley, T.D., and Doucouliagos, H. (2015). Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine, 34*(13): 2116-27.

Stanley, T.D., and Doucouliagos, H. (2017). Neither fixed nor random: weighted least squares meta-regression. *Research Synthesis Methods, 8*(1), 19-42.

Sterne, J.A.C., Sutton, A.J., Ioannidis, J.P.A., Terrin, N., Jones, D.R., Lau, J., Carpenter, J., Rucker, G., Harbord, R.M., Schmid, C.H., Tetzlaff, J., Deeks, J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D.G., Moher, D., and Higgins, J.P. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal*, *343*, d4002.

Terrin, N., Schmid, C.H., Lau, J. and Olkin, I., (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*, 2113-2126.

TABLE 1
General Framework for Simulating Meta-Analysis Studies

STAGE 0: Alternative "data environments"		
Three types of "data environments"	Fixed Effects (FE) Random Effects (RE) Panel Random Effects (PRE)	Our general framework is designed to simulate a meta-analysis study where the meta-analyst is interested in studying the effect that a variable *x* has on an outcome *y*. The framework models three types of "data environments." The simplest data environment is where all studies only report one regression and there is one true effect of *x* on *y* underlying all studies (Fixed Effects). The next data environment still assumes that studies report only one regression, but allows for heterogeneous true effects across studies (Random Effects). The last case is where studies report multiple regression results, and each regression has its own, unique, underlying true effect (Panel Random Effects). The subsequent discussion focuses on the PRE case, but is easily modified for the FE and RE cases. The ultimate parameter of interest is α, the mean true effect in the population.
STAGE 1: Generation of individual observations		
Data-generating process of individual observations for primary studies	$y_{irt} = 1 + \alpha_{ir}x_{irt} + \varepsilon_{irt}$	α_{ir} is the true effect of *x* on *y* in regression *r* of study *i*. The true effect varies for different regressions both within and across studies. Likewise, error terms ε_{irt} are constructed to have the same variance for observations belonging to the same regression (homoskedasticity), but different regressions have different error variances, which produce estimated coefficients with different standard errors. Both true effects and error variances/coefficient standard errors are generated so that they are more similar within than across studies.
Construction of samples for primary studies	(y_{irt}, x_{irt}), *t=1,2,...100*	100 observations are created for each regression *r* belonging to each primary study *i*.

STAGE 2: Estimation of effect sizes and standard errors in primary studies		
Individual regressions from primary studies	Estimate: $y_{irt} = \psi_{ir} + \alpha_{ir}x_{irt} + \epsilon_{irt}$	Each regression *r* of primary study *i* uses OLS to estimate the effect size, α_{ir}, producing the pair $\{\hat{\alpha}_{ir}, s.e.(\hat{\alpha}_{ir})\}$
STAGE 3: The Pre-Publication Selection Bias Sample		
The Pre-Publication Selection Bias Sample	$(\{\hat{\alpha}_1, SE_1\}, \ldots, \{\hat{\alpha}_{1000}, SE_{1000}\})$	The data generation and estimation process above is repeated until it produces 1000 estimated effects, $\hat{\alpha}_j$, with corresponding standard errors, SE_j. However, this "Pre-Publication Bias Sample" is not observed by the meta-analyst.
STAGE 4: The Post-Publication Selection Bias Sample		
The Post-Publication Selection Bias Sample	$(\{\hat{\alpha}_1, SE_1\}, \ldots, \{\hat{\alpha}_M, SE_M\})$ $M < 1000$	Publication selection bias filters out some of the estimated effects, so that the sample of *M* regression results collected by the meta-analyst is less than the full 1000. We separately allow two types of publication selection bias: (i) selection bias against estimates that are statistically insignificant, and (ii) selection bias against negative estimates, as we are modelling a situation where conventional belief assumes that the true effect is positive.
STAGE 5: The Meta-Analysis		
Tests and estimates using the Post-Publication Selection Bias Sample	(A) $\hat{\alpha}_j = \beta_0 + \beta_1 SE_j + e_j$ (B) $\hat{\alpha}_j = \beta_0 + e_j$	Following the FPP procedure in Figure 2, Model (A) is estimated using the WLS-FE estimator to conduct the FAT and PET. The FPP procedure continues by using WLS-FE to obtain a final estimate for β_0. For the sake of comparison, WLS-FE and WLS-RE estimate β_0 using Model (B), which does not correct for publication bias. Heteroskedasticity robust standard errors are used for the FE and RE data environments, and cluster robust standard errors are used for the PRE data environment.

STAGE 6: Iteration		
STAGES 1-5 are repeated 1,000 times	Results from individual meta-analysis: FAT, PET and estimates of β_0	The results from each of the 1,000, simulated meta-analysis studies are collected and analysed to determine how well the FPP procedure is able to (i) identify publication bias, (ii) identify a non-zero true effect, and (iii) produce a reliable estimate of α.

TABLE 2
Sample Characteristics for a Simulated Meta-Analysis Data Set: Random Effects ($\alpha = 1$)

Variable	*Median*	*Minimum*	*P5%*	*P95%*	*Maximum*
PRE-PUBLICATION SELECTION BIAS SAMPLE (100 percent of estimates):					
Estimated effect	1.00	-7.47	-2.38	4.39	9.46
t-statistic	0.79	-13.19	-1.47	5.90	42.19
Percent significant	0.26	0.22	0.24	0.28	0.30
I-squared	0.86	0.72	0.81	0.90	0.94
SAMPLE AFTER SELECTION AGAINST INSIGNIFICANCE (33.0 percent of estimates):					
Estimated effect	1.81	-7.43	-2.07	5.69	9.54
t-statistic	2.54	-13.21	-2.35	12.63	42.24
Percent significant	0.93	0.89	0.91	0.94	0.95
I-squared	0.94	0.86	0.92	0.96	0.98
SAMPLE AFTER SELECTION AGAINST NEGATIVE ESTIMATES (74.6 percent of estimates):					
Estimated effect	1.55	-5.01	0.05	4.77	9.52
t-statistic	1.28	-5.14	0.04	7.33	42.05
Percent significant	0.49	0.44	0.46	0.51	0.53
I-squared	0.81	0.65	0.73	0.88	0.91

NOTE: Values in the table are constructed by simulating 1000 meta-analysis studies given the respective conditions, and then averaging the results on the respective dimensions (e.g. median value, 5% quantile value, etc.). "Percent significant" identifies the average percent of estimates that are significant at the 5 percent level. "*I-squared*" measures the extent of effect heterogeneity ((Higgins and Thompson, 2002).

TABLE 3
Sample Characteristics for a Simulated Meta-Analysis Data Set: Panel Random Effects ($\alpha = 1$)

Variable	*Median*	*Minimum*	*P5%*	*P95%*	*Maximum*
PRE-PUBLICATION BIAS (100 percent of estimates):					
Estimated effect	0.99	-8.95	-3.51	5.51	10.89
t-statistic	0.68	-17.76	-2.90	7.05	33.43
Percent significant	0.34	0.23	0.29	0.40	0.45
I-squared	0.91	0.73	0.83	0.97	0.99
SAMPLE AFTER SELECTION AGAINST INSIGNIFICANCE (21.9 percent of estimates):					
Estimated effect	2.40	-5.34	-3.08	6.02	8.88
t-statistic	3.68	-17.57	-7.84	16.90	33.42
Percent significant	0.98	0.95	0.97	0.99	1.00
I-squared	0.97	0.87	0.94	0.99	1.00
SAMPLE AFTER SELECTION AGAINST NEGATIVE ESTIMATES (80.5 percent of estimates):					
Estimated effect	2.23	-5.36	-0.84	6.21	10.85
t-statistic	1.72	-2.93	-0.50	10.15	33.42
Percent significant	0.68	0.57	0.62	0.74	0.80
I-squared	0.83	0.51	0.69	0.94	0.98

NOTE: Values in the table are constructed by simulating 1000 meta-analysis studies given the respective conditions, and then averaging the results on the respective dimensions (e.g. median value, 5% quantile value, etc.). "Percent significant" identifies the average percent of estimates that are significant at the 5 percent level. "*I-squared*" measures the extent of effect heterogeneity ((Higgins and Thompson, 2002)

TABLE 4
Colour Coding for Tables 5, 6, and 8A through 11B

Colour	Performance Measure	Expected Value	Colour indicates
Red	FAT	0.05	Rejection rate > 0.15
Red	FAT	1.00	Rejection rate < 0.80 and more than 10% of the estimates have been censored due to publication selection bias.
Grey	FAT	1.00	Rejection rate < 0.80 and less than 10% of the estimates have been censored due to publication selection bias.
Red	PET	0.05	Rejection rate > 0.15
Red	PET	1.00	Rejection rate < 0.80
Yellow	Mean value of $\hat{\beta}_0$	α	Mean value of $\hat{\beta}_0$ is closest to α for this estimator
Yellow	MSE	Smallest value	Estimator has smallest MSE value
Yellow	Type I error rate	0.05	Rejection rates > 0.15

NOTE: The schedule above identifies the conditions that determine the colour coding of individual cells in TABLES 5, 6, and 8A – 11B. The purpose of colouring the cells is to facilitate comparison of results across tables. When printing in Black and White, the greyscale equivalents are Red - Dark Grey, Grey - Medium Grey, and Yellow - Light Grey.

TABLE 5
Funnel Asymmetry Tests (FAT) and Precision Effect Tests (PET)

	Publication Bias against Insignificance			Publication Bias against Wrong Sign		
FIXED EFFECTS (FE)						
α	***Percent***	***FAT***	***PET***	***Percent***	***FAT***	***PET***
0.0	14.4	0.07	0.16	55.0	1.00	0.09
0.5	23.1	1.00	1.00	71.7	1.00	1.00
1.0	31.9	1.00	1.00	80.6	1.00	1.00
1.5	40.0	1.00	1.00	86.5	1.00	1.00
2.0	47.6	1.00	1.00	90.6	1.00	1.00
2.5	54.6	1.00	1.00	93.5	0.98	1.00
3.0	61.1	1.00	1.00	95.5	0.81	1.00
3.5	66.9	1.00	1.00	97.0	0.52	1.00
4.0	72.2	1.00	1.00	98.0	0.29	1.00
RANDOM EFFECTS (RE)						
α	***Percent***	***FAT***	***PET***	***Percent***	***FAT***	***PET***
0.0	27.1	0.09	0.09	55.0	0.62	0.89
0.5	28.8	0.33	0.64	65.4	0.61	1.00
1.0	33.1	0.68	0.99	74.6	0.59	1.00
1.5	39.1	0.80	1.00	82.0	0.47	1.00
2.0	46.0	0.78	1.00	87.4	0.36	1.00
2.5	52.7	0.77	1.00	91.3	0.21	1.00
3.0	59.1	0.67	1.00	94.0	0.18	1.00
3.5	65.1	0.62	1.00	95.8	0.13	1.00
4.0	70.4	0.53	1.00	97.2	0.09	1.00
PANEL RANDOM EFFECTS (PRE)						
α	***Percent***	***FAT***	***PET***	***Percent***	***FAT***	***PET***
0.0	19.3	0.56	0.31	38.6	0.46	0.77
0.5	19.9	0.61	0.34	47.8	0.47	0.85
1.0	21.9	0.61	0.49	56.9	0.42	0.89
1.5	25.5	0.72	0.62	65.7	0.45	0.91
2.0	29.5	0.65	0.70	73.7	0.47	0.92
2.5	34.4	0.61	0.84	80.7	0.49	0.96
3.0	39.9	0.63	0.89	86.1	0.45	0.97
3.5	46.6	0.67	0.95	90.9	0.42	0.98
4.0	52.6	0.65	0.97	93.9	0.39	0.99

NOTE: α is the mean true effect in the simulations underlying a given experiment (see TABLE 1). "Percent" represents the percentage of estimates (out of the original 1000) that survive publication selection bias and are available to the meta-analyst for study. The values in the FAT and PET columns represent the rejection rates for the respective null hypotheses ($\beta_1 = 0$ and $\beta_0 = 0$, respectively, in Equation (1) in the text). Rejection rates are expected to be 0.05 for (i) the FAT when $\alpha = 0$ and publication selection is biased against insignificant estimates; and (ii) the PET when $\alpha = 0$ under both types of publication selection bias. Everywhere else, rejection rates are expected to be 1.00. Coloured cells indicate that the associated rejection rates represent "poor performance" (see TABLE 4 for a more detailed discussion).

TABLE 6
Comparison of FPP Estimates with WLS-FE and WLS-RE: Panel Random Effects

	Publication Bias against Insignificance			Publication Bias against Wrong Sign		
MEAN VALUE OF $\widehat{\beta}_0$						
α	*FPP*	*WLS-FE*	*WLS-RE*	*FPP*	*WLS-FE*	*WLS-RE*
0.0	0.06	0.04	- 0.01	1.69	1.77	1.88
0.5	0.58	0.67	1.04	1.93	1.97	2.08
1.0	1.25	1.36	1.92	2.13	2.17	2.29
1.5	1.71	1.84	2.58	2.36	2.40	2.53
2.0	2.17	2.34	3.13	2.62	2.67	2.80
2.5	2.70	2.83	3.58	2.93	2.96	3.09
3.0	3.18	3.32	4.02	3.30	3.33	3.44
3.5	3.69	3.78	4.40	3.72	3.75	3.83
4.0	4.07	4.17	4.77	4.09	4.12	4.21
MEAN SQUARED ERROR						
α	*FPP*	*WLS-FE*	*WLS-RE*	*FPP*	*WLS-FE*	*WLS-RE*
0.0	1.688	0.928	0.448	3.664	3.468	3.590
0.5	1.629	0.857	0.663	2.775	2.478	2.547
1.0	1.594	0.908	1.145	1.992	1.670	1.710
1.5	1.599	0.903	1.367	1.504	1.126	1.098
2.0	1.465	0.794	1.431	1.266	0.787	0.682
2.5	1.194	0.670	1.284	0.950	0.536	0.396
3.0	1.245	0.626	1.145	0.906	0.435	0.236
3.5	1.143	0.582	0.884	1.007	0.454	0.155
4.0	1.015	0.472	0.659	0.884	0.380	0.085
TYPE I ERROR RATES (H_0: $\beta_0 = \alpha$)						
α	*FPP*	*WLS-FE*	*WLS-RE*	*FPP*	*WLS-FE*	*WLS-RE*
0.0	0.27	0.19	0.06	0.77	0.99	1.00
0.5	0.31	0.18	0.14	0.79	0.95	1.00
1.0	0.33	0.22	0.39	0.62	0.77	1.00
1.5	0.32	0.25	0.63	0.47	0.58	1.00
2.0	0.31	0.22	0.78	0.40	0.39	0.98
2.5	0.27	0.21	0.87	0.29	0.23	0.85
3.0	0.26	0.21	0.91	0.26	0.18	0.57
3.5	0.30	0.21	0.89	0.30	0.19	0.37
4.0	0.26	0.16	0.84	0.26	0.15	0.16

NOTE: α is the mean true effect in the simulations underlying a given experiment (see TABLE 1). The top panel reports the mean estimated value of β_0, where $\hat{\beta}_0$ is averaged over the 1000 simulated meta-analysis studies using three different methods. "FPP" reports the estimate of β_0 in Equation (1) using the FPP procedure. "WLS-FE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. "WLS-RE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2+\tau^2}}\right)$ (cf. Equation (5) and associated discussion in the text). Yellow-coloured cells indicate the respective method produces the estimate with smallest bias. The middle panel reports the average mean squared error (MSE) value for each of the three methods. Yellow-coloured cells indicate the respective method that is most efficient. The bottom panel reports rejection rates associated with the null hypothesis $\beta_0 = \alpha$. Rejection rates are expected to be 0.05 for all experiments. "Poor performance" is identified by rejection rates greater than 0.15 and is indicated by red-coloured cells.

TABLE 7
S&D (2017) Framework for Simulating Meta-Regression Analysis Studies

STAGE 1: Generation of data for a primary study
Indirect: $Z_i = 100 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \varepsilon_i$, $\varepsilon_i \sim N(0, 100^2)$ Direct: $Z_i = 100 + (\alpha_1 + \alpha_3) X_{1i} + \alpha_2 X_{2i} + \varepsilon_i$, $\varepsilon_i \sim N(0, 100^2)$ --- $\alpha_1 = \{0, 1\}$; $\alpha_2 = 0.5$; $\alpha_3 \sim N(0, \sigma_h^2)$; $\sigma_h = \{0, 0.125, 0.25, 0.5, 1, 2, 4\}$ --- Indirect: $X_{1i} \sim U(100, 200)$; $X_{2i} = X_{1i} + N(0, 50^2)$; $X_{3i} = X_{1i} + N(0, 50^2)$ Direct: $X_{1i} \sim U(100, 200)$; $X_{2i} = X_{1i} + N(0, 50^2)$
X_{1i}, X_{2i}, and X_{3i} are randomly generated. Note that X_{2i} and X_{3i} are generated to be correlated with X_{1i}.
A sample of n_j observations of $(Z_i, X_{1i}, X_{2i}, X_{3i})$ is generated. n_j is randomly drawn from the set $\{62, 125, 250, 500, 1000\}$
STAGE 2: Estimation of effect for a primary study
One of two models are randomly chosen for estimation: (i) $Z_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$ (ii) $Z_i = \alpha_0 + \alpha_1 X_{1i} + \vartheta_i$ The coefficient of interest is α_1
Exclusion of X_{2i} from the regression model results in omitted variable bias. M_j is a dummy variable indicating whether X_{2i} was included in the primary study $M_j = 1$ if Model (i) is estimated and $M_j = 0$ if Model (ii) is estimated
STAGE 3: Publication selection bias and the creation of the MA sample
A meta-analysis observation consists of the triplet: $(\hat{\alpha}_{1j}, SE(\hat{\alpha}_{1j}), M_j)$ Publication selection bias favours estimates that are positive and significant Meta-analysis samples are constructed to have either 20 or 80 observations Publication selection bias is generated by filling 50% of the sample with positive and significant estimates; the remaining observations are randomly selected
STAGE 4: Estimation of the Meta-Regression
The following meta-regression is estimated: $\hat{\alpha}_{1j} = \beta_0 + \beta_1 M_j + \omega_i$ using a variety of estimators (e.g. fixed effects, random effects, etc.)
STAGE 5: Iteration
This process is repeated 10,000 times and the results are analysed.

TABLE 8A
FAT and PET: S&D Framework, Indirect, 50% Publication Selection Bias

MRA sample size	σ_h, *excess heterogeneity*	*True Effect*	I^2	*FAT* $(H_0: \beta_1 = 0)$	*PET* $(H_0: \beta_0 = 0)$
20	0	0	0.0724	0.1679	0.1046
20	0.125	0	0.2696	0.1170	0.0977
20	0.25	0	0.5738	0.0624	0.1030
20	0.5	0	0.8322	0.0409	0.1185
20	1.0	0	0.9419	0.0336	0.0931
20	2.0	0	0.9775	0.0589	0.0536
20	4.0	0	0.9895	0.1217	0.0458
80	0	0	0.0394	0.5200	0.2202
80	0.125	0	0.3042	0.3415	0.1096
80	0.25	0	0.6311	0.1559	0.1269
80	0.5	0	0.8598	0.0652	0.2208
80	1.0	0	0.9521	0.0884	0.1659
80	2.0	0	0.9823	0.2462	0.0682
80	4.0	0	0.9926	0.5387	0.0358
20	0	1	0.0774	0.0557	0.9997
20	0.125	1	0.2450	0.0501	0.9981
20	0.25	1	0.5356	0.0451	0.9816
20	0.5	1	0.8117	0.0377	0.8180
20	1.0	1	0.9351	0.0346	0.4772
20	2.0	1	0.9764	0.0559	0.2200
20	4.0	1	0.9899	0.1060	0.1132
80	0	1	0.0422	0.0918	1.0000
80	0.125	1	0.2781	0.0752	1.0000
80	0.25	1	0.6020	0.0607	1.0000
80	0.5	1	0.8438	0.0536	1.0000
80	1.0	1	0.9466	0.0677	0.9827
80	2.0	1	0.9815	0.2186	0.7591
80	4.0	1	0.9928	0.5321	0.4011

NOTE: This table reports FAT and PET test results using the "Indirect" heterogeneity framework of Stanley and Doucouliagos (2017). Results aggregate experimental results over 10,000 simulated meta-analysis studies. "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. I^2 measures effect heterogeneity (Higgins and Thompson, 2002). The values in the FAT and PET columns represent the rejection rates from using the FPP procedure to test the respective null hypotheses ($\beta_1 = 0$ and $\beta_0 = 0$, respectively, in Equation (1) in the text). Rejection rates are expected to be 1.00 for the FAT in all experiments. Rejection rates are expected to be 0.05 for the PET when "True Effect" = 0, and 1.00 when "True Effect" = 1. Red-coloured cells indicate that the associated rejection rates represent "poor performance" (see TABLE 4 for a more detailed discussion).

TABLE 8B
FAT and PET: S&D Framework, Direct, 50% Publication Selection Bias

MRA sample size	σ_h, *excess heterogeneity*	*True Effect*	I^2	*FAT* (H_0: $\beta_1 = 0$)	*PET* (H_0: $\beta_0 = 0$)
20	0	0	0.0724	0.1679	0.1046
20	0.125	0	0.2701	0.1201	0.0914
20	0.25	0	0.5822	0.0608	0.1100
20	0.5	0	0.8442	0.0361	0.1466
20	1.0	0	0.9535	0.0314	0.1665
20	2.0	0	0.9871	0.0346	0.1735
20	4.0	0	0.9966	0.0320	0.1779
80	0	0	0.0391	0.5193	0.2206
80	0.125	0	0.3076	0.3409	0.1113
80	0.25	0	0.6387	0.1381	0.1467
80	0.5	0	0.8702	0.0450	0.2944
80	1.0	0	0.9616	0.0265	0.3506
80	2.0	0	0.9894	0.0267	0.3505
80	4.0	0	0.9972	0.0262	0.3455
20	0	1	0.0763	0.0523	0.9998
20	0.125	1	0.2475	0.0490	0.9980
20	0.25	1	0.5455	0.0464	0.9803
20	0.5	1	0.8265	0.0383	0.7951
20	1.0	1	0.9474	0.0336	0.4739
20	2.0	1	0.9860	0.0343	0.3051
20	4.0	1	0.9964	0.0344	0.2221
80	0	1	0.0413	0.0911	1.0000
80	0.125	1	0.2801	0.0759	1.0000
80	0.25	1	0.6098	0.0631	1.0000
80	0.5	1	0.8557	0.0513	0.9998
80	1.0	1	0.9569	0.0385	0.9535
80	2.0	1	0.9886	0.0280	0.7421
80	4.0	1	0.9971	0.0282	0.5476

NOTE: This table reports FAT and PET test results using the "Direct" heterogeneity framework of Stanley and Doucouliagos (2017). Results aggregate experimental results over 10,000 simulated meta-analysis studies. "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. I^2 measures effect heterogeneity (Higgins and Thompson, 2002). The values in the FAT and PET columns represent the rejection rates from using the FPP procedure to test the respective null hypotheses ($\beta_1 = 0$ and $\beta_0 = 0$, respectively, in Equation (1) in the text). Rejection rates are expected to be 1.00 for the FAT in all experiments. Rejection rates are expected to be 0.05 for the PET when "True Effect" = 0, and 1.00 when "True Effect" = 1. Red-coloured cells indicate that the associated rejection rates represent "poor performance" (see TABLE 4 for a more detailed discussion).

TABLE 9A

MEAN VALUE OF $\widehat{\beta}_0$: S&D Framework, Indirect, 50% Publication Selection Bias

MRA sample size	σ_h, *excess heterogeneity*	*True Effect*	*FPP*	*WLS-FE*	*WLS-RE*
20	0	0	-0.0763	0.0323	0.0333
20	0.125	0	-0.0417	0.0620	0.0633
20	0.25	0	0.0503	0.1381	0.1359
20	0.5	0	0.1622	0.2625	0.2677
20	1.0	0	0.2458	0.4410	0.4878
20	2.0	0	0.2527	0.6993	0.9119
20	4.0	0	0.1997	1.0324	1.7059
80	0	0	-0.0577	0.0317	0.0322
80	0.125	0	-0.0227	0.0689	0.0699
80	0.25	0	0.0559	0.1398	0.1384
80	0.5	0	0.1719	0.2653	0.2706
80	1.0	0	0.2480	0.4365	0.4963
80	2.0	0	0.2295	0.6671	0.9072
80	4.0	0	0.1452	0.9586	1.7064
20	0	1	0.9937	1.0090	1.0100
20	0.125	1	0.9930	1.0096	1.0139
20	0.25	1	0.9953	1.0186	1.0304
20	0.5	1	0.9958	1.0626	1.0904
20	1.0	1	1.0157	1.1920	1.2570
20	2.0	1	1.0002	1.4205	1.6323
20	4.0	1	0.9731	1.7690	2.3873
80	0	1	0.9943	1.0082	1.0088
80	0.125	1	0.9938	1.0092	1.0132
80	0.25	1	0.9962	1.0150	1.0278
80	0.5	1	1.0287	1.0578	1.0900
80	1.0	1	1.1206	1.1846	1.2585
80	2.0	1	1.1647	1.3969	1.6288
80	4.0	1	1.0770	1.6838	2.3952

NOTE: This table reports the mean estimated value of β_0, using the "Indirect" heterogeneity framework of Stanley and Doucouliagos (2017) with "50% Publication Selection Bias". Results aggregate experimental results over 10,000 simulated meta-analysis studies. "50% Publication Selection Bias", "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. "FPP" reports the estimate of β_0 in Equation (1) using the FPP procedure. "WLS-FE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. "WLS-RE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2+\tau^2}}\right)$ (cf. Equation (5) and associated discussion in the text). Yellow-coloured cells indicate the respective method produces the estimate with smallest bias.

TABLE 9B
MEAN VALUE OF $\widehat{\beta}_0$: S&D Framework, Direct, 50% Publication Selection Bias

MRA sample size	*σ_h, excess heterogeneity*	*True Effect*	*FPP*	*WLS-FE*	*WLS-RE*
20	0	0	-0.0763	0.0323	0.0333
20	0.125	0	-0.0397	0.0635	0.0645
20	0.25	0	0.0549	0.1394	0.1362
20	0.5	0	0.1953	0.2722	0.2694
20	1.0	0	0.4002	0.4928	0.4929
20	2.0	0	0.7539	0.9020	0.9167
20	4.0	0	1.5020	1.7200	1.7332
80	0	0	-0.0577	0.0318	0.0323
80	0.125	0	-0.0219	0.0684	0.0693
80	0.25	0	0.0628	0.1420	0.1395
80	0.5	0	0.2052	0.2762	0.2731
80	1.0	0	0.4092	0.4946	0.5027
80	2.0	0	0.7830	0.9019	0.9189
80	4.0	0	1.4864	1.6940	1.7238
20	0	1	0.9946	1.0092	1.0104
20	0.125	1	0.9943	1.0112	1.0154
20	0.25	1	0.9934	1.0171	1.0297
20	0.5	1	1.0009	1.0669	1.0899
20	1.0	1	1.0907	1.2278	1.2599
20	2.0	1	1.4295	1.6210	1.6462
20	4.0	1	2.0603	2.3774	2.4292
80	0	1	0.9944	1.0083	1.0088
80	0.125	1	0.9934	1.0087	1.0127
80	0.25	1	0.9971	1.0155	1.0280
80	0.5	1	1.0361	1.0616	1.0896
80	1.0	1	1.1802	1.2247	1.2587
80	2.0	1	1.4863	1.6018	1.6410
80	4.0	1	2.1577	2.3883	2.4302

NOTE: This table reports the mean estimated value of β_0, using the "Direct" heterogeneity framework of Stanley and Doucouliagos (2017) with "50% Publication Selection Bias". Results aggregate experimental results over 10,000 simulated meta-analysis studies. "50% Publication Selection Bias", "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. "FPP" reports the estimate of β_0 in Equation (1) using the FPP procedure. "WLS-FE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. "WLS-RE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2+\tau^2}}\right)$ (cf. Equation (5) and associated discussion in the text). Yellow-coloured cells indicate the respective method produces the estimate with smallest bias.

TABLE 10A
MEAN SQUARED ERROR: S&D Framework, Indirect, 50% Publication Selection Bias

MRA sample size	σ_h, *excess heterogeneity*	*True Effect*	*FPP*	*WLS-FE*	*WLS-RE*
20	0	0	0.0322	0.0149	0.0152
20	0.125	0	0.0405	0.0235	0.0235
20	0.25	0	0.0643	0.0491	0.0460
20	0.5	0	0.1526	0.1246	0.1195
20	1.0	0	0.3612	0.3207	0.3542
20	2.0	0	0.7524	0.7958	1.2117
20	4.0	0	1.5520	1.8340	4.2824
80	0	0	0.0066	0.0036	0.0037
80	0.125	0	0.0077	0.0088	0.0088
80	0.25	0	0.0170	0.0259	0.0249
80	0.5	0	0.0580	0.0824	0.0836
80	1.0	0	0.1291	0.2182	0.2733
80	2.0	0	0.1939	0.5068	0.9084
80	4.0	0	0.2756	1.0624	3.2321
20	0	1	0.0067	0.0046	0.0047
20	0.125	1	0.0115	0.0074	0.0073
20	0.25	1	0.0268	0.0158	0.0143
20	0.5	1	0.0905	0.0433	0.0393
20	1.0	1	0.2582	0.1442	0.1616
20	2.0	1	0.6133	0.4366	0.7334
20	4.0	1	1.3064	1.2554	3.2016
80	0	1	0.0015	0.0011	0.0011
80	0.125	1	0.0024	0.0018	0.0018
80	0.25	1	0.0050	0.0037	0.0037
80	0.5	1	0.0138	0.0126	0.0153
80	1.0	1	0.0522	0.0580	0.0891
80	2.0	1	0.1834	0.2122	0.4745
80	4.0	1	0.4174	0.5933	2.2521

NOTE: This table reports the mean squared error (MSE) associated with estimating the mean true effect using the "Indirect" heterogeneity framework of Stanley and Doucouliagos (2017) with "50% Publication Selection Bias". Results aggregate experimental results over 10,000 simulated meta-analysis studies. "50% Publication Selection Bias", "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. "FPP" reports the estimate of β_0 in Equation (1) using the FPP procedure. "WLS-FE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. "WLS-RE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2+\tau^2}}\right)$ (cf. Equation (5) and associated discussion in the text). Yellow-coloured cells indicate the respective method produces the most efficient estimate.

TABLE 10B
MEAN SQUARED ERROR: S&D Framework, Direct, 50% Publication Selection Bias

MRA sample size	σ_h, *excess heterogeneity*	*True Effect*	*FPP*	*WLS-FE*	*WLS-RE*
20	0	0	0.0322	0.0149	0.0152
20	0.125	0	0.0404	0.0233	0.0233
20	0.25	0	0.0656	0.0494	0.0454
20	0.5	0	0.1745	0.1343	0.1202
20	1.0	0	0.5849	0.4203	0.3630
20	2.0	0	2.1702	1.4575	1.2340
20	4.0	0	8.3814	5.3850	4.4309
80	0	0	0.0066	0.0036	0.0037
80	0.125	0	0.0075	0.0087	0.0087
80	0.25	0	0.0182	0.0267	0.0252
80	0.5	0	0.0734	0.0899	0.0850
80	1.0	0	0.2635	0.2872	0.2803
80	2.0	0	0.9483	0.9614	0.9347
80	4.0	0	3.5495	3.4601	3.3076
20	0	1	0.0069	0.0047	0.0047
20	0.125	1	0.0117	0.0074	0.0073
20	0.25	1	0.0278	0.0158	0.0139
20	0.5	1	0.1024	0.0492	0.0398
20	1.0	1	0.3840	0.2087	0.1689
20	2.0	1	1.5915	0.9642	0.7690
20	4.0	1	6.9454	4.2194	3.3990
80	0	1	0.0014	0.0011	0.0011
80	0.125	1	0.0025	0.0018	0.0018
80	0.25	1	0.0053	0.0039	0.0038
80	0.5	1	0.0169	0.0146	0.0155
80	1.0	1	0.1020	0.0877	0.0902
80	2.0	1	0.5673	0.5051	0.4945
80	4.0	1	2.6810	2.5201	2.3769

NOTE: This table reports the mean squared error (MSE) associated with estimating the mean true effect using the "Direct" heterogeneity framework of Stanley and Doucouliagos (2017) with "50% Publication Selection Bias". Results aggregate experimental results over 10,000 simulated meta-analysis studies. "50% Publication Selection Bias", "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. "FPP" reports the estimate of β_0 in Equation (1) using the FPP procedure. "WLS-FE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. "WLS-RE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2+\tau^2}}\right)$ (cf. Equation (5) and associated discussion in the text). Yellow-coloured cells indicate the respective method produces the most efficient estimate.

TABLE 11A
TYPE I ERROR RATE: S&D Framework, Indirect, 50% Publication Selection Bias

MRA sample size	*σ_h, excess heterogeneity*	*True Effect*	*FPP*	*WLS-FE*	*WLS-RE*
20	0	0	0.0578	0.0856	0.0893
20	0.125	0	0.0642	0.1418	0.1673
20	0.25	0	0.0835	0.2475	0.2484
20	0.5	0	0.1068	0.3323	0.3233
20	1.0	0	0.0849	0.3294	0.3469
20	2.0	0	0.0417	0.2886	0.3482
20	4.0	0	0.0282	0.2367	0.3347
80	0	0	0.0674	0.1253	0.1353
80	0.125	0	0.0450	0.3020	0.3108
80	0.25	0	0.1119	0.5628	0.5501
80	0.5	0	0.2185	0.7650	0.7775
80	1.0	0	0.1641	0.7900	0.8383
80	2.0	0	0.0655	0.7629	0.8468
80	4.0	0	0.0249	0.7059	0.8322
20	0	1	0.0491	0.0499	0.0369
20	0.125	1	0.0730	0.0669	0.0663
20	0.25	1	0.0927	0.0918	0.0780
20	0.5	1	0.1039	0.1060	0.0890
20	1.0	1	0.1036	0.1300	0.1499
20	2.0	1	0.0928	0.1448	0.2149
20	4.0	1	0.0797	0.1556	0.2494
80	0	1	0.0544	0.0600	0.0516
80	0.125	1	0.0740	0.0740	0.0648
80	0.25	1	0.0874	0.0930	0.0775
80	0.5	1	0.0923	0.1408	0.1705
80	1.0	1	0.1453	0.2696	0.3942
80	2.0	1	0.1906	0.3993	0.5969
80	4.0	1	0.2362	0.4579	0.6957

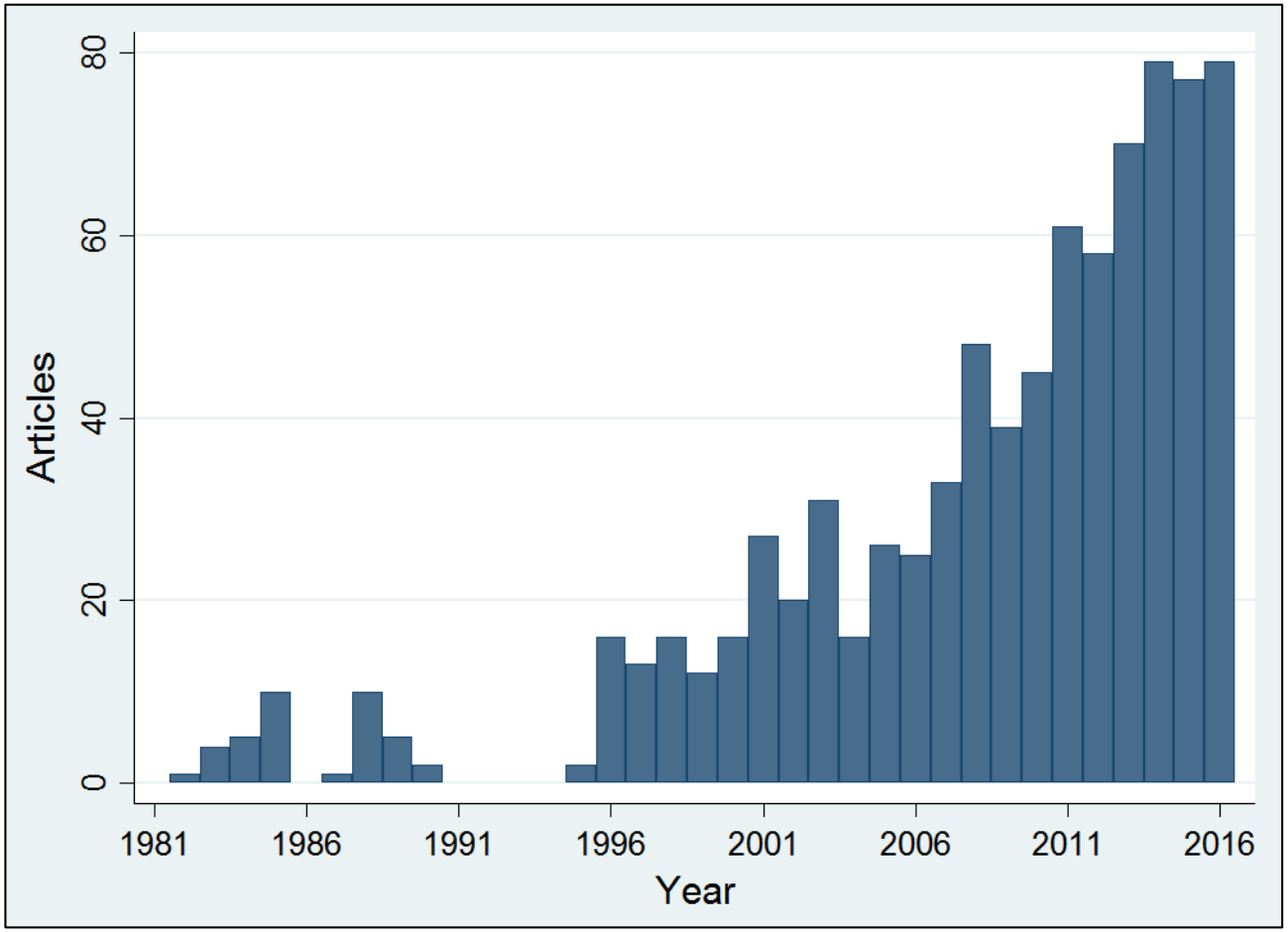
NOTE: This table reports rejection rates associated with testing the null hypothesis $\beta_0 = mean\ true\ effect$ using the "Indirect" heterogeneity framework of Stanley and Doucouliagos (2017) with "50% Publication Selection Bias". Results aggregate experimental results over 10,000 simulated meta-analysis studies. "50% Publication Selection Bias", "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. "FPP" reports the estimate of β_0 in Equation (1) using the FPP procedure. "WLS-FE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. "WLS-RE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2+\tau^2}}\right)$ (cf. Equation (5) and associated discussion in the text). Rejection rates are expected to be 0.05 for all experiments. "Poor performance" is identified by rejection rates greater than 0.15 and is indicated by red-coloured cells.

TABLE 11B
TYPE I ERROR RATE: S&D Framework, Direct, 50% Publication Selection Bias

MRA sample size	*σ_h, excess heterogeneity*	*True Effect*	*FPP*	*WLS-FE*	*WLS-RE*
20	0	0	0.0578	0.0856	0.0893
20	0.125	0	0.0609	0.1449	0.1681
20	0.25	0	0.0894	0.2507	0.2462
20	0.5	0	0.1346	0.3438	0.3209
20	1.0	0	0.1557	0.3602	0.3462
20	2.0	0	0.1607	0.3510	0.3475
20	4.0	0	0.1648	0.3346	0.3315
80	0	0	0.0673	0.1247	0.1348
80	0.125	0	0.0456	0.2975	0.3021
80	0.25	0	0.1321	0.5792	0.5574
80	0.5	0	0.2907	0.7612	0.7789
80	1.0	0	0.3486	0.7754	0.8432
80	2.0	0	0.3494	0.7570	0.8406
80	4.0	0	0.3431	0.7201	0.8204
20	0	1	0.0491	0.0525	0.0389
20	0.125	1	0.0774	0.0739	0.0687
20	0.25	1	0.0956	0.0900	0.0724
20	0.5	1	0.1145	0.1143	0.0855
20	1.0	1	0.1493	0.1625	0.1499
20	2.0	1	0.1913	0.2300	0.2128
20	4.0	1	0.1888	0.2667	0.2567
80	0	1	0.0520	0.0578	0.0490
80	0.125	1	0.0789	0.0742	0.0639
80	0.25	1	0.0975	0.1024	0.0824
80	0.5	1	0.1184	0.1569	0.1646
80	1.0	1	0.2379	0.3258	0.3875
80	2.0	1	0.3816	0.4973	0.5796
80	4.0	1	0.4576	0.5947	0.6880

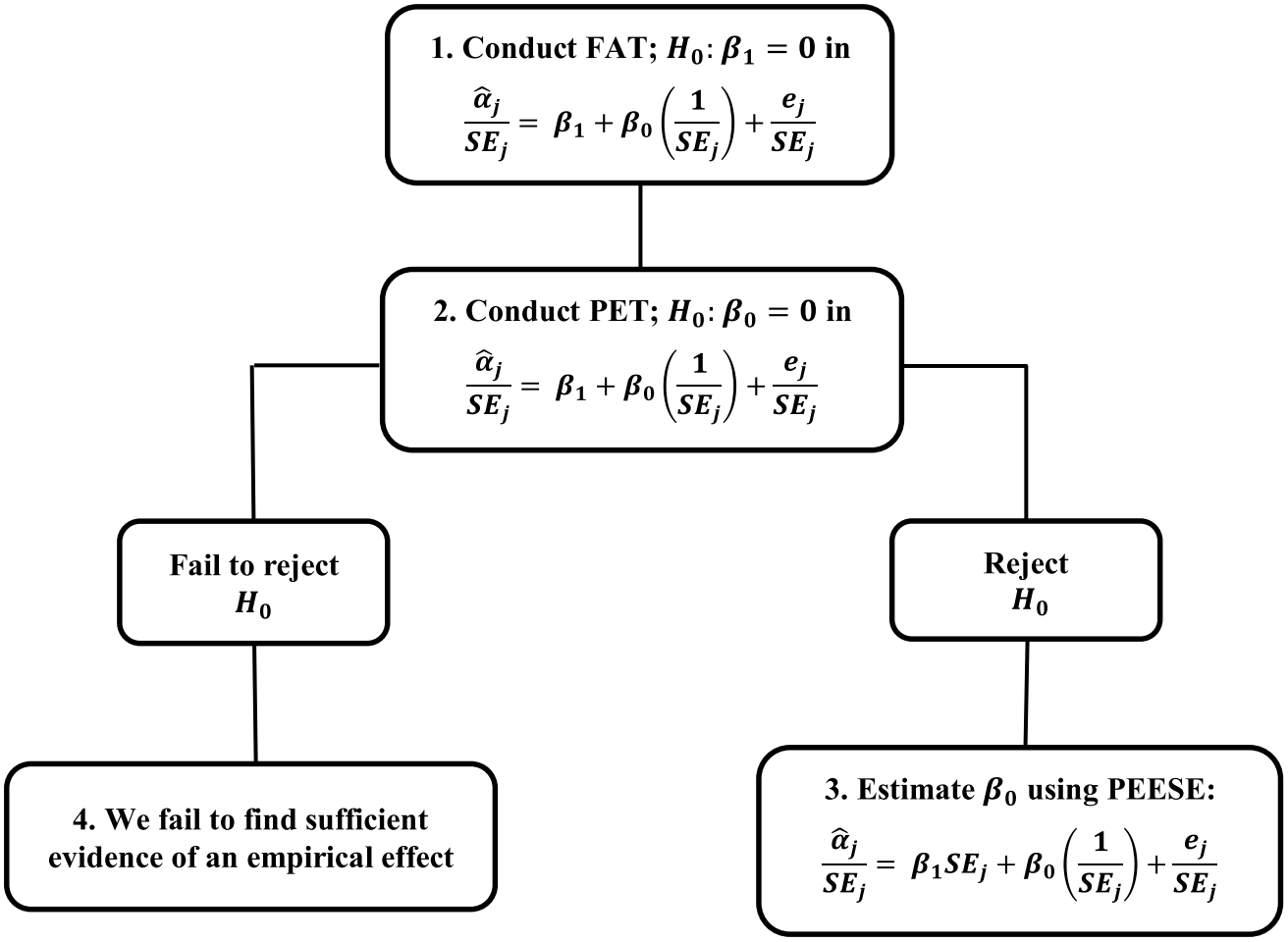
NOTE: This table reports rejection rates associated with testing the null hypothesis $\beta_0 = mean\ true\ effect$ using the "Direct" heterogeneity framework of Stanley and Doucouliagos (2017) with "50% Publication Selection Bias". Results aggregate experimental results over 10,000 simulated meta-analysis studies. "50% Publication Selection Bias", "MRA sample size", "σ_h, excess heterogeneity", and "True Effect" are all described in TABLE 7. "FPP" reports the estimate of β_0 in Equation (1) using the FPP procedure. "WLS-FE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{SE_j}\right)$. "WLS-RE" reports the estimate of β_0 in Equation (5) using Weighted Least Squares with weight $\omega_j = \left(\frac{1}{\sqrt{(SE_j)^2+\tau^2}}\right)$ (cf. Equation (5) and associated discussion in the text). Rejection rates are expected to be 0.05 for all experiments. "Poor performance" is identified by rejection rates greater than 0.15 and is indicated by red-coloured cells.

FIGURE 1
Number of Articles in Economics and Business
Listed in Web of Science with "Meta-Analysis" in the Title



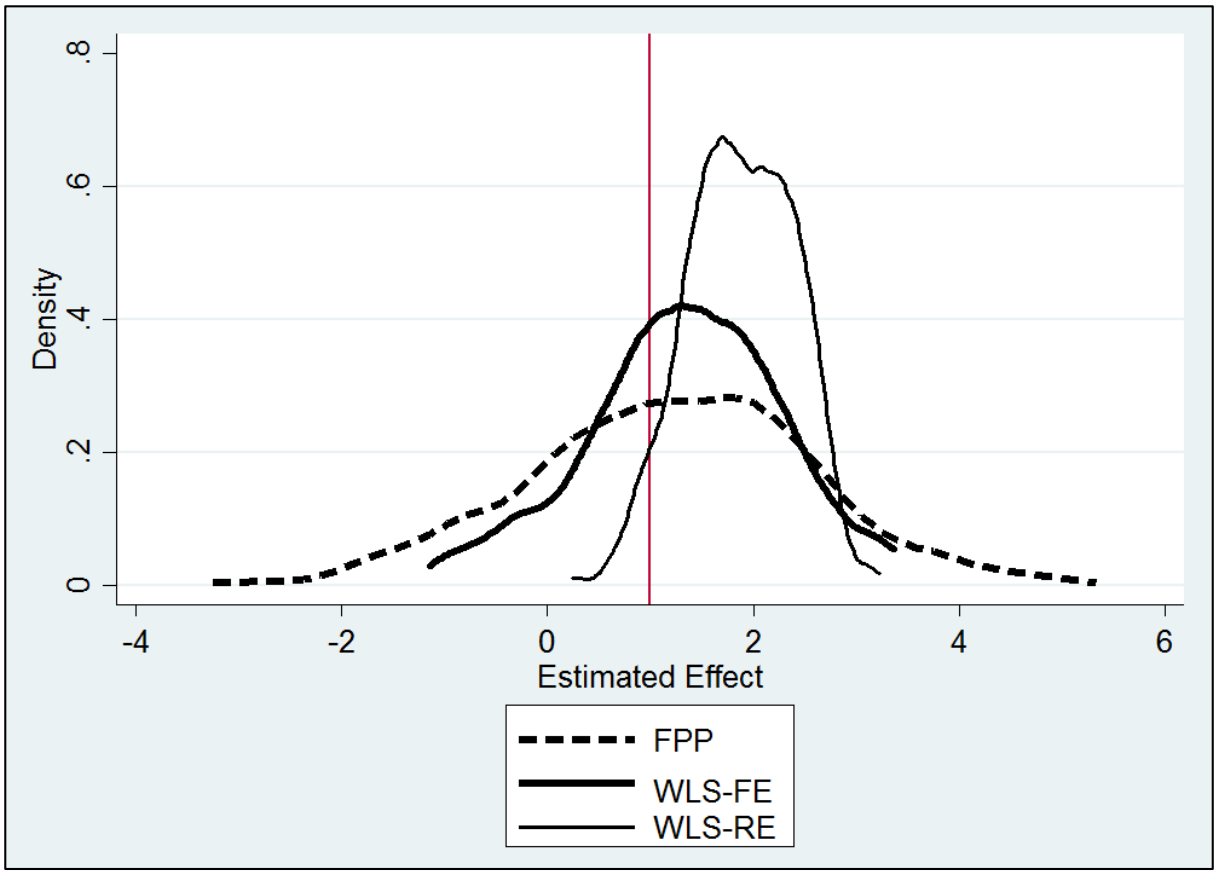
NOTE: Web of Science categories are: Economics, Business Finance, Business, Management, Criminology Penology, Urban Studies, and Social Sciences Interdisciplinary (847 articles).

FIGURE 2
The FAT-PET-PEESE Procedure



SOURCE: Stanley and Doucouliagos (2012, page 79)

FIGURE 3
Distribution of Estimated Effect Sizes by Estimator
Panel RE Case, α=1, Publication Selection Bias against Negative Estimates



About the Authors

Nazila Alinaghi is a Research Fellow in Public Finance at Victoria University of Wellington, New Zealand.
Email: nazila.alinaghi@vuw.ac.nz

W. Robert Reed is Professor of Economics at the University of Canterbury, New Zealand
Email: bob.reed@canterbury.ac.nz

VICTORIA UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA O TE ĪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School

Working Papers in Public Finance