Economics
The Open-Access, Open-Assessment E-Journal

Vol. 9, 2015-30 | October 02, 2015 | http://dx.doi.org/10.5018/economics-ejournal.ja.2015-30

A Monte Carlo Analysis of Alternative Meta-Analysis Estimators in the Presence of Publication Bias

W. Robert Reed

Abstract
A meta-analysis (MA) aggregates estimated effects from many studies to calculate a single, overall effect. There is no one, generally accepted procedure for how to do this. Several estimators are commonly used, though little is known about their relative performance. A complication arises when the sample of published studies is subject to sample selection due to "publication bias." This study uses Monte Carlo simulations to investigate the performance of five different MA estimators in the presence of publication bias. The author considers two kinds of publication bias: publication bias directed against statistically insignificant estimates, and publication bias directed against wrong-signed estimates. The experiments simulate two data environments. In the Random Effects environment, each study produces only one estimate and the true effect differs across studies. In the Panel Random Effects environment, each study produces multiple estimates, and the true effect differs both within and across studies. The simulations produce a number of findings that challenge results from previous research.

(Published in Special Issue Meta-Analysis in Theory and Practice)

JEL B41 C15 C18
Keywords Meta-analysis; random effects; fixed effects; publication bias; Monte Carlo simulations

Authors
W. Robert Reed, ✉ Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand, bob.reed@canterbury.ac.nz

Citation W. Robert Reed (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

Received January 17, 2015 **Published as Economics Discussion Paper** February 17, 2015
Revised August 29, 2015 **Accepted** September 1, 2015 **Published** October 2, 2015

1 Introduction

A meta-analysis (MA) is a systematic evaluation of a body of research that measures some "effect," such as the effect of minimum wages on unemployment, or the price elasticity of electricity demand. It is no exaggeration to say that most areas of empirical study in economics are characterized by disparate, and often conflicting, effect estimates. Meta-analysis is an attempt to summarize and "make sense" of these disparate findings. Meta-analyses have a long tradition in the medical sciences, and are increasingly popular in economics.

Meta-analyses generally have two main purposes. First, they provide an overall estimate of the size of the effect being researched. Second, they identify the reasons why estimates differ across studies. This study focuses on the first of those purposes. One way to obtain an overall estimate is to average the individual effect estimates reported by different studies. However, some effects are estimated with greater precision than others. This raises the question of how "best" to weight the respective estimates. The different answers to that question have given rise to a number of different estimators.

A major concern in the estimation of an overall effect is "publication bias." Publication bias occurs when the sample of studies available to the meta-analyst do not represent the population of all studies. Depending on the nature of the selection, the bias can cause estimates of the overall effect to either over- or under-estimate the true, underlying effect(s). Estimators have also been developed to correct publication bias in meta-analyses.

This study compares the performances of a wide variety of MA estimators. Any discussion of MA estimators is complicated by the fact that much of the terminology originates outside of economics. Terms such as "fixed effects" and "random effects" have entirely different meanings in the MA literature. In the MA literature, "fixed effects" refers to an estimation environment where different studies all estimate the same, underlying, effect. For example, it assumes that there is a single underlying price elasticity for electricity demand across all studies, even if the studies differ in time period, location, and type of consumer. Under this assumption, the only reason for studies to obtain different estimates is due to sampling error. "Random effects" refers to an environment where the "true, underlying effect" is not a single value, but a distribution of values. In this case,

the meta-analyst is interested in estimating the population mean associated with this distribution.

Another term that is misleading is "Weighted Least Squares (WLS)". In econometrics, WLS is a type of generalized least squares estimator in which observations are transformed by an individualized weight factor. Following this terminology, virtually every MA estimator is a WLS estimator. However, in the MA literature, the WLS estimator refers to a specific variant of the conventional "fixed effects" estimator that differs in how it estimates the variance of the residuals.

The term "publication bias" can also be confusing. One meaning of publication bias refers to the selection process by which research is non-randomly selected into "publication" outlets.[1] This can happen for a variety of reasons. It can happen when journals reject studies that report statistically insignificant estimates, or that are "wrong-signed." It can happen when researchers anticipate rejection by journals due to undesirable results and choose not to write a paper based on unpromising preliminary results. And it can happen when researchers only report those regressions that are most supportive of their hypothesis, or that accord with their personal biases/preferences (cf. Doucouliagos and Paldam, 2009).[2] Another meaning of publication bias is the numerical bias that arises from this selection process. Following conventional usage, this study uses "publication bias" both ways, and counts on the context to make clear which meaning is intended.

A final complication is that there is no single, universally accepted procedure for performing a meta-analysis. Many studies follow the four-step procedure advocated by Stanley and Doucouliagos (2012, pages 78–79), known as the "FAT-PET-PEESE" approach (terms to be defined below):

- STEP ONE: Test for publication bias using the FAT;
- STEP TWO: Test whether there is a zero mean effect using the PET;

[1] "Publication" means that the estimated effects appear in print. In this sense, a working paper or report can be considered as "published", even if it does not appear in a journal of book. Most meta-analyses include these types of studies in their samples.

[2] The latter kind of publication bias can be reduced by journals insisting on extensive robustness checks and meta-analysts including all such robustness checks in their analysis.

- STEP THREE: If one fails to reject the null hypothesis of no effect in STEP TWO, conclude that there is no evidence of an empirical effect;
- STEP FOUR: If one rejects the null hypothesis of no effect in STEP TWO, estimate using the PEESE.

This study contributes to an understanding of the efficacy of this procedure by investigating the performance of the associated effect estimators.

There are relatively few studies that examine the performance of MA procedures in the presence of publication bias. Stanley (2008) compares the performance of MA estimators on the dimensions of power, size, and mean squared error. His analysis produces a relatively sanguine evaluation of the ability of MA estimators to reliably detect, and estimate, variable effects: "Meta-regression methods are found to be robust against publication selection. Even if a literature is dominated by large and unknown misspecification biases, precision-effect testing and joint precision-effect and meta-significance testing can provide viable strategies for detecting genuine empirical effects" (Stanley, 2008, p. 103).

Moreno et al. (2009) compares a large number of MA estimators on the dimensions of bias, coverage rates, mean squared error, and variance. They also come to an overall positive evaluation, at least for a subset of the estimators: "In this paper we have compared some novel and existing methods for adjusting for publication bias through an extensive simulation study. Results are encouraging, with several of the regression methods displaying good performance profiles" (Moreno et al. 2009, p. 12).

Two recent papers by Stanley and Doucouliagos (2014, 2015) follow on earlier work by Koetse et al. (2010) and promote the "Weighted Least Squares" (WLS) estimator.[3] Stanley and Doucouliagos argue that this estimator outperforms both the conventional, MA "fixed effects" and "random effects" estimators in the presence of publication bias.

While all these studies make important contributions, they leave significant gaps in their coverage. For example, it is common to only consider two scenarios, "no effect" and "effect", and to ignore the interaction of effect size and publication bias. Further, they assume each study only produces one estimate. They do not consider scenarios where studies produce multiple estimates of an effect, a

[3] See Section 3.3 in Koetse et al. (2010), specifically Footnote 6.

common feature of economic studies. Finally, there is typically little effort made to ensure that the simulated samples "look like" the kinds of samples used in actual meta-analysis studies.

The main results from this study are as follows. First, while MA estimators generally outperform a simple average of estimated effects, MA estimators often struggle to completely eliminate publication bias. Second, MA estimators that do not correct for publication bias often perform as well, or better, than those that do. Third, while "random effects" estimators are often more biased than other MA estimators, they sometimes are more efficient. Fourth, hypothesis testing about the mean true effect is generally unreliable for all estimators. And fifth, caution should be exercised in applying the FAT-PET-PEESE" approach.

2 A Description of the Monte Carlo Experiments

2.1 Conceptual Framework

Figure 1 depicts the process that the Monte Carlo experiments are designed to model.[4] I conceptualize the data generating process (DGP) that produces the meta-analyst's sample as consisting of four stages. Stage 1 is the DGP that produces individual observations of *x* and *y* for a given study *i*. Let this DGP be given by:

(1) $$y_{it} = \alpha_0 + \alpha_{1i} x_{it} + \varepsilon_{it},$$

where α_{1i} is the "true" effect of *x* on *y* in study i. Note that the true effect may differ across studies. Stage 1 produces T_i observations of *x* and *y*, (y_{it}, x_{it}), $t=1,2,...T_i$.

In Stage 2, the *ith* study uses this sample of T_i observations to estimate the effect of *x* on *y*. It estimates the equation,

(2) $$y_{it} = \beta_0 + \beta_{1i} x_{it} + \epsilon_{it},\ i = 1,2,...,Ti\ ;$$

[4] Figure 1 and its discussion assume the random effects data environment, where true effects differ across studies and individual studies report only one estimate. The extension to the panel random effects data environment – where true effects differ both within and across studies, and individual studies have multiple estimates – is straightforward.

Figure 1: The Data Generating Process for the Meta-Analyst's Sample of Estimates

STAGE	*DGP/Estimates*	*Comments*
STAGE 1: Data-generating process (DGP)	$y_{it} = \alpha_0 + \alpha_{1i}x_{it} + \varepsilon_{it}$	α_{1i} is the "true" effect of *x* on *y* in study *i*. The true effect is allowed to vary across studies.
	↓	
Individual observations	(y_{it}, x_{it}), *t=1,2,...*T_i	T_i is the number of observations in the *ith* study.
	↓	
STAGE 2: Individual studies	Estimate: $y_{it} = \beta_0 + \beta_{1i}x_{it} + \epsilon_{it}$	NOTE: Observations from a given study *i* come from the same DGP
	↓	
STAGE 3 (unobserved): Pre-Publication Bias Sample	$\hat{\beta}_{11}, \hat{\beta}_{12}, \dots, \hat{\beta}_{1N}$	The different $\hat{\beta}_{1i}$ are the estimates of the effect of *x* on *y* from the different studies. Each estimate is measured with standard error, SE$(\hat{\beta}_1)_i$
	↓	
STAGE 4: "Published" studies = Post-Publication Bias Sample	$\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_M$, $M \leq N$	It is these estimates that the meta-analyst uses to obtain an estimate of the "overall effect" of *x* on *y*, given by the mean of the distribution of α_{1i} values.

producing the coefficient estimate, $\hat{\beta}_{1i}$ with standard error, SE$(\hat{\beta}_{1i})$.

Stage 3 represents the sample of estimates that would be available to the meta-analyst in the absence of publication bias. Let this sample be given by $\hat{\beta}_{11}, \hat{\beta}_{12}, \dots, \hat{\beta}_{1N}$, where *N* is the total number of estimated effects. Unfortunately, all of these estimates may not be observable to the meta-analyst. Publication bias may keep a subset of these estimates from seeing the light of day. This sample is termed the "Pre-Publication Bias Sample."

Stage 4 consists of the estimates that survive the publication selection process. This sample consists of M estimates, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_M$, $M \leq N$. This "Post-Publication Bias Sample" is what the meta-analyst uses to estimate the mean of the distribution of true effects, given by α:

(3) $\alpha = E(\alpha_{1i})$.

2.2 Publication Bias

The experiments model two types of publication bias. The first type of publication bias assumes that the publication process discriminates in favour of studies that have statistically significant estimates, indicated by *t*-statistics with absolute values greater than or equal to 2. Studies with insignificant estimates can still get "published," but only with a relatively small probability. The second type of publication bias discriminates against studies with "wrong-signed" estimated effects. Without loss of generality, it is assumed that economic theory posits that the correct sign of $E(\alpha_{1i})$ should be positive (as in value of life). Studies with negative estimates can still get "published", but, again, only with a relatively small probability.

2.2.1 Estimators

The Monte Carlo experiments compare the performances of six different estimators with regard to their ability to reliably estimate α, the mean of the distribution of true effects of *x* on *y*. These consist of the "unadjusted" average (= OLS)—included as a benchmark—and five different MA estimators, two of which are specifically designed to address publication bias. The estimators are compared on three performance measures: bias, mean-squared error (MSE), and the Type I error rates associated with testing whether the estimate of α equals its true value.

The "Unadjusted" estimator: The Unadjusted estimator of the mean true effect of *x* on *y* is given by the OLS estimate of α in the equation below:

(4) $\hat{\beta}_{i1} = \alpha + vi$, $i = 1,2,\ldots,M$,

where $\hat{\beta}_{i1}$ is the *i*th estimated effect of *x* on *y*, and *M* is the number of estimates in the "Post-Publication Bias Sample." Clearly, the unadjusted estimator simply calculates the arithmetic mean of estimated effects across studies.

The "Fixed Effects" (FE) estimator: The FE estimator weights all the observations by the inverse of the estimated standard error of $\hat{\beta}_i$, SE_i. It is motivated by the assumption that any variation in the estimated effects across studies is due solely to sampling error. The FE estimator of the mean true effect is the weighted least squares estimate of α in Equation (4):

$$(5) \qquad \frac{\hat{\beta}_{i1}}{SE_i} = \alpha\cdot\left(\frac{1}{SE_i}\right) + \frac{v_i}{SE_i}, \; i = 1,2,\ldots,M;$$

except that the residuals are standardized to produce a sample variance of 1.

The "Weighted Least Squares" (WLS) estimator. The WLS estimator is identical to the FE estimator except that the residuals remain unstandardized. Note that the FE and WLS estimators produce identical estimates of α, but the associated estimates have different standard errors.

The "Random Effects" (RE) estimator. The "RE" estimator is motivated by the assumption that differences in estimated effects across studies are due to (i) sampling variation, and (ii) genuine differences in the underlying effects. This second component is represented by τ, which is the standard deviation of underlying effects across studies. If the two error components are independent, then the s.e.($\hat{\beta}_i$) = $\sqrt{(SE_i)^2 + \tau^2} = \omega_i$. The RE estimator of the mean true effect is given by weighted least squares estimation of α in Equation (4), with weights = ω_i:

$$(6) \qquad \frac{\hat{\beta}_{i1}}{\omega_i} = \alpha\cdot\left(\frac{1}{\omega_i}\right) + \frac{v_i}{\omega_i}, \; i = 1,2,\ldots,M.$$

The PET estimator: The first of two MA estimators designed to address publication bias is the "PET" estimator. The name of this estimator derives from the fact that it is associated with a particular test within the FAT-PET-PEESE procedure known as the Precision Effect Test (PET). The PET adds the i^{th} study's estimated standard error of the estimated effect, (SE_i), as an explanatory variable to Equation (4) to control for publication bias. It then estimates the value of the mean effect in the specification below.

(7) $\hat{\beta}_{i1} = \alpha + \rho \cdot SE_i + vi,\ i = 1,2,\ldots,M.$

WLS estimation of α in Equation (7) provides an estimate of the mean true effect of *x* on *y*, adjusting for publication bias ρ:

(8) $\frac{\hat{\beta}_{i1}}{SE_i} = \alpha \cdot \left(\frac{1}{SE_i}\right) + \rho + \frac{v_i}{SE_i},\ i = 1,2,\ldots,M.$

The PEESE estimator: The second MA estimator designed to address publication bias is the "PEESE" estimator, where PEESE stands for Precision Effect Estimate with Standard Error (Stanley and Doucouliagos, 2012). This estimator is identical to the PET estimator, except that it replaces SE_i with $(SE_i)^2$ in Equation (7). This yields the following weighted least squares specification,

(9) $\frac{\hat{\beta}_{i1}}{SE_i} = \alpha \cdot \left(\frac{1}{SE_i}\right) + \rho \cdot SE_i + \frac{v_i}{SE_i},\ i = 1,2,\ldots,M.$

Note that there are no constant terms in the specifications of Equations (5), (6) and (9).

Both the PET and PEESE estimators correct for publication bias by adding some form of the effect's standard error to the regression specification. The rationale for this approach is loosely linked to Heckman-type procedures for correcting sample selection bias (see Stanley and Doucouliagos, 2012, pp. 117ff.).5

2.3 The Experiments

The experiments study estimator performance within two different data environments. In the first data environment ("Random Effects"), each study produces only one estimate, and the true effect of *x* on *y* differs across studies; perhaps because the underlying samples cover different time periods or geographical units or different types of economic agents, or because the studies use different sets of control variables. In the second data environment ("Panel Random Effects"), each study produces multiple estimates, and the true effects are

5 While it may be semantically more accurate to refer to the PET and PEESE as alternative specifications of the WLS estimator, we will refer to them as estimators for ease of exposition.

heterogeneous both across and within studies. The latter scenario is the most realistic since most MA samples include more than one estimate per study. I include the former data environment both because it provides a bridge to previous literature, and to determine whether having multiple estimates per study substantially affects the results.

In both data environments, the experiments begin by simulating a distribution of true effects that are normally distributed with mean value a. Random draws from this distribution generate study-specific "true effects", α_i. The α_i's are used to generate individual (y,x) observations, from which a single estimate is derived. This process is repeated for different draws of α_i until a total of N estimates are produced. These estimates are then put through a publication bias "filter", with the number of estimates in the Post-Publication Bias Sample, M, being determined endogenously. The respective estimators are applied to this sample to produce estimates of a, the mean of the distribution of true effects. This constitutes one meta-analysis study.

The process is repeated to produce 10,000 simulated meta-analysis studies. The estimates for each of the estimators are then aggregated over these simulated studies and compared on the dimensions of bias, MSE, and Type I error rates.

For each of the two data environments, I run experiments for nine different values of a: 0 (i.e., no overall effect), 0.5, 1, 1.5, 2, 2.5, 3, 3.5, and 4. When the distribution of true effects is centered on zero, there will be more statistically insignificant estimates, and more wrong-signed estimates, than when the distribution shifts to the right. As a result, the per cent of studies excluded by publication bias will be greatest at $\alpha = 0$. As α increases and the distribution shifts to the right, fewer studies are impacted by publication bias. Eventually, for sufficiently large α, all studies are "published", and the Post-Publication Bias Sample is identical to the Pre-Publication Bias sample. This experimental design allows me to investigate the interplay between the value of α and the consequences of publication bias. As will be demonstrated below, the consequences of increasing α will differ depending on the nature of the publication bias (statistical insignificance versus wrong-signed estimates).

3 Random Effects Data

3.1 Experimental Design

For the Random Effects data environment, I generate heterogeneity in true effects across studies by letting the true effect be normally and independently distributed with mean α and variance 1. In particular, the DGP producing individual observations for study i is given by:

(10.A) $y_{it} = 1 + \alpha_i \cdot x_{it} + \varepsilon_{it}$, $t = 1,2,\ldots,T$, where

(10.B) $\alpha_i = NID(\alpha,1)$.

All the studies have $T = 100$ observations. In order to generate different coefficient standard errors, I allow the DGP error term to have different variances across studies:

(11.A) $\varepsilon it = \lambda i \cdot NID(0,1)$, where

(11.B) $\lambda_i = 0.5 + UID(0,30)$

λ_i controls the variance of the error term. The specification in Equation (11.B) serves to set both lower and upper bounds on the values of λ_i. It is important that the error variance not be too small, lest it produce unrealistically large precision values. At the same time, the variance has to be large enough to produce realistic MA samples.

The specific parameter values used in the experiments were selected to simultaneously satisfy four criteria:

1. Produce a realistic range of *t*-values for the estimated effects
2. Produce realistic-looking funnel plots
3. Cause the per cent of studies eliminated by publication bias to range between 10 and 90 per cent (so all the MA studies are impacted by publication bias to some degree)
4. Produce realistic values of "effect heterogeneity"

"Effect heterogeneity" refers to the differences in true effects across studies. A measure of effect heterogeneity is I^2, which provides an estimate of the total per

cent of variation in estimated effects that is due to factors other than sampling error (Higgins and Thompson, 2002; Higgins et al., 2003). I^2 values of 70–95% are common in economics studies. For example, Stanley and Doucouliagos (2014, p. 14) report that "among minimum wage elasticities, I^2 is 90% (Doucouliagos and Stanley, 2009); it is 93% among estimates of the value of statistical life (Doucouliagos, Stanley and Giles, 2012) and 97% among the partial correlations of CEO pay and corporate performance (Doucouliagos, Haman and Stanley, 2012)."

Another parameterization of the experiments concerns publication bias. As discussed above, the experiments model two kinds of publication bias: selection against statistical insignificance, and selection against wrong-signed estimates. In both cases, statistically insignificant/wrong-signed estimates are allowed to be included in the Post-Publication Bias Sample, but with a relatively low probability. The experiments set this probability at 10 per cent.

Finally, for each meta-analysis study, I fix the number of Pre-Publication Bias studies/estimates at 1,000. The number of Pre-Publication Bias studies/estimates that are selected into the Post-Publication Bias Sample is determined endogenously, and will differ for different values of α. As noted above, a total of 10,000 meta-analysis studies are simulated.

3.2 Random Effects: A Representative Meta-Analysis Sample ($\alpha = 1$)

Table 1 reports the distributions of (i) estimated effects, (ii) *t*-statistics, (iii) precisions, and (iv) I^2 values for a representative[6] MA data set simulated within the Random Effects data environment.[7] The respective data characteristics (minimum value, maximum value, etc.) are averaged values over a 1,000 simulated MA data sets. The samples were constructed using the design

6 "Representative" is defined as average values across 1,000 simulations.

7 Empirical results are based on computer programs using Stata, Version 13.1. You find the associated .do files for all tables and figures here:http://dx.doi.org/10.7910/DVN/OI8XSG .

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

Variable	***Median***	***Minimum***	***P5%***	***P95%***	***Maximum***
PRE-PUBLICATION BIAS (100 per cent of estimates):					
Estimated effect	1.00	–7.40	–2.39	4.38	9.48
t-statistic	0.79	–13.24	–1.47	5.92	42.49
Precision (1/SE)	0.65	0.26	0.33	5.01	20.08
I2	0.86	0.70	0.81	0.91	0.93
PUBLICATION BIAS AGAINST INSIGNIFICANCE (33.0 per cent of estimates):					
Estimated effect	1.81	–7.42	–2.07	5.68	9.45
t-statistic	2.54	–13.62	–2.34	12.58	42.73
Precision (1/SE)	1.24	0.29	0.36	9.68	19.91
I2	0.94	0.87	0.91	0.96	0.98
PUBLICATION BIAS AGAINST NEGATIVE EFFECTS (74.7 per cent of estimates):					
Estimated effect	1.55	–5.02	0.05	4.78	9.41
t-statistic	1.28	–5.05	0.03	7.33	41.98
Precision (1/SE)	0.70	0.27	0.34	5.55	19.82
I2	0.81	0.64	0.73	0.88	0.93

parameters described above, with the mean true effect set equal to one ($\alpha = 1$). The "Pre-Publication Bias" panel of the table reports sample characteristics for the population of 1,000 studies potentially available to the meta-analyst. The next two panels in the table summarize the Post-Publication Bias samples available to the meta-analyst, depending on the type of publication bias in effect.

When $\alpha = 1$, the two types of publication bias filter out a substantial number of estimates. Only 33.0 and 74.7 per cent of all estimates appear in the meta-analyst's sample, depending on the type of the publication bias. The average minimum and maximum values of estimated effects in the Pre-Publication Bias sample are [–7.40, 9.48].[8] In the post-publication bias samples, the average ranges are [–7.42, 9.45] and [–5.02, 9.41], respectively. The average estimation bias associated with

[8] These are "average" minimum and maximum values because they are the minimum and maximum values averaged over the 1,000 simulated meta-analysis samples.

the median estimate in the two publication-biased samples is 81% and 55%, respectively.

The median *t*-statistic in the Pre-Publication Bias sample is 0.79. This compares to median *t* values of 2.54 and 1.28 in the two Post-Publication Bias samples. Note that even when publication bias discriminates against insignificant results, there are still some studies that have low *t*-statistics due to the probabilistic nature of the bias/selection process (10% are still published). Precision (1/SE) is used in a number of MA procedures to weight individual observations (see above). Relative weights across different observations can differ by a factor of 50 or more.

Table 1 also shows that the meta-analysis samples are characterized by a substantial amount of non-sampling-error-related heterogeneity, as measured by I^2. In the Pre-Publication Bias sample of 1,000 individual studies, 86% of the variation in the estimated effects is attributed to effect heterogeneity for the median MA study. The corresponding percentages for the publication-biased samples are 94% and 81%. As noted above, these values are typical for economics studies.

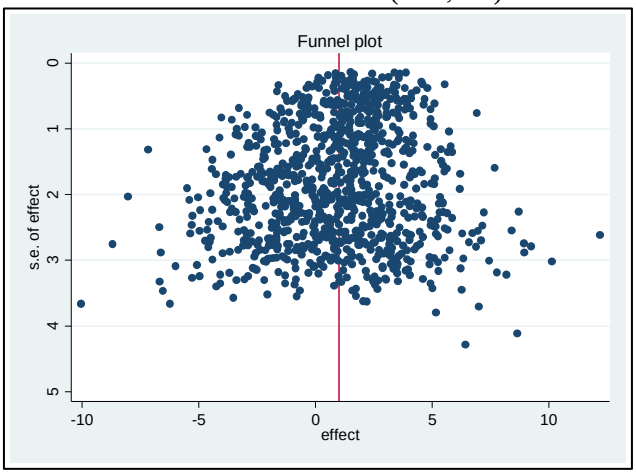
The purpose of Table 1 is to demonstrate that the parameters chosen for the Monte Carlo experiments produce samples that approximately "look like" those used in actual meta-analyses. While MA samples certainly differ, the range of estimated effects, *t*-statistics, precision, and I^2 values in Table 1 would not attract attention for being unusual. Since external validity is always an issue with Monte Carlo studies, good practice should ensure that the samples under study are similar to ones that occur in actual empirical work.

Figure 2 presents representative funnel plots both before and after publication selection. The vertical line indicates the mean true effect (=1). The top figure shows the scatter plot of estimates, graphed against their respective standard errors, with the most precise estimates plotted at the top of the funnel. The fact that there are multiple "true effects" causes the top of the "funnel" to be diffused. The same diffuse pattern is evident in both of the Post-Publication Bias funnel plots.

It is apparent from the funnel plots that publication selection distorts the distribution of estimates around the mean true effect, even when effect standard errors are very small. This induces a bias in the estimates of overall effect—a bias that the different MA estimators will be challenged to eliminate.

Figure 2: Example of Funnel Plots for a Simulated MA Data Set: Random Effects (α=1)

A. Pre-Publication Bias (N=1,000)



B. Post-Publication Bias against statistical insignificance (N=160)

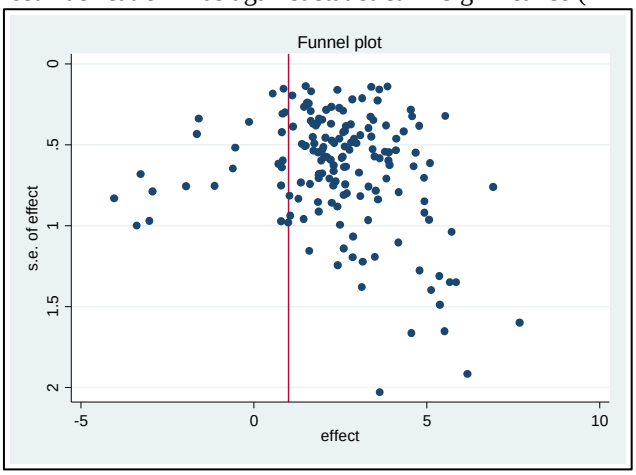
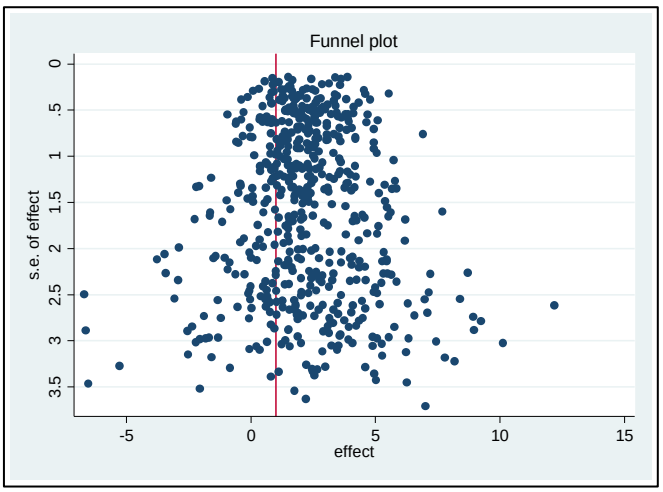


Figure 2 continued

Figure 2 continued

C. Post-Publication Bias against wrong-signed estimates (N=560)



3.3 Random Effects: Performance Tests

Tables 2 and 3 compare the six different estimators across three performance dimensions: (i) Average Estimate of Mean True Effect, (ii) Mean Squared Error (MSE), and (iii) Type I Error Rates associated with the hypothesis H_0: $true\ value\ =\ \alpha$. With respect to (iii), the significance level is set at 5 per cent, so rejection rates should likewise be equal to 5 per cent. Table 2 reports the results when publication bias is directed against statistical insignificance. Table 3 examines publication bias against wrong-signed estimates. Each of the estimators is studied for a set of mean true effect values (= α) ranging from 0.0 to 4.0 in half unit steps.

The top panel of Table 2 reports the average estimates of mean true effects for each of the respective estimators. The first two columns report the value of the true effect (α) and the average per cent of studies included in a MA study, where the average is taken over 10,000 simulated MA studies.. The first thing to note is that there is a strong relationship between the size of the true effect and the number of studies that survive publication bias against statistical insignificance. When α = 0,

an average of 27.1% of all studies appear in the meta-analyst's sample. As α increases and the mean of the distribution of estimated effects moves away from zero, more and more studies produce significant estimates. When α = 4, an average of 70.4% of estimates/studies survive publication bias and are included in the meta-analyst's sample.

The next column reports results for the Unadjusted estimator. When α = 0 and publication bias discriminates against insignificant estimates, the average estimated value of α for the Unadjusted estimator—averaged across the 10,000 MA studies—is 0.00. The Unadjusted estimator is an unbiased estimator of the true effect when α = 0 because sampling error is equally likely to produce significant estimates that are above and below the true effect. However, as α increases, publication bias disproportionately omits studies with estimates below the true effect since, ceteris paribus, studies with small estimated effects are more likely to have small *t*-values. When α = 1.0, the Unadjusted estimator overestimates the mean true effect by approximately 82%.. As α increases, fewer and fewer studies are affected by publication bias. While the table does not show this, further increases in α would eventually cause the publication bias associated with the Unadjusted estimator to disappear.

Continuing with the top panel of Table 2, we turn our attention to the performances of the five MA estimators. Of particular interest is the first two, the PET and PEESE estimators, which are specifically designed to address publication bias. With respect to estimation bias, both estimators do very well compared to the Unadjusted estimator. When α = 1, the average estimates of the mean true effect for the PET and PEESE estimators are 1.02 and 1.12, translating to biases of 2 and 12%, respectively. When α = 2, the estimates are 1.95 and 2.06, respectively. In fact, for every value of α, the PET and PEESE estimators substantially reduce bias relative to the Unadjusted estimator. This success seemingly validates the use of the PET and PEESE estimators to correct for publication bias.

The next two columns report the performance of the FE estimator and its near twin, the WLS estimator. These estimators perform almost as well as the PET and PEESE estimators, even though they do not explicitly correct for publication bias. The explanation lies in how the study estimates are weighted. In one way or another, all four of these estimators weight by the inverse of the estimated coefficient's standard error. As seen from the funnel plots in Figure 2, any

Table 2: Comparative Performance of Meta-Analysis Estimators:
Random Effects/Publication Bias against Insignificance

α	***Percent***	***Unadjusted***	***PET***	***PEESE***	***FE***	***WLS***	***RE***
Average Estimate of Mean True Effect							
0.0	27.1	0.00	0.00	0.00	0.00	0.00	0.00
0.5	28.7	1.01	0.52	0.59	0.61	0.61	0.89
1.0	33.0	1.82	1.02	1.12	1.15	1.15	1.58
1.5	39.1	2.44	1.48	1.60	1.63	1.63	2.09
2.0	45.9	2.94	1.95	2.06	2.09	2.09	2.53
2.5	52.8	3.40	2.43	2.52	2.56	2.56	2.96
3.0	59.2	3.84	2.93	3.00	3.04	3.04	3.40
3.5	65.1	4.28	3.42	3.49	3.53	3.53	3.84
4.0	70.4	4.71	3.93	3.99	4.02	4.02	4.29
Mean Squared Error							
0.0	27.1	0.026	0.059	0.037	0.036	0.036	0.012
0.5	28.7	0.286	0.056	0.043	0.044	0.044	0.164
1.0	33.0	0.693	0.049	0.046	0.050	0.050	0.340
1.5	39.1	0.888	0.043	0.036	0.041	0.041	0.352
2.0	45.9	0.893	0.042	0.028	0.032	0.032	0.285
2.5	52.8	0.815	0.046	0.026	0.027	0.027	0.216
3.0	59.2	0.711	0.044	0.024	0.024	0.024	0.160
3.5	65.1	0.609	0.044	0.024	0.023	0.023	0.120
4.0	70.4	0.511	0.042	0.024	0.022	0.022	0.089
Type I Error Rates							
0.0	27.1	0.05	0.08	0.07	0.89	0.47	0.03
0.5	28.7	0.92	0.09	0.12	0.90	0.55	0.95
1.0	33.0	1.00	0.08	0.16	0.92	0.64	1.00
1.5	39.1	1.00	0.08	0.13	0.91	0.65	1.00
2.0	45.9	1.00	0.09	0.09	0.89	0.61	1.00
2.5	52.8	1.00	0.10	0.07	0.88	0.59	1.00
3.0	59.2	1.00	0.10	0.07	0.87	0.60	1.00
3.5	65.1	1.00	0.10	0.07	0.87	0.59	1.00
4.0	70.4	1.00	0.10	0.07	0.87	0.61	1.00

estimator that heavily weights precise estimates is likely to produce estimated effects close to the mean true effect.. In other words, while the PET and PEESE estimators are successful in greatly reducing publication bias, their success has little to do with the inclusion of an *SE* term in the specification (see Equations 7 through 9).

The last column reports the estimates for the RE estimator. The RE estimator is the estimator specifically designed to address heterogeneity in true effects. It is tailored to match the data environment in which the simulations are conducted. Despite that fact, it is the most biased of the five MA estimators. This seemingly paradoxical result has been noted by other researchers (Doucouliagos and Paldam, 2013, p. 586; Stanley and Doucouliagos, 2012, p. 83).

The middle panel of Table 2 focuses on MSE performance, with smaller MSE values indicating greater efficiency. The Unadjusted estimator performs poorly compared to the MA estimators for all values of $\alpha > 0$. Among MA estimators when $\alpha > 0$, the PEESE and FE/WLS estimators generally perform best. Interestingly, the FE/WLS estimators are almost always more efficient than the PET estimator, despite sometimes having greater bias (e.g., when $\alpha = 0.5$). This shall be discussed in greater detail below.

Finally, when it comes to hypothesis testing, the bottom panel of Table 2 suggests that caution is in order. The FE, WLS, and RE estimators all produce Type I error rates that are unacceptably large. For example, when $\alpha = 0.0$, the FE and WLS estimators reject the hypothesis that $\alpha = 0.0$ in 89% and 47% of the tests, despite the fact that the hypothesis is true. This compares with an expected rejection rate of 5% given the 5% significance level employed in the tests. The PEESE estimator is substantially better, though it also produces Type I error rates substantially larger than 5% when $0.5 \leq \alpha \leq 1.5$. Given these unattractive choices, one might use be tempted to conclude that the PET estimator is serviceable for hypothesis testing. However, subsequent results will render this option less tempting.

Table 3 repeats the preceding analysis for the case when publication bias discriminates against negative effect estimates. The Unadjusted estimator again produces substantially biased estimates of the mean true effect, now even when $\alpha = 0$. Unlike the previous case, the MA estimators also produce biased estimates when α is relatively small. For example, when $\alpha = 1$, the associated biases range

Table 3: Comparative Performance of Meta-Analysis Estimators:
Random Effects/Publication Bias against Wrong Sign

α	***Percent***	***Unadjusted***	***PET***	***PEESE***	***FE***	***WLS***	***RE***
Mean Estimate of True Effect							
0.0	55.0	1.26	0.61	0.66	0.69	0.69	0.91
0.5	65.4	1.52	0.90	0.95	0.97	0.97	1.18
1.0	74.7	1.81	1.21	1.26	1.29	1.29	1.49
1.5	82.0	2.12	1.59	1.63	1.65	1.65	1.85
2.0	87.4	2.48	2.01	2.05	2.07	2.07	2.24
2.5	91.3	2.86	2.49	2.51	2.53	2.53	2.66
3.0	94.0	3.27	2.98	3.00	3.02	3.02	3.11
3.5	95.9	3.70	3.48	3.50	3.51	3.51	3.58
4.0	97.2	4.15	3.99	4.00	4.01	4.01	4.06
Mean Squared Error							
0.0	55.0	1.602	0.405	0.461	0.498	0.498	0.828
0.5	65.4	1.053	0.184	0.218	0.241	0.241	0.467
1.0	74.7	0.654	0.073	0.087	0.099	0.099	0.245
1.5	82.0	0.392	0.038	0.037	0.041	0.041	0.122
2.0	87.4	0.229	0.032	0.024	0.025	0.025	0.060
2.5	91.3	0.133	0.033	0.023	0.022	0.022	0.030
3.0	94.0	0.078	0.035	0.023	0.022	0.022	0.016
3.5	95.9	0.045	0.035	0.024	0.022	0.022	0.009
4.0	97.2	0.026	0.035	0.024	0.022	0.022	0.006

Table3 continued

Table 3 continued

Type I Error Rates							
0.0	55.0	1.00	0.89	0.96	1.00	1.00	1.00
0.5	65.4	1.00	0.74	0.91	1.00	1.00	1.00
1.0	74.7	1.00	0.29	0.53	0.98	0.96	1.00
1.5	82.0	1.00	0.10	0.18	0.91	0.79	1.00
2.0	87.4	1.00	0.08	0.09	0.87	0.68	1.00
2.5	91.3	1.00	0.08	0.07	0.87	0.65	0.92
3.0	94.0	1.00	0.08	0.07	0.87	0.65	0.63
3.5	95.9	0.94	0.08	0.07	0.87	0.66	0.36
4.0	97.2	0.70	0.08	0.07	0.87	0.66	0.20

from 21% to 49%. These biases get smaller as α increases and the proportion of included studies becomes larger.

Table 3 tells a story for MSE performance that is similar to Table 2. The FE/WLS estimator often performs as well, and sometimes slightly better, than the PET and PEESE estimators. Interestingly, when $\alpha \geq 3$, the RE estimator is most efficient, despite being the most biased. The explanation has to do with the fact that RE estimates have generally smaller variances than other MA estimators.[9] This is illustrated in Figure 3, which plots the distribution of RE and PEESE estimates when $\alpha = 3$. Figure 3 makes the general point that the MA estimators can have substantially different variances, so that a single focus on biasedness can be insufficient when comparing estimator performance.[10,11] Finally, as in Table 2,

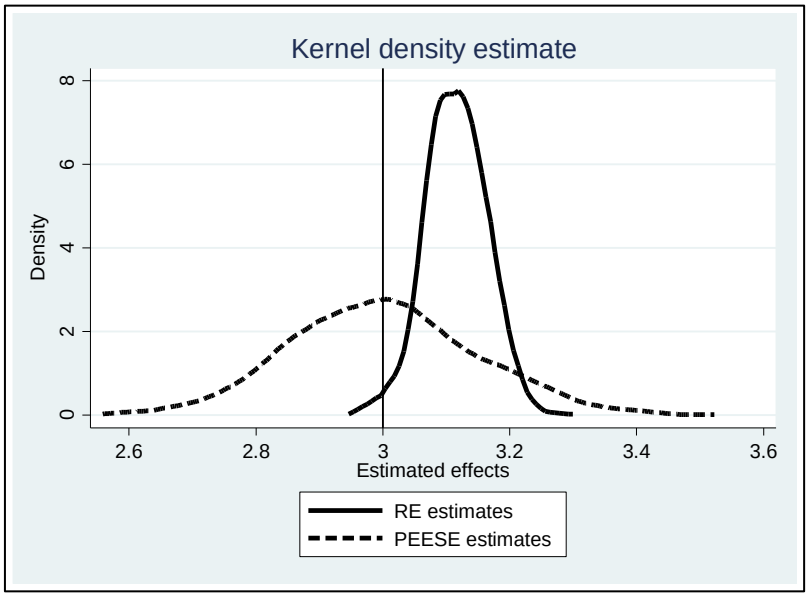
[9] The RE estimator divides the estimated effects by $(SE_i + \tau^2)$, while the other MA estimators divide by SE_i. The effect of adding a large, constant value to SE_i serves to reduce the variation in the weighting term. As a result, the RE estimator will generally have smaller variance than the other MA estimators.

[10] For example, in their response to Mekasha and Tarp (2013), Doucouliagos and Paldam (2013, p. 586) state, "One fundamental difference is that M&T13 strongly advocate the use of the random effects model, whereas D&P08 draw statistical inferences from the fixed effects models. Stanley (2008) and Stanley and Doucouliagos (2012) show that while both fixed and random effects

Type I error rates for the FE, WLS, and RE estimators are unacceptably large. Unlike Table 2, both the PET and PEESE estimators have unacceptably large Type I error rates for small values of α. As we shall see, as disappointing as these results are, they get worse.

Summarizing the results for the Random Effects data environment, we find that the MA estimators that do not explicitly correct for publication bias often perform as well, if not better, than those that do. While the MA estimators always reduce estimation bias in our experiments, they do not always eliminate it. And

Figure 3: Distribution of RE and PEESE Estimates in the RE Case, α=3, Publication Bias against Wrong Sign



weighted averages are biased in the presence of publication selection, fixed effects averages are less biased (this explains why M&T13 find significantly larger meta-averages with the random effects weighted average)." Our results show that just because the FE estimator is less biased than the RE estimator, that does not imply that the associated estimates are "better."

11 The same phenomenon can be observed for the Unadjusted estimator. For large values of α, the Unadjusted estimator has lower MSE than all but the RE estimator.

while the RE estimator is generally the most biased, it sometimes offers efficiency gains over the other MA estimators. Finally, the results with respect to hypothesis testing are generally very poor.

Monte Carlo simulations and external validity. One problem with Monte Carlo analyses is that it is difficult to determine if the associated artificial data represent empirical situations that a researcher would actually encounter. I have tried to address that concern in reporting representative sample characteristics for the simulated data in Table 1. Another approach to determine whether the funnel plots of Figure 2 are realistic is to compare them with funnel plots published in the literature.

Figure 4 presents examples of funnel plots from the MA literature. These can be compared with the funnel plots of Figure 2. Panel A shows little evidence of publication bias and is most comparable to Panel A of Figure 2. Panel B shows some evidence of bias against statistical significance and is most comparable to Panel B of Figure 2. Panel C shows clear evidence of publication bias against wrong-signed estimates (since this is a price elasticity, positive coefficients would be regarded as "wrong-signed") and is therefore comparable to Panel C of Figure 2. Together, these funnel plots provide additional evidence that the artificial data analyzed in this study are similar to actual data encountered by MA researchers.

Figrue 4: Examples of Funnel Plots from the Meta-Analysis Literature (Random Effects)

A. Mekasha and Tarp (2013, Figure 1, p. 571)

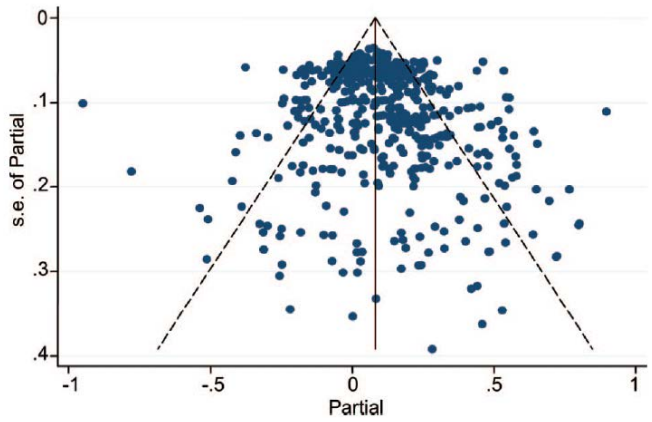
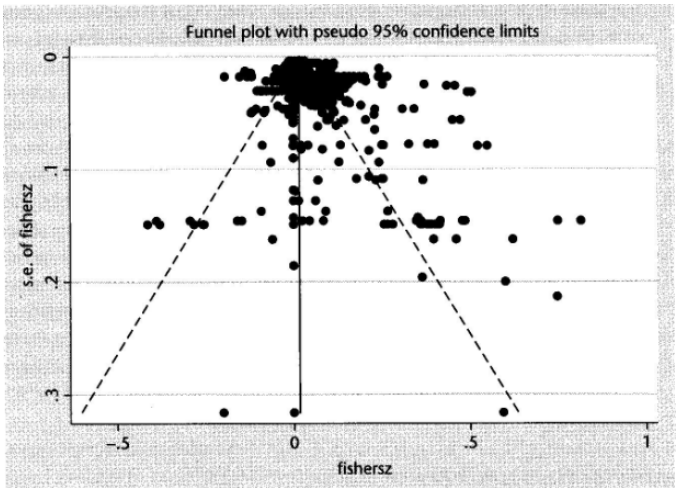


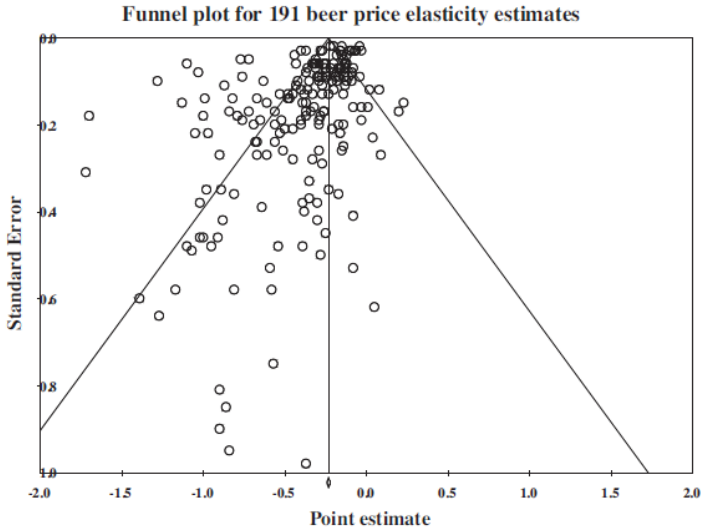
Figure 4 continued

Figure 4 continued

B. Ringquist (2013, Figure 6.7, page 252)



C. Nelson (2014, Figure 1, p. 184)



4 Panel Random Effects Data Environment

4.1 Panel Random Effects: Design of Monte Carlo Experiments

SIMULATING THE STUDIES. The last set of experiments examines the performance of the respective MA estimators when studies have multiple estimates, and true effects differ both across and within studies.[12]

There is a debate in the literature as to whether MA studies should include all estimates from a study, or just one, or a selected few. To the extent a consensus exists, it is that MA estimators should include all the estimates, but correct for error correlation across estimates within studies (Stanley and Doucouliagos, 2012; Ringquist, 2013).

The Monte Carlo experiments fix the number of Pre-Publication Bias studies at 100, each with 10 estimates per study, where each estimate is based upon 100 observations. True effects are modelled as differing both within and across studies, with the differences within studies being smaller than the difference across studies. Equations (10.A) and (10.B) from the Random Effects data environment are modified accordingly to be:

(10'.A) $y_{ijt} = 1 + \alpha_{ij} \cdot x_{ij,t} + e_{ijt}$, $t = 1,2,\ldots,100$, where

(10'.B1) $\alpha_{ij} = \alpha_i + 0.5*N(0,1)$, $j = 1,2,\ldots,10.$, and

(10'.B2) $\alpha_i = \alpha + 2*N(0,1)$, $i = 1,2,\ldots,100.$

The different weights on the standard normal variates in (10'.B1) and (10'.B2) are designed to capture the idea that effects are more likely to be similar within a study than across studies.

The error terms are modelled similarly, with error variances again differing both within and across studies, but with most of the variation occurring across studies. Equations (11.A) and (11.B) are modified from the Random Effects data environment to be:

---

[12] This study is unique in analysing estimator performance in the presence of publication bias when studies have multiple estimates. While Bijmolt and Pieters (2001) use Monte Carlo simulations to analyze meta-analysis estimators under single and multiple sampling, they do not incorporate publication bias.

(11.A) $e_{ijt} = \lambda_{ij} \cdot NID(0,1)$, where

(11.B1) $\lambda_{ij} = \lambda_i + UID(0,1)$, and

(11.B2) $\lambda_i = 0.5 + 30*UID(0,1)$.

As in the Random Effects data environment, these DGP parameters are designed to simultaneously satisfy the four criteria listed above.

Publication bias is also treated differently in the panel random effects environment. The experiments assume that the bias works at the level of the study, and not the individual estimate. In the case of bias against statistical insignificance, I assume that in order to be published, a study must have most of its estimates (7 out of 10, or more) be statistically significant. If the study meets that selection criterion, all the estimates from that study are "published." If the study does not meet that criterion, none of the estimates from that study are published. An identical "7 out of 10, or more" rule applies to publication bias against wrong-signed estimates.

Another difference has to do with the specification of the MA regressions. I modify Equation (4) to include multiple estimates per study:

(12) $\hat{\beta}_{ij1} = \alpha + \nu_{ij}$,

Dividing through by the appropriate standard error (either SE_{ij} or $\omega_{ij} = \sqrt{(SE_{ij})^2 + \tau^2}$) produces the FE, WLS, and RE estimators as described above. The PET estimator follows the recommendation of Stanley and Doucouliagos (2012, see (i) Equation 5.5, p. 85, and (ii) Equation 5.9, p. 101):

(13) $\hat{\beta}_{ij1} = \alpha + \sum_i \gamma_i SE_{ij} D_i + \nu_{ij}$,

where D_i is a dummy variable that takes the value 1 for study i and 0 for other studies. Dividing through by SE_{ij} produces the following specification:

(14) $\frac{\hat{\beta}_{ij1}}{SE_{ij}} = \alpha\left(\frac{1}{SE_{ij}}\right) + \sum_i \gamma_i D_i + \frac{(\nu_{ij})}{SE_{ij}}$.

The panel version of the PEESE estimator is given by:

$$(15) \qquad \frac{\widehat{\beta}_{ij1}}{SE_{ij}} = \alpha\left(\frac{1}{SE_{ij}}\right) + \sum_i \gamma_i SE_{ij} D_i + \frac{(v_{ij})}{SE_{ij}}.$$

For all estimators except the FE estimator, coefficient standard errors are calculated using a clustered robust procedure to allow for within-study correlation of error terms.[13]

4.2 Panel Random Effects: A Representative Meta-Analysis Sample ($\alpha = 1$)

Table 4 is similar to Table 1, except that it characterizes a representative MA sample within the Panel Random Effects data environment. It reports average characteristics of a representative MA sample with a mean true effect of one (α = 1). A total of 100 studies (= 1,000 effect estimates since each study has 10 estimates) is included in the Pre-Publication Bias sample. An average of 21.9% and 56.6% of these survive to the Post-Publication Bias samples, respectively.

The range of estimated effects and *t*-statistics is somewhat broader than their analogues in Table 1. The average minimum and maximum values of estimated effects in the Pre-Publication Bias sample are [–8.95, 10.89]. In the Post-Publication Bias samples, the average ranges are [–5.34, 8.88] and [–5.36, 10.85], respectively. For the α = 1 case, the respective publication biases generate numerical effect biases of 140% when publishing is biased against statistical insignificance, and 123% when the bias is against "wrong-signed" (= negative) estimates.

The median *t*-statistic in the Pre-Publication Bias sample is 0.68. This compares to median *t* values of 3.68 and 1.72 in the two Post-Publication Bias samples. The range of precision values, and measures of heterogeneity I^2 are similar to those reported in Table 1.

[13] Unlike the other estimators, the FE estimator imposes the assumption of a constant variance of 1 on the variance of the residuals in Equations (4) and (12). Allowing a robust form of error correlations is at odds with this assumption.

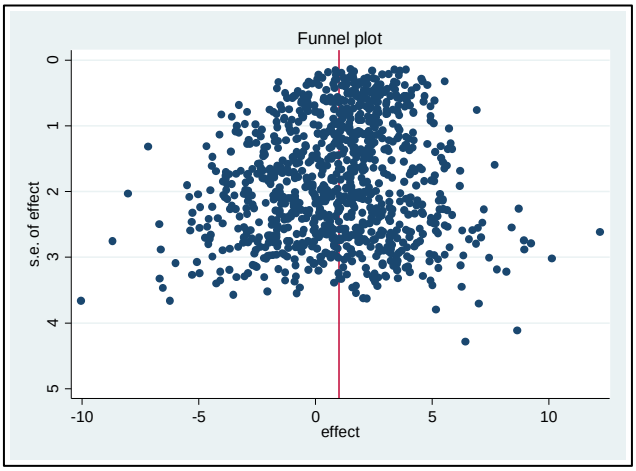
Table 4: Sample Characteristics for a Simulated Meta-Analysis Data Set: Panel Random Effects Case ($\alpha = 1$)

Variable	***Median***	***Minimum***	***P5%***	***P95%***	***Maximum***
PRE-PUBLICATION BIAS (100% 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
Precision (1/SE)	0.63	0.26	0.33	4.01	12.99
I2	0.91	0.73	0.83	0.97	0.99
PUBLICATION BIAS AGAINST INSIGNIFICANCE (21.9% 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
Precision (1/SE)	1.83	0.45	0.60	7.21	12.69
I2	0.97	0.87	0.94	0.99	1.00
PUBLICATION BIAS AGAINST NEGATIVE EFFECTS (56.6% 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
Precision (1/SE)	0.69	0.27	0.34	4.29	11.64
I2	0.83	0.51	0.69	0.94	0.98

Figure 5 shows corresponding funnel plots for the Pre- and Post-Publication Bias samples. The vertical line reports the overall mean of true effects ($\alpha = 1$). The diffusion at the top of the funnels in each panel reflects the heterogeneous nature of true effects. The distorting effects of publication bias are clearly evident in the two Post-Publication Bias funnel plots.

Figure 5: Example of Funnel Plots for a Simulated MA Data Set:
Panel Random Effects (α=1)

A. Pre-Publication Bias (N=1,000)



B. Post-Publication Bias against statistical insignificance (N=160)

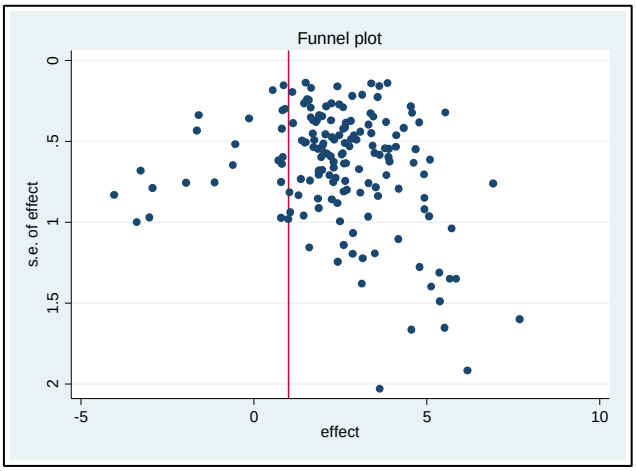
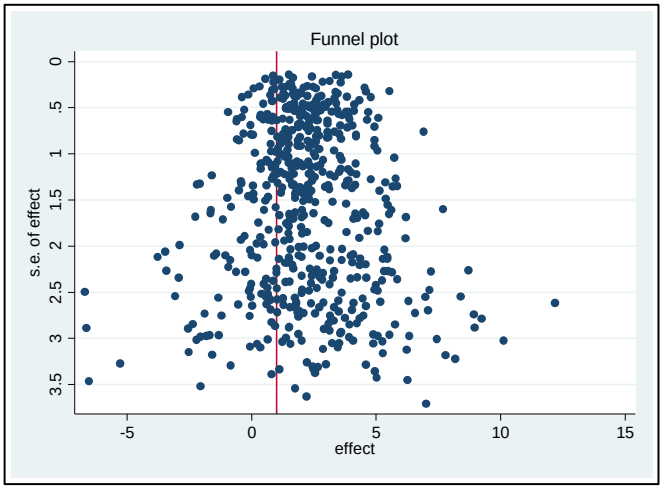


Figure 5 continued

Figure 5 continued

C. Post-Publication Bias against wrong-signed estimates (N=560)



4.3 Panel Random Effects: Performance Tests

As we have seen in previous cases, the Unadjusted estimator provides an unbiased estimate of the mean true effect when α = 0 and publication bias discriminates against statistical insignificance. As α increases, publication bias at first worsens, then eventually starts to improve as more studies are "published." The numerical bias for the Unadjusted estimator can be quite substantial. For example, when α = 2.0, the Unadjusted estimator estimates an average value for α of 3.36 (Table 5).

With respect to bias, the PET and PEESE estimators perform best of all MA estimators. For example, when α = 2.0, the PET and PEESE estimators produce a mean estimate of α equal to 2.24, compared to 2.37 and 3.13 for the MA estimators that do not correct for publication bias.

Superiority on the dimension of bias does not necessarily translate into superiority in MSE performance. While the PET and PEESE estimators are least biased, they are also least efficient among the MA estimators, and sometimes even less efficient than the Unadjusted estimator. This seeming anomaly was previously addressed in the discussion around Figure 3. As Figure 6 demonstrates, this

Table 5: Comparative Performance of Meta-Analysis Estimators:
Panel Random Effects/Publication Bias against Insignificance

α	*Percent*	*Unadjusted*	*PET*	*PEESE*	*FE*	*WLS*	*RE*
Mean Estimate of True Effect							
0.0	19.2	0.01	0.01	0.01	0.01	0.01	0.01
0.5	19.9	1.09	0.61	0.61	0.66	0.66	1.01
1.0	22.0	2.05	1.20	1.21	1.29	1.29	1.90
1.5	25.2	2.78	1.73	1.74	1.85	1.85	2.59
2.0	29.5	3.36	2.24	2.24	2.37	2.37	3.13
2.5	34.7	3.84	2.74	2.75	2.86	2.86	3.60
3.0	40.4	4.26	3.21	3.21	3.31	3.31	4.00
3.5	46.4	4.65	3.66	3.67	3.76	3.76	4.39
4.0	52.8	5.03	4.11	4.12	4.20	4.20	4.77
Mean Squared Error							
0.0	19.2	0.506	1.765	1.553	0.874	0.874	0.443
0.5	19.9	0.796	1.767	1.554	0.879	0.879	0.655
1.0	22.0	1.435	1.700	1.506	0.880	0.880	1.111
1.5	25.2	1.866	1.673	1.465	0.851	0.851	1.387
2.0	29.5	2.000	1.531	1.341	0.782	0.782	1.428
2.5	34.7	1.916	1.461	1.277	0.722	0.722	1.312
3.0	40.4	1.671	1.415	1.231	0.652	0.652	1.094
3.5	46.4	1.397	1.335	1.159	0.577	0.577	0.874
4.0	52.8	1.126	1.287	1.107	0.527	0.527	0.670
Type I Error Rates							
0.0	19.2	0.05	0.29	0.28	0.97	0.17	0.05
0.5	19.9	0.15	0.29	0.29	0.97	0.17	0.14
1.0	22.0	0.43	0.29	0.29	0.97	0.19	0.37
1.5	25.2	0.71	0.30	0.30	0.97	0.22	0.62
2.0	29.5	0.89	0.29	0.29	0.97	0.23	0.80
2.5	34.7	0.95	0.29	0.29	0.97	0.23	0.88
3.0	40.4	0.98	0.29	0.29	0.96	0.21	0.90
3.5	46.4	0.98	0.27	0.26	0.96	0.18	0.89
4.0	52.8	0.98	0.28	0.27	0.96	0.17	0.84

time it is the PET estimator that has greatest variance, which explains its poor efficiency performance. Among MA estimators, the FE/WLS estimators are generally most efficient, though the RE estimator is best for low values of α.

Finally, when it comes to hypothesis testing, the lesson from the bottom panel of Table 5 could perhaps be summarized as "don't." In almost every case, the Type I error rates are so much larger than 5% that any results from hypothesis testing about the mean true effect should be regarded as highly dubious.

Table 6 shows that all five of the MA estimators produce estimates of α that are positively biased when publication bias is directed against wrong-signed (negative) estimates. While the MA estimators always produce estimates that are less biased than the Unadjusted estimator, the reduction in estimation bias is often quite small. For example, when $\alpha = 2.0$, the Unadjusted estimator produces a mean estimate of 2.89, while the PET, PEESE, and FE/WLS estimators produce estimates of 2.66, 2.66, and 2.68, respectively. The bias in estimated effects disappears only as publication bias becomes less severe and more studies are included in the MA sample. The RE estimator is most biased of all MA estimators across all values of α.

Figure 6: Distribution of All Estimators: Panel RE Case, α=1, Publication Bias against Insignificance

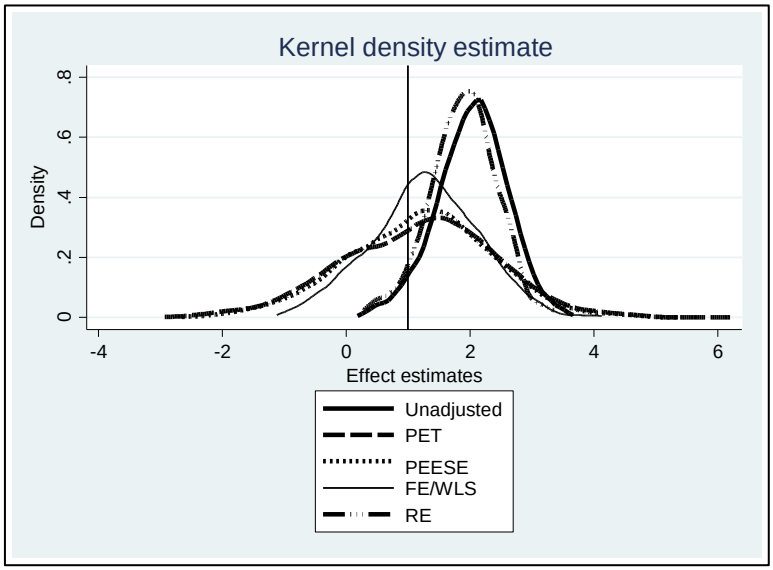


Table 6: Comparative Performance of Meta-Analysis Estimators:
Panel Random Effects /Publication Bias against Wrong Sign

α	*Percent*	*Unadjusted*	*PET*	*PEESE*	*FE*	*WLS*	*RE*
Mean Estimate of True Effect							
0.0	38.4	2.01	1.74	1.74	1.77	1.77	1.88
0.5	47.7	2.19	1.92	1.92	1.94	1.94	2.07
1.0	56.8	2.40	2.14	2.15	2.17	2.17	2.29
1.5	65.6	2.63	2.40	2.40	2.41	2.41	2.53
2.0	73.6	2.89	2.66	2.66	2.68	2.68	2.80
2.5	80.6	3.19	3.00	3.00	3.01	3.01	3.11
3.0	86.2	3.51	3.35	3.35	3.36	3.36	3.45
3.5	90.6	3.87	3.73	3.73	3.73	3.73	3.82
4.0	93.9	4.26	4.14	4.14	4.14	4.14	4.22
Mean Squared Error							
0.0	38.4	4.090	3.897	3.664	3.414	3.414	3.592
0.5	47.7	2.900	2.884	2.672	2.388	2.388	2.513
1.0	56.8	2.002	2.176	1.999	1.672	1.672	1.709
1.5	65.6	1.312	1.703	1.522	1.143	1.143	1.106
2.0	73.6	0.830	1.362	1.194	0.796	0.796	0.689
2.5	80.6	0.507	1.231	1.061	0.609	0.609	0.418
3.0	86.2	0.299	1.173	1.004	0.502	0.502	0.245
3.5	90.6	0.172	1.143	0.980	0.445	0.445	0.144
4.0	93.9	0.105	1.142	0.973	0.423	0.423	0.092
Type I Error Rates							
0.0	38.4	1.00	0.78	0.90	1.00	0.99	1.00
0.5	47.7	1.00	0.67	0.77	1.00	0.92	1.00
1.0	56.8	1.00	0.54	0.61	1.00	0.77	1.00
1.5	65.6	1.00	0.44	0.47	0.99	0.56	1.00
2.0	73.6	1.00	0.36	0.38	0.97	0.38	0.98
2.5	80.6	0.97	0.32	0.33	0.96	0.28	0.87
3.0	86.2	0.79	0.30	0.30	0.96	0.21	0.60
3.5	90.6	0.49	0.28	0.27	0.95	0.17	0.34
4.0	93.9	0.26	0.28	0.27	0.96	0.16	0.18

Table 6 provides further support that inferiority on the dimension of biasedness does not imply inferiority on efficiency. The RE estimator is now either best or close to best on the dimension of MSE for all values of α It is also worth noting that for large values of α, the Unadjusted estimator is more efficient than every MA estimator except the RE estimator. Reliability in hypothesis testing for all the estimators continues to be abysmal across the full range of α values.

5 Possible Directions for Future Research

While this study has focussed on the performance of MA estimators in a variety of experimental settings, there are many performance issues deserving study that it has not addressed. One issue concerns the weighting of individual estimates. As discussed above, most MA estimators weight either by the standard error of the estimated effect, SE_i, or by a term that expands this to include unobserved heterogeneity in true effects across estimates, $\sqrt{(SE_i)^2 + \tau^2}$. Another possibility is to expand the weighting to incorporate correlation across estimates from the same study. Currently, most MA studies adjust standard errors for correlation of estimates from the same study, but ignore this correlation in calculating coefficient estimates. To address this omission, Ringquist (2013, pp. 218ff.) suggests using generalized estimating equations (GEE).

A related issue concerns how best to handle different numbers of estimates per study. If all coefficient estimates had the same estimated standard errors, conventional MA estimators would weight all estimates the same. Thus, a study that reported 100 estimates would implicitly receive 100 times the weight of a study that reported just one estimate. An alternative is to give equal weight to studies, as opposed to (standardized) estimates. Examples include Havranek and Irsova (2015) and Reed and Sidek (in press).

Another issue is the potential endogeneity of the standard error. Endogeneity arises whenever a study characteristic systematically affects both the estimated effect from that study, and the standard error of that effect. For example, a study that employs GLS methods to correct for nonspherical errors will produce both different coefficient estimates and different standard errors than a study that uses OLS methods applied to the same data. The potential consequence of this endogeneity is that it biases coefficient estimates in any specification that includes

SE as an explanatory variable. Havranek (in press) suggests using the number of observations as an instrument.

A final issue concerns the handling of outliers. On the one hand, "outliers", as in exceptionally precise effect estimates, are the key to improving estimates of overall effects (Stanley et al., 2010). On the other hand, if these highly precise coefficient estimates are biased by omitted study characteristics, they can produce misleading estimates. A variety of methods exist to measuring, and correcting, the influence of outliers (Williams, 2015).

6 Conclusion

This study uses Monte Carlo simulations to investigate the performances of five different meta-analysis (MA) estimators in the presence of publication bias: the Precision Effect Test (PET) estimator, the Precision Effect Estimate with Standard Errors (PEESE) estimator, the Fixed Effects (FE) estimator, the Weighted Least Squares (WLS) estimator, and the Random Effects (RE) estimator. Two types of publication bias are analyzed: publication bias directed against statistically insignificant estimates, and publication bias directed against wrong-signed estimates.

The simulated experiments are conducted within two different data environments. In the Random Effects data environment, each study produces only one estimate, and the true effect of *x* on *y* differs across studies, perhaps because the underlying samples cover different time periods or geographical units or different types of economic agents, or because the studies use different sets of control variables. In the Panel Random Effects data environment, each study produces multiple estimates, and the true effects are heterogeneous both across and within studies. The latter scenario is the most realistic since most MA samples include more than one estimate per study.

Table 7 summarizes the main findings of this study. It also identifies the specific experimental results in this study that support those findings. First, while MA estimators generally outperform a simple average of estimated effects, MA estimators often struggle to eliminate publication bias. Second, MA estimators that do not correct for publication bias often perform as well, or better, than those that

Table 7: Main Results from Simulation Experiments

	RESULT
1	***While MA estimators generally outperform a simple average of estimated effects, MA estimators often struggle to eliminate publication bias.*** EVIDENCE FROM PANEL RE: See top panel (*Mean Estimate of True Effect*) of Table 6. EVIDENCE FROM RE: See top panel of Table 3 for low values of α.
2	***The FE/WLS estimators—which do not correct for publication bias—often are equally efficient, and sometimes more efficient, than the PET and PEESE estimators, which do correct for publication bias.*** EVIDENCE FROM PANEL RE: See middle panel *(MSE*) of Table 6. EVIDENCE FROM RE: See middle panel of Table 3 for large values of α.
3	***The RE estimator is often more biased than other MA estimators, but sometimes more efficient.*** EVIDENCE FROM PANEL RE: See top and middle panels of Table 6. EVIDENCE FROM RE: See top and middle panels of Table 3 for large values of α.
4	***Hypothesis testing about the mean true effect is generally unreliable for all estimators.*** EVIDENCE FROM PANEL RE: See bottom panel (*Type I Error Rates*) of Table 5 EVIDENCE FROM RE: See bottom panel of Table 3 for small values of *α*.
5	***With respect to the four-step FAT-PET-PEESE procedure advocated by Stanley and Doucouliagos (2012), the simulation results suggest caution about two of the steps.***
5A	• ***STEP TWO: Test whether there is a zero mean effect using the PET. Simulation results indicate that hypothesis testing using the PET when α=0 is not reliable.*** EVIDENCE FROM PANEL RE: See bottom panel of Table 5. EVIDENCE FROM RE: See bottom panel of Table 3.
5B	• ***STEP FOUR: If one rejects the null hypothesis of no effect in STEP TWO, estimate using the PEESE.*** **Simulation results indicate that the FE/WLS and RE estimators are sometimes more efficient than the PEESE estimator.** EVIDENCE FROM PANEL RE: See FE/WLS results in middle panel of Table 5. EVIDENCE FROM RE: See RE results in middle panel of Table 3 for large values of α.

do. Third, while the RE estimator is often more biased than other MA estimators, it is sometimes more efficient. Fourth, hypothesis testing about the mean true effect is generally unreliable for all estimators. And fifth, caution should be exercised when using the four-step FAT-PET-PEESE procedure (Stanley and Doucouliagos, 2012, pp. 78–79).

These findings challenge results from previous research, which are also based on Monte Carlo simulations. The intended takeaway from this study is not that previous research is necessarily wrong. Rather, it highlights the need for additional research in order to develop best practice for meta-analyses. One way to reconcile conflicting results from different studies is to allow researchers access to the programming code used to produce those results. You can download the Stata programming code used to produce all the results in this study here: http://dx.doi.org/10.7910/DVN/OI8XSG. This should make it easy for researchers to check for robustness from modifications of the experimental design.

Until such time that a consensus can be developed, I suggest two recommendations for meta-analysts interested in identifying "true effects" from publication-biased samples. The first is that meta-analysts should use a variety of MA estimators in their research. In particular, they should include results from MA estimators that do not correct for publication bias even if there is evidence that publication bias exists in their sample. The second recommendation is that hypothesis testing about the true mean effect should be viewed with substantial scepticism.

On the positive side, this study provides further evidence that current MA procedures generally offer improvements on simple averaging of published estimates. Nevertheless, there is much room for improvement. It is hoped that this research will stimulate future efforts in this direction.

Acknowledgements This research has benefitted from comments made by participants at the 2014 New Zealand Econometric Study Group Meetings, the 2014 New Zealand Association of Economists Meetings, and the 2014 MAER-Net Colloquium. Nazila Alinaghi provided excellent research assistance. I am especially indebted to Jacques Poot and Raymond Florax for many wide-ranging discussions that have greatly improved both the manuscript, and my understanding of the subject. I retain full property rights for all remaining errors.

References

Bellavance, R., Dionne, G., and Lebeau, M. (2009). The value of a statistical life: A meta-analysis with a mixed effects regression model. *Journal of Health Economics* 28: 444–464. http://www.sciencedirect.com/science/article/pii/S0167629608001549

Bijmolt, T.H.A. and Pieters, R.G.M. (2001). Meta-analysis in marketing when studies contain multiple measurements. *Marketing Letters* 12: 157–169. http://www.jstor.org/stable/40216595

Dalhuisen, J.M., Florax, R.J.G.M., de Groot, H.L.F., and Nijkamp, P. (2003). Price and income elasticities of residential water demand: A meta-analysis. *Land Economics* 79: 292–308. http://www.jstor.org/stable/3146872

Doucouliagos, H. and Paldam, M. (2009). The aid effectiveness literature: The sad results of 40 years of research. *Journal of Economic Surveys* 23: 433–461. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-6419.2008.00568.x/abstract

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. http://www.tandfonline.com/doi/abs/10.1080/00220388.2013.764595

Doucouliagos, H. and Stanley, T.D. (2009). Publication selection bias in minimum wage research? A meta-regression analysis. *British Journal of Industrial Relations* 47(2): 406–28. https://ideas.repec.org/a/bla/brjirl/v47y2009i2p406-428.html

Doucouliagos, H., Haman, J., and Stanley, T.D. (2012). Pay for performance and corporate governance reform. *Industrial Relations*, 51(3): 670–703. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-232X.2012.00695.x/abstract

Doucouliagos, C., Stanley, T.D., and Giles, M. (2012). Are estimates of the value of a statistical life exaggerated? *Journal of Health Economics*, 31(1): 197–206. http://www.sciencedirect.com/science/article/pii/S0167629611001342

Havranek, T. (2015). Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association*, in press. http://onlinelibrary.wiley.com/doi/10.1111/jeea.12133/abstract

Havranek, T. and Irsova, Z. (2015). Do borders really slash trade? A meta-analysis. William Davidson Institute Working Papers Series wp1088, William Davidson Institute at the University of Michigan. https://ideas.repec.org/p/wdi/papers/2015-1088.html

Higgins, J.P.T., and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21(11): 1539–1558. http://www.ncbi.nlm.nih.gov/pubmed/12111919

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* 327(7414): 557–560. http://www.ncbi.nlm.nih.gov/pubmed/12958120

Koetse, M.J., Florax, R.J.G.M., and de Groot, H.L.F. (2010). Consequences of effect size heterogeneity for meta-analysis: a Monte Carlo study. *Statistical Methods and Applications* 19(2): 217–236. http://link.springer.com/article/10.1007%2Fs10260-009-0125-0

Mekasha, T.J. and Tarp, F. (2013). Aid and growth: What meta-analysis reveals. *The Journal of Development Studies* 49(4): 564–583. http://www.tandfonline.com/doi/abs/10.1080/00220388.2012.709621

Moreno, S.G., Sutton, A.J., Ades, A.E. Stanley, T.D., Abrams, K.R., Peters, J.L., and Cooper, N.J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology* 9:2. http://www.ncbi.nlm.nih.gov/pubmed/19138428

Nelson, J.P. (2014). Estimating the price elasticity of beer: Meta-analysis of data with heterogeneity, dependence, and publication bias. *Journal of Health Economics* 33: 180–187.

Reed, W.R., and Sidek, N.N. (2015). A replication of 'Meta-analysis of the effect of fiscal policies on long-run growth' (European Journal of Political Economy, 2004). *Public Finance Review*, in press. http://pfr.sagepub.com/content/early/2015/02/10/1091142114568659.refs

Ringquist, E.J. (2013). *Meta-analysis for public management and policy.* San Francisco: Jossey-Bass.

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. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0084.2007.00487.x/abstract

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

Stanley, T.D., and Doucouliagos, H. (2014). Better than random: Weighted least squares meta-regression analysis. Working paper, Deakin University, School of Accounting, Economics, and Finance, Economics Series, SWP 2013/2, updated February 2014. http://www.deakin.edu.au/__data/assets/pdf_file/0010/408655/2013_2.pdf

Stanley, T.D., and Doucouliagos, H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Stat Med*. 34(13):2116–2127. http://www.ncbi.nlm.nih.gov/pubmed/25809462

Stanley, T.D., Jarrell, S.B. and Doucouliagos, C. (2010). Could it be better to discard 90% of the data? A statistical paradox. *American Statistician* 64(1): 70–77. http://www.tandfonline.com/doi/abs/10.1198/tast.2009.08205

Williams, R. (2015). "Outliers."12 August. Retrieved from: https://www3.nd.edu/~rwilliam/stats2/l24.pdf