DEPARTMENT OF ECONOMICS AND FINANCE

COLLEGE OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Meta-Analysis and Publication Bias: How Well Does the FAT-PET-PEESE Procedure Work?

Nazila Alinaghi
W. Robert Reed

WORKING PAPER

No. 26/2016

Department of Economics and Finance
College of Business and Economics
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 26/2016

Meta-Analysis and Publication Bias: How Well Does the FAT-PET-PEESE Procedure Work?

Nazila Alinaghi[1]
W. Robert Reed[1†]

5 November 2016

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 overall mean 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 overall, mean effect cannot be trusted.

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

JEL Classifications: B41, C15, C18

Acknowledgements: We are grateful for comments received from Tom Stanley, Chris Doucouliagos, and participants at the 2016 MAER-Net Colloquium.

[1] Department of Economics and Finance, University of Canterbury, Christchurch, NEW ZEALAND

† Corresponding author is W. Robert Reed. Email: bob.reed@canterbury.ac.nz

I. INTRODUCTION

Meta-analysis is the statistical analysis of estimates from multiple studies that are all concerned with measuring a similar "effect." Two 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 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 bias.

A common procedure for doing this in the economics and business literature is the FAT-PET-PEESE procedure (Stanley and Doucouliagos, 2012; 2014a). FIGURE 2 depicts the associated four steps. The first is the Funnel Asymmetry Test (FAT). It uses Weighted Least Squares (WLS) to regress the estimated effects ($\hat{\alpha}_j$**)** on a constant term (β_0**)** and the standard errors of the estimated effects (SE_j**)**; where weights $\omega_j = \left(\frac{1}{SE_j}\right)$ are applied to correct for heteroskedasticity in the estimates. If the estimated coefficient on the standard error variable, $\hat{\beta}_1$, is significant, that indicates the estimates suffer from publication bias.

The next step is the Precision Effect Test (PET). It uses the same equation as the FAT, but tests whether $\beta_0 = 0$. If the SE_j variable were not included in the equation, and if OLS was

used rather than WLS, then the estimate of β_0 would simply be the arithmetic average of the estimated effects in the literature. Thus, $\hat{\beta}_0$ is an estimate of the overall effect, and the PET tests $\hat{\beta}_0$ for statistical significance, correcting for publication bias.

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).

This study examines 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.

We use Monte Carlo experiments to demonstrate that the FPP procedure does not perform well in the kind of data environments likely to be encountered in economics and business. Section II describes our experimental design. 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 presents the main conclusions from this research.

II. EXPERIMENTAL DESIGN

The data generating process. The conceptual design for our Monte Carlo experiments is based on Reed (2015). There is an infinite population of "studies" *i.* Each study contains multiple regression equations *r*. The data generating process (DGP) that produces the individual observations *t* for regression equation *r* in study *i* is given by:

(1) $y_{irt} = \mu_{ir} + \alpha_{ir} x_{irt} + \varepsilon_{irt}.$

The population effect of *x* on *y* in this regression equation is represented by α_{ir}. We assume that the population effects for the different regression equations in a given study *i* share a common component, α_i, but differ according to a normally distributed random component.

(2) $\alpha_{ir} \sim N(\alpha_i, \sigma_1^2)$.

Likewise, the study-specific, common components α_i are randomly drawn from a population having mean α, so that:

(3) $\alpha_i \sim N(\alpha, \sigma_2^2)$.

We set values for σ_1^2 and σ_2^2 such that $var(\alpha_{ir}|\alpha_i) < var(\alpha_i)$.[1] In other words, while population effects differ both within and across studies, they are more similar within studies. Note that the "overall mean effect" is given by α. This is the parameter that researchers attempt to estimate via meta-analysis.

The above experimental design intends to capture the fact that studies typically contain more than one estimate of a given "effect," perhaps because separate regressions are estimated for different subsamples of the data, or because the regression equations differ in their specifications or econometric procedures used. Thus, a realistic study of meta-analysis performance should incorporate this feature. We call this experimental design "Panel Random Effects" (PRE). This nomenclature is admittedly confusing for those who are familiar with the terms "fixed" and "random" effects from panel data econometrics.

[1] In our experiments, σ_1^2 and σ_2^2 are set equal to 0.25 and 4, respectively.

In meta-analysis, "Fixed Effects" (FE) is the assumption that all studies are characterized by the same population effect. This can be represented by the following variant on Equations (2) and (3):

(4a) Fixed Effects: $\alpha_{ir} \sim N(\alpha_i, 0)$, $\alpha_i \sim N(\alpha, 0)$.

In contrast, "Random Effects" embodies the idea that the population effects underlying different studies are different:

(4b) Random Effects: $\alpha_{ir} \sim N(\alpha_i, 0)$, $\alpha_i \sim N(\alpha, \tau^2)$,

where τ^2 is the variance of the population effect across studies. In keeping with the meta-analysis literature, we adopt this nomenclature, and append "Panel" to "Random Effects" to indicate that studies have more than one estimate.

The pre-publication bias sample. From the infinite population of studies *i*, our design conceptualizes that there is a finite number of studies (*N*) that are actually undertaken. Each of these studies *i* has R_i regression equations, each of which has an associated estimated effect, $\hat{\alpha}_{ir}$, so that there are $\sum_{i=1}^{N} R_i$ total estimated effects. However, the full sample of total estimated effects is not generally observable to the meta-analyst.

The post-publication bias sample. Due to publication bias, many of the estimates in the pre-publication bias sample never see the light of day. Maybe the respective researcher never writes up the results in a formal paper because the results are statistically insignificant or wrong-signed. Maybe he/she does write up the results, but the paper never gets published. In any case, the meta-analyst only observes estimates $\{\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_M\}$, $M \leq \sum_{i=1}^{N} R_i$. This post-publication bias sample is the sample that the meta-analyst uses when implementing the FPP procedure.

Monte Carlo experiments. We design six classes of Monte Carlo experiments to study the performance of the FPP procedure. These are comprised of three types of DGPs and two types of publication bias. The three types of DGPs are (i) Panel Random Effects (PRE), where

studies have more than one estimated effect and the underlying population effect differs both within and across studies, (ii) Random Effects (RE), where each study only has one estimate and the population effect differs across studies, and (iii) Fixed Effects (FE), where each study only has one estimate and the population effect is the same for all studies.[2]

The two types of publication bias are (i) publication bias against statistical insignificance, and (ii) publication bias against wrong-signed estimates. In the latter case, we assume that the correct sign is positive, so negative estimated effects have a harder time getting published. Publication bias isn't absolute. We allow some estimates that are statistically insignificant/wrong-signed to get "published" with positive probability of approximately 10%.[3]

The six classes of experiments are created by forming all possible pairs of DGP and publication bias types (Fixed Effects/bias against insignificance; Fixed Effects/bias against wrong signs; etc.). Within each class of experiments there are nine individual experiments, corresponding to nine different values of the overall mean effect, $\alpha = 0.0, 0.5, 1.0, 1.5, \ldots, 4.0$. Each individual experiment creates 10,000 meta-analysis ("post-publication bias") samples. Each of these samples is subjected to the FPP procedure described above. Thus, for each class of experiment, and for each value of α, we have 10,000 tests for publication bias (FAT), 10,000 tests for the existence of a non-zero overall true effect (PET), and 10,000 estimates of α.

[2] Reed, Florax, and Poot (2015) provide greater detail about the three types of DGPs. In addition to differences in the variances of the population effects α_{ir} in Equation (1), there are also differences in the variances of the error terms, ε_{irt}, across the three types of DGPs. For example, $var(\alpha_{ir}) = 0$ in the FE case. In order to keep the distribution of estimates for the post-publication bias samples approximately the same across DGPs, the $var(\varepsilon_{irt})$ must be larger in the FE DGP relative to the RE DGP. Reed, Florax, and Poot (2015) also provide detail about the distribution of effects for the post-publication bias samples for each of the three DGPs, including example funnel plots. DGP characteristics were chosen so that the post-publication bias samples had the following 4 characteristics to make them "realistic": (1) a "realistic" range of t-values for the estimated effects; (2) "realistic"-looking funnel plots; (3) the per cent of studies eliminated by publication bias to range between 10 and 90 per cent (so all the meta-analysis samples were impacted by publication bias to some degree); and (4) have "realistic" values of "effect heterogeneity".

[3] For the FE and RE cases, an insignificant/wrong-signed estimate was given a 0.10 probability of surviving to the post-publication bias sample. For the PRE case, 7 of the study's 10 estimates had to be significant/correctly signed.

These 10,000 estimates of α are then compared to the estimates from two common meta-analysis estimators. The first is a WLS estimator that uses weight $\omega_j = \left(\frac{1}{SE_j}\right)$ like the FPP procedure, but only includes the constant term (β_0) in the regression equation, without adding SE_j as an explanatory variable. It does not correct for publication bias. The second estimator also does not correct for publication bias, but uses 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.

The first estimator assumes there is only one population effect underlying all the estimates in the meta-analyst's sample. Thus the only source of variation across estimates is sampling error, represented by SE_j. The second estimator allows for the additional variance due to heterogeneous population effects. We denote these two estimators by WLS-FE and WLS-RE, as these accord with the meta-analytic concepts of "Fixed Effects" and "Random Effects".

The subsequent analysis evaluates the performance of the FPP procedure as α increases. Increasing α has different consequences for the two types of publication bias. With publication bias against insignificance, there is no bias when $\alpha = 0$ because an equal share of positive and negative estimates are eliminated. As α increases, a disproportionate share of negative estimates are eliminated, producing a positive bias. In contrast, when publication bias targets negative coefficients, the bias is maximized when $\alpha = 0$. As α increases and the distribution of estimates shifts to the right, fewer negative estimates are eliminated, and the bias gets smaller. In both cases, publication bias disappears as α becomes sufficiently large and all estimates become significant/correctly signed.

III. CHARACTERISTICS OF THE SIMULATED DATA

Implicit in the conduct of Monte Carlo experiments is the assumption that the results have external validity. As a result, it is important to demonstrate that the artificial meta-analysis datasets created by our experimental design "look like" the kinds of samples actually encountered in practice.

This section presents a statistical picture of an "average" simulated meta-analysis sample. We do this for each of the three DGPs: Fixed Effects (FE), Random Effects (RE), and Panel Random Effects (PRE). In each case, we simulate 1000 meta-analysis samples and average the sample characteristics across the samples. To facilitate comparison, we set the same value of α for each of the three DGPs ($\alpha = 1$).

TABLE 1 reports average sample characteristics for the FE DGP. The top panel reports the characteristics of the full sample of 1000 estimated effects before publication bias is imposed ("Pre-Publication Bias"). The next two panels report the characteristics for meta-analysis samples after the estimates have been filtered for statistical significance and having the correct coefficient sign, respectively.

As would be expected in the absence of publication bias, when $\alpha = 1$, the (average) median value of estimated effect in the full sample is 1.00. Estimated effects range from an average minimum of -6.85 to an average maximum of 8.92. *t*-statistics range from an average minimum of -2.69 to an average maximum of 45.62. The median *t*-value in the full sample is, on average, statistically insignificant. This sample, however, is unobserved.

The meta-analyst only observes estimated effects after they have passed through the publication bias filter. When $\alpha = 1$ and publication bias targets statistical insignificance, the average meta-analysis sample shrinks to 318 estimated effects. The corresponding median estimated effect is 1.18 (representing a bias of 18%), and the average median *t*-statistic has gone from 0.94 in the full (unobserved) sample to 2.60, and is now statistically significant.

Similar results obtain when publication bias is directed against wrong-signed coefficients, though the median *t*-statistic is, on average, not so large as to be significant.

Note that when $\alpha = 1$, both types of publication bias disproportionately weed out negative estimated effects, inducing a positive bias in both estimated effects and *t*-statistics in the post-publication bias samples. Further, both post-publication bias samples look "reasonable." The estimated range of *t*-statistics/precisions are comparable to those researchers encounter in economics and business subject areas.

TABLES 2 and 3 give average sample characteristics for the RE and PRE DGPs, respectively. The associated parameter values have been chosen to produce a similar range of estimated effects and *t*-statistics. The additional sample characteristic added to the tables is a measure of effect heterogeneity, I^2.

I^2 takes values between 0 and 1 and measures the share of variation in the estimated effects that is not attributed to sampling error (Higgins and Thompson, 2002). For the RE and PRE DGPs, the average median I^2 value when publication bias is against insignificance is 93% and 92%, respectively. The corresponding values in the simulated meta-analysis samples when publication bias targets wrong-signed estimates are 79% and 74%, respectively. As discussed below, I^2 values of 70-95% are commonly encountered in economics and business research. In summary, the simulated meta-analysis samples that we use for analysing the performance of the FAT-PET-PEESE "look like" the kinds of meta-analysis samples that researchers apply these procedures to in practice.

IV. RESULTS

TABLE 4 reports the results of the Funnel Asymmetry and Precision Effect Tests.[4] As noted above, there are six classes of experiments based on the pairing of (i) type of DGP (FE, RE,

[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 .

PRE), and (ii) type of publication bias. The table is divided vertically into three panels according to type of DGP, from least (FE) realistic 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 true overall effect, α.

We start with the basic case of FE, where each study produces only one estimate and there is one population effect underlying all studies. Each cell in the table reports the results of testing the 10,000, simulated meta-analysis samples. Each meta-analysis sample starts with 1,000 estimates, but not all of these are observed by the meta-analyst due to publication bias.[5] For example, when $\alpha = 0$ and publication bias is directed against insignificance (cf. left side of the table), the average meta-analysis sample contains 143 studies/estimates (14.3 percent).

Each of these 10,000 meta-analysis samples is tested for publication bias (FAT). As discussed above, under publication bias against insignificance, when $\alpha = 0$, there is no bias in the estimate of the overall effect, so that the null hypothesis is true. The FAT performs very well in this case, producing a rejection rate of 6 percent -- close to its 5 percent significance level. In contrast, the PET is oversized with a 16 percent rejection rate. Both the FAT and the PET show excellent power. Rejection rates for the null hypotheses of no publication bias and no effect are 100% whenever $\alpha > 0$.

Continuing with publication bias against insignificance (left side of the table), we move down a panel to the more realistic case of RE. While the rejection rates of 0.08 for both the FAT and PET are close to their significance levels when $\alpha = 0$, the tests do not perform as well when $\alpha > 0$. For example, when $\alpha = 0.5$, the FAT rejects the (false) null of no publication bias only about 33 percent of the time. The PET fails to reject the (false) null of no effect

[5] For the FE and RE DGPs, there is one estimate per study. For the PRE DGP, there are 100 studies, each containing 10 estimates.

approximately 35 percent (=1 – 0.65) of the time. While the performances of the FAT and PET generally improve as α increases, the tests are not as reliable as they were in the FE case.

The bottom panel reports results for the most realistic case, PRE, where studies contain more than one estimate and there is heterogeneity in true effects both within and across studies. Both the FAT and the PET perform substantially worse.[6] When $\alpha = 0$ and publication bias is directed towards statistical insignificance, the FAT rejects the (true) null of no publication bias over half the time (55 percent). The PET rejects the true null 29 percent of the time, and that rejection rate increases only slowly as α gets larger.

Moving to the right side of the table and beginning again with the top panel, we see that the FAT again does well with FE. When publication bias is directed against negatively signed estimates and $\alpha = 0$, so that approximately half of all estimates are wrong signed, the FAT rejects the null of no publication bias 100 percent of the time. As α increases, fewer and fewer estimates are eliminated via publication bias, so that publication bias diminishes. Correspondingly, the rejection rate from the FAT also falls.

The PET also performs well. When $\alpha = 0$, 45 percent (= 100 - 55) of the estimates are eliminated because of negative signs. This causes the remaining estimates to have a strong positive bias. Even so, the PET is not fooled, and generally leads to the correct conclusion of no effect: The rejection rate of 8 percent is close to its 5 percent significance level. Further, the PET accurately identifies the existence of a nonzero effect 100 percent of the time for all $\alpha > 0$.

As before, the performances of the FAT and PET decline as the experimental designs become more realistic. Compared to the 100 percent rejection rate for the FAT when $\alpha = 0$ in the FE case, the FAT falls to 62 percent for the same scenario when the DGP is RE. Likewise,

[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.

the PET finds evidence of an effect 90 percent of the time under RE when there is no effect ($\alpha = 0$). Things decline further still in the most realistic case of PRE. The FAT is largely insensitive to changes in the degree of publication bias, and the ability of the PET to identify an effect when there really is one is worse. In summary, while the FPP procedure does very well in the basic, unrealistic case of FE, its performance declines substantially in more realistic data environments.

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 overall effect. In the context of our experiments, that means using $\hat{\beta}_0$ to estimate α.

TABLE 2 reports the results of estimating β_0 following the FPP procedure represented in FIGURE 2. It also reports the results from estimating β_0 using two WLS estimators that do not correct for publication bias (WLS-FE, WLS-RE). As we are interested in how the respective estimators perform in realistic data environments, we focus on the PRE experiments, and study both types of publication bias. Each experiment produces three sets of 10,000 $\hat{\beta}_0$ values, one set for each of the three estimators.

Three measures of performance are calculated: (i) mean value of $\hat{\beta}_0$, (ii) mean-squared error (MSE) for the 10,000 $\hat{\beta}_0$ values, and (iii) Type I error rates, where the null hypothesis is $H_0: \beta_0 = \alpha$. Good performance is measured, respectively, by (i) a mean value for $\hat{\beta}_0$ close to α; (ii) a relatively small MSE value, and (iii) Type I error rates close to their 5 percent significance levels.[7]

[7] Clustered robust standard errors were used when testing hypotheses with the FPP estimator. Heteroskedasticity-robust standard errors were used for the WLS-FE and WLS-RE estimators.

The results can be easily summarized. First, FPP produces $\hat{\beta}_0$ values that are as close or closer to α than those produced by the two WLS estimators that do not correct for publication bias. For example, when $\alpha = 4.0$, the FPP procedure produces a mean $\hat{\beta}_0$ value of 4.11, compared to 4.20 and 4.77 for the WLS-FE and WLS-RE estimators. However, while the FPP procedure produces the "best" estimates according to this measure of performance, it still suffers from a substantial bias in some cases.

Interestingly, superior performance on the first moment of the distribution does not translate into greater efficiency.[8] When $\alpha = 0$ and publication bias is targeted towards insignificance, the MSE associated with the FPP estimates is 1.629 versus 0.874 and 0.443 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 ($\alpha = 0$/publication bias against wrong signs).

Finally, as was foreshadowed by the PET results in TABLE 1, the FPP procedure does not do well when testing hypotheses about α. Sometimes it performs better than the WLS estimators, and sometimes worse. But the results are sufficiently poor that hypothesis testing about α should not be relied upon for any of the estimators.

V. COMPARISON TO PREVIOUS STUDIES

Several other studies have examined various aspects of FPP performance in the presence of publication bias. The closest to this study in design is Reed, Florax, and Poot (2015). RFM use the same simulation DGPs as this paper but do not examine the performance of FAT and PET, and do not study the conditional PET-PEESE approach for estimating average effect sizes.

Stanley (2008) studies the performance of the FAT and PET, estimates mean bias and MSE using a hybrid PET approach, and calculates coverage rates for the hybrid PET approach.

[8] Figure 3 in Reed (2015) illustrates a case where the estimator with mean closest to the true value is less efficient.

He considers publication bias related to statistical significance, but not due to wrong signs. His "headline" conclusion is that meta-regression, which includes the FAT and PET, generally performs well in the presence of publication bias:

> 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).

Stanley and Doucouliagos (2014a) use the same simulation DGPs as Stanley (2008) but specifically examine the FPP procedure. They compare the PET-PEESE approach to various Taylor polynomial approximations to the conditional mean of a truncated distribution, which is the distribution of effect sizes one would observe if insignificant effect sizes were not "published" and went unobserved. The approaches are compared on the dimensions of Mean and Mean Squared Error (MSE). They also compare the PEESE and PET-PEESE approaches with the unadjusted OLS, Fixed Effects (FE) and Random Effects (RE) estimators. They conclude that the FAT-PET-PEESE approach is superior.

> A quadratic approximation without a linear term, precision effect estimate with standard error (PEESE), is shown to have the smallest bias and mean squared error in most cases and to outperform conventional meta-analysis estimators, often by a great deal. Monte Carlo simulations also demonstrate how a new hybrid estimator that conditionally combines PEESE and the Egger regression intercept can provide a practical solution to publication selection bias. (Stanley and Doucouliagos, 2014a, p. 60).

Moreno et al. (2009) compare a large number of estimators of effect sizes consisting of various regression-based methods along with nonparametric, "Trim and Fill" methods. They generally conclude that regression methods, which include Fixed Effects, Random Effects, and PET and PEESE approaches, perform better than "Trim and Fill" methods.[9] In comparing the PET and PEESE approaches with the FE and RE estimators, the former generally have both

[9] The PET and PEESE approaches are identified as FE-se and FE-var in their paper.

better coverage rates and smaller MSE values than the latter, unadjusted meta-analysis estimators.

Thus, on the face of it, it looks as if our results stand in stark contrast to earlier studies. However, on closer examination, our results are not so much in conflict. A key issue is that of "excess heterogeneity", the variation in effect sizes that is not attributable to sampling error, and which is measured by I^2. All three studies above emphasize that the successful performance of the FPP procedure assumes that there is not "excessive heterogeneity."

Moreno et al. (2009, p. 12) write: "The overall performance of all the methods deteriorates as I^2 exceeds 50%." And Stanley (2008, p. 123) writes:

> Naturally, there are limits to the robustness of these tests…The relative magnitude of misspecification bias [i.e., heterogeneity] is the key parameter in assessing the vulnerability of these meta-regression methods to type I error inflation. … Thus, FAT and PET need to be interpreted carefully when there remains large unexplained variation…"

Additionally, Stanley and Doucouliagos (2014a, p. 75) write:

> …overwhelming unexplained heterogeneity can invalidate the underlying meta-regression tests (i.e., the PET) (Stanley, 2008). … when unexplained heterogeneity is responsible for more than 90% of the observed variation among reported research results, uncorrected publication biases will expand greatly."[10]

In economics and business, "excessive heterogeneity" is the norm. While studies only infrequently report I^2 values, when they do, the reported values are almost always larger than 50%, and frequently larger than 90%. In this context, Stanley and Doucouliagos (2014b) write:

> For example, among minimum wage elasticities, I^2 is 90% (Doucouliagos and Stanley, 2009); it is 93% among estimates of a statistical life (Doucouliagos, Stanley and Giles, 2012) and 97% among the partial correlations of CEO pay and corporate performance (Doucouliagos, Haman, and Stanley, 2012).

Recent articles that report $I^2 > 90\%$ come from a wide diversity of areas in economics and business: returns to education (Rodriguez and Muro, 2015); determinants of innovation

[10] Though it should be noted that they go on to say that "the methods advanced here will remain a marked improvement over conventional meta-analytic summary statistics" even in the presence of substantial heterogeneity.

(Sarooghi et al., 2015); marketing (Eisend, 2015); racial discrimination and the workplace (Triana et al., 2015); microcredit and development (Chliova et al., 2015); FDI and economic growth (Gunby et al., 2016); team training in health care (Hughes et al., 2016); the adoption and use of technology in the workplace (Khechine et al., 2016); and utility-based models of pulmonary disease (Moayeri et al., 2016).[11]

This study also notes a substantial decline in the performance of the FPP approach as heterogeneity increases. The simulated, FE post-publication bias samples are similar in most characteristics to the RE and PRE samples (compare TABLES 1 and 2), except that the only heterogeneity across estimated effects is due to sampling error, so that $I^2 = 0$. This increase in heterogeneity from FE to RE is the likely explanation for the decline in FPP performance.

However, excessive heterogeneity is not the only explanation for the subsequent poor performance of the FPP procedure. The median I^2 values in the post-publication bias samples in the RE and PRE cases are quite close (compare TABLES 2 and 3), and yet FPP performance is distinctly worse in the PRE samples. The reason for this is not clear and is a topic for future research.

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.

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,

[11] In some cases, Q-statistics are reported rather than I^2 values. These can be converted to the latter using the formula $I^2 = \left(\frac{Q-df}{Q}\right) \times 100\%$.

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 cannot not be trusted.

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. The second conclusion we draw from our study is that future research should more intensively explore the conditions under which FPP performs well. As noted elsewhere (Stanley, 2008; Moreno et al., 2009; Stanley and Doucouliagos, 2014a), publication bias is a very serious problem and the FPP procedure has shown great promise in mitigating its deleterious consequences. Having a better understanding of where the FPP procedure can be successfully applied is an important topic for future research.

REFERENCES

Chliova, M., Brinckmann, J., and Rosenbusch, N. (2015). Is microcredit a blessing for the poor? A meta-analysis examining development outcomes and contextual considerations. *Journal of Business Venturing, 30*(3), 467-487.

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.

Doucouliagos, H., Haman, J., and Stanley, T.D. (2012). Pay for performance and corporate governance reform. *Industrial Relations*, *51*, 670- 703.

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. and Stanley, T.D. (2009). Publication selection bias in minimum-wage research? A meta-regression analysis, *British Journal of Industrial Relations, 47*, 406-29.

Doucouliagos, H., Stanley, T.D., and Giles, M. (2012). Are estimates of the value of a statistical life exaggerated? *Journal of Health Economics, 31,* 197-206.

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.

Eisend, M. (2015). Have we progressed marketing knowledge? A meta-analysis of effect sizes in marketing research. *Journal of Marketing, 79*(3), 23-40.

Gunby, P., Jin, Y., and Reed, W.R. (2016). Did FDI really cause Chinese economic growth? A meta-analysis. *World Development, forthcoming*.

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.

Hughes, A.M., Gregory, M.E., Joseph, D.L., Sonesh, S.C., Marlow, S.L., Lacerenza, C.N., Benishek, L.E., King, H.B., and Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology*, *101*(9), 1266-1304.

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

Khechine, H., Ndjambou, P., and Lakhal, S. (2016) A meta-analysis of the UTAUT model: Eleven years later. *Canadian Journal of Administrative Sciences, 33,* 138-152.

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

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.

Moayeri, F., Hsueh, Y.S., Clarke, P., and Dunt, D. (2016). Do model-based studies in chronic obstructive pulmonary disease measure correct values of utility? A meta-analysis. *Value in Health, 19*(4), 363-373.

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

Rodríguez, J. and Muro, J. (2015). On the size of sheepskin effects: A meta-analysis. *Economics: The Open-Access, Open-Assessment E-Journal, 9,* (2015-37): 1—18. http://dx.doi.org/10.5018/economics-ejournal.ja.2015-37

Sarooghi, H., Libaers, D., and Burkemper, A. (2015). Examining the relationship between creativity and environmental factors. *Journal of Business Venturing, 30*(5), 714-731.

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

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

Stanley, T.D., and Doucouliagos, H. (2014b). 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.

Triana,., Jayasinghe, M., and Pieper, J. (2015). Perceived workplace racial discrimination and its correlates: A meta-analysis. *Journal of Organizational Behavior, 36*(4), pages 491-513.

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

Variable	*Median*	*Minimum*	*P5%*	*P95%*	*Maximum*
PRE-PUBLICATION BIAS (100 percent of estimates):					
Estimated effect	1.00	-6.85	-1.99	3.99	8.92
t-statistic	0.94	-2.69	-0.96	6.08	45.62
PUBLICATION BIAS AGAINST INSIGNIFICANCE (31.8 percent of estimates):					
Estimated effect	1.18	-6.74	-0.84	5.19	8.86
t-statistic	2.60	-2.69	-0.45	14.95	45.44
PUBLICATION BIAS AGAINST NEGATIVE EFFECTS (80.5 percent of estimates):					
Estimated effect	1.20	-4.61	0.11	4.26	8.92
t-statistic	1.27	-1.89	0.07	7.33	45.45

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

Variable	*Median*	*Minimum*	*P5%*	*P95%*	*Maximum*
PRE-PUBLICATION BIAS (100 percent of estimates):					
Estimated effect	1.00	-6.24	-2.38	4.37	8.25
t-statistic	0.79	-8.12	-1.48	5.98	31.79
I2	0.84	0.65	0.74	0.92	0.95
PUBLICATION BIAS AGAINST INSIGNIFICANCE (32.9 percent of estimates):					
Estimated effect	1.81	-5.90	-2.09	5.64	8.26
t-statistic	2.55	-8.15	-2.29	12.77	31.50
I2	0.93	0.72	0.87	0.97	0.99
PUBLICATION BIAS AGAINST NEGATIVE EFFECTS (74.7 percent of estimates):					
Estimated effect	1.55	-3.50	0.01	4.76	8.30
t-statistic	1.28	-2.93	0.01	7.40	31.57
I2	0.79	0.41	0.63	0.90	0.95

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

Variable	*Median*	*Minimum*	*P5%*	*P95%*	*Maximum*
PRE-PUBLICATION BIAS (100 percent of estimates):					
Estimated effect	0.96	-7.50	-3.47	5.47	9.43
t-statistic	0.67	-10.16	-3.17	7.44	20.69
I2	0.86	0.46	0.68	0.97	0.99
PUBLICATION BIAS AGAINST INSIGNIFICANCE (21.8 percent of estimates):					
Estimated effect	2.24	-3.03	-2.18	5.73	7.03
t-statistic	3.62	-8.81	-6.08	15.56	20.66
I2	0.92	0.00	0.71	0.99	1.00
PUBLICATION BIAS AGAINST NEGATIVE EFFECTS (56.2 percent of estimates):					
Estimated effect	2.22	-3.99	-0.84	6.17	9.33
t-statistic	1.75	-2.11	-0.50	11.28	20.66
I2	0.74	0.11	0.43	0.94	0.98

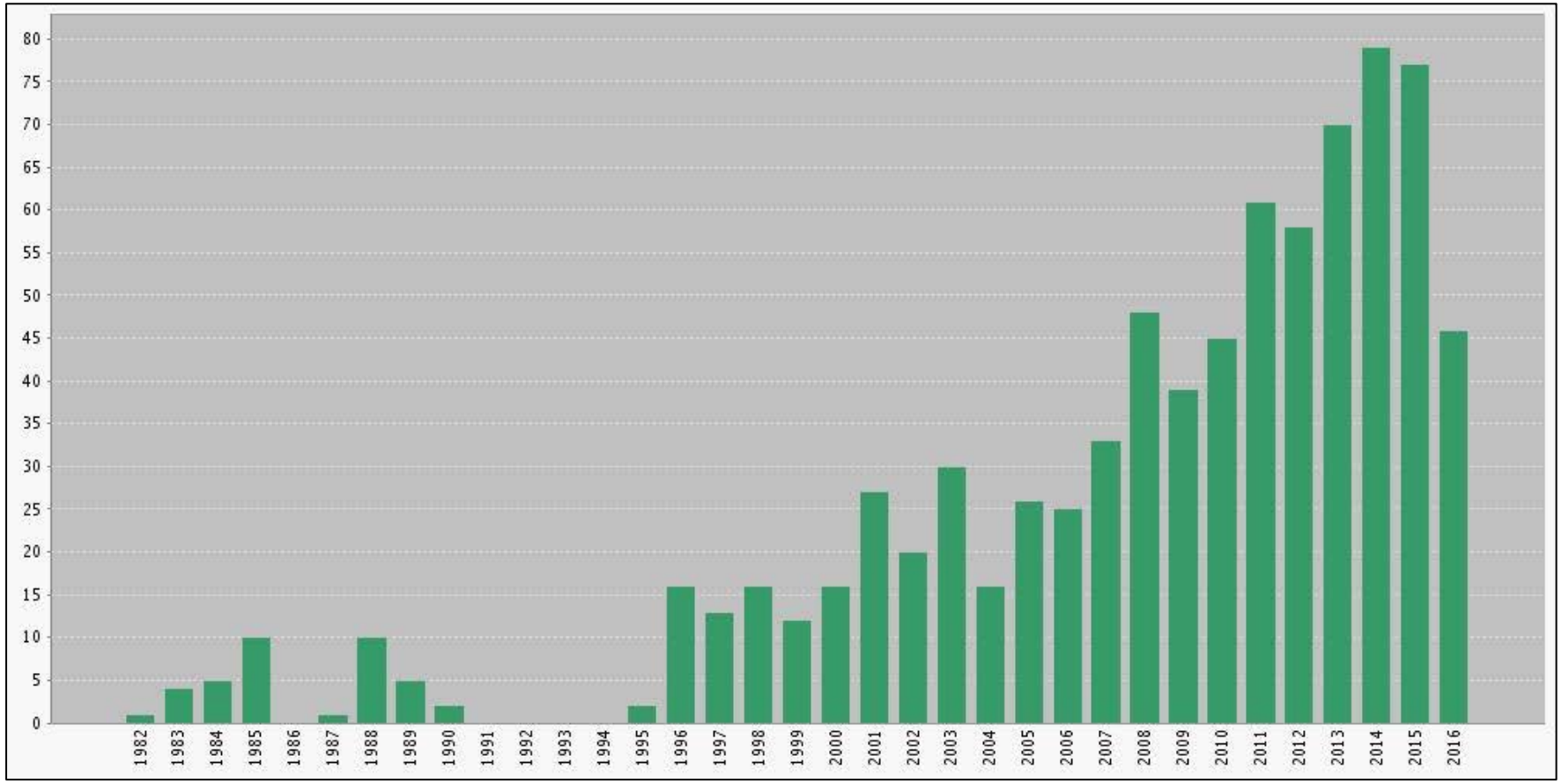
TABLE 4
Funnel Asymmetry Tests (FAT) and Precision Effect Tests (PET)

	Publication Bias against Insignificance			Publication Bias against Wrong Sign		
FIXED EFFECTS (FE)						
α	***Percent***	***FAT***	***PET***	***Percent***	***FAT***	***PET***
0.0	14.3	0.06	0.16	55.0	1.00	0.08
0.5	23.0	1.00	1.00	71.7	1.00	1.00
1.0	31.8	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	67.0	1.00	1.00	97.0	0.53	1.00
4.0	72.2	1.00	1.00	98.0	0.30	1.00
RANDOM EFFECTS (RE)						
α	***Percent***	***FAT***	***PET***	***Percent***	***FAT***	***PET***
0.0	27.1	0.08	0.08	55.0	0.62	0.90
0.5	28.7	0.33	0.65	65.4	0.62	1.00
1.0	33.0	0.67	0.99	74.7	0.56	1.00
1.5	39.1	0.79	1.00	82.0	0.48	1.00
2.0	45.9	0.79	1.00	87.4	0.35	1.00
2.5	52.8	0.75	1.00	91.3	0.24	1.00
3.0	59.2	0.69	1.00	94.0	0.17	1.00
3.5	65.1	0.61	1.00	95.9	0.13	1.00
4.0	70.4	0.55	1.00	97.2	0.10	1.00
PANEL RANDOM EFFECTS (PRE)						
α	***Percent***	***FAT***	***PET***	***Percent***	***FAT***	***PET***
0.0	19.2	0.55	0.29	38.4	0.45	0.78
0.5	19.9	0.58	0.34	47.7	0.47	0.84
1.0	22.0	0.66	0.46	56.8	0.44	0.87
1.5	25.2	0.72	0.60	65.6	0.43	0.91
2.0	29.5	0.66	0.73	73.6	0.46	0.93
2.5	34.7	0.59	0.83	80.6	0.50	0.95
3.0	40.4	0.66	0.90	86.2	0.46	0.97
3.5	46.4	0.69	0.94	90.6	0.45	0.98
4.0	52.8	0.65	0.97	93.9	0.41	0.99

TABLE 5
Comparison of FPP Estimates with WLS-FE and WLS-RE: Panel Random Effects

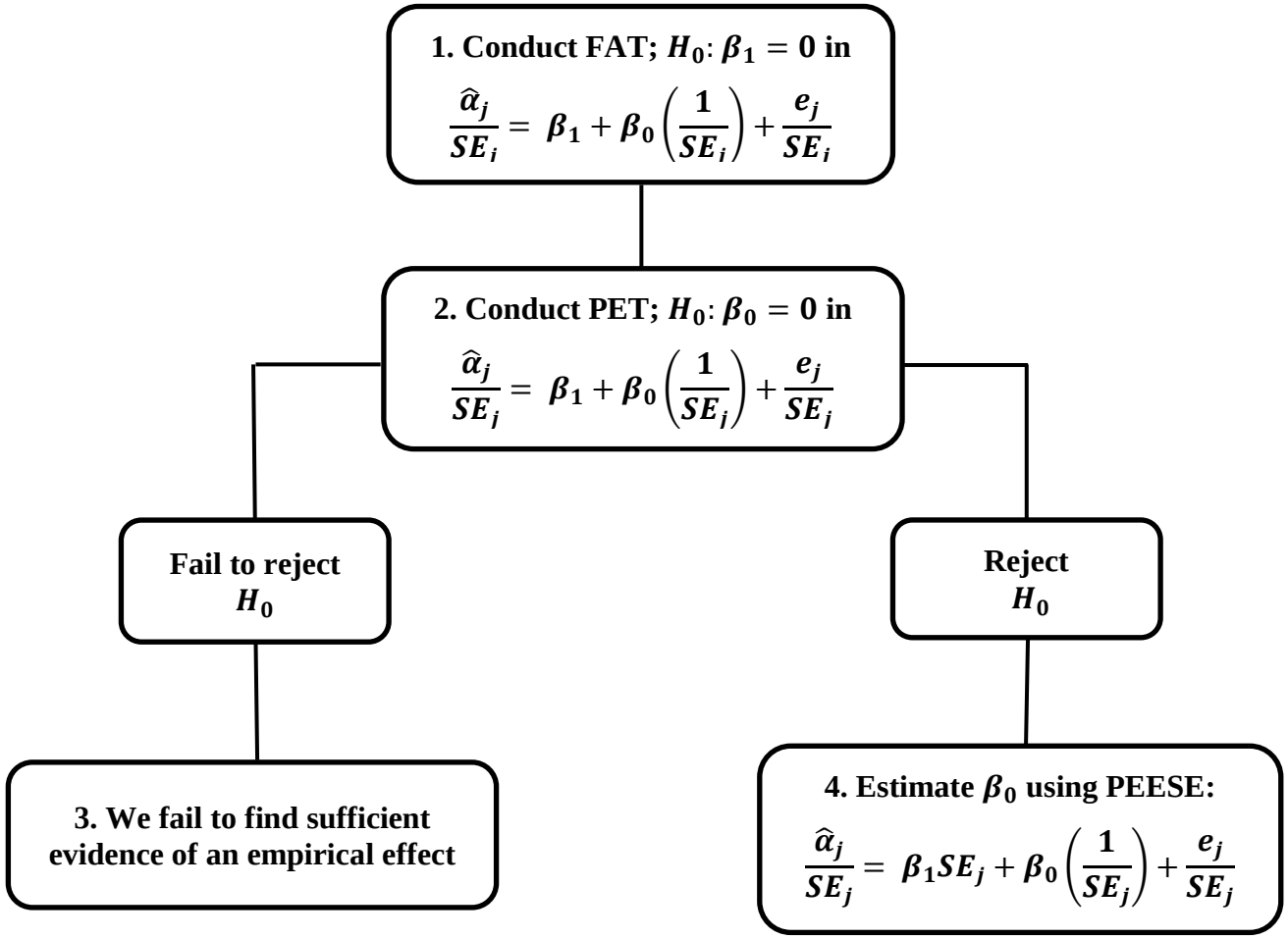
	Publication Bias against Insignificance			**Publication Bias against Wrong Sign**		
MEAN VALUE OF $\widehat{\beta}_0$						
α	***FPP***	***WLS-FE***	***WLS-RE***	***FPP***	***WLS-FE***	***WLS-RE***
0.0	0.01	0.01	0.01	1.69	1.77	1.88
0.5	0.59	0.66	1.01	1.89	1.94	2.07
1.0	1.18	1.29	1.90	2.12	2.17	2.29
1.5	1.71	1.85	2.59	2.38	2.41	2.53
2.0	2.21	2.37	3.13	2.64	2.68	2.80
2.5	2.73	2.86	3.60	2.98	3.01	3.11
3.0	3.20	3.31	4.00	3.34	3.36	3.45
3.5	3.66	3.76	4.39	3.72	3.73	3.82
4.0	4.11	4.20	4.77	4.13	4.14	4.22
MEAN SQUARED ERROR						
α	***FPP***	***WLS-FE***	***WLS-RE***	***FPP***	***WLS-FE***	***WLS-RE***
0.0	1.629	0.874	0.443	3.591	3.414	3.592
0.5	1.628	0.879	0.655	2.648	2.388	2.513
1.0	1.577	0.880	1.111	1.997	1.672	1.709
1.5	1.548	0.851	1.387	1.540	1.143	1.106
2.0	1.415	0.782	1.428	1.221	0.796	0.689
2.5	1.338	0.722	1.312	1.084	0.609	0.418
3.0	1.281	0.652	1.094	1.027	0.502	0.245
3.5	1.198	0.577	0.874	1.000	0.445	0.144
4.0	1.138	0.527	0.670	0.989	0.423	0.092
TYPE I ERROR RATES ($H_0: \beta_0 = \alpha$)						
α	***FPP***	***WLS-FE***	***WLS-RE***	***FPP***	***WLS-FE***	***WLS-RE***
0.0	0.25	0.17	0.05	0.78	0.99	1.00
0.5	0.29	0.17	0.14	0.75	0.92	1.00
1.0	0.30	0.19	0.37	0.61	0.77	1.00
1.5	0.31	0.22	0.62	0.48	0.56	1.00
2.0	0.29	0.23	0.80	0.38	0.38	0.98
2.5	0.29	0.23	0.88	0.33	0.28	0.87
3.0	0.29	0.21	0.90	0.30	0.21	0.60
3.5	0.27	0.18	0.89	0.27	0.17	0.34
4.0	0.27	0.17	0.84	0.27	0.16	0.18

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 (813 articles). 2016 articles are current through August, 2016.

FIGURE 2
The FAT-PET-PEESE Procedure



SOURCE: Stanley and Doucouliagos (2012, page 79)