OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 70, 1 (2008) 0305-9049
doi: 10.1111/j.1468-0084.2007.00487.x

Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection*

T. D. Stanley

Department of Economics, Hendrix College, Conway, AR 72032 USA
(e-mail: Stanley@Hendrix.edu)

Abstract

This study investigates the small-sample performance of meta-regression methods for detecting and estimating genuine empirical effects in research literatures tainted by publication selection. Publication selection exists when editors, reviewers or researchers have a preference for statistically significant results. 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. Publication biases are greatly reduced by combining two biased estimates, the estimated meta-regression coefficient on precision (1/*Se*) and the unadjusted-average effect.

> [P]ublication bias is leading to a new formulation of Gresham's law – like bad money, bad research drives out good. – Bland (1988, p. 450)

I. Empirical economics and its publication selection bias

This paper offers a statistical approach to estimating and testing empirical effects in the presence of publication selection and simulates its properties under realistic

*I wish to thank Chris Doucouliagos, Stephen Jarrell, Randall Rosenberger, Alex Sutton, and an anonymous referee for their helpful comments. I also gratefully acknowledge the support of a US Environmental Protection Agency STAR (Science To Achieve Results) grant #RD-832-421-01. Although the research has been funded in part by the US-EPA, it has not been subjected to the Agency's peer and policy review and therefore does not necessarily reflect the views of the Agency. Any remaining error or omission is solely my responsibility.

JEL Classification numbers: C12, C13, B40.

research conditions. Publication bias has long been recognized as another serious threat to empirical economics (De Long and Lang, 1992). More recently, Card and Krueger (1995), Ashenfelter, Harmon and Oosterbeek (1999), Görg and Strobl (2001), Doucouliagos, Laroche and Stanley (2005), Abreu, de Groot and Florax (2005), Doucouliagos (2005), Nijkamp and Poot (2005), Rose and Stanley (2005), and Stanley (2005a) have all used meta-regression analysis (MRA) to uncover evidence of publication bias in specific areas of economic research. Publication bias, or the 'file drawer problem', is the consequence of choosing research papers for the statistical significance of their findings. 'Statistically significant' results are often treated more favourably by researchers, reviewers and/or editors; hence, larger, more significant, effects are over-represented. Studies with small, 'insignificant' effects will tend to remain in the 'file drawer' (Rosenthal, 1979). Publication selection biases a literature's average reported empirical effect away from zero.[1] This bias is a problem for any summary of empirical research, including narrative literature reviews (Laird and Mosteller, 1988; Phillips and Goss, 1995; Sutton *et al.*, 2000a; Stanley, 2001).[2]

Econometric estimates can easily be overwhelmed by publication selection because there are so many plausible econometric models to choose from. Conventional literature reviews and econometric techniques are powerless to address publication bias. If, for example, only half the studies select the results that they report, the average estimates across a literature and the proportion of studies that find a significant effect will be dominated by publication bias, irrespective of the underlying empirical 'truth'. Such an empirical literature is indistinguishable, by conventional econometric methods, from a literature that contains an authentic effect and yet refrains from publication selection. Therefore, current econometric methodology cannot reliably assess the empirical merit of any economic hypothesis. Issues of publication selection, its identification and circumvention are crucial to a genuinely empirical economics.

Without some correction for publication bias, a literature that appears to contain a large empirical effect offers little, if any, reason for accepting this effect. At best, conventional narrative reviews serve as vote-counts of the number of studies that find a significant effect vs. those that do not (Stanley, 2001, pp. 144–146). When there is no authentic effect, but only publication selection, the expected proportion of research studies that will report a significant effect is: $\phi + (1 - \phi)\alpha$; where ϕ = the incidence of publication selection (i.e. the proportion of studies that choose to report only significant effects), and α is a conventional significance level (0.05). Thus, even if a minority of the reported effects is the result of selection, the majority of the literature can be expected to report a significant effect. In particular, for a research

[1] Stanley (2005a) offers a more detailed discussion of publication selection bias and its effects on empirical economics.

[2] As an anonymous referee points out, publication bias also makes the interpretation from a single study problematic, no matter how well this study is conducted. The advantage of a summary, of course, is that random misspecification and sampling errors are averaged and thereby lessened.

area with $\phi \geq 0.474$ and no actual empirical effect, the probability that a majority vote-count comes to the *wrong* conclusion *increases* with added research.[3]

Other areas of social science widely acknowledge the importance of publication selection and its associated bias (Sterling, 1959; Rosenthal, 1979; Begg and Berlin, 1988). 'Many other commentators have addressed the issue of publication bias.... All agree that it is a serious problem' (Begg and Berlin, 1988, p. 421). Publication selection can make small, practically insignificant effects or random, yet selected, misspecification biases appear to be an authentic empirical phenomenon.[4] At a minimum, the magnitudes of the reported effects found in the literature are likely to be exaggerated. If we are to rely on our assessment of the efficacy of a given social program or on the validity of an economic theory, it is necessary to have access to methods that correct, test or circumvent publication bias.

This paper offers and evaluates several meta-regression methods for testing whether an empirical literature contains reliable evidence of a genuine empirical effect beyond potential contamination from publication selection. Although meta-regression methods of publication bias identification have been used in economics (Card and Krueger, 1995; Ashenfelter *et al.*, 1999; Görg and Strobl, 2001; Abreu *et al.*, 2005; Doucouliagos, 2005; Doucouliagos and Laroche, 2003; Nijkamp and Poot, 2005; Rose and Stanley, 2005; Stanley, 2005a), there has been no systematic evaluation of their properties. This study provides evidence that MRA can help researchers see through the murk of random errors and selected misspecification biases to identify and magnify the underlying statistical structure that characterizes genuine empirical effects.[5] However, if uncorrected, meta-analysis is itself susceptible to the distortion of publication selection. With meta-analysis, at least, statistical methods can be employed to identify and/or accommodate publication selection.

The purpose of this study is to assess the size and the power of several meta-regression tests for empirical effects, to 'correct' the estimated magnitudes for publication selection, and to compare the statistical performance of these meta-regression methods to Hedges' (1992) maximum likelihood, publication selection estimator (MLPSE). The central goal is to develop methods that identify the traces of statistical structure associated with genuine empirical effects, irrespective of publication selection and misspecification bias. In order to ensure that our empirical inferences are reliable, likely sources of error must first be eliminated. In empirical economics, as well as other areas of non-experimental research, published findings are often the result of statistical bias (omitted-variables bias or the misspecification of dynamics, functional forms, etc.) compounded by publication selection (choosing results that support conventional theory or are statistically significant) rather than

[3]Use $\phi + (1 - \phi)\alpha = 0.5$ to solve for ϕ. Similarly, when individual tests have low power, Hedges and Oklin (1985) show that the probability of the majority vote-counting giving the wrong impression increases as research accumulates.

[4]This notion of 'random, yet selected, misspecification biases' is defined in detail in the Appendix.

[5]With publication selection, effects are reported only if they are statistically significant. When insignificant effects are produced, models will be respecified and re-estimated. As a result of this selection and re-estimation, reported results are likely to be little more than random errors and misspecification biases.

any authentic underlying phenomenon. Yet, science progresses as the result of severe empirical testing (Mayo, 1996). To 'learn from error' requires that we eliminate other plausible sources of our observed effects, ensuring that resulting empirical phenomena are more than the artifact of some statistical bias or poor experimental design. I do not explicitly address the publication bias that may arise from ideological or theoretical commitment. However, it is unlikely that the MRA methods offered here will identify a predilection for a given theory as a genuine effect.

This study reports Monte Carlo simulations of the small-sample properties of two meta-regression tests for identifying genuine effects in the presence of publication selection and a third MRA test for publication selection itself: meta-significance testing (MST), precision-effect testing (PET) and funnel-asymmetry testing (FAT). In likely applications, these tests for detecting genuine effects perform quite well even when the incidence of publication selection is severe. However, as the relative magnitude of pervasive misspecification bias increases, MST becomes biased and vulnerable to an inflation of type I error rates. Nonetheless, PET and the combined PET/MST joint test have acceptable type I error rates. Likewise, combining two biased estimates of effect (the unadjusted average reported effect and the meta-regression coefficient on precision) can greatly reduce bias.

II. The meta-regression of publication selection and statistical power

> [L]et $Y(X, Z)$ be the response surface giving the expected treatment effect for an outcome Y as a function of scientifically interesting factors X... and design factors Z.... Letting $Z = Z_0$ represent perfect studies (e.g. infinitely large, perfectly randomized), the objective should be to estimate $Y(X, Z = Z_0)$ by estimating $Y(X, Z)$ from observed studies and extrapolating to Z_0. The required statistical modelling efforts. . . address the underlying *scientific* questions as opposed to the peripheral *publication process* questions. – Rubin (1988, pp. 457–458)

The purpose of MRA is to model, estimate and explain the excess variation among reported empirical results (Stanley, 2001). MRA offers a methodology in which to understand the research process itself and to map the effect of the researchers' choices about data, estimation techniques, and econometric models onto a research literature (Stanley and Jarrell, 1989). If MRA were only to identify which research choices are responsible for the excess variation routinely found among empirical economic estimates (Roberts and Stanley, 2005), it would be a great success. Yet, MRA offers more, an explicit econometric model of statistical power and tests for publication selection and genuine effect corrected for the contaminating effects of publication selection.

Funnel plots

> The simplest and most commonly used method to detect publication bias is an informal examination of a funnel plot. – (Sutton *et al.*, 2000b, p. 1574)

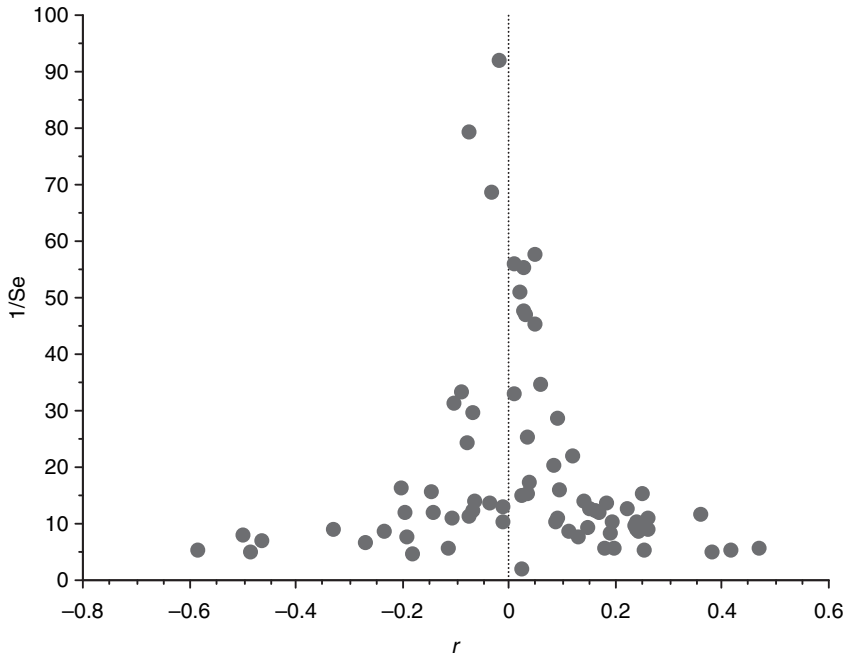

Figure 1. Funnel plot, union-productivity partial correlations

A funnel graph plots precision, the inverse of the standard error (1/Se), vs. non-standardized estimates. Figure 1 is such a funnel graph and plots data from Doucouliagos and Laroche's (2003) meta-analysis of 73 studies of union-productivity effects. In the absence of publication selection, estimates will vary randomly, hence *symmetrically*, around the 'true' effect. Thus, it is the graph's asymmetry that is the key to identifying publication selection.[6] Typically, selection is for negative price elasticity (Dalhuisen *et al.*, 2003; Stanley, 2005a), stationary unemployment rates (Stanley, 2004), positive trade effects from the adoption of a common currency (Rose and Stanley, 2005), or a positive correlation between economic freedom and economic growth (Doucouliagos, 2005). However, heteroscedasticity dictates the expected inverted funnel shape. Studies with less precision and hence larger standard errors are at the bottom of the graph and will produce estimates that are more spread out.

The funnel graph in Figure 1 appears to provide a nearly ideal depiction of the expected inverted funnel shape of a research literature absent publication selection.[7]

[6]When statistically significant effects are selected regardless of their direction, a symmetric, but hollow, funnel graph may result. Though interesting, this type of publication selection bias is not as pernicious because selection biases tend to cancel one another out. In any case, by taking the absolute value of effects in equation (1) below, this type of publication bias may also be tested and corrected (Stanley, 2005a, pp. 317–320).

[7]Yet appearances can be deceiving. This is the reason that more formal statistical tests are needed. In this case, pockets of publication selection bias can be statistically identified even though the graph appears symmetric, more or less (Doucouliagos *et al.*, 2005).

Its inverted funnel shape is unmistakable, but its left side has a somewhat thinner midsection. Visual inspections of funnel graphs are inherently ambiguous and subjective. Fortunately, MRA offers a formal test of the asymmetry of funnel graphs.

FAT-PET: MRA tests for publication selection and genuine effects

> Publication bias, the phenomenon in which studies with positive results are more likely to be published than studies with negative results, is a serious problem in the interpretation of scientific research. – Begg and Berlin (1988, p. 419)

In economics, the standard model of publication selection is the simple MRA between a study's estimated effect and its standard error (Card and Krueger, 1995; Ashenfelter *et al*., 1999; Görg and Strobl, 2001).

$$effect_i = \gamma_1 + \gamma_0 Se_i + \varepsilon_i. \quad (1)$$

Estimated effects will vary randomly around the 'true' effect, γ_1, when there is no publication selection. In contrast, publication bias is correlated with the standard error – $\gamma_0 Se_i$. Begg and Berlin (1988, pp. 431–432) show that publication bias is proportional to the inverse of the square root of the sample size, $n^{-1/2}$, and thus proportional to the standard error. Equation (1) may be derived from statistical theory when all research studies are subjected to publication selection.[8]

Although equation (1) is clearly heteroscedastic, a measure of the heteroscedasticity (Se_i) is readily available. Thus, weighted least squares (WLS) is the obvious method of obtaining efficient estimates.

$$t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i. \quad (2)$$

Note that the independent variable is precision ($1/Se_i$) and that the intercept and slope coefficients are reversed. Equation (2) may now be estimated by ordinary least squares (OLS) and provides a basis for testing both the funnel graph's asymmetry (FAT – funnel-asymmetry test) and also whether there is a genuine effect beyond publication selection (PET – precision-effect test).

Egger *et al.* (1997) offer the conventional *t*-test of the intercept of equation (2), γ_0, as a test for publication bias. The sign of g_0, γ_0's estimate, indicates the direction of this bias. Testing $H_0 : \gamma_0 = 0$ becomes a test of the funnel graph's asymmetry (FAT) (Sutton *et al.*, 2000a). Unfortunately, this test for publication bias is known to have low power (Egger *et al.*, 1997).

[8]Consider the conventional *t*-statistic: $t_i = (effect_i - \gamma_1)/Se_i$, for γ_1 representing the true effect. This will be approximately normal in large samples. If there is strict selection for significance yet no genuine effect, then a study is published only when its t_i exceeds the critical value, t_c. Thus, observed t_is will have a truncated non-central, *t*-distribution. We may define γ_0 as the mean of the large sample approximation to this truncated non-central *t*-distribution. Hence, conditional on this strict selection, $\gamma_0 = [E(effect_i - \gamma_1]/Se_i$, implying equation (1).

However, the MRA model defined by equation (2) has the unexploited potential to identify genuine empirical effects, regardless of publication selection bias. Here, we propose the testing of γ_1 in equation (2) as a test for authentic empirical effects, 'corrected' for publication selection – precision-effect test (PET). According to this model of publication selection, equation (1), as n approaches infinity and as Se goes to zero (recall Rubin's 'perfect studies'), observed effects approach γ_1. Testing $H_0: \gamma_1 = 0$ in equation (2) becomes a test for genuine effects beyond systematic contamination from publication selection. Investigating the properties of this proposed test is the central focus of this paper.

Unfortunately, the MRA regression model (2) has several statistical problems. Because the standard errors (Se_i) are themselves estimates, the MRA estimates may be biased (Macaskill, Walter and Irwig, 2001). Publication selection is likely to cause additional problems for the estimation of MRA model (2). When only significant effects are reported, the sampling errors of the observed effects are drawn from a truncated distribution. These errors will be thereby skewed and no longer normal. Lastly, equation (2) will misspecify the relationship between observed t-values and standard errors when some studies do not engage in publication selection.[9] Thus, it will be a great challenge for these MRA methods to see through publication selection, sampling error, and likely misspecification biases to identify underlying empirical effects, reliably. Nonetheless, the Monte Carlo simulations below reveal that PET is surprisingly effective in separating the wheat from the chaff.

Hedges' MLPSE

Hedges (1992) offers a more sophisticated econometric model of the publication selection process. This method assumes that the likelihood of publication is an increasing step function of the complement of a study's p-value.

$$w(effect_i, \sigma_i) = \begin{cases} \omega_1 & \text{if } -\sigma_i\Phi^{-1}(a_1) < effect_i \leq \infty \\ \omega_j & \text{if } -\sigma_i\Phi^{-1}(a_j) < effect_i \leq -\sigma_i\Phi^{-1}(a_{j-1}) \\ \omega_k & \text{if } -\infty \leq effect_i \leq -\sigma_i\Phi^{-1}(a_{k-1}) \end{cases}$$

where $1 < j < k$, and $\Phi^{-1}(a_1)$ is the inverse cumulative normal (Hedges and Vevea, 1996, p. 304). The weights of these arbitrary cut points, a_j, can be estimated from the data. After fully parameterizing this selection model, Hedges (1992) derives the joint likelihood and uses a multivariate Newton–Raphson method to find its maximum.

Hedges' MLPSE has been applied to six areas of economic research but with mixed success – Ashenfelter *et al.* (1999), Florax (2002), Abreu *et al.* (2005), and

[9]If a literature is composed of a mix of studies that select for significant effects and others that do not, equation (1) becomes: $effect_i = \gamma_1 + S_i\gamma_0 Se_i + \varepsilon_i$; where $S_i = 1$ if study i is engaged in selection, 0; otherwise. However, S_i is unobservable, and studies where $S_i = 1$ are typically over-sampled. Of course, if we had full information about the characteristics of published and unpublished studies alike, Heckman's correction for selection bias could be used.

Nijkamp and Poot (2005). Of the six sets of economic research literature to which MLPSE has been applied, three are problematic. Florax (2002) finds that MLPSE does not converge for estimates of the price elasticity of water demand. He also unearths the 'awkward' implication that the probability of publishing an insignificant stringency elasticity is greater than the probability of publishing a statistically significant one. Likewise, Abreu *et al.* (2005) obtain implausible weights for publishing estimates of economic convergence. They find that studies with insignificant *p*-values between 0.05 and 0.10 are more likely to be published than statistically significant ones.

Meta-significance testing

A second meta-regression approach to identify genuine effects beyond publication selection is based on statistical power. Statistical power causes the magnitude of a standardized test statistic (e.g. its *t*-value) to vary positively with degrees of freedom when there is, in fact, an overall genuine empirical effect. Suppose researchers wish to know whether some parameter, β_1, is equal to zero. When this hypothesis is true, estimates of β_1 will vary randomly around zero, and the *t*-value will be independent of its degrees of freedom. Because the probability of the type I error, α, is constant for all sample sizes, standardized test statistics adjust for any effect caused by differences in degrees of freedom alone. Therefore, when $H_0 : \beta_1 = 0$ is true, large values of the standardized test statistic will be observed rarely and randomly, regardless of degrees of freedom. Alternatively, should H_0 be false, statistical power will cause the observed magnitude of the standardized test statistic to be positively associated with the square root of its degrees of freedom. This positive relationship can be expected regardless of the size of the effect, $\beta_1 \neq 0$, and irrespective of contamination from random misspecification biases. This trace of statistical power identifies a genuine empirical effect when found across a given research literature.

More precisely, statistical theory predicts that the t-ratio will be related to the square root of degrees of freedom, or

$$E(\log |t_i|) = \alpha_0 + \alpha_1 \log(df_i) \tag{3}$$

where $\alpha_1 = 0$ if there is no effect (i.e., $H_0 : \beta_1 = 0$ is true), and $\alpha_1 = \frac{1}{2}$ when H_0 is false (Stanley, 2005a).

This meta-significance test provides evidence of a genuine empirical effect if the corresponding MRA rejects $H_0 : \alpha_1 \leq 0$. In this way, a genuine empirical effect creates a power trace. If there is a genuine underlying effect, there will also be a logarithmic relationship between a study's *t*-statistic and its degrees of freedom. If we find a positive association between *df* and the standardized test statistic across a given empirical literature, the authenticity of the effect in question is confirmed.

However, statistical theory is only one consideration; practical application might prove quite different.[10] How well can we expect these tests to work in practice? Do these meta-regression strategies for publication bias possess desirable statistical properties? Are they biased or inefficient? Next, the small-sample properties of FAT, MST and PET are investigated when selection and misspecification biases infest the research literature.

III. Simulation results

Funnel-asymmetry testing

Table 1 reports size and power of 10,000 replications of the funnel-asymmetry test – see the Appendix for a description of the details of the simulation design. Recall that FAT tests $H_0: \gamma_0 = 0$ in equation (2). The numbers reported in the tables below are the proportions of these 10,000 replications that produced a statistically significant test result ($\alpha = 0.05$). Depending on the particular design condition, these proportions represent either the level or the power of FAT in detecting publication selection. When the incidence of publication selection is 0%, the reported proportion represents the

TABLE 1

Power and level of FAT

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of FAT*		*Effect, i.e. ($\beta_1 = 1$) Power of FAT*	
		n = 20	*n = 80*	*n = 20*	*n = 80*
None	0	0.0479	0.0449	0.0469	0.0489
	25	0.0474	0.1822	0.1034	0.2070
	50	0.1434	0.7315	0.2251	0.5692
	75	0.5870	0.9999	0.4189	0.9101
Random bias:	0	0.0504	0.0476	0.0508	0.0458
$\sigma_{bias} = 0.25$	25	0.0529	0.1538	0.0945	0.1812
	50	0.1209	0.5956	0.1958	0.4959
	75	0.4706	0.9935	0.3562	0.8412

Notes: Reported frequencies are based on 10,000 replications. FAT tests $H_0: \gamma_0 = 0$ in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. FAT: funnel-asymmetry testing.

[10]Publication selection can be expected to affect equation (3) and make MST less powerful. Recall that publication bias is expected to be proportional to $n^{-1/2}$, inducing an inverse relation between effect and sample size. In the absence of publication selection, the standardized effect (e.g. the *t*-statistic) will be *positively* associated with degrees of freedom and proportional to the square root of degrees of freedom when there is an effect–recall equation (3). Thus, this inverse relation caused by selection attenuates, or lessens, the expected positive relationship between *t* and *df*. Publication selection makes it more difficult to reject $H_0: \alpha_1 \leq 0$ and thereby reduces MST's power.

observed frequency of a type I error (i.e. the test's level).[11] Table 1 reports type I error rates that are all near their 0.05 nominal levels, making FAT a valid test for the presence of publication selection.

In contrast, the power of FAT in detecting publication bias (publication selection incidences of 25%, 50% and 75%) is disappointing. Unless half or more of the research literature is selected for statistical significance, FAT is unlikely to detect it. For example when there are 80 studies in the literature ($n = 80$), our chance of catching selection bias exceeds 50% only if the incidence of selection is also at least 50%. In smaller literatures ($n = 20$), FAT is not at all reliable in detecting existing publication selection.

Yet, such is the nature of trying to distinguish the existence of publication selection from a potentially genuine effect. Hedges and Vevea (1996) also report low power levels for Hedges' MLPSE. 'The significance tests for selection performed poorly. . . . The likelihood ratio test. . . was abysmally nonrobust' (Hedges and Vevea, 1996, p. 330).

All methods for detecting the presence of publication selection bias have low power. Thus, it would be unwise to use the absence of evidence of publication bias as a reason not to take measures to adjust or to correct for it. Even when publication bias greatly exaggerates the magnitude of the reported effects, statistical tests are likely to miss it. And, conventional reviews will mistake these undetected publication biases for some genuine empirical effect. The unreliability of testing for the presence of publication selection bias forces us to focus upon the more important scientific questions. Is there is a genuine empirical effect irrespective of publication selection? What is the magnitude of this effect? Researchers and policy makers need answers to these central questions. Thus, we need a reliable method to identify genuine empirical effects that is robust to potential contamination from publication selection. In the next section, MRA tests for genuine effects are investigated for resilience to publication selection bias. Afterwards, the statistical properties of corrected estimates are simulated and discussed.

MRA tests for empirical effects in the presence of publication selection

Table 2 reports size and power of 10,000 replications of MST – equation (3). Recall that MST tests $H_0 : \alpha_1 = 0$ in the meta-regression model (3). When $H_0 : \beta_1 = 0$ is forced to be true, there is no genuine effect to uncover, and the reported proportion is the test's level, or the observed type I error rate. Alternatively, when $\beta_1 = 1$, the proportions reported in these tables (Tables 2–8) measure the power (or frequency) that the MRA tests will correctly detect a genuine underlying empirical effect.

[11]All tables report the *observed levels* of these MRA tests rather than their nominal levels, which is always set at 0.05. Reported levels are the observed relative frequency of type I errors for the data generating process (DGP) described in the Appendix. Technically, the reported levels need not be the 'size' of these tests, because the size 'is the supremum of the rejection probability over all DGPs that satisfy the null hypothesis' (Davidson and MacKinnon, 2004, p. 125). I wish to make no claim that the reported levels are the supremum for all possible DGPs.

TABLE 2

Power and level of MST

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of MST*		*Effect, i.e. ($\beta_1 = 1$) Power of MST*	
		n = 20	*n = 80*	*n = 20*	*n = 80*
None	0	0.0479	0.0420	0.9037	1.000
	25	0.0189	0.0201	0.8766	1.000
	50	0.0059	0.0101	0.8512	1.000
	75	0.0010	0.0020	0.7909	0.9993
Random bias:	0	0.0999	0.1790	0.7726	0.9981
$\sigma_{\text{bias}} = 0.25$	25	0.0534	0.1113	0.7412	0.9951
	50	0.0260	0.0652	0.7098	0.9935
	75	0.0168	0.0386	0.7070	0.9928

Notes: Reported frequencies are based on 10,000 replications. MST tests $H_0 : \alpha_1 = 0$ in equation (3), $E(\log |t_i| = \alpha_0 + \alpha_1 \log(df_i)$. MST: meta-significance testing.

Note first the simulation findings when there are no misspecification biases (Table 2). Regardless of the sample size or of the incidence of publication selection, the type I error rates are always under the nominal 5% level. Yet, MST is quite powerful, correctly detecting the effect 80–90% of the time even when there are only 20 observations. With 80 observations, MST is virtually guaranteed to find even a small empirical effect. The effect that we are attempting to identify in these simulations is rather small, $R^2 = 9\%$, and the sample sizes used are also quite modest $\{30, 50, 75, 100, 200\}$ – see the Appendix for a detailed description of the simulation design. When conventional statistical assumptions are valid, MST possesses very desirable properties, regardless of the incidence of publication selection. Hence, MST exhibits robustness against publication selection. In most experimental applications where misspecification bias is not an important concern (e.g. medical research or experimental economics), MST can provide a solid basis upon which to infer an effect even in the presence of dominant publication selection.

Most Monte Carlo studies of econometric methods assume that the underlying models are correctly specified. Variations in design are investigated for different levels of effects, variation, etc., as long as the estimation model is not fundamentally compromised. Here, however, it is fully recognized that reported findings, which constitute our MRA data, may be affected by many different types of misspecification biases. Thus, literatures thoroughly infested with unknown misspecification biases are simulated. Obviously, such cases are more problematic to any statistical method but, in turn, are more likely to reflect the actual nature of economic research.

Even when all studies contain random misspecification bias, $\sigma_{\text{bias}} = 0.25$ (see the Appendix), MST's power remains quite high – 70% or larger when n is 20 and over 99% for 80 observations – Table 2. Note how publication selection obscures a genuine effect. Typically, a larger incidence of publication selection lowers MST's power by attenuating its estimated MRA coefficient. This attenuation and lower power is exactly

what is expected. The only problem is that there is an inflation of type I error rates in some cases. In particular, when there is no publication selection, the level of MST rises to 10 or 18% for $n = 20$ and 80, respectively.

This inflation of type I error rates represents a serious problem for MST. The purpose of these meta-regression tests is to harden our empirical economic inferences. Thus, the probability that findings of genuine empirical effects are the result of misspecification bias, error, or any other artifact must be kept low. In this context, even though the level of MST is many times smaller than its associated power, the type I error rates can be unacceptably high. Not only should we not ignore this problem, meta-regression testing strategy must somehow accommodate or remedy these inflated αs. As discussed below, joint testing, which should be conducted in any case by a comprehensive meta-analysis, provides a readily available remedy.[12]

Table 3 reports the same simulation design for PET, $H_0 : \gamma_1 = 0$ in equation (2). Note that the pattern of power and level of PET is very similar to MST. When there is no large-sample misspecification bias, precision testing has very desirable properties. Thus, it too is robust against publication selection. Like MST, the power of PET declines as the incidence of publication selection rises. It simply becomes harder

TABLE 3

Power and level of PET

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of PET*		*Effect, i.e. ($\beta_1 = 1$) Power of PET*	
		$n = 20$	*$n = 80$*	*$n = 20$*	*$n = 80$*
None	0	0.0502	0.0481	0.9557	1.000
	25	0.0060	0.0089	0.9368	1.000
	50	0.0011	0.0032	0.8990	1.000
	75	0.0011	0.0012	0.8778	1.000
Random bias:	0	0.0630	0.0633	0.8488	1.000
$\sigma_{bias} = 0.25$	25	0.0135	0.0192	0.8152	0.9999
	50	0.0066	0.0105	0.7842	0.9999
	75	0.0078	0.0152	0.7490	0.9996

Notes: Reported frequencies are based on 10,000 replications. PET tests $H_0 : \gamma_1 = 0$ in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. PET: precision-effect testing.

[12] It might be instructive to consider why MST experiences this inflation of α. The random, omitted-variable misspecification biases that are added to each estimate in this simulation do not lessen with larger samples. However, because the standard error of the estimator does decline predictably with the square root of the study's sample size, the effect of the random omitted-variable bias on the reported t-value rises with the sample size. Unlike PET, MST takes the absolute value, forcing both large positive and large negative t-values to become large and positive. Therefore, with random misspecification bias, the typical reported $|t|$ will increase with sample size. Unfortunately, MST is looking for exactly such a positive association between sample size and reported t-values as the signature of statistical power and hence of genuine empirical effect. In this way, MST may be fooled by pervasive misspecification and publication bias. When there are large-sample misspecification biases but no genuine empirical effect, MST is, itself, upwardly biased.

TABLE 4

Power and level of the joint PET–MST test

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of joint test*		*Effect, i.e. ($\beta_1 = 1$) Power of joint test*	
		$n = 20$	*$n = 80$*	*$n = 20$*	*$n = 80$*
None	0	0.0042	0.0019	0.8866	1.000
	25	0.0004	0.0004	0.8513	1.000
	50	0.0001	0.0002	0.8249	1.000
	75	0.0000	0.0000	0.7909	0.9993
Random bias:	0	0.0117	0.0121	0.7314	0.9981
$\sigma_{\text{bias}} = 0.25$	25	0.0018	0.0024	0.6958	0.9951
	50	0.0011	0.0019	0.6638	0.9935
	75	0.0020	0.0019	0.6590	0.9927

Notes: Reported frequencies are based on 10,000 replications. PET tests $H_0 : \gamma_1 = 0$ in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. MST tests $H_0 : \alpha_1 = 0$ in equation (3), $E(\log|t_i| = \alpha_0 + \alpha_1 \log(df_i)$. PET: precision-effect testing; MST: meta-significance testing.

to see the message through the noise. PET is more powerful in detecting genuine effects and less vulnerable to an inflation of α. PET seems to provide a viable method for testing the presence of a genuine empirical effect, irrespective of the extent of publication selection.

PET's high power is surprising because the meta-regression of t-values on precision ($1/Se_i$) is known to contain errors-in-variables (EV) bias. The downward bias for g_1 revealed by these simulations is more benign than MST's bias; this bias reduces power but does not increase PET's level. On the other hand, the way that publication selection is simulated here by selecting t-values that are statistically significant regardless of the sample size is more consistent with equation (1) and hence with PET. Thus, PET's MRA model better reflects the publication selection mechanism used in these simulations. In the real world, publication selection is apt to be more complex; hence, Table 4 reports the power and level of a joint PET–MST test for genuine empirical effects. This PET–MST test is passed only in the case that both tests reject the null hypothesis of no effect. In all cases, the observed type I error rates are much less than the nominal 0.05 level. But, fortunately, modest power is sacrificed as the result of combining these tests. Even in literatures riddled with serious misspecification bias and dominated by publication selection, precision-effect and joint PET–MST testing may be relied upon to determine correctly whether the alleged effect is authentic.

Not satisfied with these postive findings, it remains to subject the previous simulations to even larger ubiquitous misspecification baises (i.e. $\sigma_{\text{bias}} = 0.5$ and 1.0) and to other potential problems. Tables 5, 6 and 7 report the power and level of MST, PET and the joint PET–MST tests, respectively, when a study's typical misspecification bias is doubled to $\sigma_{\text{bias}} = 0.5$ and then redoubled. Missing are the results for the case of no misspecification selection because they are identical to those reported in

TABLE 5

Power and level of MST

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of MST*		*Effect, i.e. ($\beta_1 = 1$) Power of MST*	
		$n = 20$	*$n = 80$*	*$n = 20$*	*$n = 80$*
Random bias:	0	0.1950	0.4566	0.5275	0.9369
$\sigma_{\text{bias}} = 0.50$	25	0.1368	0.3517	0.5047	0.9279
	50	0.0944	0.2873	0.5059	0.9246
	75	0.0912	0.2620	0.5401	0.9408
Random bias:	0	0.3170	0.7395	0.3790	0.8070
$\sigma_{\text{bias}} = 1.0$	25	0.2859	0.6848	0.3697	0.8091
	50	0.2640	0.6438	0.3963	0.8388
	75	0.2776	0.6708	0.4841	0.9040

Notes: Reported frequencies are based on 10,000 replications. MST tests $H_0 : \alpha_1 = 0$ in equation (3), $E(\log |t_i| = \alpha_0 + \alpha_1 \log(df_i)$. MST: meta-significance testing.

TABLE 6

Power and level of PET

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of PET*		*Effect, i.e. ($\beta_1 = 1$) Power of PET*	
		$n = 20$	*$n = 80$*	*$n = 20$*	*$n = 80$*
Random bias:	0	0.0700	0.0689	0.6111	0.9940
$\sigma_{\text{bias}} = 0.50$	25	0.0227	0.0297	0.5919	0.9838
	50	0.0169	0.0270	0.5763	0.9930
	75	0.0275	0.0606	0.5795	0.9931
Random bias:	0	0.0650	0.0635	0.3369	0.8594
$\sigma_{\text{bias}} = 1.0$	25	0.0389	0.0387	0.3408	0.8673
	50	0.0330	0.0418	0.3383	0.8935
	75	0.0548	0.1052	0.3764	0.9187

Notes: Reported frequencies are based on 10,000 replications. PET tests $H_0 : \gamma_1 = 0$ in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. PET: precision-effect testing.

Tables 2–4. As seen clearly in Table 5, MST's type I error rates become drastically inflated in most design conditions. This is especially true when the MRA uses 80 observations. In contrast, PET exbibits little α-inflation (Table 6). Even if the typical misspecification bias greatly exceeds the sampling error, PET has acceptable size in nearly all cases. When combined, the joint PET–MST's type I error rates remain below their nominal level (α), except in one case – $\sigma_{\text{bias}} = 1.0$, $n = 80$ with 75% of the studies selected (Table 7).

However, some price must be paid for combining tests and for attempting to confuse these MRA tests with large and pervasive biases. Unsurprisingly, these larger

TABLE 7

Power and level of the joint PET–MST test

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of joint test*		*Effect, i.e. ($\beta_1 = 1$) Power of joint test*	
		$n = 20$	*$n = 80$*	*$n = 20$*	*$n = 80$*
Random bias:	0	0.0195	0.0346	0.4518	0.9354
$\sigma_{\text{bias}} = 0.50$	25	0.0064	0.0140	0.4334	0.9272
	50	0.0063	0.0121	0.4304	0.9230
	75	0.0120	0.0283	0.4602	0.9334
Random bias:	0	0.0270	0.0503	0.2249	0.7331
$\sigma_{\text{bias}} = 1.0$	25	0.0177	0.0296	0.2241	0.7453
	50	0.0172	0.0319	0.2411	0.7820
	75	0.0386	0.0891	0.3022	0.8546

Notes: Reported frequencies are based on 10,000 replications. PET tests $H_0 : \gamma_1 = 0$ in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. PET: precision-effect testing; MST: meta-significance testing.

misspecification biases cause power to decline. With only 20 observations, the joint test correctly detects a genuine effect less than half the time. In contrast, when a larger number of observed effects are available ($n = 80$), joint testing and PET remain rather powerful. The critical dimension is the relative magnitude of the misspecification bias. As the noise dominates the signal, or as random bias overwhelms declining sampling error, it becomes harder to detect effects. If the observed variance among reported effects is too many times larger than the typical sampling error variance of these effects, then the reviewer should exercise added caution in interpreting these meta-regression tests.[13] Among past meta-analyses in economics, the ratio of observed variance among effects to their sampling error variance ranges from 1.6 to 3.4 (Dalhuisen *et al.*, 2003; Doucouliagos and Laroche, 2003; Rose and Stanley, 2005). Ratios generated by these simulations encompass the observed range of variance ratios.

Lastly, the statistical performance of these tests is examined when the original studies use much larger sample sizes. For these simulations, sample size is chosen as a random number uniformly distributed between 100 and 1,000. Otherwise, the previous design conditions are unchanged. Table 8 reports the size and power of PET–MST. By increasing the average sample size five-fold, the relative magnitude of misspecification bias to sampling error becomes much larger (hence 'worse'). As before, in all cases but one, type I error rates are within their nominal 0.05 level. However, there is also a positive side to these larger sample sizes. This larger sample design contains greater variation among sample sizes and thereby among standard

[13] Extreme heterogeneity (i.e. very large σ_{bias}) accompanied by a high incidence of publication selection can invalidate PET by inflating the type I error rate. However, this potential type I inflation can be controlled by adding a test for heterogeneity. Simulations show that the failure to reject $H_0 : \sigma_e^2 \leq 2$ serves as an effective means to limit PET's type I errors (Stanley, 2005b), where σ_e^2 is the error variance from equation (2).

TABLE 8

Power and level of PET/MST: Original samples are uniform (100, 1,000)

Misspecification	*Incidence of publication selection (%)*	*No effect, i.e. ($\beta_1 = 0$) Level of joint test*		*Effect, i.e. ($\beta_1 = 1$) Power of joint test*	
		n = 20	*n = 80*	*n = 20*	*n = 80*
None	0	0.0014	0.0021	0.9996	1.000
	25	0.0066	0.0038	0.9998	1.000
	50	0.0111	0.0074	1.000	1.000
	75	0.0168	0.0102	1.000	1.000
Random bias:	0	0.0082	0.0116	0.9187	1.000
$\sigma_{\text{bias}} = 0.25$	25	0.0103	0.0172	0.9192	0.9999
	50	0.0189	0.0292	0.9245	0.9999
	75	0.0388	0.0789	0.9263	0.9998

Notes: Reported frequencies are based on 10,000 replications. PET tests $H_0 : \gamma_1 = 0$ in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. PET: precision-effect testing; MST: meta-analysis testing.

errors across studies. Larger variation of the MRA independent variable, in turn, causes statistical power of these MRA tests to increase, dramatically. Even though this design presents a serious challenge to these MRA tests, the joint PET–MST test has quite desirable statistical properties.

Thus far, a strategy for testing underlying empirical effects has been proposed, and its properties have been investigated using simulations. Next, we consider how best to estimate the magnitude of the effect when one has been identified.

Mean-PET: Estimating empirical effects in the presence of publication selection

Beyond tests of significance and parameter restrictions, researchers are often interested in estimating the magnitude of the empirical effect at issue. Obviously, the magnitude of key economic parameters has important practical and policy implications. How can we best estimate the size of these effects when a literature may contain substantial publication selection? Such a question forces the researcher to face a fundamental dilemma: all evaluations of empirical estimates are biased in the presence of publication selection.

Recall that the MRA estimate of the coefficient on precision, $1/Se_i$ (equation 2) serves as a useful test for authentic effects. Although the conventional t-test of this coefficient often provides a valid test of effects (PET), simulations show that the estimate of precision's coefficient, g_1 from equation (2), is biased downward when there is a genuine effect. Simple or weighted averages of reported effects are no better. These average reported estimates consistently overestimate the magnitude of the effect whenever there is directional publication selection. Perhaps, the predictability of these biases offers a means to reduce estimation bias?

TABLE 9

Mean effects of mean-PET vs. the unadjusted average

Misspecification	*True effect* β_1	*Publication selection (%)*	*Sample size n*	*Simple average*	*Mean-PET*
Random bias:	0	0	20	−0.0005	−0.0002
$\sigma_{bias} = 0.25$	0	0	80	−0.0012	−0.0012
	0	25	20	0.2334	0.1383
	0	25	80	0.2345	0.1368
	0	50	20	0.4676	0.2630
	0	50	80	0.4677	0.2648
	0	75	20	0.7004	0.3847
	0	75	80	0.7016	0.3877
	1	0	20	1.0012	1.0008
	1	0	80	1.0001	1.0001
	1	25	20	1.0660	0.9981
	1	25	80	1.0653	0.9974
	1	50	20	1.1320	0.9905
	1	50	80	1.1318	0.9911
	1	75	20	1.1988	0.9885
	1	75	80	1.1971	0.9844
Random bias:	0	0	20	0.0027	0.0022
$\sigma_{bias} = 0.50$	0	0	80	0.0011	0.0004
	0	25	20	0.2701	0.1529
	0	25	80	0.2706	0.1532
	0	50	20	0.5393	0.3102
	0	50	80	0.5399	0.3114
	0	75	20	0.8115	0.4740
	0	75	80	0.8095	0.4733
	1	0	20	0.9996	1.0024
	1	0	80	0.9991	0.9984
	1	25	20	1.0962	1.0217
	1	25	80	1.0966	1.0200
	1	50	20	1.1958	1.0401
	1	50	80	1.1933	1.0360
	1	75	20	1.2901	1.0531
	1	75	80	1.2909	1.0516

Notes: Reported frequencies are based on 10,000 replications. Mean-PET is the average of the simple unadjusted average effect and the estimate of γ_1 in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. PET: precision-effect testing.

The approach investigated here combines the simple average of all reported estimates in a literature (i.e. $\sum b_{1i}/L$, where L is the number of studies in a literature) with the MRA estimate of γ_1, g_1, in equation (2). By taking the midpoint of these two biased estimators, $(g_1 + \sum b_{1i}/L)/2$, estimation accuracy can be greatly improved. Table 9 reports the average values of 10,000 simulations for both the unadjusted sample mean of the reported effects and the combined average of these two estimates – 'mean-PET'. In most cases, mean-PET has a greatly reduced bias. Under likely conditions, this

TABLE 10

Mean square errors of mean-PET vs. the unadjusted average

Misspecification	*True effect* β_1	*Publication selection (%)*	*Sample size n*	*Simple average*	*Mean-PET*
Random bias:	0	0	20	0.0134	0.0273
$\sigma_{\text{bias}} = 0.25$	0	0	80	0.0035	0.0070
	0	25	20	0.0652	0.0435
	0	25	80	0.0577	0.0246
	0	50	20	0.2267	0.0877
	0	50	80	0.2208	0.0746
	0	75	20	0.4959	0.1605
	0	75	80	0.4936	0.1534
	1	0	20	0.0137	0.0284
	1	0	80	0.0034	0.0069
	1	25	20	0.0164	0.0263
	1	25	80	0.0072	0.0066
	1	50	20	0.0276	0.0254
	1	50	80	0.0199	0.0065
	1	75	20	0.0478	0.0245
	1	75	80	0.0410	0.0062
Random bias:	0	0	20	0.0238	0.0583
$\sigma_{\text{bias}} = 0.50$	0	0	80	0.0059	0.0144
	0	25	20	0.0918	0.0733
	0	25	80	0.0780	0.0355
	0	50	20	0.3054	0.1369
	0	50	80	0.2952	0.1064
	0	75	20	0.6685	0.2531
	0	75	80	0.6578	0.2308
	1	0	20	0.0237	0.0584
	1	0	80	0.0059	0.0139
	1	25	20	0.0298	0.0531
	1	25	80	0.0144	0.0132
	1	50	20	0.0557	0.0486
	1	50	80	0.0417	0.0129
	1	75	20	0.0983	0.0490
	1	75	80	0.0882	0.0136

Notes: Reported frequencies are based on 10,000 replications. Mean-PET is the average of the simple unadjusted average effect and the estimate of γ_1 in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. PET: precision-effect testing.

combined estimate drastically reduces the absolute magnitude of the bias – by more than 90% in many cases.

Although mean-PET has much to offer in reduced bias when the researcher has reason to suspect publication selection, biased estimates may sometimes be more efficient. It remains to be seen whether the added variation caused by combining separate estimates overwhelms the reduction of bias. Table 10 compares the mean square error (MSE) of mean-PET to the average of observed effect. Here too,

mean-PET is superior in the majority of cases. Relative efficiencies are as low as 0.15, implying that mean-PET out-performs the average by better than 6 to 1. As expected, however, the simple average turns the tables when a literature exhibits no selection – out-performing mean-PET by better than 2 to 1. Without selection, nothing is to be gained by combining estimates and thereby increasing sampling variation. When there is light selection (25%), the results are mixed. Mean-PET is better in larger samples ($n = 80$) but sometimes worse in small ones ($n = 20$). Clearly, combining these two biased estimators is prudent only when publication selection is likely.

It should be noted that mean-PET is quite robust to publication selection where there is an authentic effect. In fact, MSE decreases somewhat with higher incidence of selection. When there is an effect, mean-PET is remarkably accurate. On the other hand, these desirable properties vanish when the effect likewise vanishes.

Table 11 reports the observed frequency that the true effect, β_1, falls in a conventionally constructed 95% confidence interval around the combined estimate. Before constructing the confidence interval, it is assumed that the literature passes several tests for effect. This stringent criterion for effect is defined as having an average reported t-value equal to 2 or greater and passing three separate hypothesis tests – MST, PET and a simple t-test of the mean reported effect. If any of these tests are not passed, the researcher should assume that the effect in question is zero.

To be conservative, cases where there are no effects are mixed equally with cases where there are genuine effects. When all of the above tests are passed, a confidence interval is constructed. In practice, the problem is to differentiate between falsely identified effects (i.e. type I errors) and authentic effects. To address this issue, Table 11 reports the results from 10,000 replications of genuine effects (either $\beta_1 = 1$ or $\beta_1 = 2$) that are mixed together with 10,000 replications where there is no effect ($\beta_1 = 0$). Confidence intervals are constructed by: $(g_1 + \sum b_{1i}/L)/2 \pm t_{\alpha/2} S_c$; where $S_c = (Sb_1^2/n + Sg_1^2)^{1/2}$, Sb_1^2 is the observed variance among reported effects ($b_1 s$) and Sg_1^2 is the variance of the estimated PET coefficient.[14] As reflected in Table 11, the confidence levels largely conform to their nominal levels. The frequency that the true effect falls in the 95% confidence levels rises with the size of the genuine effect but declines as the magnitude of misspecification bias increases.

IV. Conclusions

This study offers a simple meta-regression approach to test and estimate empirical effects in research literatures dominated by publication selection and investigates their small-sample properties. Both MST and PET are quite powerful and possess type I error rates within their nominal levels, regardless of the incidence of publication selection, when conventional statistical assumptions are satisfied. For many

[14]This formula for the standard error of mean-PET assumes that the two estimates are independent. Table 11 shows that confidence intervals constructed under this assumption are largely valid.

TABLE 11

Proportion of 95% confidence intervals containing the true effect

Misspecification	*True effect* β_1	*Publication selection (%)*	*Sample size n*	*Mean-PET*
Random bias:				
$\sigma_{\text{bias}} = 0.25$	1	0	20	0.9615
	1	0	80	0.9528
	1	25	20	0.9593
	1	25	80	0.9539
	1	50	20	0.9593
	1	50	80	0.9528
	1	75	20	0.9557
	1	75	80	0.9367
	2	0	20	0.9506
	2	0	80	0.9524
	2	25	20	0.9497
	2	25	80	0.9500
	2	50	20	0.9487
	2	50	80	0.9404
	2	75	20	0.9403
	2	75	80	0.9427
Random bias:	1	0	20	0.9330
$\sigma_{\text{bias}} = 0.50$	1	0	80	0.9467
	1	25	20	0.9200
	1	25	80	0.9417
	1	50	20	0.8976
	1	50	80	0.9275
	1	75	20	0.8611
	1	75	80	0.8888
	2	0	20	0.9367
	2	0	80	0.9407
	2	25	20	0.9404
	2	25	80	0.9313
	2	50	20	0.9369
	2	50	80	0.9310
	2	75	20	0.9261
	2	75	80	0.9167

Notes: Reported frequencies are based on 10,000 replications. Mean-PET is the average of the simple unadjusted average effect and the estimate of γ_1 in equation (2), $t_i = effect_i/Se_i = \gamma_0 + \gamma_1(1/Se_i) + e_i$. PET: precision-effect testing.

areas of social and medical research, both MST and PET can provide a viable remedy for publication selection. Combining them only makes their findings more resilient to gross misspecification biases. As long as an area of research is free from large and pernicious estimation biases, both of these meta-regression testing strategies provide valid tests for genuine empirical merit, regardless of the incidence of publication selection.

Naturally, there are limits to the robustness of these tests for genuine empirical merit. Random, large-sample misspecification biases can cause MST to identify a genuine effect much too frequently (i.e. an inflation of the type I error rates). Even in cases where random misspecification biases are somewhat larger than conventional sampling error, PET and joint testing possess acceptable, if less powerful, small-sample properties. The relative magnitude of misspecification bias to sampling error is the key parameter in assessing the vulnerability of these meta-regression methods to type I error inflation. Combining PET with a test for 'excess' remaining heterogeneity ($H_0 : \sigma_e^2 \leq 2$) in equation (2) can be used to manage potential type I error inflation should a literature contain dominating misspecification biases – $\sigma_{\text{bias}} > 1.0$ (Stanley, 2005b). Thus, FAT and PET need to be interpreted carefully when there remains large unexplained variation in equation (2).

However not all misspecification biases will be random or selected to produce a statistically significant result. Variations in reported research results will often be systematically related to the modeling choices made by researchers concerning which independent variables, functional forms, data sets, or estimation techniques they use. Explaining the effect of such research choices on research outcomes has been the primary focus of many applications of MRA in economics (Roberts and Stanley, 2005). The MRA methods developed in this paper may also be employed in a multivariate context to explain potential systematic variation in reported research, aside from what might arise from publication selection. See Doucouliagos (2005), Rose and Stanley (2005) and Stanley (2005a, b) for examples of multivariate explanatory FAT–PET–MRAs. 'It is our view that meta-regression analysis while no panacea, no magic elixir, is a helpful framework to integrate and explain disparate empirical economic literature.… MRA provides a mechanism through which one can more objectively ask questions about economic research, offer explanatory hypotheses, and rigorously test those conjectures by confronting them with the actual research record' (Stanley and Jarrell, 1989, p. 169). By explicitly modelling and correcting publication selection, the methods developed in this paper can help researchers better understand and explain the wide variation routinely found in published economic research.

Final Manuscript Received: April 2006

References

Abreu, M., de Groot, H. L. F. R. and Florax, R. G. M. (2005). 'A meta-analysis of beta-convergence: the legendary two-percent', *Journal of Economic Surveys*, Vol. 19, pp. 389–420.

Ashenfelter, O., Harmon, C. and Oosterbeek, H. (1999). 'A review of estimates of the schooling/earnings relationship, with tests for publication bias', *Labour Economics*, Vol. 6, pp. 453–470.

Begg, C. B. and Berlin, J. A. (1988). 'Publication bias: a problem in interpreting medical data', *Journal of the Royal Statistical Society (Series A)*, Vol. 151, pp. 419–445.

Bland, J. M. (1988). 'Discussion of the paper by Begg and Berlin', *Journal of the Royal Statistical Society (Series A)*, Vol. 151, pp. 450–451.

Card, D. and Krueger, A. B. (1995). 'Time-series minimum-wage studies: a meta-analysis', *American Economic Review*, Vol. 85, pp. 238–243.
Dalhuisen, J., Florax, R. J. G. M., deGroot, H. L. F. and Nijkamp, P. (2003). 'Price and income elasticities of residential water demand: a meta-analysis', *Land Economics*, Vol. 79, pp. 292–308.
Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*, Oxford University Press, Oxford.
De Long, J. B. and Lang, K. (1992). 'Are all economic hypotheses false?', *Journal of Political Economy*, Vol. 100, pp. 1257–1272.
Doucouliagos, C. and Laroche, P. (2003). 'What do unions do to productivity: a meta-analysis', *Industrial Relations*, Vol. 42, pp. 650–691.
Doucouliagos, C. (2005). 'Publication bias in the economic freedom and economic growth literature', *Journal of Economic Surveys*, Vol. 19, pp. 367–388.
Doucouliagos, C., Laroche, P. and Stanley, T. D. (2005). 'Publication bias in union-productivity research', *Relations Industrielles/Industrial Relations*, Vol. 60, pp. 320–346.
Egger, M., Smith, G. D., Scheider, M. and Minder, C. (1997). 'Bias in meta-analysis detected by a simple, graphical test', *British Medical Journal*, Vol. 316, pp. 629–634.
Florax, R. J. G. M. (2002). 'Methodological pitfalls in meta-analysis: publication bias', in Florax R., Nijkamp P. and Willis K. (eds), *Comparative Environmental Economic Assessment*, Edward Elgar, Cheltenham.
Fuller, W. A. (1987). *Measurement Error Models*, John Wiley and Sons, New York.
Görg, H. and Strobl, E. (2001). 'Multinational companies and productivity spillovers: a meta-analysis', *Economic Journal*, Vol. 111, pp. F723–F740.
Green, W. H. (1990). *Econometric Analysis*, MacMillan, New York.
Hedges, L. V. (1992). 'Modelling publication selection effects in meta-analysis', *Statistical Science*, Vol. 7, pp. 246–255.
Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-analysis*, Academic Press, Orlando.
Hedges, L. V. and Vevea, J. L. (1996). 'Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model', *Journal of Educational and Behavioral Statistics*, Vol. 21, pp. 299–332.
Laird, N. and Mosteller, F. (1988). 'Discussion of the paper by Begg and Berlin', *Journal of the Royal Statistical Society (Series A)*, Vol. 151, p. 456.
Macaskill, P., Walter, S. D. and Irwig, L. (2001). 'A comparison of methods to detect publication bias in meta-analysis', *Statistics in Medicine*, Vol. 20, pp. 641–654.
Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago.
Nijkamp, P. and Poot, J. (2005). 'The last word on the wage curve? A meta-analytic assessment', *Journal of Economic Surveys*, Vol. 19, pp. 421–450.
Phillips, J. M. and Goss, E. P. (1995). 'The effects of state and local taxes on economic development: a meta-analysis', *Southern Economic Journal*, Vol. 62, pp. 2–29.
Roberts, C. J. and Stanley, T. D. (eds) (2005). *Meta-Regression Analysis: Issues of Publication Bias in Economics*, Blackwell, Oxford.
Rose, A. K. and Stanley, T. D. (2005). 'A meta-analysis of the effect of common currencies on international trade', *Journal of Economic Surveys*, Vol. 19, pp. 347–365.
Rosenthal, R. (1979). 'The "file drawer problem" and tolerance for null results', *Psychological Bulletin*, Vol. 86, pp. 638–641.
Rubin, D. B. (1988). 'Discussion of the paper by Begg and Berlin', *Journal of the Royal Statistical Society (Series A)*, Vol. 151, pp. 457–458.
Stanley, T. D. (2001). 'Wheat from chaff: meta-analysis as quantitative literature review', *Journal of Economic Perspectives*, Vol. 15, pp. 131–150.

Stanley, T. D. (2004). 'Does unemployment hysteresis falsify the natural rate hypothesis? A meta-analysis', *Journal of Economic Surveys*, Vol. 18, pp. 589–612.
Stanley, T. D. (2005a). 'Beyond publication bias', *Journal of Economic Surveys*, Vol. 19, pp. 309–345.
Stanley, T. D. (2005b). 'Integrating the empirical tests of the natural rate hypothesis: a meta-regression analysis,' *Kyklos*, Vol. 58, pp. 587–610.
Stanley, T. D. and Jarrell, S. B. (1989). 'Meta-regression analysis: a quantitative method of literature surveys', *Journal of Economic Surveys*, Vol. 3, pp. 161–170.
Sterling, T. D. (1959). 'Publication decisions and their possible effects on inferences drawn from tests of significance', *Journal of the American Statistical Association*, Vol. 54, pp. 30–34.
Sutton, A. J., Abrama, K. R., Jones, D. R., Sheldon, T. A. and Song, F. (2000a). *Methods for Meta-analysis in Medical Research*, John Wiley and Sons, Chichester.
Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R. and Jones, D. R. (2000b). 'Empirical assessment of the effect of publication bias on meta-analyses', *British Medical Journal*, Vol. 320, pp. 1574–1577.

Appendix: Simulation design

These simulations are all based on research literatures that test a given regression coefficient (i.e. $H_0 : \beta_1 = 0$). Such simple regression tests are meant only as a paradigm for testing other effects, in general. Similar statistical properties for these MRA tests should be found when an empirical literature uses other specific statistical tests.[15]

The basic structure of these meta-regression simulations may be sketched as:

1. Generate the regression variables randomly.
2. Use OLS to estimate and test $H_0 : \beta_1 = 0$. Select significant test results. Each selected test of $H_0 : \beta_1 = 0$ comprises one study's reported result in our hypothetical empirical literature.
3. Simulate these meta-regression tests by repeating the previous steps either 20 or 80 times. At this stage, meta-regression models (2) and (3) are estimated to provide one realization of the corresponding MRA test.
4. Repeat all of the above steps 10,000 times while tracking various outcomes for FAT, MST and PET.

The first step defines the data-generating process. The independent variable (X_1) for each study is simulated by a random uniform variable (100, 200). As long as the independent variable is stationary, its distribution will not matter. Y is then generated from:

$$Y_i = 100 + \beta_1 X_{1i} + \beta_2 X_{2i} + 100 e_i \quad i = 1, 2, \ldots, n \tag{4}$$

$e_i \sim NID(0, 1)$. The effect in question, β_1, is assumed to be either 0 or 1. When $\beta_1 = 1$, the average R^2 is approximately 9%, and the correlation coefficient is about 0.3. Changing the values of β_1 only affects the power of these tests, not their type I error rates. The larger one makes β_1, the higher the power. The $\beta_2 X_{2i}$ term induces misspecification bias, in general, and omitted-variable bias, in particular.

[15]However, additional limitations are found for F and χ^2 restriction tests. Because both positive and negative misspecification biases can cause larger F and χ^2 values, large and ubiquitous misspecification biases can invalidate these MRA tests. Stanley (2005b) simulates these MRA methods when applied to restriction testing and suggests how they might be adapted.

As is widely known, omitting a relevant variable from a regression model causes the estimate of β_1 to be biased and inconsistent. Because this bias remains in large samples, it can be mistaken for a genuine effect, potentially causing problems for our MRA tests. Other types of bias (i.e. small-sample bias that diminish with larger samples) cause little or no difficulty for these MRA tests. Although only omitted-variable biases are simulated here, they serve to represent any type of large-sample, misspecification bias (or inconsistency). Because such large-sample bias is the most problematic for these MRA tests for effect, only this type of bias is used in these simulations. To establish a baseline, these simulations include cases where there are no misspecification biases.

Random misspecification bias is induced by making β_2 in equation (4) a random normal variable, $N(0, 0.25)$. This random misspecification bias acts as 'heterogeneity,' which has been recognized as a key parameter by meta-analysts (Hedges and Vevea, 1996; Sutton *et al.*, 2000a). The difference is largely a matter of interpretation. Here, we assume that variation in the expected estimated effects is caused by weaknesses in our empirical methods; whereas, heterogeneity is usually viewed as variation in the 'true' effect.

The most influential magnitude for the performance of these meta-regression methods is the size of the typical misspecification bias relative to the sampling error. The larger the ratio of the standard deviation of these misspecification biases (σ_{bias}) to the standard deviation of the sampling errors (σ_{b_1}), the worse the size and power of our MRA tests can become. Both MST and PET investigate the expected effects that sample size (or standard error) has on the observed standardized test statistic. Random misspecification bias obscures the underlying statistical structure, inducing bias in MST. The larger the typical omitted-variable bias, the poorer the performance of these MRA tests.

In order to calibrate the simulations, relevant magnitudes from previous economic meta-analyses are investigated. The average observed variation among reported effects found among three previous meta-analyses implies a standard deviation of this misspecification bias of approximately 0.1.[16] To be conservative, this value is multiplied by two and half times, $\sigma_{\text{bias}} = 0.25$. To further explore the robustness of these MRA tests, this value is again doubled and re-doubled in some simulations. In the worst case explored here, σ_{bias} is assumed to be more than two and a half times the

[16] σ_{bias} is calculated from previous meta-analyses assuming that observed misspecification bias and sampling errors are independent. However, our simulation shows that they will be somewhat positively correlated ($r = 0.2$) when there is publication selection. Assuming independence, therefore, will tend to overestimate the size of σ_{bias}. Given the observed variation of reported effects across studies and also the reported sampling variance (Sb_1^2), σ_{bias}^2 can be calculated roughly as the difference. Because standard errors decrease with larger samples, these estimates were extrapolated to the typical sample size used by the simulation. The meta-analyses used for the calibration of σ_{bias} concern estimates of union-productivity effects (Doucouliagos and Laroche, 2003), common currency effects (Rose and Stanley, 2005), and income elasticities of water demand (Dalhuisen *et al.*, 2003). Estimates of unemployment persistence are not used because non-stationarity induces non-standard distributions (Stanley, 2004). Also price elasticities of water demand are not used because they give a negative σ_{bias}^2. Simulations reveal that the observed variance of estimated effects can be less than its typical error variance when there is 100% publication selection.

size of the typical sampling error and the same magnitude as the true effect when one is assumed to exist. These larger values conform to the relative magnitudes of heterogeneity to the genuine effect assumed by Hedges and Vevea (1996) in their simulations of Hedges' maximum likelihood corrections for publication selection.

In order to induce omitted-variable bias, X_{2i} is made equal to X_{1i} plus a random normal error. Thus, β_2, itself, becomes the omitted-variable bias. Given these magnitudes and the randomness of both the omitted-variable bias and the sampling error, bias and error will often overwhelm a study's results. And, this is all the more true when there is also publication selection.

The meta-regression models are assumed to be estimated using either 20 or 80 studies. Twenty is chosen because it is a rather small sample size for any regression estimate, while eighty is both practically feasible and, as we shall see, gives these tests power to spare. Power depends in the expected way on the MRA sample size, σ_{bias}, the magnitude of the effect, and σ_{b_1}. Sample sizes chosen for the original studies and used to test $H_0 : \beta_1 = 0$ are $\{30, 50, 75, 100, 200\}$. Simulations based upon randomly selected sample sizes give similar results, and distributions of sample sizes containing larger samples are also simulated and reported in Table 8.

Publication bias is simulated as selecting a statistically significant positive b_1. That is, if the random estimate does not provide a significantly positive t-value, a new sample is taken and the original regression is run again with different random errors and random biases until a significant t-value is obtained by chance. For example, the 50% publication selection condition assumes that exactly half of the studies estimate and re-estimate their regression models until a random, yet statistically significant, estimate is found and reported. For the other half, the first random estimate, significant or not, is reported and used.

In practice, not all reported results that are published will have been selected for statistical significance. Among previous economic meta-analyses for which I have sufficient data, the proportion of statistically significant results varies from 29% to 79%. These proportions (minus a portion of 5%) set an upper limit to the incidence of publication selection. 100% publication selection can be eliminated as very unlikely, if for no other reason, economic research is too contentious to permit unanimous agreement. Therefore, it is assumed that the incidence of publication selection is either: 0%, 25%, 50% or 75%. The latter three correspond roughly to what Hedges and Vevea (1996) call 'light', 'moderate', and 'extreme'.