DEAKIN
UNIVERSITY AUSTRALIA

Faculty of Business and Law
School of Accounting, Economics and Finance

School Working Paper

ECONOMICS SERIES

SWP 2011/4

Meta-Regression Approximations to Reduce Publication Selection Bias

T.D. Stanley and Hristos Doucouliagos

Meta-Regression Approximations to Reduce Publication Selection Bias

T.D. Stanley* and Hristos Doucouliagos**

Abstract

Publication selection bias represents a serious challenge to the integrity of all empirical sciences. We develop meta-regression approximations that are shown to reduce this bias and outperform conventional meta-analytic methods. Our approach is derived from Taylor polynomial approximations to the conditional mean of a truncated distribution. Monte Carlo simulations demonstrate how a new hybrid estimator provides a practical solution. These meta-regression methods are applied to several policy-relevant areas of research including: antidepressant effectiveness, the value of a statistical life and the employment effect of minimum wages and alter what we think we know.

Keywords: meta-regression; publication selection bias; systematic reviews, truncation

* Bill and Connie Bowen Odyssey Professor of Economics, Hendrix College, 1600 Washington St., Conway, AR, 72032 USA. Email: Stanley@hendrix.edu. Phone: 1-501-450-1276; Fax: 1-501-450-1400.

** Professor of Economics, School of Accounting, Economics and Finance, Deakin University, 221 Burwood Highway, Burwood, 3125, Victoria, Australia. Email: douc@deakin.edu.au. Phone: 61 03 9244 6531.

Meta-Regression Approximations to Reduce Publication Selection Bias

1. INTRODUCTION

> Many other commentators have addressed the issue of publication bias. . . . All agree that it is a serious problem— Begg and Berlin (1988, p. 421).

The bias that arises from the preferential reporting of statistically significant or 'positive' scientific results has long been a focus and concern of statisticians (Sterling 1959; Rosenthal 1979; Hedges and Olkin 1985; Begg and Berlin 1988; Sterling, Rosenbaum, and Weinkam 1995; Copas 1999; Senn 2008; Mandel and Rinott 2009, to mention a few). This 'publication bias' is widely recognized to exaggerate the effectiveness of pharmaceuticals (Friedman 2003; Cary 2008; Turner et al. 2009).[1] Others have found publication selection to be widespread in the natural sciences and economics (Sterling, Rosenbaum, and Weinkam 1995; Doucouliagos and Stanley 2008). For example, the adverse employment effect that is often claimed to be the result of raising the minimum wage is shown to be nothing more than the selective reporting of negative employment effects (Doucouliagos and Stanley 2009).[2]

To minimize publication selection bias, the leading medical journals require the prior registration of clinical trials as a condition of their later publication (Krakovsky, 2004). Nonetheless, a recent systematic review found that publication selection is quite common in medical research (Hopewell et al. 2009). The problem created by publication bias can be so severe that it would be better, statistically, to discard 90% of empirical research (Stanley, Jarrell and Doucouliagos 2010). Without some way correct or minimize this bias, the validity of science itself comes into question (Lehrer 2010).

In this paper, we offer a practical solution to the exaggerated scientific record. Simple meta-regression models can greatly reduce publication selection bias. Following

[1] 'Publication bias' is somewhat a misnomer; 'reporting bias' would be a more accurate reflection of this threat to scientific validity. Because the preference for statistical significance is widely known among researchers, they will tend to select statistically significant findings even in their unpublished working papers and theses. On the other hand, funders may choose not to submit less than strongly positive results of the randomized clinical trials (RCT) of medical treatments to a journal —hence 'publication selection.'

[2] An adverse employment effect from minimum wage raises is the expected effect according to conventional economic theory—see any economics textbook or Doucouliagos and Stanley (2009). These findings are discussed in greater detail in Section 4 below.

the seminal work of Begg and Berlin (1988) and Copas (1999), we recognize that it may not be feasible to estimate all the needed parameters of a fully specified statistical model of publication selection. "It is difficult to conceive of a correction methodology which would be universally credible" (Begg and Berlin, 1988, p. 440). Nonetheless, we identify an approximate meta-regression model from considerations of limiting cases and a quadratic Taylor polynomial for the expected value of a truncated normal distribution. Furthermore, this meta-regression model easily accommodates research heterogeneity from different methods, data, populations, controls, *etc*. and can thereby distinguish publication selectivity from more substantive research differences.[3]

Simulations show that a quadratic meta-regression approximation can greatly decrease publication selection bias found in the conventional meta-analytic summary statistics of reported research results. This approach has already been successfully applied to correct highly exaggerated research on: the employment consequences of raising the minimum wage (Doucouliagos and Stanley, 2009), health care and income (Costa-Font et al. 2011), the trade effects of joining the Euro (Havranek 2010), and the relation of foreign investments and taxes (Feld and Heckmeyer 2011). The purpose of this paper is to provide a theoretical basis for our meta-regression model, to offer an improved combined estimator, and to compare the bias and efficiency of alternative approaches through Monte Carlo simulation.

2. MODELS OF PUBLICATION SELECTION

2.1 Publication Selection as Truncation

When all results are selected to be statistically significant in the desirable direction, reported effects may be regarded as 'incidentally' truncated.[4] It is 'incidental' truncation

[3] In some cases, it may be impossible to distinguish fully between genuine heterogeneity and publication selection bias. For example, assume that there is a drug that is very effective in a small sub-population, but not very effective in general. Further assume that the drug's producer chooses to publish trials which target this sub-population and suppress findings from broad populations of patients. In this scenario, the exaggerated effects found in the research record are both a result of publication bias and also a genuine biological phenomenon about which the scientific community needs to know. Our approach allows for both effects and lets others assess their meanings. Section 4.1 illustrates these multiple meta-regression methods.

[4] Here, we are only interested in directional publication selection. It is directional selection that is the main threat to medical research, favoring results that are 'positive.' In the social science, selection is typically in the direction of the currently favored theory. However, over time favored theory will likely change causing

because the magnitude of the reported effect, itself, is not selected but rather some related variable, for example the calculated z- or t-value (Wooldridge 2002, p. 552). With publication selection for directional statistical significance, we observe an estimated effect only if $effect_i/\sigma_i > a$ (assuming that these estimated effects have a normal distribution).[5]

By referring to the well-known conditional expectation of a truncated normal distribution, it is easy to show that observed effects will depend on the population's 'true' effect, μ, plus a term that reflects selection bias.

$$(1) \qquad E(effect_i \mid truncation) = \mu + \sigma_i \cdot \lambda(c)$$

Where $\lambda(c)$ is the inverse Mills' ratio, μ is the 'true' effect, which is the expected value of the original distribution, σ_i is the standard error of the estimated effect, $c = a - \mu/\sigma_i$, and a is the critical value of the standard normal distribution (Johnson and Kolz 1970, p. 278; Green 1990, Theorem 21.2).

When we replace reported sample estimates for the population values in (1),

$$(2) \qquad effect_i = \mu + SE_i \cdot \lambda(c) + \varepsilon_i$$

Equation (2) may be interpreted as a meta-regression of observed effect on its standard error. Unfortunately, $\lambda(c)$ is not generally constant with respect to μ and σ_i, and this complication causes considerable difficulty in finding an unbiased corrected estimate.

To clarify the context of this selection problem, we briefly digress. In econometrics, there is the well-known Heckman two-step solution to the analogous problem of sample selection (Heckman 1979; Wooldridge 2002; Davidson and MacKinnon 2004). However, in this empirically tractable case of sample selection,

---

a predictable lessening of publication bias (Kuhn 1962; Stanley, Doucouliagos and Jarrell 2008; Leher 2010). When selection is genuinely in both directions, publication bias will likely be smaller and much less problematic.

[5] The below argument will also hold in large samples if the estimates are asymptotically normal, such as regression coefficients under rather weak assumptions— i.i.d. residuals and X′X/n is a finite positive definite matrix (Greene 1990).

characteristics of the unselected individuals are observed and used to estimate a selection equation, by logit or probit. The estimated values of the inverse Mills' ratio, $\lambda(c)$, from this selection relation are then used to estimate the Heckman regression, which is similar to our equations above. What makes the Heckman approach feasible is the additional information contained in the selection variables that are observed whether the individual is selected or not. We do not have the luxury of extra relevant information in the case of publication selection. In general, nothing further is known about the unreported empirical research results. Thus, this well-worn avenue is unavailable for the problem at hand.

Rather than give up altogether, let us approximate the publication bias term, $SE_i \cdot \lambda(c)$, by other means. Recall that the inverse Mills' ratio is the normal probability density function evaluated at $c = a\text{-}\mu/\sigma_i$, $\phi(c)$, divided by one minus its cumulative density, $[1\text{-}\Phi(c)]$. As a consequence, this term is a complex function of μ and σ_i. To survey this complexity, we take the derivative of equation (1) with respect to σ_i.

$$(3) \qquad \partial E(effect_i \mid truncation)/\partial\sigma_i = \lambda(c) + \sigma_i \cdot \partial\lambda(c)/\partial\sigma_i$$
$$= \lambda(c) + \sigma_i \cdot \partial\lambda(c)/\partial c \cdot (\partial c/\partial\sigma_i)$$

However, $\partial\lambda(c)/\partial c = \lambda(c)^2 - c\lambda(c)$ (Heckman 1979, p. 159), which gives:

$$(4) \qquad \partial E(effect_i \mid truncation)/\partial\sigma_i = \lambda(c) + (\mu/\sigma_i) \cdot (\lambda(c)^2 - c\lambda(c))$$

This derivative suggests that the conditional mean is, in general, a rather complex, nonlinear function of σ_i; thus, some approximation such as the Taylor polynomial (or power series) will need to be employed to estimate the expected empirical relation between a reported estimate and its standard error.[6]

[6] There is a rich, two hundred year history of constructing limits and approximations for the Mills' ratio; hence also for inverse Mills' ratio (Laplace 1812; Johnson and Kotz 1970). Some of these approximations are in fact power series (Abramowitz and Stegun, 1964). For our application, all of these approximations will involve complex functions of μ/σ_i and thereby involve the very parameter, μ, we wish to estimate. Unfortunately, we find no specific estimation model that can be derived from these approximations. A

(5) $$effect_i = \beta_1 + \sum_{k=1}^{K} \alpha_k SE_i^k + \varepsilon_i$$

Estimates of β_1 from this Taylor polynomial approximation, equation (5), will then serve as estimates of the 'true' effect, μ. Econometricians typically employ linear or quadratic approximations in similar applications. In our simulations, below, we investigate quadratic (*i.e.*, K=2), cubic (*i.e.*, K=3), as well as linear approximations (*i.e.*, K=1). However, before we turn to these simulations, we need to make several relevant observations.

2.2 Examining Limit Cases of Publication Selection

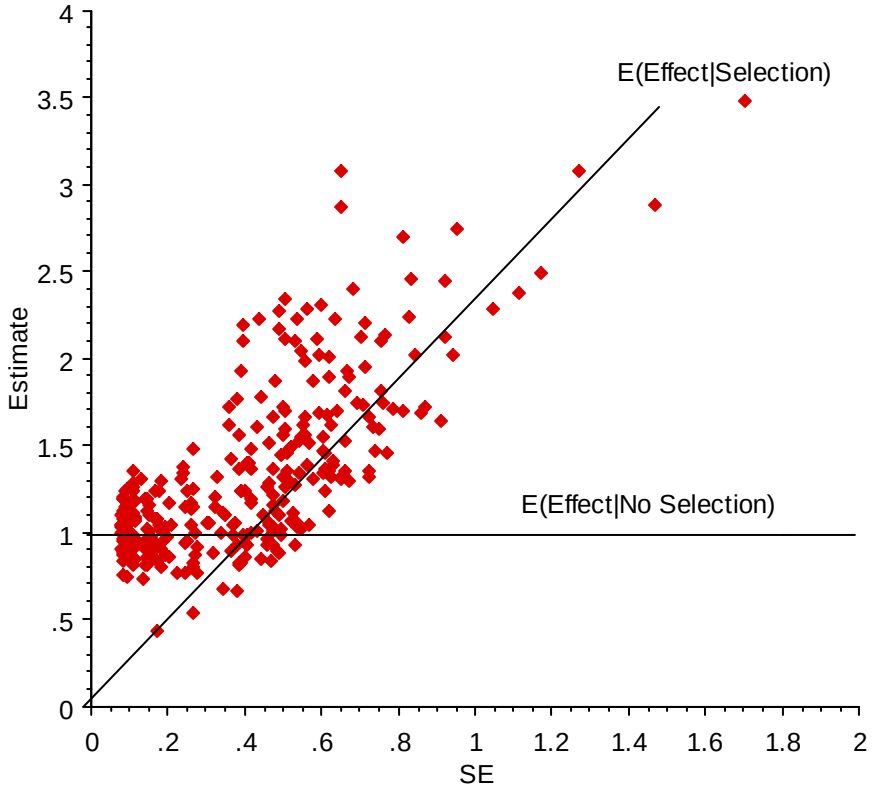
Examining limit cases reveal how a parabola in SE_i might provide an adequate approximation to the relation between the effect size and its standard error. Figure 1 plots 300 randomly generated yet selected effects when there is strict selection of significantly positive effects and the true effect is one (μ=1). These randomly generated values come from the same data generating processes used by the simulations reported and discussed in the next section. However, for our present purposes, the limiting cases of publication bias represented by the two lines in Figure 1 are much more informative than any random scatter of selected results. These limit cases give shape to the relationship between the expected reported effect and the standard errors. As we discuss below, this shape is known *a priori* from statistical theory.

To understand the shaping forces of these simple lines, first consider the horizontal line, E(Effect| No Selection). When all empirical findings are reported with no selection, they will be randomly distributed, by definition, around the true effect — μ=1 for this illustration. Without selection, the magnitude of the reported effect will be constant and independent of SE_i ; hence the horizontal line. Next, note the upwardly sloping line in Figure 1, E(Effect| Selection). This second line represents the conditional

possible exception is Gordon's (1941) upper bound for the Mills' ratio. When applied to our equation (1) gives $E(effect_i \mid truncation) = a\sigma_i$ as a lower bound. However, our limit cases, especially E(Effect| Selection), in Figure 1 below are more informative and useful.

expectation, equation (1), when the true effect is zero, μ=0. This upward-sloping line represents the worse case scenario for publication bias. The slope of this line will be equal to the inverse Mills' ratio evaluated at the critical value, a.[7] To a greater or lesser extent, these two polar cases shape the reported effects.

FIGURE 1: LIMITING CASES OF PUBLICATION SELECTION



Beginning with the most precise studies (those with small SE_i), researchers will find no need to report anything other than the first observed effect. When the true effect is many times larger than the standard error, the probability of finding an insignificant effect is virtually zero. Thus, even when there is selection for a statistically positive effect, very precise studies will not be biased, assuming of course that there is some genuine positive effect to begin with. As SE increases, occasionally an estimated effect

[7] To see this, substitute μ = 0 into equation (4). This is discussed further in Section 2.3.

will not be statistically significant and will need to be re-estimated to become so.[8] Thus, for the 'middle' range of SE, expected observed effects will be gradually pulled up above the horizontal line. Notice the scatter for .3 ≤ SE ≤ .5 in Figure 1. As SE grows larger still, the standardized true effect, μ / SE_i, will play a weaker and weaker role, while the ray from the origin presents a greater attraction for reported effects. In the limit, expected reported effects and their standard errors will be linearly related, $\lambda(a)SE_i$.

As the above discussion clearly illustrates, a simple thought experiment identifies rather clearly the approximate shape of expected reported effects and their standard errors. Equally apparent is that the right half of a parabola ($E(effect_i) = \alpha_2 SE_i^2$) can approximate this relationship.[9] Note further that a parabola will also approximate this relationship when μ is increased or decreased. Changing μ lengthens or shortens the horizontal line segment before it intersects the ray from the origin. Making α_2 smaller in $E(effect_i) = \alpha_2 SE_i^2$ allows for a more gradual increase initially and a wider parabola. Of course, the fit will not be exact, but then we need only to estimate the minimum of this the parabola (*i.e.,* its vertex).

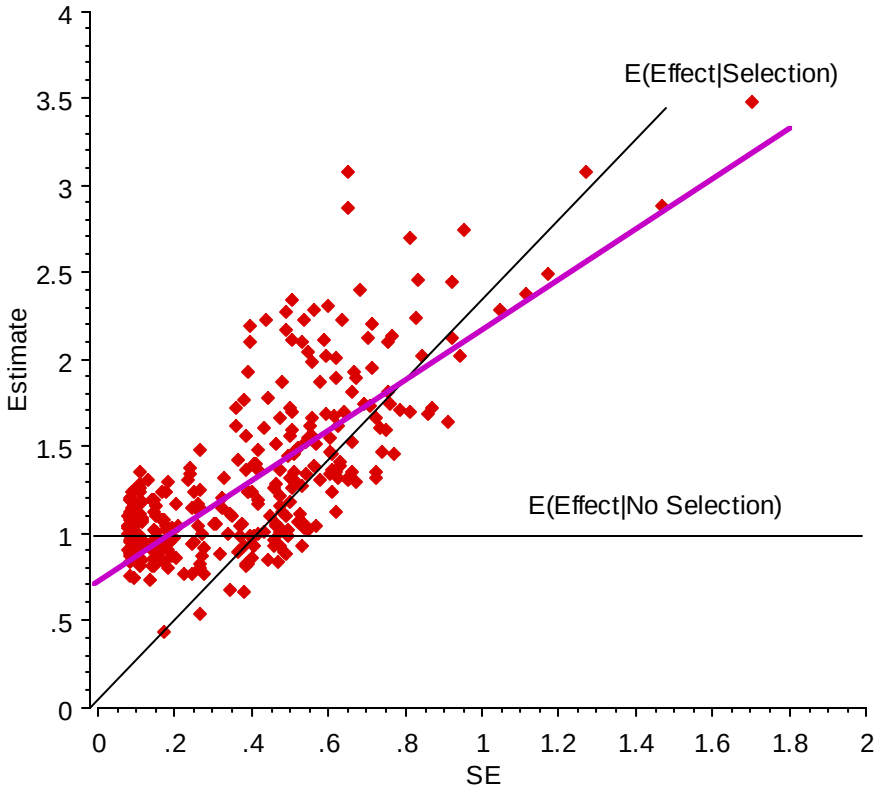
Our purpose for estimating this relationship between reported effect and its standard error is merely to find an adequate corrected estimate of effect, and we know that there will be no publication bias when SE is small, approaching zero. Recall that publication bias will be practically zero when SE is much smaller than μ and that the expected relationship will be a horizontal line for such small SEs. Thus, the slope of the fitted relationship will also need to be zero around SE =0. Figure 1 also makes this point clear. To force the slope a second-order polynomial (or a quadratic approximation) to be zero at SE =0 requires that the linear term be omitted from equation (5); that is, α_1=0.

[8] How multiple estimates for the same effect are generated depends on the discipline and the type of data used. In economics, where the data are observational, model specifications (independent variables and functional forms of the relations) are routinely varied. If this does not produce the needed statistical significance, econometricians are free to use different econometrics techniques and subsets of the data (perhaps by removing 'outliers'). For experimental data such as RCTs, different outcome measures can be investigated. Or, the entire clinical trial can be suppressed when significantly positive effects are not found. This is what is seen when one compares the phase II and phase III clinical trials of antidepressants that were reported in the FDA registry to those published in the medical journals (Turner et al. 2008).

[9] An exception occurs when there is no genuine effect. Then, the relation is linear. This special case is denoted as E(Effect| Selection) in Figure 1.

Our below simulations demonstrate that constraining α_1 to be zero in the quadratic approximation of equation (5) is critical. As discussed above, very precise estimates will vary around μ and contain negligible publication bias.[10] Thus our ideal corrected estimate is where this relation crosses the vertical axis. The trick, of course, is to estimate this intersection well from the statistical results typically reported in empirical studies.

FIGURE 2: LEAST SQUARES FIT OF SELECTED ESTIMATES



Using a linear approximation to the Taylor polynomial would be one approach, but not a very good one. Previous simulations show that this leads to an underestimate of the true effect when there is an effect (Stanley, 2008), and this is easily seen in Figure 2. Figure 2 places the least squares line (upward sloping with a positive intercept) through

[10] This simple observation serves as the starting point for an alternative estimate of the corrected effect—'*Top 10*.' *Top10* is the simple mean of the most precise ten percent of a research literature. This estimator has been shown to greatly reduce publication bias (Stanley, Jarrell and Doucouliagos, 2010).

this scatter of reported effects, the intercept of which underestimates the true effect by 25%. This illustration is no isolated incident, but is robustly confirmed by the simulations reported in the next section. In spite of this bias, there is an important special case where the expected reported effect and its standard will be linearly related.

2.3 Egger Regression and the Precision-Effect Test

Egger et al. (1997) uses the linear approximation to this complex relation of reported effect to its standard error as a test for the presence of publication bias.

$$effect_i = \gamma_1 + \alpha_1 SE_i + \varepsilon_i \quad (6)$$

Testing H_0: $\alpha_1 = 0$ in this simple meta-regression model is widely used in medical research to investigate whether a research literature is contaminated by publication selection. This Egger test serves as a valid if low power test for publication selection (Egger et al. 1997; Stanley 2008). This test is related to the symmetry of the associated funnel graph. A funnel graph is a plot of precision ($1/SE_i$) vs. $effect_i$, and it is widely used in systematic reviews as a visual indicator of publication selection (Stanley 2005; Stanley 2008; Stanley and Doucouliagos 2010).[11] Because this meta-regression relation contains obvious heteroscedasticity, equation (6) is almost never estimated using ordinary least squares (OLS), but rather weighted least squares (WLS). WLS can be obtained by dividing the entire equation (6) by an estimate of the standard deviation of this heteroscedasticity (*i.e.*, SE_i).[12]

$$t_i = \alpha_1 + \gamma_1 (1/SE_i) + u_i \quad (7)$$

[11] See Stanley and Doucouliagos (2010) for a primer on funnel graphs.

[12] Statistical packages also routinely provide WLS estimates. To obtained these WLS results, meta-regression model (6) may used if the weights are specified as $1/SE_i^2$.

Where t_i is the commonly reported t-value and $1/SE_i$ is the precision of an estimate. Note that the intercept and slope coefficients are reversed from the OLS version, equation (6). Testing H_0: γ_1=0 (the 'precision-effect test' or PET) from (7) proves a valid basis for determining whether there is a genuine empirical effect beyond publication selection bias (Stanley 2008). The weakness of this linear approximation becomes apparent when one attempts to use $\hat{\gamma}_1$ as the corrected estimate of the true effect. Although PET provides a valid test for the presence of a genuine non-zero effect, $\hat{\gamma}_1$ is downwardly biased, as seen in Figure 2. The reason for this apparent discrepancy is easily explained when one realizes that the linear relation between reported effects and their standard errors can be derived as a special case of the conditional mean of a truncated distribution when the underlying true effect, μ, is in fact zero.

When the underlying empirical effect is zero (*i.e.,* μ=0), equation (4) simplifies to $\sigma_i \cdot \lambda(c)$, and the slope of the expected effect relation reduces to this inverse Mills' ratio. Further recall that $c = a - \mu/\sigma_i$. Thus, $\partial E(effect_i \mid truncation)/\partial \sigma_i$ reduces to the inverse Mills' ratio evaluated at critical value of the standard normal distribution, $\lambda(a)$, which, of course, is just a constant. Thus, when there is no genuine empirical effect, the slope of expected reported effect is a constant, and the expected reported effect and its standard error will be linearly related— illustrated by line E(Effect| Selection) in Figures 1 and 2. This observation is important because it further validates the precision-effect test (H_0: γ_1 =0). Because the null hypothesis assumes that there is no underlying effect, μ=0, a linear relation of reported effect and the standard error provides a valid basis for testing whether there is a genuine non-zero empirical effect.

To recap, the above discussion and past simulations demonstrate that a simple linear relation between an estimate and its standard error may be used to test both for the presence of publication selection bias and genuine true effect beyond publication bias (Egger et al. 1997; Stanley 2008). However, this linear approximation is also known to give biased estimates of the underlying true effect, μ (Stanley 2008). Our approach is to appeal to a higher order. In particular, we recommend using the WLS estimate of β_1 from a quadratic approximation:

(8) $$effect_i = \beta_1 + \alpha_2 SE_i^2 + \varepsilon_i$$ or

(9) $$t_i = \alpha_2 SE_i + \beta_1 (1/SE_i) + u_i$$

when meta-regression equation (8) uses $1/SE_i^2$ for weights. Note that this quadratic model of publication selection is constrained to have α_1 =0. Elsewhere, $\hat{\beta}_1$ has been called the 'precision-effect estimate with standard error' (PEESE) (Stanley and Doucouliagos 2007; Doucouliagos and Stanley 2009; Costa-Font et al. 2011; Havranek 2010). Next, we report simulations of PEESE's bias and mean squared error (MSE) and compare them to alternative approximations and estimates, including $\hat{\gamma}_1$ from the linear approximation to this relation, equation (7).

3. SIMULATION

The design of our simulations closely follows Stanley (2008) and Stanley, Jarrell and Doucouliagos (2010). The range of parameters employed is selected to mirror observed properties from several published meta-analyses. Briefly, random data are generated and used to test whether a regression coefficient is zero. Random heterogeneity and residuals are drawn from independent normal distributions. See Stanley (2008) for more complete details. Regression is chosen because it is the most common statistical technique employed in the social sciences, and it encompasses many other statistical tests, including ANOVA, t-tests, and tests of fixed-effects (Stanley, Jarrell and Doucouliagos 2010; Moore 1997).

Publication selection is modeled as the repeated sampling from these distributions until a statistically positive regression coefficient is obtained. If a given set of generated data, errors, and random heterogeneity does not produce a significant regression coefficient, an entirely new set of data, errors, and random heterogeneity are generated. This process continues until a statistically positive regression coefficient is found by chance. However, we know that not all reported scientific findings are the result of publication selection because almost all areas of research report at least a few

insignificant estimates. To ensure that our simulations are realistic and robust, varying incidences of publication selection are modeled (0%, 25%, 50%, 75%, and 100%). For example, when the incidence of publication selection is 75%, exactly three fourths of the reported values have been chosen to be statistically significant, while the first estimate generated, significant or not, is reported for the remaining 25% of the reported values.

Meta-regression sample sizes are either 20 or 80. In economics, most areas of empirical research have many times more estimates. Among 87 areas of economics research, the average number of reported estimates exceeds 200 (Doucouliagos and Stanley, 2008). In medical research, there tend to be fewer RCTs on a given topic. But some areas of medical research have more than enough estimates. For example, Turner *et al.* (2008) reports findings on 74 antidepressant trials, and Stead et al. (2008) report 42 RCT of nicotine replacement therapy using the 'patch and 112 when other delivery systems are included. The meta-regression sample size of twenty is chosen because it is a rather small sample size for any regression estimate, while eighty is both practically feasible in many cases and gives these meta-regression tests power to spare. Needless to say, regression-based estimators may not be appropriate if only a handful of comparable empirical estimates exist.[13]

In addition to the incidence of publication selection, the statistical properties of these alternative estimators are most influenced by the relative magnitude of the unexplained heterogeneity relative to the sampling errors. We use Higgins and Thompson's (2002) $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$ as the indicator of the size of the relative heterogeneity. σ_h^2 is the between-study heterogeneity variance, and σ_ε^2 is the within-study sampling variance. I^2 is analogous to R^2 in regression analysis. It reflects the proportion of the total variation due to unexplained heterogeneity. Simulations are conducted over a wide range of heterogeneity and publication selection and reported in Tables 1 through 4.[14] Although the exact calculated value of I^2 varies for each random sample, these tables state its population value when there is no publication selection.

[13] Such small samples are even more problematic for the *Top10*, which begins by discarding 90% of the reported research.

[14] Simulations for n =20 are reported in Appendix Tables 1-4.

TABLE 1: MEANS OF THE INTERCEPT OF POLYNOMIAL APPROXIMATIONS (n=80)

Hetero-geneity*	True effect	Selection Incidence	*Linear* $\hat{\gamma}_1$	*Quadratic*	*Cubic*	*PEESE*, $\hat{\beta}_1$ from (9)
	0	0%	0.00	-0.01	0.00	0.00
	0	25%	0.04	-0.08	0.04	0.13
	0	50%	0.06	-0.11	0.07	0.25
	0	75%	0.07	-0.07	0.14	0.36
I^2=25%	0	100%	0.07	0.05	0.12	0.46
	1	0%	1.00	1.00	1.01	1.00
	1	25%	0.92	0.94	1.08	0.99
	1	50%	0.85	0.89	1.11	0.97
	1	75%	0.77	0.88	1.15	0.96
	1	100%	0.68	0.87	1.12	0.94
	0	0%	0.00	0.00	0.00	0.00
	0	25%	0.04	-0.05	-0.05	0.14
	0	50%	0.08	-0.03	-0.06	0.29
	0	75%	0.14	0.04	-0.02	0.44
I^2=58%	0	100%	0.20	0.15	0.07	0.60
	1	0%	1.00	0.99	0.99	1.00
	1	25%	0.94	0.94	1.01	1.01
	1	50%	0.88	0.89	1.00	1.02
	1	75%	0.81	0.85	0.91	1.02
	1	100%	0.74	0.84	0.83	1.03
	0	0%	0.00	0.01	0.01	0.00
	0	25%	0.04	0.03	-0.09	0.18
	0	50%	0.10	0.10	-0.10	0.38
	0	75%	0.22	0.19	-0.09	0.61
I^2=85%	0	100%	0.37	0.34	0.09	0.86
	1	0%	1.00	1.00	1.02	1.00
	1	25%	0.97	0.96	0.87	1.07
	1	50%	0.93	0.90	0.76	1.12
	1	75%	0.87	0.86	0.62	1.17
	1	100%	0.80	0.82	0.47	1.21

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. *Linear*, *Quadratic*, *Cubic*, and *PEESE* refer to different estimates of the intercept of the polynomial approximation to the conditional mean of a truncated distribution—equation (5). $\hat{\gamma}_1$ is estimated from equation 7, and $\hat{\beta}_1$ is estimated from equation 8.

Table 1 reports the average of 10,000 replications for alternative polynomial approximations, equation (5), and Table 2 the associated mean squared errors (MSE) of these approximations. In all cases, the estimated intercept, β_1, is used as the corrected estimate in a WLS version of equation (5). The first column of simulation results reports the 'linear' approximation (*i.e.,* K=1) of (5) is which is equivalent to $\hat{\gamma}_1$ from equation (7). Next is the 'quadratic' approximation (*i.e.,* K=2), followed by the 'cubic' approximation (*i.e.,* K=3). Lastly, our recommended PEESE estimator which is the quadratic

approximation with the further constraint that α_1=0. PEESE is the same as estimating $\hat{\beta}_1$ in equation (9). True effects (μ) are either 0 or 1.

TABLE 2: MEAN SQUARE ERRORS OF THE INTERCEPT OF POLYNOMIAL APPROXIMATIONS (times 1,000 with n=80)

Hetero-geneity*	True effect	Selection Incidence	*Linear* $\hat{\gamma}_1$	*Quadratic*	*Cubic*	*PEESE*, $\hat{\beta}_1$ from (9)
	0	0%	27	195	1808	8
	0	25%	25	206	2148	24
	0	50%	22	193	2224	68
	0	75%	17	134	1749	135
I^2=25%	0	100%	10	42	422	214
	1	0%	27	200	1837	8
	1	25%	30	192	1783	8
	1	50%	45	191	1700	8
	1	75%	74	182	1627	8
	1	100%	115	157	1385	10
	0	0%	51	344	2862	16
	0	25%	45	341	3357	35
	0	50%	43	303	3103	97
	0	75%	44	219	2353	204
I^2=58%	0	100%	51	115	708	359
	1	0%	50	347	2927	15
	1	25%	50	317	2684	15
	1	50%	56	312	2529	13
	1	75%	73	284	2266	12
	1	100%	99	232	1841	11
	0	0%	116	616	3826	37
	0	25%	104	598	3952	63
	0	50%	94	525	3414	168
	0	75%	108	400	2466	385
I^2=85%	0	100%	172	281	955	745
	1	0%	114	621	3819	36
	1	25%	101	552	3414	35
	1	50%	93	477	3088	41
	1	75%	88	421	2716	52
	1	100%	95	330	2111	65

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. *Linear, Quadratic, Cubic*, and *PEESE* refer to different estimates of the intercept of the polynomial approximation to the conditional mean of a truncated distribution—equation (5). $\hat{\gamma}_1$ is estimated from equation 7, and $\hat{\beta}_1$ is estimated from equation 8.

Although the shape of bias (Table 1) is rather complex, a few clear patterns emerge, especially when one considers both bias and efficiency as measured by MSE. First, PEESE ($\hat{\beta}_1$) has the smallest MSE in the great majority (70%) of cases, often by a wide margin (Table 2), and it also has the smallest bias in a plurality of simulations. However, PEESE is upwardly biased when the true effect is zero. Second, the PET

coefficient, $\hat{\gamma}_1$, dominates PEESE as expected when μ=0. Recall that the linear approximation is correctly specified when true effect is zero. Nonetheless, in a few incidences either the quadratic or the cubic approximation has a smaller bias than $\hat{\gamma}_1$. Like PEESE, it is easy to see that $\hat{\gamma}_1$ is upwardly biased when μ=0. Perhaps, this upward bias is a reflection of attenuation bias (or, equivalently, 'errors-in-variables' bias) that will result from using a fallible estimate, SE_i, in the place of σ_i? Third, the unconstrained quadratic and cubic approximations are clearly inferior to either $\hat{\beta}_1$ or $\hat{\gamma}_1$. Their MSEs are typically many times larger than these other approximations. The few cases where they have a slightly smaller bias seem random and unpredictable unless we were to know the exact incidence of publication selection. In practice, we have no way to know the percent of estimates that have been selected.

The unreliability of the unconstrained quadratic and cubic approximations is likely caused by multicollinearity among powers of SE. Technically, SE, SE^2 and SE^3 cannot be 'multicollinear', but the unreliability of the estimated regression coefficients caused by multicollinearity depends only on the correlations among the independent variables. These powers of SE are highly correlated. For example, the variance inflation factor for the unconstrained quadratic approximation using the data shown in Figures 1 and 2 is 6.5 and 166 for the cubic approximation.[15] Obviously, unreliability in estimating slope coefficients will be transferred to the estimates of the intercept, which is our corrected estimate of effect. This multicollinearity-induced unreliability is clearly seen in the huge MSEs of the cubic model (Table 2). The MSEs of the cubic model get much worse still for n=20 (Appendix Table 2). Our constrained quadratic, equation (9), as well as the linear approximation, has no multicollinearity; hence, the resulting estimators are much more reliable and efficient.

Several implications and suggestions can be drawn from the relative bias and efficiency of these alternative approximations. First, both unconstrained polynomial approximations are distinctly inferior and can thereby be eliminated from further consideration. Secondly, PEESE dominates the linear approximation, $\hat{\gamma}_1$, when there is

[15] Recall that the variance inflation factor (VIF) is the conventional way to measure multicollinearity, and VIF=1-R_x^2; where R_x is the multiple correlation coefficient among the x variables.

a genuine nonzero effect. Third, the opposite is largely true when there is no genuine effect. This suggests that a combined estimator may be better than either PEESE or $\hat{\gamma}_1$, individually. We propose that meta-analysts use the PEESE estimator only when there is evidence of a nonzero effect (reject H_0: γ_1=0) in equation (7). When PET is not passed (*i.e.*, accept H_0: γ_1=0), $\hat{\gamma}_1$ should be used as the corrected estimate. We call this conditional estimator, 'PET-PEESE,' and its bias and MSE are reported in Tables 3 and 4 along with alternative conventional meta-analysis summary estimates.

TABLE 3: MEANS OF ALTERNATIVE RESEARCH SUMMARY ESTIMATORS (n=80)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	*PEESE*, $\hat{\beta}_1$ from (9)	PET-PEESE
	0	0%	0.00	0.00	0.00	0.00	0.00	-0.01
	0	25%	0.23	0.20	0.22	0.13	0.13	0.04
	0	50%	0.47	0.39	0.43	0.28	0.25	0.07
	0	75%	0.70	0.59	0.63	0.41	0.36	0.08
I^2=25%	0	100%	0.93	0.78	0.78	0.55	0.46	0.16
	1	0%	1.00	1.00	1.00	1.00	1.00	1.00
	1	25%	1.07	1.04	1.04	1.00	0.99	0.99
	1	50%	1.13	1.08	1.09	1.01	0.97	0.98
	1	75%	1.20	1.11	1.13	1.02	0.96	0.96
	1	100%	1.26	1.15	1.16	1.02	0.94	0.94
	0	0%	0.00	0.00	0.00	0.00	0.00	-0.01
	0	25%	0.27	0.23	0.25	0.14	0.14	0.03
	0	50%	0.54	0.45	0.51	0.30	0.29	0.09
	0	75%	0.81	0.68	0.75	0.47	0.44	0.16
I^2=58%	0	100%	1.08	0.91	0.92	0.66	0.60	0.43
	1	0%	1.00	1.00	1.00	1.00	1.00	1.00
	1	25%	1.10	1.07	1.08	1.02	1.01	1.01
	1	50%	1.19	1.13	1.16	1.04	1.02	1.02
	1	75%	1.29	1.19	1.23	1.07	1.02	1.02
	1	100%	1.39	1.26	1.30	1.08	1.03	1.03
	0	0%	0.00	0.00	0.00	0.00	0.00	-0.02
	0	25%	0.36	0.29	0.34	0.18	0.18	0.02
	0	50%	0.72	0.58	0.68	0.38	0.38	0.10
	0	75%	1.09	0.88	1.02	0.62	0.61	0.26
I^2=85%	0	100%	1.45	1.20	1.29	0.88	0.86	0.72
	1	0%	1.00	1.00	1.00	1.00	1.00	0.98
	1	25%	1.18	1.13	1.17	1.06	1.07	1.04
	1	50%	1.36	1.27	1.33	1.12	1.12	1.09
	1	75%	1.54	1.39	1.49	1.17	1.17	1.15
	1	100%	1.73	1.52	1.63	1.22	1.21	1.20

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. *Top10* is the simple average of the most precise 10% of the observations. $\hat{\beta}_1$ is estimated from equation 8.

TABLE 4: MEAN SQUARE ERRORS OF ALTERNATIVE RESEARCH SUMMARY ESTIMATORS (times 1,000 with n=80)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	*PEESE*, $\hat{\beta}_1$ from (9)	PET-PEESE
	0	0%	3	3	3	14	8	22
	0	25%	58	41	49	33	24	23
	0	50%	221	155	186	88	68	23
	0	75%	494	344	396	180	135	24
I^2=25%	0	100%	875	603	603	310	214	73
	1	0%	3	3	3	14	8	8
	1	25%	7	4	5	13	8	8
	1	50%	20	8	10	13	8	8
	1	75%	41	15	18	13	8	8
	1	100%	71	25	28	13	10	10
	0	0%	6	6	6	27	16	42
	0	25%	78	55	69	49	35	42
	0	50%	295	207	260	115	97	45
	0	75%	658	464	560	243	204	65
I^2=58%	0	100%	1168	830	839	447	359	255
	1	0%	6	6	5	27	15	16
	1	25%	15	9	11	26	15	16
	1	50%	42	22	30	25	13	14
	1	75%	88	42	58	25	12	13
	1	100%	152	70	92	26	11	11
	0	0%	16	14	14	64	37	96
	0	25%	145	94	129	97	63	96
	0	50%	535	344	477	206	168	95
	0	75%	1087	788	1040	422	385	159
I^2=85%	0	100%	2100	1436	1674	792	745	619
	1	0%	16	14	14	63	36	59
	1	25%	46	30	39	59	35	58
	1	50%	144	80	120	63	41	64
	1	75%	305	161	245	71	52	72
	1	100%	534	273	406	82	65	74

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. *Top10* is the simple average of the most precise 10% of the observations. $\hat{\beta}_1$ is estimated from equation 8.

Tables 3 and 4 display the bias and efficiency of PEESE, the combined estimator, PET-PEESE, and several conventional summary meta-estimates. The fixed- and random-effects estimators (FEE and REE) are weighted averages or the reported effects, where the weights are the inverse of the estimates' variances. REE employs a more complex variance estimate that includes the between-study variance, σ_h^2 (Cooper and Hedges, 1994). In our simulation, excess unexplained heterogeneity is always included; thus, by conventional practice, REE should be preferred over FEE. However, conventional practice is wrong when there is publication selection. With selection for statistical

significance, REE is always more biased than FEE (Table 3). This predictable inferiority is due to the fact that REE is itself a weighted average of the simple mean, which has the largest publication bias, and FEE. Both weighted averages are less biased than the simple mean because they give greater weights to the less selected and smaller biased estimates, which tend to be the most precise (recall our discussion in Section 2).

The simple average is included in Table 3 and 4 to document how large the publication biases are when there is selective reporting of scientific results. The magnitude of this bias can be especially severe when there is no genuine underlying empirical effect. *Top10* is a more radical weighted average introduced by Stanley, Jarrell and Doucouliagos (2010) to emphasize the importance of publication bias for scientific inference. *Top10* is the simple average of the most precise 10% (smallest standard errors) of the reported research results. That is, 90% of research results have a weight of 0, while the most precise 10% are given a weight of 1. Publication bias is such a serious threat to the integrity of scientific inference that it is often better to just throw out 90% of the reported research (Stanley, Jarrell and Doucouliagos, 2010).

For all incidences of selection, *Top10* has smaller bias than any of the conventional summary statistics that use all the research results. Surprisingly, throwing away 90% of the research is more efficient in the majority of cases (Table 4). In spite of this amazing performance, the meta-regression estimators derived here are clearly better than the *Top10* and the more conventional summary statistics. We do not report the statistical properties for the popular nonparametric 'trim-and-fill' correction strategy because previous 'comprehensive simulations' reveal that its statistical performance is unacceptable, especially when compared to meta-regression methods (Duval and Tweedie 2000; Moreno *et al*. 2009).

For ease of comparison, we report the simulation results for PEESE ($\hat{\beta}_1$) in Tables 3 and 4 along with our new hybrid estimator, PET-PEESE. First, notice how PEESE dominates all of the conventional summary estimators and *Top10*. Table 4 shows very clearly that PEESE has smaller MSE when there is publication selection. Even when there is no selection, PEESE has only slightly larger variance. Otherwise, there is little reason to use any of the better known summary statistics in a systematic review.

Only *Top10* has smaller bias in any of these simulation combinations, and this occurs only in a small minority of cases.

Lastly, note that our conditional estimator ($\hat{\beta}_1$ when we reject H_0: γ_1=0 and $\hat{\gamma}_1$ when we fail to reject it) improves upon PEESE. If there is any selection for statistical significance, PET-PEESE has equal or smaller bias, in some cases by several times. When there is no publication selection the conditional estimator has a very small downward bias. Overall, however, PET-PEESE has the smallest average bias among any of these estimators. When it comes to efficiency, the simulations are less favorable to PET-PEESE. Nonetheless, it has equal or smaller MSE than PEESE in the majority of cases, and recall that PEESE is more efficient than any of these other estimates in the great majority of cases (Table 4). Thus, our new conditional estimator is the best choice whenever a research literature is suspected to contain publication selection, and such a suspicion will be warranted for most empirical literatures across the social, medical and natural sciences.

These meta-regression methods do not perform quite as strongly when there are only 20 estimates available (n=20)—see Appendix Tables 1-4. Nonetheless, they still have lower average bias and MSE than the conventional alternatives. Even when there are only twenty estimates, PEESE has the lowest average MSE, and PET-PEESE has the lowest bias.

In spite of these favorable findings, we would be remiss if we did not recommend some caution. The largest threat to these meta-regression methods of publication bias reduction occurs when there is no genuine underlying empirical effect (*i.e.*, μ=0). In these cases, all estimators are biased if there is selection for statistical significance. In the unlikely case that all studies are prepared to report only statistically positive effects, very large biases are manufactured. However, even under such worse case scenarios, PET-PEESE has a much smaller bias than the other alternatives, reducing the publication bias seen in the simple mean by at least half and often much more. When there is evidence of publication bias (reject H_0: α_1=0 in equation 7) but no evidence of an underlying empirical effect (accept H_0: γ_1=0), caution might suggest that we offer no summary estimate of effect. Secondly, meta-regression methods (and *Top10)* are unlikely to be

reliable when there are only a handful of comparable research results in a given area. In such cases, FEE is likely to provide the best summary of a systematic review. However, when there are as few as 20 estimates these meta-regression methods still fare rather well relative to alternative methods.

In sum, when there are sufficient reported estimates, we advocate that meta-analysts first run meta-regression model (7). If they find evidence of a genuine empirical effect (reject H_0: $\gamma_1=0$), then use $\hat{\beta}_1$ from MRA (9) as the corrected estimate of effect.[16] Otherwise, $\hat{\gamma}_1$ should be employed. To be conservative, one should always use either $\hat{\beta}_1$ or $\hat{\gamma}_1$ even if there is insufficient evidence of publication selection (*i.e.*, accept H_0: $\alpha_1=0$ in equation 7) because the Egger test is known to have low power (Egger et al. 1997; Stanley 2008).

4. PRACTICAL SIGNIFICANCE[17]

In many areas of empirical science, correcting for publication bias will make an important practical difference to our understanding. For example, the magnitude of the value of a statistical life (VSL) is a critical parameter for many public health and safety initiatives. These statistical estimates may be derived from hedonic wage equations that gage how workers choose between higher wages and less job safety (Viscusi, 1993). A meta-analysis of 39 separate hedonic wage estimates reveals an average value of a statistical life to be $9.5 million (Bellavance et al. 2009). Table 5 reports the meta-regression findings for these value estimates using meta-regression models (7) and (9). The value of a statistical life is reduced by 82% when publication selection is considered; PEESE = $1.67mil. Needless to say, there is clear evidence of publication bias (reject H_0: $\alpha_1=0$; $p<.01$). Which researcher would be willing to report that the value of life is negative? Also, there strong evidence that VSL is genuinely larger than zero (reject

[16] Of course, MRA models (6) and (8) may be used in place of (7) and (9), respectively, when a WLS statistical routine is also employed.

[17] For illustrative purposes, we have *selected* four areas of research where there is clear evidence of publication bias. We do not wish to imply that all areas of research have evidence of such large publication bias. Here, we wish only to show that the methods advanced in this paper can actually make a large practical difference for some important applications.

$H_0: \gamma_1 = 0$; p<.01); thus, PET-PEESE would also be $1.67mil. Needless to say, reducing VSL by 82% greatly reduces the number of health and safety projects or regulations that are socially beneficial (or cost effective).

The adverse employment effect from a raise in the minimum wage is another important dimension for public policy. Raising the minimum wage always engenders a public controversy that is often stated in terms of harm to workers. When we apply these methods to 1,474 estimates of the effect of minimum wage on employment, a small adverse employment effect, -0.19, is reduced to one that is both statistically and practically insignificant, -0.009 (Doucouliagos and Stanley, 2009). Because we accept $H_0: \gamma_1 = 0$, $\hat{\gamma}_1 = -0.009$ is our preferred estimate. These effects are measured in terms of elasticity, which in this case measures the percent decrease in teen employment that results from a one percent increase in the minimum wage. Our corrected estimate of effect, -0.009, implies that a doubling of the minimum wage would cause a less than one percent reduction of teen employment.

TABLE 5: CORRECTED ESTIMATES AND META-REGRESSION MODEL (7)
Dependent Variable = t

Variable	*Minimum Wage*	*Statistical Life*	*NRT Patch*	*Anti-Depressants*
Intercept ($\hat{\alpha}_1$)	-1.60(-17.36)*	3.20 (6.67)*	1.09 (2.38)*	1.84 (5.47)*
$\hat{\gamma}_1$	-.0009 (-1.09)	0.81 (3.56)	.197 (1.29)	.13 (2.50)
Simple Mean	-.19	$9.5 mil	.657	.47
PEESE	-.036	$1.67mil	.314	.29
n	1474	39	42	50

*t-values are reported in parenthesis.

No doubt, some economists will fail to believe such a large correction of minimum wage's adverse employment effect. However, we find that a negligible practical effect from minimum wage is a very robust summary of this extensive empirical literature. This employment effect remains practically insignificant whether one uses PEESE=-0.036, *Top10*=-0.0217, or multiple meta-regression results that use dozens of

moderator variables (Doucouliagos and Stanley, 2009). In actual applications, the simple meta-regression models of publication selection bias advanced here need to be embedded within more complex, multiple meta-regression models that also account for observed systematic heterogeneity.[18] Conservatively, the modest average adverse employment effect found in the minimum-wage literature is reduced by a factor of 6 when observed publication selection is accommodated.

Or, take a medical example with public health policy implications. Stead et al. (2008) systematically review all of the clinical trials of nicotine replacement therapy (NRT) for smoking cessation, 42 of which involve the 'patch.' Table 5 reports the meta-regression findings for these clinical trails and indicates publication selection (reject H_0: α_1=0; p<.05). The average log risk ratio is .657, which implies that smokers who use the 'patch' are 93% more likely to quit smoking. Because these clinical trials do not pass the precision-effect test (*i.e.*, accept H_0: γ_1=0), the PET coefficient, $\hat{\gamma}_1$ =.197, is our preferred corrected estimate. Such a correction reduces the efficacy of the patch to only 22%.

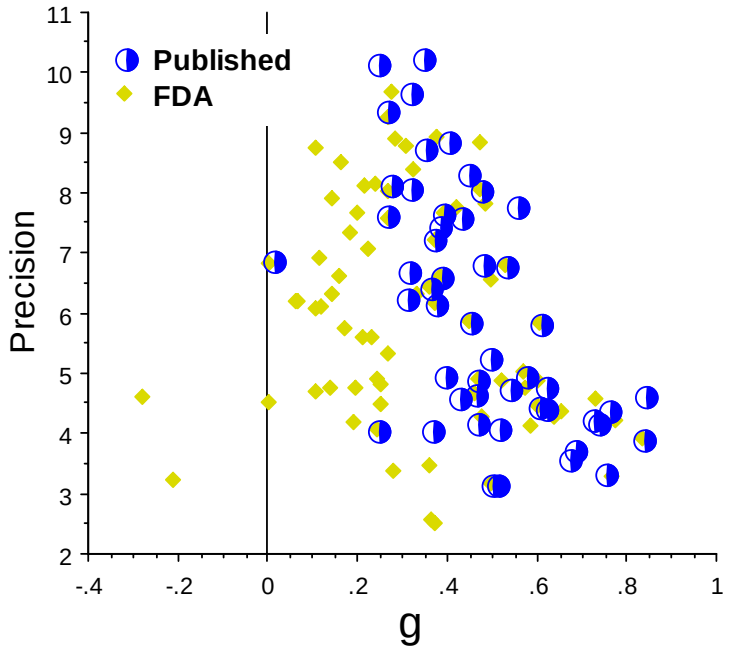
Lastly, we use apply these meta-regression methods to the controversial issue of the effectiveness of antidepressants. Turner et al. (2008) tracked down all of the phase II and phase III trials of antidepressants registered at the US Department of Food and Drug (FDA) and those that were also published. To sell pharmaceuticals in the US, RCTs of their safety and efficacy must be reported to the FDA. Thus, the FDA registry of clinical trials is considered the 'gold standard.' Of these 74 RCTs of antidepressants, only 50 are published in the journals (Turner et al., 2008).

Figure 3 displays the funnel graph for the FDA gold standard, and Figure 4 plots only the published trials. In Figure 3, published trails are shown twice. First, as they were reported to the FDA, 'diamond', and secondly as published, 'half moon.' It is difficult to imagine a clearer depiction of selective reporting. A funnel graph is a plot of an estimate's precision (1/SE) vs. the magnitude of the reported effect, measured here by effect size (Glass's g). Cochrane and Campbell reviews often use a visual inspection of a

[18] See the next section, 4.1, for a brief illustration. Current space does not permit a detailed discussion of the conventional econometric practice of using moderator variables in a multiple meta-regression model to explain observed variation among research results. See Doucouliagos and Stanley (2009), Costa-Font et al. (2010), Havranek (2010), and Feld and Heckmeyer (2011).

funnel graph for asymmetry as their test of publication selection. However, the associated funnel-asymmetry test (FAT; $H_0: \alpha_1=0$; Table 5 and Equation 7) is a more objective and reliable statistical test for publication bias (Egger et al. 1997; Stanley, 2008). Among published antidepressant trials, FAT agrees with a visual inspection of Figure 3 and finds significant publication selection for positive effects (t=5.47; p< .01).

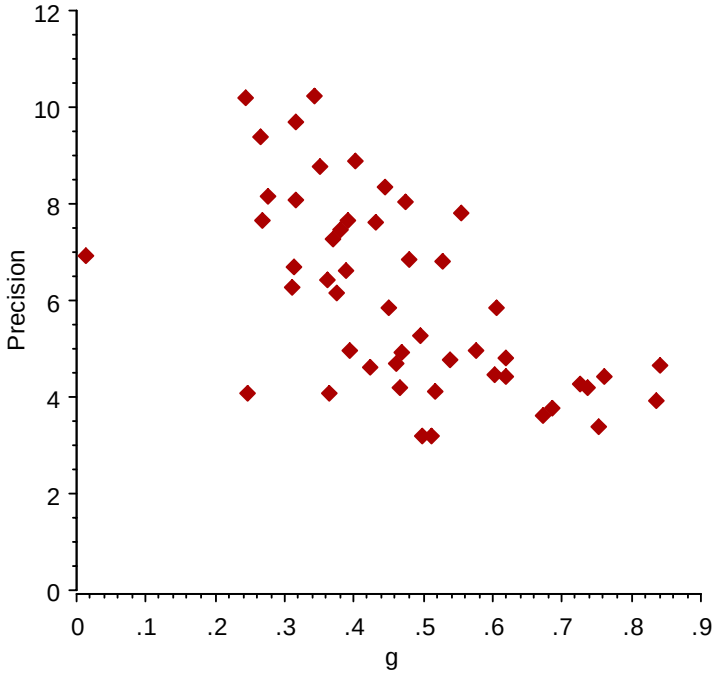
FIGURE 3: FUNNEL OF FDA REGISTRY OF ANTIDEPRESSANTS TRIALS



Fortunately, there is also evidence of a genuine positive clinical effect from taking antidepressants (t=2.30; p< .05). However, the modest average effect size of 0.47 is exaggerated by over 60% when compared to PEESE = 0.29. Note how $\hat{\beta}_1$ is twice as large as $\hat{\gamma}_1$ here and also for our other example where PET is passed. Thus, using the right approximation can make an important practical difference. Our corrected meta-regression estimate for the effect size of antidepressants is almost exactly equal to the weighted average, 0.31 (FEE and REE), of those trials reported to the FDA (Turner et al., 2008). Knowing that antidepressants have a smaller effect might change clinical practice

and thereby affect millions of patients.[19] This would be even more likely if doctors were to factor in the well-documented risks from taking antidepressants.

FIGURE 4: FUNNEL PLOT PUBLISHED ANTIDEPRESSANTS TRIALS



4.1 Explaining Systematic Heterogeneity through Multiple Meta-Regression Analysis

Perhaps the best thing about these meta-regression approximations for publication selection is that they easily accommodate systematic heterogeneity. Nearly all areas of empirical research contain excess systematic heterogeneity. That is, empirical effects depend or the population being treated, the severity of the subjects' prior conditions, dosage, the exact treatment protocol, etc. Among hundreds of meta-regression analyses of economics research, none have found the absence of excess heterogeneity as measure by the conventional Cochran's Q-test (Cooper and Hedges 1994). In all cases, meta-analysts have found that the choice of variables, econometric model, and methods makes a huge practical difference to reported research results. Thus, statistically valid meta-

[19] Recall that the conventional Cohen guideline suggests that effect sizes between .2 and .5 are 'small.' PEESE= .29. Less that .2, are considered negligible.

analyses must also accommodate systematic heterogeneity. Other publication correction strategies do not (Moreno et al. 2009).

Explaining reported research variation can easily be accomplished by expanding these meta-regression models of publication selection. For example, meta-regression model (8) becomes:

$$effect_i = \beta_1 + \sum_k \delta_k Z_k + \alpha_0 SE_i^2 + \sum_j \alpha_j K_j SE_i^2 + \varepsilon_i \quad (10)$$

where Z_k are moderator variables that may help to explain genuine systematic variation among reported findings, and K_j are selection variables that are related to publication bias.

To illustrate the power of this multiple meta-regression model, we estimate the WLS version of (10) using the Turner et al.'s (2008) data on antidepressants effectiveness. In addition to effect size and standard errors, they also note which antidepressants were used. Meta-regression model (10) allows different drugs to have different levels of effectiveness (Z_k) and also different propensities to selectively publish their findings (K_j).[20] Next, we employ a general-to-specific strategy where insignificant variables were removed one at a time starting with the one with the largest p-value and report the results in Table 6.[21]

Only floxetine (popularly know as prozac) seems to have a differential (yet negligible) level of effectiveness. Four antidepressants (including floxetine) exhibit a greater tendency to select which results to report. This multiple meta-regression model estimates floxetine's corrected effect size to be only 0.145, and the average for the remaining antidepressants is larger but still small, 0.366, by Cohen's guideline. Another systematic review also finds that floxetine is less effective (Cipriani et al. 2005).

[20] Obviously, it is not the drugs themselves that are doing the selection. However, drug manufactures (or research funders) may exert differential pressures, implicitly or explicitly.

[21] This is an accepted modeling strategy in econometrics to minimize the threats from data-mining (Charemza and Deadman, 1997).

TABLE 6: MULTIPLE META-REGRESSION OF ANTIDEPRESSANT PUBLISHED TRIALS (10)

Dependent Variable— g	MRA Coefficient	t	p-value
(Constant)	.366	18.078	.000
fluoxetine	-.221	-2.948	.005
fluoxetineSE2	4.813	2.369	.022
mirtazapineSE2	3.843	2.964	.005
paroxetineSE2	3.434	3.621	.001
venlafaxineSE2	6.016	3.385	.002

Based on equation (10). Variables ending with SE2 are K variables.

5. CONCLUSION

Publication selection bias is a widely recognized threat to the validity of empirical scientific inquiry. This threat is often so severe that a balanced assessment of the efficacy of medical treatment is difficult or impossible. This threat remains even when there have been clear findings reported from the 'gold standard' of empirical science— double-blind, placebo-controlled randomized clinical trials. In the social sciences where empirical inquiry often uses observational data, this bias is routinely much worse still (Doucouliagos and Stanley, 2008). Fortunately, there is a long history of statistical interest in this problem. Unfortunately, corrections for publication selection bias have not been widely adopted, and their performance and reliability has been wanting.

In this paper, we offer meta-regression methods that are easy to apply and are likely to greatly reduce publication selection bias in most applications. Although these methods are based on imperfect approximations to the statistical model of the conditional mean of a truncated distribution, they offer a practical solution to this important threat to modern science. As a side effect of investigating the theoretical foundation for our meta-regression model of publication selection, we are able to explain both the success and the bias of the Egger meta-regression model, which is based on a linear approximation to a complex nonlinear function. Nonetheless, this linear approximation provides adequate tests of both the existence of selection and the presence of a genuine nonzero empirical

effect beyond publication bias (Stanley, 2008). Unfortunately, the linear approximation does not offer a suitable corrected estimate when there is nonzero 'true' effect.

For these cases, we demonstrate how a constrained quadratic approximation, PEESE, to the conditional expected value of a truncated distribution is considerably less biased and often more efficient. Furthermore, simulations demonstrate how a hybrid between these two approximations improves the correction for publication selection bias yet further. Both approximations are very simple to apply, merely ordinary least squares of common statistics (t-values, standard errors, and precision) or, equivalently, weighted least squares of reported effects, their standard errors and variances. To date, no better strategy for correcting publication bias has been offered.

Needless to say, these methods have limitations. First, being based on regression analysis, they require more than a handful of estimates on the same empirical phenomenon. Second, overwhelming unexplained systematic heterogeneity can invalidate the underlying meta-regression tests (*i.e.,* the precision-effect test) (Stanley, 2008). However, when unexplained heterogeneity is responsible for more than 90% of the observed variation among reported research results, publication biases will expand greatly. Thus, balanced scientific assessment does not have the luxury to do nothing. Even in these extreme cases, the methods advanced here will remain a marked improvement over conventional meta-analytic summary statistics.

References

Abramowitz, M., and I. A. Stegun 1964. Handbook of mathematical functions with formulas, graphs and mathematical tables. Washington, DC: U.S. Department of Commerce.

Begg, C. B. and Berlin, J.A. 1988. Publication bias: A problem in interpreting medical data,*Journal of the Royal Statistical Society A*, 151:419-445.

Bellavance, F. Dionne, G. Lebeau, M. 2009. The value of a statistical life: A meta-analysis with a mixed effects regression model. *Journal of Health Economics* 28: 444–464.

Bland, J.M. 1988. Discussion of the paper by Begg and Berlin. *Journal of the Royal Statistical Society A* 151:450-451.

Bowland, B.J. and J.C. Beghin. 2001. Robust estimates of value of a statistical life for developing economies. *Journal of Policy Modeling* 23:385-96.

Card, D. and Krueger A.B. 1995. Time-series minimum-wage studies: A meta-analysis. *American Economic Review* 85: 238-43.

Cary, B. 2008. Researchers find a bias towards upbeat findings on antidepressants. *New York Times*, January 17.

Charemza, W. and D. Deadman. 1997. *New Directions in Econometric Practice*, 2nd edition. Cheltenham: Russell Edward Elgar.

Cipriani A., Brambilla P., Furukawa T.A., Geddes J., Gregis M., Hotopf M., Malvini L., Barbui C. 2005. Fluoxetine versus other types of pharmacotherapy for depression. *The Cochrane Library*, Issue 4. Available at http://www.thecochranelibrary.com.

Cooper, H.M. and Hedges, L. V. (eds.) 1994. *Handbook of Research Synthesis.* New York: Russell Sage

Copas, J. 1999. "What works? Selectivity models and meta-analysis," *Journal of the Royal Statistical Society*, A , 161:95-105.

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

Davidson, R. and MacKinnon, J.G. 2004. *Econometric theory and methods.* Oxford: Oxford University Press.

De Long, J.B. and Lang, K. 1992. Are all economic hypotheses false? *Journal of Political Economy* 100:1257-72.

Duval S. and Tweedie R. L 2000. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* 95:89-98.

Doucouliagos, C. (H.) and Stanley T.D. 2008. Theory competition and selectivity. Working Paper, Economics Series 2008, Deakin University.

Doucouliagos, C.(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.

Egger, M., Smith, G.D., Scheider, M., and Minder, C. 1997. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 316: 629-34.

Feld, L.P. and Heckemeyer, J.H. 2011. FDI and taxation: A meta-study. *Journal of Economic Surveys*, forthcoming.

Friedman, R.A. 2003. What you do know can't hurt you. New York Times, August 12.

Gemmill, M.C., Costa-Font, J., and McGuire, A. 2007. In search of a corrected prescription drug elasticity estimate: A meta-regression approach. *Health Economics* 16: 627-43.

Gordon, R.D. 1941. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Annals of mathematical Statistics* 12: 364-66.

Havranek, T. 2010. Rose effect and the Euro: Is the magic gone?" *Review of World Economics* 146: 241-261.

Heckman, J.J. 1979. Sample selection bias as a specification error," *Econometrica*, 47:153-61.

Hedges, L. V. and Olkin, I. 1985. *Statistical Methods for Meta-Analysis*, Orlando: Academic Press.

Higgins J.P.T. and Thompson S.G. 2002. Quantifying heterogeneity in meta-analysis. *Statistics in Medicine* 21: 1539-1558.

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., and Dickersin, K. 2009. Publication bias in clinical trials due to statistical significance or direction of trial result," *Cochrane Review*, Issue 1. Available at http://www.thecochranelibrary.com.

Hunter, J.E. and Schmidt F.L. 2004. *Methods of meta-analysis: Correcting error and bias in research findings*. 2nd ed. Newbury Park: Sage Publications.

Johnson, N., and S. Kotz 1970. *Distributions in Statistics: Continuous Univariate Distribution*. New York: Wiley.

Kuhn, T.S. 1962. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Krakovsky, M. 2004. Register or perish. *Scientific American* 291:18-20.

Krassoi-Peach, E. and Stanley, T. D. 2009. Efficiency wages, productivity and simultaneity: A meta-regression analysis. *Journal of Labor Research* 30: 262-8.

Laplace, P.S. 1812. *Théorie Analytique des Probabilitiés.* Paris: Gauthier Villars.

Lehrer, J. 2010. The truth wears off: Is there something wrong with the scientific method? *The New Yorker*, December 13.

Mandel, M., and Rinott, Y. 2009. A selection bias conflict and frequentist versus Bayesian viewpoints, *American Statistician* 63: 211–217.

Moore, D. S. 1997.Bayes for beginners? Some reasons to hesitate. *The American Statistician*, 51: 254–261.

Moreno, S.G., Sutton, A.J., Ades, A., Stanley, T.D, Abrams, K.R., Peters, J.L., and Cooper, N.J. 2009. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, 9:2. http://www.biomedcentral.com/1471-2288/9/2.

Rosenthal, R. 1979. The 'file drawer problem' and tolerance for null results. *Psychological Bulletin* 86:638-41.

Senn, S. 2008. A note concerning a selection 'paradox' of Dawid's," *American Statistician*, 62: 206–210.

Skloot, R. 2006. Publication probity. *New York Times*, December 10.

Stanley, T.D. 2001. Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives* 15:131-150.

Stanley, T.D. 2005. Beyond publication bias. *Journal of Economic Surveys* 19:309-45.

Stanley, T.D. 2008. Meta-regression methods for detecting and estimating empirical effect in the presence of publication selection. *Oxford Bulletin of Economics and Statistics* 70:103-27.

Stanley, T.D. and Jarrell, S.B. 1989. Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys* 3:54-67.

Stanley, T.D. and Doucouliagos, C. 2007. Identifying and correcting publication selection bias in the efficiency-wage literature: Heckman meta-regression. School Working Paper, Economics Series 2007-11, Deakin University.

Stanley, T.D., and Doucouliagos, H(C) 2010. Picture this: A simple graph that reveals much ado about research. *Journal of Economic Surveys* 24: 170-191.

Stanley, T.D., Jarrell, S. B.and Doucouliagos, H(C). 2010. Could it be better to discard 90% of the data? A statistical paradox. *American Statistician* 64:70-77.

Stead, L. F., Perera, R., Bullen, C., Mant, D., and Lancaster, T. 2008) Nicotine replacement therapy for smoking cessation, *The Cochrane Library*, Issue 2. Available at *http://www.thecochranelibrary.com*.

Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance. *Journal of the American Statistical Association*, 54: 30–34.

Sterling, T. D., Rosenbaum, W. L., and Weinkam, J. J. 1995. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician*, 439: 108–112.

Turner, E.H., Matthews, A.M., Linardatos, E., Tell, R.A., and Rosenthal, R. 2008. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine* 358:252-60.

Viscusi, W.K. 1993. The value of risks to life and health. *Journal of Economic Literature* 31: 1912-46.

Wooldridge,J.M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.

APPENDIX

APPENDIX TABLE 1: MEANS OF THE INTERCEPT OF POLYNOMIAL APPROXIMATIONS (n=20)

Hetero-geneity*	True effect	Selection Incidence	*Linear* $\hat{\gamma}_1$	*Quadratic*	*Cubic*	*PEESE*, $\hat{\beta}_1$ from (9)
	0	0%	0.00	0.01	-0.01	0.00
	0	25%	0.04	-0.09	0.01	0.13
	0	50%	0.06	-0.10	0.15	0.25
	0	75%	0.07	-0.08	0.19	0.36
I^2=25%	0	100%	0.07	0.07	0.15	0.46
	1	0%	1.00	1.01	1.01	1.00
	1	25%	0.92	0.97	1.14	0.99
	1	50%	0.86	0.91	1.14	0.98
	1	75%	0.77	0.89	1.18	0.96
	1	100%	0.69	0.88	1.10	0.94
	0	0%	0.00	0.02	0.01	0.00
	0	25%	0.05	-0.06	0.02	0.15
	0	50%	0.09	-0.06	-0.13	0.30
	0	75%	0.14	0.04	-0.03	0.44
I^2=58%	0	100%	0.20	0.19	0.09	0.59
	1	0%	1.00	1.00	1.08	1.00
	1	25%	0.94	0.93	1.03	1.01
	1	50%	0.88	0.87	0.98	1.02
	1	75%	0.82	0.87	0.95	1.03
	1	100%	0.74	0.86	0.79	1.03
	0	0%	0.00	0.02	0.04	0.00
	0	25%	0.02	0.03	-0.10	0.17
	0	50%	0.10	0.08	-0.11	0.38
	0	75%	0.23	0.20	-0.19	0.61
I^2=85%	0	100%	0.39	0.36	0.07	0.86
	1	0%	1.00	0.98	0.96	1.00
	1	25%	0.98	0.99	0.90	1.07
	1	50%	0.93	0.92	0.82	1.12
	1	75%	0.91	0.88	0.63	1.19
	1	100%	0.82	0.84	0.42	1.22

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. *Linear*, *Quadratic*, *Cubic*, and *PEESE* refer to different estimates of the intercept of the polynomial approximation to the conditional mean of a truncated distribution—equation (5). $\hat{\gamma}_1$ is estimated from equation 7, and $\hat{\beta}_1$ is estimated from equation 8.

APPENDIX TABLE 2: MEAN SQUARE ERRORS OF POLYNOMIAL APPROXIMATIONS
(times 1,000 with n=80)

Hetero-geneity*	True effect	Selection Incidence	*Linear* $\hat{\gamma}_1$	*Quadratic*	*Cubic*	*PEESE*, $\hat{\beta}_1$ from (9)
	0	0%	105	910	12805	31
	0	25%	93	960	15507	44
	0	50%	76	837	15772	84
	0	75%	55	629	13268	144
I^2=25%	0	100%	25	196	2889	217
	1	0%	105	960	12749	32
	1	25%	108	909	12117	31
	1	50%	114	861	11916	29
	1	75%	138	778	10694	28
	1	100%	177	677	9236	28
	0	0%	208	1705	21298	64
	0	25%	182	1719	25583	78
	0	50%	154	1506	24863	133
	0	75%	126	1082	18969	231
I^2=58%	0	100%	89	452	5468	362
	1	0%	208	1690	21122	64
	1	25%	196	1553	19738	60
	1	50%	186	1432	19300	55
	1	75%	182	1242	16471	48
	1	100%	193	1015	13080	41
	0	0%	482	3350	33902	150
	0	25%	447	3429	37066	167
	0	50%	377	3074	34960	256
	0	75%	316	2190	24800	454
I^2=85%	0	100%	281	1033	8735	778
	1	0%	494	3416	34339	153
	1	25%	418	3014	32328	134
	1	50%	373	2553	28468	131
	1	75%	320	2196	22638	133
	1	100%	270	1645	17405	124

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. *Linear*, *Quadratic*, *Cubic*, and *PEESE* refer to different estimates of the intercept of the polynomial approximation to the conditional mean of a truncated distribution—equation (5). $\hat{\gamma}_1$ is estimated from equation 7, and $\hat{\beta}_1$ is estimated from equation 8.

APPENDIX TABLE 3: MEANS OF ALTERNATIVE RESEARCH SUMMARY ESTIMATORS (n=20)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	*PEESE*, $\hat{\beta}_1$ from (9)	PET-PEESE
	0	0%	0.00	0.00	0.00	0.00	0.00	-0.02
	0	25%	0.23	0.20	0.21	0.14	0.13	0.03
	0	50%	0.47	0.39	0.43	0.28	0.25	0.06
	0	75%	0.70	0.59	0.63	0.41	0.36	0.07
I^2=25%	0	100%	0.93	0.78	0.78	0.55	0.46	0.10
	1	0%	1.00	1.00	1.00	1.00	1.00	0.98
	1	25%	1.07	1.04	1.05	1.01	0.99	0.95
	1	50%	1.13	1.08	1.09	1.01	0.98	0.93
	1	75%	1.20	1.11	1.13	1.02	0.96	0.90
	1	100%	1.27	1.15	1.17	1.03	0.94	0.86
	0	0%	0.00	0.00	0.00	0.00	0.00	-0.03
	0	25%	0.27	0.23	0.26	0.15	0.15	0.03
	0	50%	0.54	0.45	0.51	0.31	0.30	0.09
	0	75%	0.81	0.68	0.75	0.48	0.44	0.15
I^2=58%	0	100%	1.08	0.91	0.92	0.67	0.59	0.27
	1	0%	1.00	1.00	1.00	1.00	1.00	0.93
	1	25%	1.09	1.07	1.08	1.03	1.01	0.93
	1	50%	1.19	1.13	1.16	1.05	1.02	0.92
	1	75%	1.29	1.19	1.23	1.07	1.03	0.91
	1	100%	1.39	1.26	1.30	1.09	1.03	0.90
	0	0%	0.00	0.00	0.00	0.00	0.00	-0.04
	0	25%	0.36	0.28	0.34	0.19	0.17	0.00
	0	50%	0.72	0.58	0.68	0.39	0.38	0.10
	0	75%	1.09	0.89	1.02	0.64	0.61	0.23
I^2=85%	0	100%	1.44	1.20	1.29	0.90	0.86	0.47
	1	0%	1.00	1.00	1.00	1.00	1.00	0.88
	1	25%	1.18	1.13	1.17	1.07	1.07	0.91
	1	50%	1.36	1.26	1.33	1.14	1.12	0.92
	1	75%	1.54	1.40	1.49	1.19	1.19	0.95
	1	100%	1.72	1.52	1.63	1.23	1.22	0.96

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. *Top10* is the simple average of the most precise 10% of the observations. $\hat{\beta}_1$ is estimated from equation 8.

APPENDIX TABLE 4: MEAN SQUARE ERRORS OF ALTERNATIVE ESTIMATORS
(times 1,000 with n=20)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	*PEESE*, $\hat{\beta}_1$ from (9)	PET-PEESE
	0	0%	13	11	11	56	31	89
	0	25%	65	47	55	83	44	86
	0	50%	228	161	191	135	84	74
	0	75%	497	347	397	214	144	57
I^2=25%	0	100%	878	606	608	323	217	53
	1	0%	13	11	11	55	32	53
	1	25%	16	12	12	54	31	63
	1	50%	27	15	17	51	29	69
	1	75%	47	22	25	50	28	85
	1	100%	77	31	35	49	28	103
	0	0%	24	22	20	112	64	173
	0	25%	92	69	81	143	78	161
	0	50%	307	220	270	202	133	144
	0	75%	667	474	566	307	231	128
I^2=58%	0	100%	1171	834	858	476	362	167
	1	0%	23	23	21	112	64	132
	1	25%	30	25	26	103	60	139
	1	50%	54	35	41	97	55	147
	1	75%	97	53	68	92	48	150
	1	100%	161	81	103	87	41	155
	0	0%	62	58	56	270	150	400
	0	25%	182	130	162	336	167	393
	0	50%	564	376	500	428	256	343
	0	75%	1211	819	1060	614	454	307
I^2=85%	0	100%	2113	1461	1692	906	778	425
	1	0%	63	58	55	284	153	322
	1	25%	87	68	76	257	134	304
	1	50%	178	113	150	237	131	306
	1	75%	330	195	273	218	133	297
	1	100%	554	298	427	208	124	291

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. *Top10* is the simple average of the most precise 10% of the observations. $\hat{\beta}_1$ is estimated from equation 8.