Social Psychological and
Personality Science
1-11

DOI: 10.1177/1948550617693062
spps.sagepub.com

Limitations of PET-PEESE and Other Meta-Analysis Methods

T. D. Stanley[1,2]

Abstract

A novel meta-regression method, precision-effect test and precision-effect estimate with standard errors (PET-PEESE), predicts and explains recent high-profile failures to replicate in psychology. The central purpose of this article is to identify the limitations of PET-PEESE for application to social/personality psychology. Using typical conditions found in social/personality research, our simulations identify three areas of concern. PET-PEESE performs poorly in research areas where there are only a few studies, all studies use small samples, and where there is very high heterogeneity of results from study to study. Nonetheless, the statistical properties of conventional meta-analysis approaches are much worse than PET-PEESE under these same conditions. Our simulations suggest alterations to conventional research practice and ways to moderate PET-PEESE weaknesses.

Keywords

meta-analysis, publication bias, PET-PEESE, random effects, weighted least squares

Recently, there have been high-profile failures to replicate psychological phenomenon (e.g., Hagger & Chatzisarantis, 2016; Open Science Collaboration, 2015). Yet, reproducibility by independent researchers has long been regarded as the "hallmark of science" (Popper, 1959). In at least one case, novel meta-regression methods, precision-effect test and precision-effect estimate with standard errors (PET-PEESE), anticipated the failures to replicate psychological phenomenon, the ego-depletion effect (Carter, Kofler, Forster, & McCullough, 2015; Stanley & Doucouloagos, 2014). These new meta-analysis methods for accommodating "publication bias" can do much to address one source of the current credibility and replication "crises" across the social sciences. Our simulations demonstrate that PET-PEESE typically reduces publication bias to practical insignificance and provides a valid Neyman–Pearson procedure to identify a genuine effect from selective reporting biases. However, PET-PEESE too has shortcomings. The central purpose of this article is to identify those limitations when applied to typical areas of social/personality psychology. In the process, we also show that conventional meta-analytic methods are of little use in identifying an authentic effect when there is selective reporting of statistically significant results.

Selective Reporting and Publication Bias

For decades, researchers have been acutely aware that the selective reporting of statistically significant results (also known as [aka] the file-drawer problem, publication bias, small sample bias, and p-hacking) poses a major threat to the scientific validity of psychology and other social sciences (Begg & Berlin, 1988; Glass, McGaw, & Smith, 1981; Hedges & Oklin, 1985; Rosenthal, 1979; Schmidt & Hunter, 2014; Sterling, 1959, to cite a few). When even a portion of reported findings have been selected to be statistically significant and "positive," average effect sizes can be greatly exaggerated or made to appear to be important when there is no genuine effect (Stanley & Doucouliagos, 2014).

Some researchers, referees, or editors may suppress insignificant findings, leaving them in the proverbial "file drawer" (Rosenthal, 1979). Others might "p-hack" their statistical analysis by employing questionable statistical practices such as data-peaking, choosing which of multiple dependent measures to report, and selectively omitting "outliers" (Simonsohn, Nelson, & Simmons, 2014). Regardless, the effect on the research record will be much the same; reported effects will be larger than the underlying "true" effect size. The simulations reported below are constructed in a way that makes the exact mechanism of selective reporting bias immaterial, encompassing research practices called: the "file-drawer problem," publication bias, and "p-hacking."

[1] Hendrix College, Conway, AR, USA
[2] School of Business and Law, Deakin University, Burwood, Victoria, Australia

Corresponding Author:
T. D. Stanley, Hendrix College, 1600 Washington St., Conway, AR 72032, USA.
Email: stanley@hendrix.edu

Meta-Analysis

Meta-regression models of selective reporting and publication bias. When only statistically significant, positive results are reported, selective reporting bias is equal to a complex function of the reported estimate's standard error (Copas & Shi, 2000; Greene, 1990; Johnson & Kotz, 1970; Stanley & Doucouliagos, 2014). Medical researchers sometimes use a linear approximation to this bias as the basis for a test of selective reporting bias—H_0: $\beta_1 = 0$ in (Egger, Smith, Schneider, & Minder, 1997; Stanley, 2008; Stanley & Doucouliagos, 2014):

$$\hat{d}_i = \beta_0 + \beta_1 SE_i + u_i \qquad i = 1, 2, \ldots, m, \tag{1}$$

where $\hat{d}_i$ is the estimated effect size, SE_i is its standard error, and m is the number of estimates in the research record. Equation 1 is estimated by weighted least squares (WLS), using $1/SE_i^2$ as the weights.

The conventional Neyman–Pearson t-test of β_0 (H_0: $\beta_0 = 0$) provides a statistical test for a genuine empirical effect beyond the reach of selective reporting bias called the "precision-effect test" or PET (Stanley, 2008). As SE_i approaches 0, studies become objectively better and better, and meta-regression (equation 1) implies that estimated effect sizes approach β_0, on average. Simulations of estimated regression coefficients demonstrate that PET is often a powerful test for the presence of an authentic effect beyond selective reporting bias (Stanley, 2008). However, $\hat{\beta}_0$ from equation 1 tends to underestimate the true mean effect when there is a nonzero treatment effect. In these cases, Stanley and Doucouliagos (2014) find that replacing the effect size's standard error, SE_i, in equation 1 by its variance, SE_i^2, reduces the bias of the estimated meta-regression intercept.

$$\hat{d}_i = \gamma_0 + \gamma_1 SE_i^2 + \upsilon_i \qquad i = 1, 2, \ldots, m, \tag{2}$$

with $1/SE_i^2$ as the WLS weight. $\hat{\gamma}_0$ is the PEESE.

Stanley and Doucouliagos (2014) claim only that this PET-PEESE conditional estimator reduces selective reporting bias. When there is evidence of a genuine treatment effect, PEESE ($\hat{\gamma}_0$) from equation 2 is used; otherwise, the corrected effect is best estimated by $\hat{\beta}_0$ from equation 1. For the purpose of deciding which meta-regression accommodation for selective reporting bias to employ, we recommend testing H_0: $\beta_0 \leq 0$ at the 10% significance level.

These simple meta-regression models of selective reporting need to be embedded within more complex, multiple meta-regression that also account for observed systematic heterogeneity. Social/personality psychological effects often vary greatly by socioeconomic status, age, gender, culture, or the passage of time. Or, alternative instruments used to measure social/personality effects may have low reliability and/or biases, further causing systematic heterogeneity among the reported findings. Dealing with heterogeneity (systematic and random) is the raison d'être of PET-PEESE meta-regression (Stanley & Doucouliagos, 2012, 2014). Simulations demonstrate that these meta-regression models can simultaneously and successfully accommodate systematic and random heterogeneity, selective reporting, and sampling error (Stanley & Doucouliagos, 2016).

In contrast, several papers demonstrate that p-curve approach is not valid (biased and unreliable) when there is heterogeneity, misspecification biases, or when some insignificant studies are reported (Bruns & Ioannidis, 2016; McShane, Böckenholt, & Hansen, 2016; van Aert, Wicherts, & van Assen, 2016), conditions ubiquitous in the social sciences. Nor do we simulate the popular "Trim and Fill." "Comprehensive simulations" demonstrate that both PET and PEESE "consistently outperformed the Trim & Fill estimators... With respect to the popular Trim & Fill method, we find it hard to recommend over the regression-based alternatives due to its potentially misleading adjustments and poor coverage probabilities" (Moreno, Sutton, Ades, et al., 2009, p. 1, 12). Prior advocates of Trim and Fill have subsequently embraced PEESE (Moreno et al., 2012; Moreno, Sutton, Turner, et al., 2009).

Conventional meta-analysis. The role of conventional meta-analysis estimators, "fixed effect" (*FE*) and "random effects" (RE), is to integrate and summarize all comparable estimates found in the research record. They assume that the individual reported effect sizes, $\hat{d}_i$, are randomly and normally distributed around some common overall mean effect, μ. Each estimates μ using a weighted average,

$$\hat{\mu} = \frac{\sum \omega_i y_i}{\sum \omega_i}, \tag{3}$$

but they employ different weights and thereby have different variances. *FE* uses weights $w_i = 1/SE_i^2$ and has variance $1/\Sigma w_i$. RE has weights $w_i' = 1/(SE_i^2 + \hat{\tau}^2)$ with variance $1/\Sigma w_i'$; where $\hat{\tau}^2$ is the estimated heterogeneity variance.

An alternative weighted average—WLS. The *unrestricted* weighted least squares weighted average, WLS, makes use of the multiplicative invariance property implicit in all WLS approaches (Stanley & Doucouliagos, 2015). It is calculated by running a simple meta-regression, with no intercept, of t-statistics versus precision:

$$t_i = \frac{\hat{d}_i}{SE_i} = \alpha\left(\frac{1}{SE_i}\right) + u_i \qquad i = 1, 2, \ldots, m \tag{4}$$

Ordinary least squares using any standard statistical software will calculate this WLS weighted average, $\hat{\alpha}$, its standard error, and confidence interval. WLS's point estimate is identical to *FE*, but its standard errors and confidence intervals are always superior when there is excess heterogeneity (Stanley & Doucouliagos, 2015). Thus, *FE* is not reported in the below simulations.

Comprehensive simulations show that WLS is as good as and often better than RE *when the RE model is true.* When there is no selective reporting bias, WLS's properties are practically equivalent to RE; if there is selective reporting, WLS has

consistently smaller bias and mean squared error [MSE] than RE (Stanley & Doucouliagos, 2015).

Simulations

We simulate randomized controlled experiments over a wide variety of conditions typically found in social/personality psychological research. Past simulations of PET and PET-PEESE concerned estimated regression coefficients from observational studies (Stanley, 2008; Stanley & Doucouliagos, 2014). Thus, the properties of these meta-regression methods may differ when applied to standardized mean differences from social/personality experiments. In particular, the well-known dependence of the standard error of Cohen's d upon the value of Cohen's d may cause special difficulties for the PET meta-regression model (equation 1).

Design. The average reported Cohen's d in social psychology is approximately .4 (Richard & Bond, 2003). We round this up to .5 in our simulations to allow for potential "medium"-size effect as defined by Cohen's guidelines. Because there is evidence of selective reporting in at least some areas of social/personality psychology, true effects are likely to be smaller. After replicating 100 psychological experiments, the Open Science Collaboration (2015) found that average effects were only one-half the size reported in the original studies. Such a 100% "research inflation" has also been found in a survey of over 6,700 studies in economics (Ioannidis, Stanley, & Doucouliagos, in press). Given this 100% exaggeration and Richard and Bond's (2003) survey, $d = .2$ may be more representative of social/personality psychology. We also investigate $d = 0$ to bracket typical effect sizes.

Our simulation experiments allow different numbers of studies in different areas of research, $m = \{10, 20, 40, 80\}$. All these results are reported in the Online Appendix, but only those for $m = \{20, 80\}$ are reported here.

To be more specific, these simulations first involved the generation of individual subject outcomes as:

$$y_{cj} = x_{cj} + u_{cj} \quad j = 1, 2, \ldots, n, \tag{5}$$

for individuals in the control group; where $u_{cj} \sim N(0, 50^2)$ and $x_{cj} \sim N(300, 86.6^2)$. Outcomes in the experimental group are generated in the exact same, yet independent, manner, with the single exception that they add the treatment effect, $T_e = \mu + \theta_i$ and $\theta_i \sim N(0, \sigma_h^2)$, to equation 5.

Our simulations fix the mean of true effects, μ, as either 0, 20, or 50. The values of the other parameters make the mean true value of Cohen's d equal to either .0, .2, or .5. Fraley and Vazire (2014) find that the median combined sample size is 100 in social/personality psychology's top journals. We also follow Fraley and Vazire's (2014) posted distribution of sample sizes across these top journals, giving $n = \{15, 35, 50, 100, \text{or } 200\}$ as our distribution of sample sizes per group across studies. To be comprehensive, we also generate other distributions of sample sizes representing worse-case scenarios (very small samples with a compact distribution of sample sizes) and better-case scenarios (larger samples with more dispersed sample sizes).

Past simulation studies found that the magnitude of excess heterogeneity is the most important research dimension that drives selective reporting bias and the statistical properties of alternative meta-analysis methods (Stanley, 2008; Stanley & Doucouliagos, 2014, 2015, 2016; Stanley, Doucouliagos, & Ioannidis, in press; Stanley, Jarrell, & Doucouliagos, 2010). Following these other studies, we investigate a wide range of heterogeneity by varying the standard deviation of random between-study heterogeneity, θ_i, from 0 to 50, $\sigma_h = \{0, 6.25, 12.5, 25, 50\}$. It is important to recognize that such heterogeneity means that there is no single true effect size. Instead, there is a distribution of true effects that are normally distributed around their mean, $\mu_d = \{.0, .2, .5\}$. This heterogeneity causes the relative measure of observed heterogeneity, I^2, to vary from near 0 to over 95%. I^2 is easy to calculate. $I^2 = \{(\text{MSE} - 1)/\text{MSE}\}$ from the simple WLS meta-regression (equation 4; Higgins & Thompson, 2002, pp. 1546–1547). I^2 values are computed empirically for each simulated meta-analysis and reported in the tables.

Cohen's d and its standard error are calculated for each simulated study. This is repeated $m = \{10, 20, 40, 80\}$ times to represent one meta-analysis, and everything is again repeated 10,000 times to calculate various averages and statistics across 10,000 meta-analyses.

We simulate areas of research that do not have any selective reporting and others in which half of the reported results have undergone a process of selection to be statistically significant and positive. For the remaining 50%, each randomly generated result is reported, statistically significant or not. This choice of 50% selective reporting is chosen to reflect what is generally seen in the psychological research record. When the true mean effect is $\mu_d = .2$ and there is 50% selective reporting, the average reported effect will be .4046, quite close to the average effect found in social psychology by Richard and Bond (2003).

Results

Figure 1 and Table 1 report the biases of RE, the unrestricted WLS and the conditional meta-regression estimator PET-PEESE when there are either 20 or 80 studies, $m = \{20, 80\}$, and 50% of the studies are selected to be statistically significant. Bias is calculated empirically by the average of an estimator across 10,000 replications minus the known true effect. Results for all $m = \{10, 20, 40, 80\}$ are placed in the Online Appendix (https://www.hendrix.edu/maer-network/default.aspx?id=15206). Figure 2 and the last three columns of these tables report the observed frequency in which RE, WLS, and PET reject the null hypothesis of no effect (H_0: $\mu_d = 0$; $\alpha = .05$). When the mean true effect is 0 (i.e., $\mu_d = 0$), these proportions represent the observed frequency of a type I error (aka "level"). About one third of the way down Tables 1–4, the average type I error rates (or levels) are displayed in the last three columns. When the true effect is not 0 (i.e., $\mu_d = .2$ or $\mu_d = .5$), these proportions represent the power of these alternative estimators to identify a nonzero

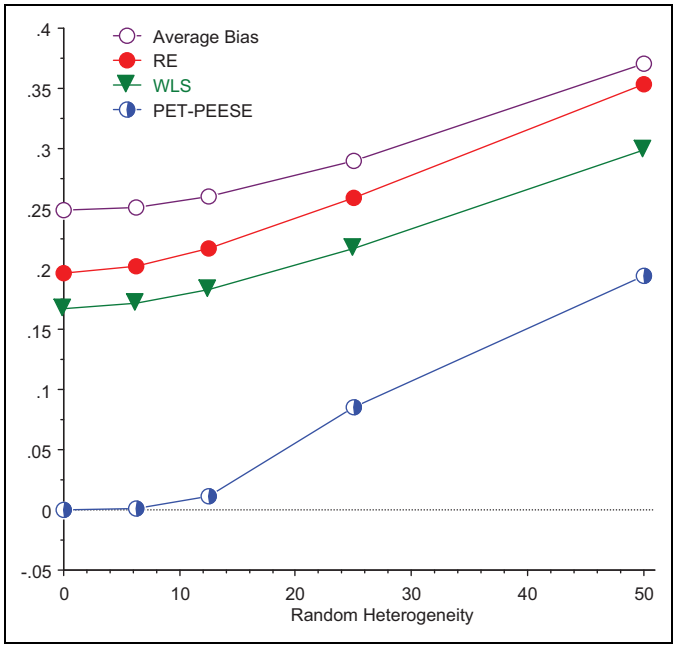

Figure 1. Bias of alternative estimates with true mean effect = 0, 50% selection, and 80 studies. Bias is measured on the vertical axis; random heterogeneity, σ_h, is plotted on the horizontal.

overall effect. At the bottom of Tables 1–4, the average powers are displayed along with the average biases and I^2.

The simulations revealed in Table 1 assume that the distribution of sample sizes is $n = \{15, 35, 50, 100, \text{or } 200\}$ per group following Fraley and Vazire (2014) and that 50% of the reported results are selected to be statistically significant and positive. With 50% selective reporting biases can be substantial. Over all three true mean effect sizes, the average selective reporting bias is .2016. However, this bias is larger (.2840) when there is no true effect, $\mu_d = 0$. Although RE reduces this bias somewhat (to .1577, overall), RE can give the appearance of a small effect (.2448) when there is none ($\mu_d = 0$). Worse still, RE makes a type I error nearly 98% of the time (.9792). Thus, conventional RE meta-analysis cannot provide a basis for valid statistical inference when there is selective reporting bias. The unrestricted WLS weighted average dominates RE *in all cases* (smaller biases and lower type I error rates; see Table 1). However, it too tends to have large type I error inflation (92%, on average).

Only the PET has acceptable type I error rates (4% on average), less than the nominal 5% level. Likewise, the related PET-PEESE conditional meta-regression estimator successfully reduces average *absolute* bias to practical insignificance (.0397). Its average bias is .024 because a few small negative biases cancel out some of the larger positive ones. Nonetheless, PET and PET-PEESE too have limitations (see section "Discussion and Comments").

Both RE and WLS have quite high power to reject H_0: $\mu_d = 0$ when there is either a small ($\mu_d = .2$) or a medium-size effect ($\mu_d = .5$). However, this is neither surprising nor meaningful, because both have very high rates of falsely rejecting H_0:$\mu_d = 0$ when there is no genuine effect (i.e., $\mu_d = 0$). Only PET has acceptable levels, so only its statistical power is relevant. For a small effect, $\mu_d = .2$, PET's power is only 33% when there are 20 estimates of effect size but rises to 83% when we have 80 estimates. However, when there is a medium-size effect, $\mu_d = .5$, PET's power is almost always greater than 85% or even 90%. The only exceptions to this positive evaluation of PET and PET-PEESE for these typical social/personality psychology conditions (Table 1; Figures 1 and 2) occur when there is very high heterogeneity (for the meaning of these limitations and how they might be mitigated, see section "Discussion and Comments").

Table 5 displays MSE and coverage rates for 95% confidence intervals under the same conditions as those reported in Table 1 (also see Figure 3). Here too, PET-PEESE dominates. Its MSE is less than half RE's, and only PET-PEESE's confidence intervals are at or even close to the 95% nominal level. However, PET-PEESE again has limitations. For example, its slight bias cause coverage rates to be too small (83%, on average).

Simulations reported in Table 2 calculate the same statistics for the exact same design parameters as those that generate Table 1 results, except that none of the simulated study results have been selected for statistical significance. When there is no selective reporting bias, all three meta-analysis approaches have practically insignificant bias, small type I errors and large powers. All three have average rates of type I errors 1–3% higher than the nominal 5% level, with RE closest to 5%. All three generally have high power to detect a genuine nonzero effect, but, as before, their powers decrease at the highest levels of heterogeneity. PET's power is the lowest of the three, when there is no selective reporting, and PET's power can be rather low for small meta-regression samples, high heterogeneity and small effects (see Table 2). PET-PEESE has a small negative bias at the highest level of heterogeneity. Although PET-PEESE's underestimate is worthy of note, it is not large enough to be practically relevant. In all cases, RE's has superior properties when there is no publication or selective reporting bias. Unfortunately, researchers can never rule out the potential presence of selective reporting bias in practice, because all tests for publication bias have low power (Egger et al., 1997; Stanley, 2008).

To explore other potential weaknesses, we also simulate cases where the studies in the primary research literature use different distributions of sample sizes. The simulation results reported in Tables 3 and 4 are identical in every way to those reported in Table 1, except they rely on different distributions of sample sizes in the primary literature. The simulations displayed in Table 3 assume that the sample size, n, in each group is either 32, 64, 125, 250, or 500. Larger sample sizes with greater dispersion between studies are quite common in other areas of research, especially economics and medical research. Overall, the results are quite similar to those reported in Table 1. With these larger samples sizes, average selective reporting bias decreases along with the biases of both RE and WLS. Nonetheless, notable biases will still persist, on average, when

Table 1. Bias, Power, and Level of Alternative Meta-Methods With 50% Reporting Selection.

Design			Average		Bias			Power/Type I Error		
d	*m*	α_h	Bias	I^2	RE	WLS	PET-PEESE	RE	WLS	PET
0	20	0	.2482	.5140	.1958	.1668	.0008	0.9942	0.9503	0.0002
0	20	6.25	.2517	.5409	.2015	.1714	.0086	0.9931	0.9290	0.0005
0	20	12.5	.2603	.6020	.2158	.1824	.0254	0.9825	0.8833	0.0052
0	20	25	.2902	.7367	.2581	.2177	.0714	0.9469	0.7913	0.0342
0	20	50	.3683	.8818	.3502	.2977	.1455	0.8654	0.6761	0.0914
0	80	0	.2487	.5162	.1964	.1673	.0019	1.0000	1.0000	0.0005
0	80	6.25	.2516	.5450	.2019	.1714	.0097	1.0000	1.0000	0.0009
0	80	12.5	.2604	.6131	.2166	.1828	.0307	1.0000	1.0000	0.0109
0	80	25	.2902	.7561	.2587	.2168	.0829	1.0000	1.0000	0.0853
0	80	50	.3703	.8958	.3530	.2994	.1739	1.0000	0.9990	0.1941
Average type I error rate (level)								0.9782	0.9229	0.0423
.2	20	0	.1659	.2254	.0986	.0868	−.0345	1.0000	1.0000	0.3056
.2	20	6.25	.1704	.2759	.1065	.0919	−.0297	1.0000	1.0000	0.3268
.2	20	12.5	.1809	.4034	.1266	.1060	−.0126	1.0000	1.0000	0.3423
.2	20	25	.2136	.6509	.1756	.1418	.0167	1.0000	0.9983	0.3378
.2	20	50	.2973	.8632	.2762	.2218	.0669	0.9973	0.9592	0.2857
.2	80	0	.1663	.2198	.0982	.0864	.0178	1.0000	1.0000	0.9244
.2	80	6.25	.1706	.2839	.1070	.0921	.0239	1.0000	1.0000	0.9136
.2	80	12.5	.1814	.4301	.1275	.1055	.0370	1.0000	1.0000	0.8927
.2	80	25	.2136	.6843	.1772	.1414	.0676	1.0000	1.0000	0.8078
.2	80	50	.2970	.8819	.2768	.2201	.1207	1.0000	1.0000	0.6334
.5	20	0	.0793	.0786	.0252	.0222	−.0239	1.0000	1.0000	0.9997
.5	20	6.25	.0834	.1282	.0300	.0249	−.0224	1.0000	1.0000	0.9978
.5	20	12.5	.0894	.2786	.0401	.0289	−.0209	1.0000	1.0000	0.9854
.5	20	25	.1193	.5905	.0819	.0552	−.0107	1.0000	1.0000	0.8578
.5	20	50	.2020	.8492	.1771	.1197	−.0151	1.0000	0.9992	0.5567
.5	80	0	.0804	.0395	.0241	.0227	−.0243	1.0000	1.0000	1.0000
.5	80	6.25	.0826	.1012	.0280	.0244	−.0231	1.0000	1.0000	1.0000
.5	80	12.5	.0906	.3028	.0416	.0300	−.0193	1.0000	1.0000	1.0000
.5	80	25	.1208	.6370	.0846	.0556	.0029	1.0000	1.0000	0.9998
.5	80	50	.2027	.8694	.1800	.1208	.0514	1.0000	1.0000	0.9568
Average			.2016	.5132	.1577	.1291	.0240	0.9999	0.9978	0.7562

Note. RE and WLS denote the random effects and unrestricted weighted least squares meta-analysis averages, respectively; PET-PEESE is the meta-regression publication bias corrected estimate; and average bias is the difference between the simple mean of the reported effects and the mean true effect size, *d*. *m* is the number of studies meta-analyzed and σ_h is standard deviation of the heterogeneity among true effects. PET-PEESE = precision-effect test and precision-effect estimate with standard errors.

there is no overall true effect ($\mu_d = 0$); average bias = .2094, .1849 for RE, WLS's bias is .1550, and PET-PEESE's absolute bias is only .0608. Here too, only PET produces type I error rates even close to their nominal 5% level. With access to these larger studies, PET's power improves. Table 3 shows that PET has high power to detect even small effects when there are sufficient estimates. Both PET and PET-PEESE dominate RE and WLS and have generally desirable properties. However, as before, both PET and PET-PEESE have difficulties at the highest levels of heterogeneity (see section "Discussion and Comments").

The simulations displayed in Table 4 assume yet another sample size distribution, $n = \{10, 18, 25, 33, \text{or } 40\}$ per group. We believe that these small sample sizes represent the worst-case scenario for all meta-analysis methods. Nonetheless, these sample sizes are found in at least one psychological meta-analysis on the transfer of working memory to fluid intelligence (Au et al., 2015; Boggs & Lasecki, 2015). As before, when there is selective reporting bias, there are large biases for conventional meta-analysis, and their type I error rates are unacceptably large, 93% and 87% for RE and WLS, respectively. Although PET's type I errors are very low, .001, its power to detect nonzero effects is now unacceptably low, 16% on average. Also, PET-PEESE consistently underestimates true average effect when it has access to only small sample studies. When all research studies use small samples and if some results are selected to be statistically significant, all meta-analysis methods have unacceptable statistical properties.

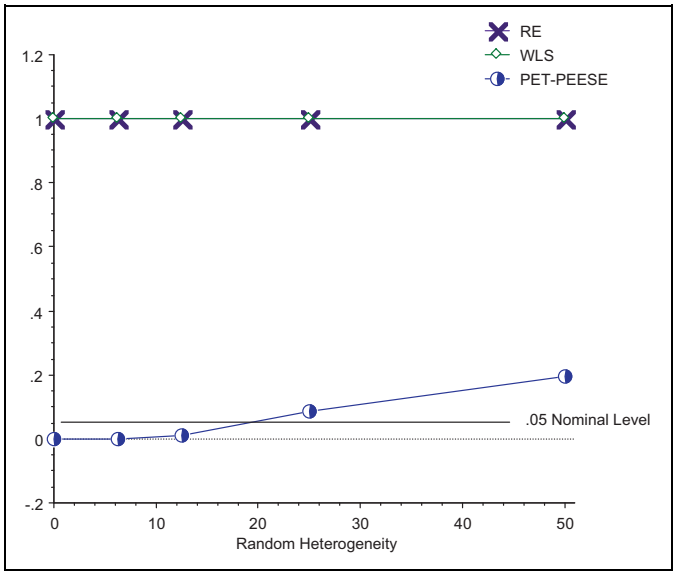

Figure 2. Type I error rates with true mean effect = 0, 50% selection, and 80 studies. Type I probabilities are plotted on the vertical axis; random heterogeneity, σ_h, is plotted on the horizontal.

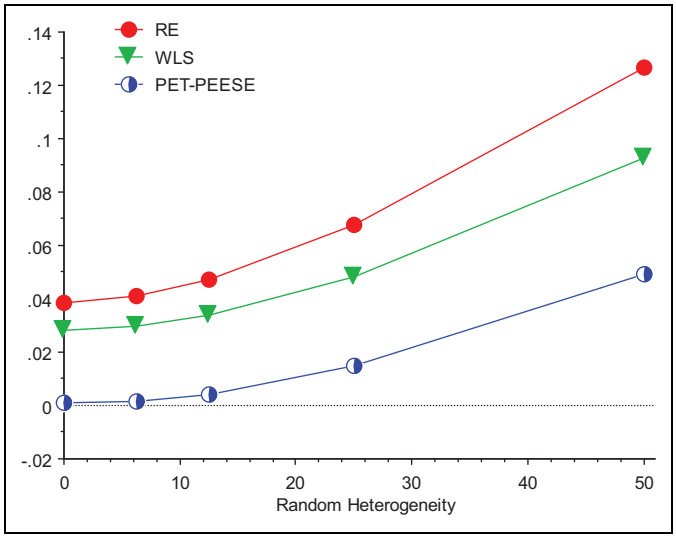

Figure 3. Mean squared errors (MSE) with true mean effect = 0, 50% selection, and 80 studies. MSEs are plotted on the vertical axis; random heterogeneity, σ_h, is plotted on the horizontal.

Discussion and Comments

The central purpose of this study is to identify limitations of recently developed meta-regression methods to accommodate and reduce publication bias—PET and PET-PEESE. These simulations succeed in uncovering several important limitations and weaknesses. First, PET sometimes has low power in identifying a genuine nonzero effect when there are only 10 or 20 estimates available in an area of research. This is especially true if the true effect is small (i.e., $\mu_d = .2$)—recall Table 1. This limitation is not especially surprising, because PET is based on a regression that tries to find evidence that power is increasing as research studies have access to larger samples (or smaller *SE*s). Nonetheless, researchers should be very cautious when applying PET to 10, 20, or fewer results. Under realistic assumptions, PET's power to detect a small effect may be less than 50% in small meta-samples.

Second, when there are very high levels of heterogeneity, the properties of both PET and PET-PEESE worsen. At the highest level of heterogeneity, $\sigma_h = 50$, PET's level becomes inflated, higher than the nominal 5% level, confidence intervals are too narrow, and MSEs are larger than idea. This type I error inflation actually worsens as the meta-analysis sample increases. Although serious problems, they are minor compared with very high type I error inflation rates, large MSEs and completely invalid confidence intervals that are typical of conventional meta-analysis: RE and WLS—recall Tables 1 and 5. When there are 20 or more estimates in an area of research, it is nearly certain that RE will find that an effect is present when, in fact, there is no overall effect. Conventional meta-analysis is entirely invalid as a test for the presence of social–psychological phenomena if there is selective reporting bias (or publication bias or p-hacking). Also, with the highest level of heterogeneity, $\sigma_h = 50$, PET-PEESE tends to exaggerate the size of the effect, by as much as .17, which explains PET's type I error inflation. Nonetheless, PET-PEESE is much better than RE in these same cases. RE's bias is at least twice as large as PET-PEESE's and often much larger.

Although extreme heterogeneity poses an important challenge for all meta-analysis methods, this is to be expected when one understands what such high heterogeneity implies about the underlying social/personality psychological phenomenon. With $\sigma_h = 50$, the typical variation of true effects from their mean true effect is $\pm .5$, in terms of Cohen's *d*. This implies that nearly 16% of the time, the *true* effect is actually negative when the mean true effect, μ_d, is positive and medium size ($\mu_d = .5$). Heterogeneity means that there is no single true effect, but rather true effects vary from study to study by the equivalent of $d = \pm .5$ for $\sigma_h = 50$. At this highest level of heterogeneity, *true* positive and negative small effects will in fact exist 69% of the time when the true mean effect is 0. From nothing, medium-sized effects (positive and negative) will occur 32% of the time. The point is that such high levels of heterogeneity obscure the very meaning of what the true social/personality phenomenon is or is not.

When the underlying true phenomenon is so highly variable and random, it would be unrealistic to expect any statistical method to be able to see reliably through this fog of truth without access to many highly reliable study results. Add selective reporting bias and sampling error to this mix of truth, and it would be remarkable if any statistical method could provide a reliable basis for inference.

So what can be done? Is reliable inference under realistic conditions impossible? Because tests of heterogeneity have low power, formal hypothesis testing of I^2 or its related variance is unlikely to be useful in practice. Thus, we recommend that a 80% cutoff be used as an application guideline for the reliable use of PET-PEESE. When applied to these simulation results, PET-PEESE would not be calculated for the great majority of the instances where heterogeneity is at its highest level, $\sigma_h = 50$. As a result, most of the worrisome cases we report

Table 2. Bias, Power, and Level of Alternative Meta-Methods With No Reporting Selection.

Design			Average		Bias			Power/Type I Error		
d	*m*	α_h	Bias	I^2	RE	WLS	PET-PEESE	RE	WLS	PET
0	20	0	.0004	.0824	.0000	−.0000	−.0082	0.0354	0.0459	0.0522
0	20	6.25	−.0002	.1388	.0000	.0000	−.0082	0.0554	0.0587	0.0534
0	20	12.5	−.0002	.3120	.0004	.0004	−.0097	0.0775	0.0816	0.0745
0	20	25	−.0009	.6514	−.0006	−.0004	−.0173	0.0787	0.1105	0.0965
0	20	50	−.0001	.8803	.0000	.0013	−.0255	0.0768	0.1152	0.0967
0	80	0	.0003	.0428	.0005	.0005	−.0031	0.0422	0.0527	0.0482
0	80	6.25	−.0000	.1188	−.0001	−.0001	−.0046	0.0533	0.0602	0.0587
0	80	12.5	−.0006	.3519	−.0004	−.0002	−.0051	0.0569	0.0823	0.0740
0	80	25	.0003	.6909	.0005	.0006	−.0079	0.0574	0.1056	0.0922
0	80	50	.0003	.8957	.0003	.0001	−.0155	0.0590	0.1149	0.0935
Average type I error rate (level)								0.0593	0.0828	0.0740
.2	20	0	.0019	.0811	−.0005	−.0005	−.0139	0.9997	0.9999	0.7048
.2	20	6.25	.0027	.1402	.0000	−.0002	−.0165	0.9990	0.9990	0.6403
.2	20	12.5	.0015	.3105	−.0010	−.0015	−.0249	0.9922	0.9906	0.5242
.2	20	25	.0036	.6516	.0006	−.0023	−.0427	0.8800	0.8460	0.3240
.2	20	50	.0027	.8790	−.0002	−.0103	−.0774	0.5337	0.4905	0.1799
.2	80	0	.0022	.0440	−.0003	−.0004	−.0026	1.0000	1.0000	0.9978
.2	80	6.25	.0023	.1200	−.0003	−.0004	−.0028	1.0000	1.0000	0.9931
.2	80	12.5	.0018	.3499	−.0006	−.0012	−.0051	1.0000	1.0000	0.9521
.2	80	25	.0023	.6914	−.0004	−.0031	−.0199	0.9999	0.9997	0.7069
.2	80	50	.0009	.8953	−.0022	−.0130	−.0620	0.9485	0.9020	0.3281
.5	20	0	.0059	.0819	−.0002	−.0003	−.0048	1.0000	1.0000	1.0000
.5	20	6.25	.0056	.1337	−.0002	−.0007	−.0052	1.0000	1.0000	0.9996
.5	20	12.5	.0049	.3074	−.0015	−.0030	−.0092	1.0000	1.0000	0.9923
.5	20	25	.0050	.6454	−.0014	−.0069	−.0293	1.0000	1.0000	0.8457
.5	20	50	.0048	.8772	−.0028	−.0251	−.1168	0.9925	0.9680	0.4332
.5	80	0	.0053	.0433	−.0006	−.0007	−.0056	1.0000	1.0000	1.0000
.5	80	6.25	.0049	.1161	−.0009	−.0012	−.0063	1.0000	1.0000	1.0000
.5	80	12.5	.0053	.3438	−.0010	−.0026	−.0089	1.0000	1.0000	1.0000
.5	80	25	.0054	.6858	−.0013	−.0079	−.0188	1.0000	1.0000	0.9999
.5	80	50	.0057	.8932	−.0019	−.0273	−.0685	1.0000	1.0000	0.8500
Average			.0025	.4152	−.0005	−.0036	−.0216	0.9673	0.9598	0.7736

Note. RE and WLS denote the random effects and unrestricted weighted least squares meta-analysis averages, respectively; PET-PEESE is the meta-regression publication bias corrected estimate; and average bias is the difference between the simple mean of the reported effects and the mean true effect size, *d*. *m* is the number of studies meta-analyzed and σ_h is standard deviation of the heterogeneity among true effects. PET-PEESE = precision-effect test and precision-effect estimate with standard errors.

in these tables and figures for PET-PEESE would be eliminated. Besides, the very meaning of social/personality psychological phenomenon is questionable if heterogeneity is higher than 80%. When $\mu_d = +.2$ and $\sigma_h = 50$, *true effect* will in fact be *negative* 34% of the time.

The third limitation of PET-PEESE and PET revealed by this study is that the viability of these meta-analysis methods depends on the distribution of sample sizes (or statistical power) found among the primary studies in the social/personality psychological research literature. For typical sample sizes found in social/personality psychology (Fraley & Vazire, 2014), these methods work rather well with the exceptions of small meta-analysis sample sizes and very high heterogeneity as discussed above. However, in those rare cases where an entire research literature contains very small studies, PET becomes virtually impotent, unlikely to identify a genuine effect when it exists. In this worst-case scenario, the average power is only 16%, but the type I error rate is practically 0 (.001; see Table 4). Even here, as a referee points out, should PET find evidence of an effect, it can, in fact, be trusted. If all the sample sizes in a research literature are small, PET's statistical properties would improve notably if α were increased to 20%. With this modest adjustment, PET's type I error rate remains less than 5% (.0359 on average), and average power more than doubles, climbing as high as 90%. PET can provide a valid test even under these most demanding circumstances. Nonetheless, great caution should be used in interpreting *any meta-analysis*, regardless of the methods used when all studies are highly underpowered.

It is important to put PET-PEESE's limitations in context. First, in all these cases where the use of PET-PEESE is compromised: Small meta-analysis samples, high heterogeneity

Table 3. Bias, Power, and Level: 50% Reporting Selection for Larger Sample Sizes.

Design			Average		Bias			Power/Type I Error		
d	*m*	α_h	Bias	I^2	RE	WLS	PET-PEESE	RE	WLS	PET
0	20	0	.1663	.5309	.1276	.1063	.0001	0.9957	0.9458	0.0001
0	20	6.25	.1710	.5838	.1364	.1132	.0114	0.9885	0.8990	0.0017
0	20	12.5	.1837	.6864	.1561	.1291	.0343	0.9711	0.8273	0.0201
0	20	25	.2191	.8405	.2034	.1693	.0789	0.9085	0.7181	0.0784
0	20	50	.3071	.9428	.2999	.2564	.1524	0.8065	0.6187	0.1168
0	80	0	.1665	.5314	.1281	.1066	.0006	1.0000	1.0000	0.0007
0	80	6.25	.1713	.5929	.1370	.1134	.0134	1.0000	1.0000	0.0032
0	80	12.5	.1837	.7047	.1568	.1288	.0406	1.0000	1.0000	0.0575
0	80	25	.2192	.8578	.2042	.1699	.0971	1.0000	0.9994	0.2025
0	80	50	.3065	.9512	.2999	.2566	.1788	0.9998	0.9934	0.2704
Average type I error rate (level)								0.9670	0.9002	0.0751
.2	20	0	.0882	.1615	.0393	.0336	−.0100	1.0000	1.0000	0.8610
.2	20	6.25	.0942	.2841	.0513	.0416	−.0051	1.0000	1.0000	0.7956
.2	20	12.5	.1073	.5200	.0748	.0573	.0028	1.0000	1.0000	0.6677
.2	20	25	.1466	.7971	.1285	.0985	.0268	1.0000	0.9980	0.4997
.2	20	50	.2354	.9355	.2269	.1825	.0792	0.9930	0.9426	0.3567
.2	80	0	.0883	.1446	.0390	.0341	−.0036	1.0000	1.0000	1.0000
.2	80	6.25	.0941	.3045	.0511	.0411	.0042	1.0000	1.0000	0.9998
.2	80	12.5	.1079	.5630	.0764	.0575	.0224	1.0000	1.0000	0.9952
.2	80	25	.1462	.8241	.1295	.0990	.0633	1.0000	1.0000	0.9438
.2	80	50	.2371	.9457	.2293	.1828	.1264	1.0000	1.0000	0.7590
.5	20	0	.0283	.0698	.0060	.0050	−.0110	1.0000	1.0000	1.0000
.5	20	6.25	.0308	.2041	.0085	.0055	−.0120	1.0000	1.0000	1.0000
.5	20	12.5	.0367	.4967	.0166	.0074	−.0127	1.0000	1.0000	0.9999
.5	20	25	.0652	.7933	.0499	.0263	−.0052	1.0000	1.0000	0.9430
.5	20	50	.1500	.9323	.1401	.0905	.0043	1.0000	0.9994	0.6506
.5	80	0	.0288	.0325	.0056	.0052	−.0112	1.0000	1.0000	1.0000
.5	80	6.25	.0305	.2095	.0079	.0052	−.0124	1.0000	1.0000	1.0000
.5	80	12.5	.0370	.5490	.0170	.0074	−.0131	1.0000	1.0000	1.0000
.5	80	25	.0650	.8200	.0508	.0261	−.0007	1.0000	1.0000	1.0000
.5	80	50	.1495	.9425	.1407	.0890	.0468	1.0000	1.0000	0.9842
Average			.1354	.5917	.1113	.0882	.0296	0.9997	0.9970	0.8728

Note. RE and WLS denote the random effects and unrestricted weighted least squares meta-analysis averages, respectively; PET-PEESE is the meta-regression publication bias corrected estimate; and average bias is the difference between the simple mean of the reported effects and the mean true effect size, *d*. *m* is the number of studies meta-analyzed and σ_h is standard deviation of the heterogeneity among true effects. PET-PEESE = precision-effect test and precision-effect estimate with standard errors.

and research literatures comprised of only small sample studies, the statistical properties of conventional meta-analysis (RE and *FE*) are much worse. Thus, the limitations identified by our simulations are not challenges for PET-PEESE, alone, but apply to all conventional meta-analysis methods.

Furthermore, the above meta-analysis's limitations are not of their making but are caused by serious inadequacies in the primary research base. If all studies in an area of research are greatly underpowered, this can only be seen as weakness of this area of research and nothing else. For over 30 years, psychologists have been acutely aware of the critical importance of statistical power (Cohen, 1988; Fraley & Vazire, 2014). Without adequate power,

> the published literature is likely to contain a mixture of apparent results buzzing with confusion . . . Not only do underpowered studies lead to a confusing literature but they also create a literature that contains biased estimates of effect sizes. (Maxwell, 2004, p. 161)

Conventional meta-analysis can *effectively* increase statistical power by combining many underpowered primary results only if they are also known to be unbiased. With selective reporting bias, some adequately powered studies are required to distinguish the genuine signal of a psychological effect reliably from bias and noise. Small meta-analysis samples are another limitation that stems from the primary research record. If an area of research is relatively new and/or underresearched, then there will insufficient research knowledge to be confident about anything. Lastly is the issue of very high levels of heterogeneity. The source of such complex and confused effects is not

Table 4. Bias, Power, and Level: 50% Reporting Selection and Smaller Sample Sizes.

Design			Average		Bias			Power/Type I Error		
d	*m*	α_h	Bias	I^2	RE	WLS	PET-PEESE	RE	WLS	PET
0	20	0	.3754	.4760	.3365	.3172	−.0739	0.9939	0.9756	.0000
0	20	6.25	.3778	.4837	.3397	.3200	−.0678	0.9936	0.9712	.0003
0	20	12.5	.3828	.5108	.3462	.3249	−.0642	0.9895	0.9592	.0002
0	20	25	.4052	.5956	.3728	.3470	−.0427	0.9737	0.9205	.0005
0	20	50	.4752	.7490	.4508	.4119	−.0126	0.9254	0.8235	.0025
0	80	0	.3757	.4806	.3372	.3178	−.0718	1.0000	1.0000	.0014
0	80	6.25	.3776	.4911	.3398	.3198	−.0688	1.0000	1.0000	.0019
0	80	12.5	.3837	.5202	.3473	.3259	−.0628	1.0000	1.0000	.0010
0	80	25	.4052	.6080	.3734	.3471	−.0427	1.0000	1.0000	.0015
0	80	50	.4737	.7645	.4500	.4102	−.0134	1.0000	1.0000	.0024
Average type I error rate (level)								0.9876	0.9650	.0012
.2	20	0	.2917	.2696	.2441	.2333	−.1547	1.0000	1.0000	.0059
.2	20	6.25	.2952	.2874	.2479	.2364	−.1531	1.0000	1.0000	.0076
.2	20	12.5	.3022	.3351	.2572	.2436	−.1464	1.0000	1.0000	.0100
.2	20	25	.3254	.4716	.2859	.2646	−.1405	1.0000	0.9998	.0176
.2	20	50	.3984	.7060	.3684	.3275	−.1441	0.9988	0.9928	.0208
.2	80	0	.2922	.2741	.2440	.2333	−.1539	1.0000	1.0000	.0141
.2	80	6.25	.2947	.2924	.2471	.2355	−.1535	1.0000	1.0000	.0197
.2	80	12.5	.3014	.3483	.2562	.2422	−.1473	1.0000	1.0000	.0204
.2	80	25	.3252	.4978	.2866	.2649	−.1310	1.0000	1.0000	.0273
.2	80	50	.3968	.7265	.3678	.3254	−.1481	1.0000	1.0000	.0192
.5	20	0	.1820	.0794	.1277	.1242	−.2202	1.0000	1.0000	.1729
.5	20	6.25	.1856	.0930	.1317	.1278	−.2175	1.0000	1.0000	.1739
.5	20	12.5	.1925	.1374	.1399	.1341	−.2206	1.0000	1.0000	.1716
.5	20	25	.2173	.3129	.1708	.1563	−.2281	1.0000	1.0000	.1277
.5	20	50	.2950	.6448	.2573	.2141	−.3123	1.0000	1.0000	.0636
.5	80	0	.1821	.0374	.1255	.1240	−.1107	1.0000	1.0000	.6159
.5	80	6.25	.1845	.0526	.1283	.1262	−.1121	1.0000	1.0000	.6013
.5	80	12.5	.1918	.1068	.1374	.1330	−.1169	1.0000	1.0000	.5560
.5	80	25	.2172	.3349	.1697	.1546	−.1524	1.0000	1.0000	.3977
.5	80	50	.2942	.6766	.2578	.2128	−.2889	1.0000	1.0000	.1201
Average			.3133	.4121	.2715	.2519	−.1324	0.9999	0.9996	.1582

Note. RE and WLS denote the random effects and unrestricted weighted least squares meta-analysis averages, respectively; PET-PEESE is the meta-regression publication bias corrected estimate; and average bias is the difference between the simple mean of the reported effects and the mean true effect size, *d*. *m* is the number of studies meta-analyzed and σ_h is standard deviation of the heterogeneity among true effects. PET-PEESE = precision-effect test and precision-effect estimate with standard errors.

meta-analysis, but rather some combination of the social/personality psychological phenomenon itself and/or the research methods used to study it. Before meta-analysis can reliably reduce ubiquitous selective reporting biases, the research record must contain some adequately powered studies with genuine information about the phenomenon in question.

Conclusion

> It is difficult to conceive of a correction methodology which would be universally credible. (Begg & Berlin, 1988, p. 440)

We investigate the statistical properties and limitations of the PET-PEESE approach to identifying and estimating a genuine effect in the presence of selective reporting bias. For decades, publication bias has been widely acknowledged to be a serious, yet intractable, problem for empirical science (Begg & Berlin, 1988; Glass et al., 1981; Hedges & Oklin, 1985; Rosenthal, 1979; Schmidt & Hunter, 2014; Sterling, 1959). It would be naïve to suggest that an ideal solution has been found. Yet, if no action is taken, small effects can be doubled and small to medium effect sizes can be manufactured from nothing in typical social/personality areas of research. We do not claim that PET-PEESE is a perfect estimator of the underlying true effect size, corrected for selective reporting. Aside from the limitations singled out above, its MSE can be too large to give reliable estimates of the exact effect sizes in question. We claim only that PET-PEESE is better than the alternatives: conventional meta-analysis and alternative corrections for publication

Table 5. MSE and Coverage of Alternative Meta-Methods With 50% Reporting Selection.

Design			MSE			Coverage (95%)		
d	*m*	α_h	RE	WLS	PET-PEESE	RE	WLS	PET-PEESE
0	20	0	.0390	.0285	.0043	.0056	.0506	.9915
0	20	6.25	.0414	.0303	.0055	.0070	.0654	.9815
0	20	12.5	.0478	.0348	.0096	.0162	.1181	.9482
0	20	25	.0694	.0514	.0261	.0481	.2055	.8413
0	20	50	.1316	.1017	.0782	.1362	.3269	.7742
0	80	0	.0387	.0281	.0012	.0000	.0000	.9894
0	80	6.25	.0410	.0296	.0017	.0000	.0000	.9769
0	80	12.5	.0472	.0338	.0042	.0000	.0000	.9072
0	80	25	.0676	.0480	.0149	.0000	.0001	.7257
0	80	50	.1268	.0928	.0490	.0000	.0007	.5639
.2	20	0	.0105	.0084	.0103	.3221	.4811	.9173
.2	20	6.25	.0123	.0095	.0111	.2982	.4693	.9080
.2	20	12.5	.0175	.0130	.0135	.2641	.4439	.8773
.2	20	25	.0339	.0244	.0249	.2485	.4521	.8073
.2	20	50	.0861	.0635	.0675	.2840	.4957	.7641
.2	80	0	.0098	.0077	.0014	.0000	.0012	.9059
.2	80	6.25	.0117	.0087	.0020	.0000	.0020	.8333
.2	80	12.5	.0166	.0116	.0037	.0000	.0055	.6858
.2	80	25	.0322	.0211	.0109	.0002	.0150	.4997
.2	80	50	.0790	.0519	.0337	.0004	.0351	.4270
.5	20	0	.0018	.0017	.0027	.9253	.9148	.9133
.5	20	6.25	.0023	.0021	.0031	.8999	.9030	.9119
.5	20	12.5	.0037	.0031	.0054	.8553	.8776	.8953
.5	20	25	.0108	.0083	.0183	.7468	.8141	.8709
.5	20	50	.0421	.0302	.0704	.6037	.7418	.8442
.5	80	0	.0009	.0008	.0011	.7714	.7679	.8135
.5	80	6.25	.0011	.0010	.0012	.7230	.7601	.8272
.5	80	12.5	.0022	.0015	.0015	.5820	.7114	.8598
.5	80	25	.0082	.0044	.0025	.2722	.5633	.8856
.5	80	50	.0351	.0186	.0128	.0833	.3988	.8043
Average			.0356	.0257	.0164	.2698	.3540	.8317

Note. RE and WLS denote the random effects and unrestricted weighted least squares meta-analysis averages, respectively; PET-PEESE is the meta-regression publication bias corrected estimate; and average bias is the difference between the simple mean of the reported effects and the mean true effect size, *d*. *m* is the number of studies meta-analyzed, and σ_h is standard deviation of the heterogeneity among true effects. PET-PEESE = precision-effect test and precision-effect estimate with standard errors; MSE = mean squared error.

bias under most realistic circumstances. Our simulations reveal that PET is valid and PET-PEESE reduces bias in typical social/personality psychological areas of research, but they have important limitations.

First, very large heterogeneity ($I^2 > 80\%$) can reduce power, increase MSE, and raise the probability of a type I error beyond its nominal level. Thus, systematic reviewers should regard $I^2 < 80\%$ as a guideline for the reliable application of PET-PEESE. Second, reliability and statistical power depend on the distribution of sample sizes found in the research record in question. If all studies are small, PET-PEESE has little power to identify a genuine empirical effect. However, this weakness is easily addressed by raising the nominal level to 20% when all studies are highly underpowered. Third, recent or sparse areas of research with only 10, 20, or fewer studies may also pose a challenge to PET-PEESE because this approach is based upon regression. Thus, reviewers and meta-analysts should use caution when applying these meta-regression methods. Nonetheless, even under these unfavorable conditions, PET-PEESE is likely to be more reliable than conventional meta-analysis, which is almost always invalid when there is selective reporting (or publication) bias.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: I acknowledge support from the Czech Science Foundation (grant 15-02411S).

Supplemental Material

The online data supplements are available at http://journals.sagepub.com/doi/suppl/10.1177/1948550617693062.

References

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychological Bulletin Review*, *22*, 366–377.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society. Series A*, *151*, 419–463.

Boggs, T., & Lasecki, L. (2015). Reliable gains? Evidence for substantially underpowered designs in studies of working memory training transfer to fluid intelligence. *Frontiers in Psychology*. doi:10.3389/fpsyg.2014.01589

Bruns, S. B., & Ioannidis, J. P. A. (2016). p-Curve and p-hacking in observational research. *PLoS One*, *11*, e0149144. doi:10.1371/journal. pone.0149144

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*, 796–815.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, MI: Lawrence Erlbaum.

Copas, J., & Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, *1*, 247–262.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Fraley, R. C., & Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One, 9*, e109019. doi:10.1371/journal. pone.0109019

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Greene, W. E. (1990). *Econometric analysis*. New York, NY: Macmillan.

Hagger, M. S., & Chatzisarantis, N. L. D. (2016). A multi-lab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine*, *21*, 1539–1558.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, C. (H). (in press). The power of bias in economics research. *The Economic Journal*. Also as *SWP*, Economics Series 2016/2, Deakin University, Australia. Retrieved July 21, 2016, from http://www.deakin.edu.au/__data/assets/pdf_file/0007/477763/2016_1.pdf.

Johnson, N., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distribution*. New York, NY: Wiley.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & 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. Retrieved November 23, 2016, from http://www.biomedcentral.com/1471-2288/9/2

Moreno, S. G., Sutton, A. J., Thompson, J. R., Ades, A. E., Abrams, K. R., & Cooper, N. J. (2012). A generalized weighting regression derived meta-analysis estimator robust to small-study effects and heterogeneity. *Statistics in Medicine*, *31*, 1407–1417.

Moreno, S. G., Sutton, A. J., Turner, E. H., Abrams, K. R., Cooper, N. J., Palmer, T. M., & Ades, A. E. (2009). Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *British Medical Journal*, *339*, b298. Retrieved November 23, 2016, from http://dx.doi.org/10.1136/bmj.b2981

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716–aac4716. doi:10.1126/science.aac4716

Popper, K. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Richard, F. D., & Bond, C. F. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

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*, 103–127.

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

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

Stanley, T. D., & Doucouliagos, C. H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine*, *34*, 2116–2127.

Stanley, T. D., & Doucouliagos, C. H. (2016). Neither fixed nor random: Weighted least squares meta-regression. *Research Synthesis Methods*. doi:10.1002/jrsm.1211

Stanley, T. D., & Doucouliagos, C. (H), & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*. Advance online publication. doi:10.1002/sim.7228

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

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

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, *11*, 713–729.

Author Biography

T. D. Stanley is the Julia Mobley Professor of Economics at Hendrix College; where he teaches several economics, statistics, and research classes. Since the 1980s, his research has focused on the development and application of meta-regression methods. Prof. Stanley is the Associate Editor of the Journal of Economic Surveys and the convener of the Meta-Analysis of Economics Research Network, https://www.hendrix.edu/maer-network/default.aspx?id=13678.

Handling Editor: Simine Vazire