Limitations of PET-PEESE and other meta-analysis methods

T.D. Stanley*

September 2016

Rough draft, please do not quote or cite without written permission.

Abstract

A novel meta-regression method, PET-PEESE, predicts and explains recent high-profile failures to replicate in psychology. The central purpose of this paper 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

* Julia Mobley 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.

Limitations of PET-PEESE and other meta-analysis methods

Recently, there have been high-profile failures to replicate psychological phenomenon (*e.g.*, Open Science Collaboration, 2015; Hagger et al., 2016). 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 failure to replicate psychological phenomenon, the ego depletion effect (Stanley and Doucouloagos, 2014; Carter et al.; 2015). These new meta-analysis methods for accommodating 'publication bias' can do much to address the source of the current credibility and replication 'crises' across the social sciences. However, they too have shortcomings. The central purpose of this paper is to identify the limitations of *PET-PEESE* when applied to typical areas of social/personality psychology. In the process, we also show that conventional meta-analytic methods (fixed- and random-effects weighted averages) are of little use in identifying an authentic effect when there is selective reporting of statistical significant results (aka, 'publication bias').

Selective Reporting and Publication Bias

For decades, researchers have been acutely aware that the selective reporting of statistically significant results (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 (Sterling, 1959; Rosenthal, 1979; Glass, McGaw and Smith, 1981; Hedges and Oklin, 1985; Begg and Berlin, 1988; Schmidt and Hunter, 2014, 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, 2008; Stanley et al, 2010; Stanley and 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 et al, 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.'

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 the reported estimate's standard error times the inverse Mill's ratio (Stanley and Doucouliagos, 2014, p. 61). Medical researchers sometimes use a linear approximation to the inverse Mill's ratio as the basis for a test of selective reporting bias—H_0:β_1 =0 in:

$$\hat{d}_i = \beta_0 + \beta_1 SE_i + u_i \qquad \text{i=1, 2, . . . , m} \qquad (1)$$

(Egger et al, 1997; Stanley, 2008; Stanley and Doucouliagos, 2014). 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 t-test of β_0 (H_0:β_0 =0) in the *WLS* estimate of equation (1) 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 (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 (1) tends to underestimate the true 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.

$$\hat{d}_i = \gamma_0 + \gamma_1 SE_i^2 + \upsilon_i \qquad \text{i=1, 2, . . . , m} \qquad (2)$$

with $1/SE_i^2$ as the *WLS* weight. $\hat{\gamma}_0$ from (2) is the precision-effect estimate with the standard error (*PEESE*).

To reduce the bias in estimating the 'true' average effect from either meta-regression model (1) or (2), Stanley and Doucouliagos (2014) recommend a conditional estimator. When there is evidence of a genuine treatment effect, *PEESE* from equation (2) should be 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.

Conventional meta-analysis

The role of conventional meta-analysis estimators, 'fixed'- and 'random-effects', 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,

$$\hat{\mu} = \Sigma \omega_i y_i / \Sigma \omega_i \ , \tag{3}$$

but they employ different weights and thereby have different variances. Fixed effect (*FE*) uses weights, $w_i = 1/SE_i^2$, and has variance, $1/\Sigma w_i$. Random effects (*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 weighted least squares approaches (Stanley and Doucouliagos, 2015). It is calculated by running a simple meta-regression, with no intercept, of t-statistics vs. precision:

$$t_i = \hat{d}_i / SE_i = \alpha(1/SE_i) + u_i \qquad \text{i=1, 2, . . . , m} \tag{4}$$

(Stanley and Doucouliagos, 2015). Ordinary least squares using any standard statistical software will calculate this *WLS* weighted average, $\hat{\alpha}$, its standard error and confidence interval.

Comprehensive simulations demonstrate that the *unrestricted* weighted least squares estimator's statistical properties are as good as and often better than random-effects *when the random-effects model is true* (Stanley and Doucouliagos, 2015). When there is no selective reporting (or publication) bias, *WLS*'s properties are practically equivalent to *RE*. However, if there is selective reporting, *WLS* has consistently smaller bias than *RE*. The simulations reported in this paper do not report fixed-effect (*FE*) to conserve space and because *WLS* gives the exact same point estimate but always has superior standard errors when there is heterogeneity.

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 and 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 FAT-PET meta-regression model (1).

Design

The average reported Cohen's *d* in social psychology is approximately .4 (Richard and 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. While replicating 100 psychological experiments, the Open Science Collaboration (2015) found that average effects were one-half the magnitude as those 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 et al., 2016). Combining this 100% selective reporting bias with Richard and Bond's (2003) survey suggests that a 'true' effect of 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}. For those few areas of research which have more than 80

comparable estimates, the relative statistical properties reported below will differ little from what we find for m=80.

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

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

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, σ_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: 0.0, 0.2 or 0.5. 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, or 200} as 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 et al., 2010; Stanley and Doucouliagos, 2014; Stanley and Doucouliagos, 2015; Stanley and Doucouliagos, 2016). 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, σ_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, μ_d = {0.0, 0.2, 0.5}. This heterogeneity causes the relative measure of observed heterogeneity, I^2, to vary from near 0 to over 95% (Higgins and Thompson, 2002). I^2 is easy to calculate. I^2 = {(MSE-1)/MSE} from the simple WLS meta-regression, equation (4). See Tables 1-4 and note that these I^2 values are computed empirically for each simulated meta-analysis.

Table 1: Bias, power and level of alternative meta-methods with 50% reporting selection

Design			Average		Bias			Power/Type I Error		
d	**m**	**σ_h**	**Bias**	**I^2**	***RE***	***WLS***	*PET-PEESE*	***RE***	***WLS***	***PET***
0	10	0	.2489	.5113	.1957	.1674	.0014	.7977	.4828	0.0000
0	10	6.25	.2506	.5317	.2004	.1707	.0082	.7784	.4955	.0001
0	10	12.5	.2609	.5847	.2156	.1828	.0239	.7481	.4836	.0029
0	10	25	.2917	.7082	.2580	.2171	.0622	.6921	.4717	.0245
0	10	50	.3701	.8602	.3503	.2989	.1367	.6037	.4192	.0574
0	20	0	.2482	.5140	.1958	.1668	.0008	.9942	.9503	.0002
0	20	6.25	.2517	.5409	.2015	.1714	.0086	.9931	.9290	.0005
0	20	12.5	.2603	.6020	.2158	.1824	.0254	.9825	.8833	.0052
0	20	25	.2902	.7367	.2581	.2177	.0714	.9469	.7913	.0342
0	20	50	.3683	.8818	.3502	.2977	.1455	.8654	.6761	.0914
0	40	0	.2484	.5154	.1958	.1667	.0006	1.0000	1.0000	.0002
0	40	6.25	.2515	.5427	.2016	.1712	.0089	1.0000	.9999	.0009
0	40	12.5	.2614	.6102	.2170	.1832	.0275	1.0000	.9988	.0068
0	40	25	.2899	.7486	.2581	.2167	.0746	.9995	.9840	.0544
0	40	50	.3697	.8917	.3521	.2981	.1555	.9935	.9286	.1144
0	80	0	.2487	.5162	.1964	.1673	.0019	1.0000	1.0000	.0005
0	80	6.25	.2516	.5450	.2019	.1714	.0097	1.0000	1.0000	.0009
0	80	12.5	.2604	.6131	.2166	.1828	.0307	1.0000	1.0000	.0109
0	80	25	.2902	.7561	.2587	.2168	.0829	1.0000	1.0000	.0853
0	80	50	.3703	.8958	.3530	.2994	.1739	1.0000	.9990	.1941
Average type I error rate (size)								**.9198**	**.8247**	**.0342**
0.2	10	0	.1665	.2365	.0993	.0863	-.0616	1.0000	.9999	.1262
0.2	10	6.25	.1696	.2780	.1070	.0923	-.0499	.9998	.9993	.1459
0.2	10	12.5	.1808	.3780	.1252	.1045	-.0378	.9991	.9926	.1859
0.2	10	25	.2134	.6038	.1747	.1433	.0037	.9884	.9290	.2150
0.2	10	50	.2970	.8353	.2731	.2191	.0451	.9207	.7783	.2003
0.2	20	0	.1659	.2254	.0986	.0868	-.0345	1.0000	1.0000	.3056
0.2	20	6.25	.1704	.2759	.1065	.0919	-.0297	1.0000	1.0000	.3268
0.2	20	12.5	.1809	.4034	.1266	.1060	-.0126	1.0000	1.0000	.3423
0.2	20	25	.2136	.6509	.1756	.1418	.0167	1.0000	.9983	.3378
0.2	20	50	.2973	.8632	.2762	.2218	.0669	.9973	.9592	.2857
0.2	40	0	.1667	.2206	.0984	.0866	-.0043	1.0000	1.0000	.6265
0.2	40	6.25	.1701	.2797	.1066	.0920	.0019	1.0000	1.0000	.6271
0.2	40	12.5	.1810	.4212	.1269	.1055	.0147	1.0000	1.0000	.6121
0.2	40	25	.2130	.6754	.1763	.1409	.0397	1.0000	1.0000	.5390
0.2	40	50	.2974	.8761	.2768	.2202	.0863	1.0000	.9998	.4152
0.2	80	0	.1663	.2198	.0982	.0864	.0178	1.0000	1.0000	.9244
0.2	80	6.25	.1706	.2839	.1070	.0921	.0239	1.0000	1.0000	.9136
0.2	80	12.5	.1814	.4301	.1275	.1055	.0370	1.0000	1.0000	.8927
0.2	80	25	.2136	.6843	.1772	.1414	.0676	1.0000	1.0000	.8078
0.2	80	50	.2970	.8819	.2768	.2201	.1207	1.0000	1.0000	.6334
0.5	10	0	.0806	.1071	.0277	.0236	-.0243	1.0000	1.0000	.9528
0.5	10	6.25	.0824	.1429	.0301	.0238	-.0282	1.0000	1.0000	.9251
0.5	10	12.5	.0912	.2573	.0421	.0305	-.0339	1.0000	1.0000	.8259
0.5	10	25	.1188	.5344	.0797	.0545	-.0431	1.0000	.9995	.6137
0.5	10	50	.2033	.8102	.1769	.1247	-.0368	.9970	.9599	.3785
0.5	20	0	.0793	.0786	.0252	.0222	-.0239	1.0000	1.0000	.9997
0.5	20	6.25	.0834	.1282	.0300	.0249	-.0224	1.0000	1.0000	.9978
0.5	20	12.5	.0894	.2786	.0401	.0289	-.0209	1.0000	1.0000	.9854
0.5	20	25	.1193	.5905	.0819	.0552	-.0107	1.0000	1.0000	.8578
0.5	20	50	.2020	.8492	.1771	.1197	-.0151	1.0000	.9992	.5567
0.5	40	0	.0799	.0567	.0247	.0226	-.0240	1.0000	1.0000	1.0000
0.5	40	6.25	.0833	.1102	.0290	.0247	-.0229	1.0000	1.0000	1.0000
0.5	40	12.5	.0909	.2892	.0415	.0302	-.0190	1.0000	1.0000	.9998
0.5	40	25	.1192	.6227	.0824	.0542	.0007	1.0000	1.0000	.9838
0.5	40	50	.2021	.8624	.1789	.1196	.0234	1.0000	1.0000	.7905
0.5	80	0	.0804	.0395	.0241	.0227	-.0243	1.0000	1.0000	1.0000
0.5	80	6.25	.0826	.1012	.0280	.0244	-.0231	1.0000	1.0000	1.0000
0.5	80	12.5	.0906	.3028	.0416	.0300	-.0193	1.0000	1.0000	1.0000
0.5	80	25	.1208	.6370	.0846	.0556	.0029	1.0000	1.0000	.9998
0.5	80	50	.2027	.8694	.1800	.1208	.0514	1.0000	1.0000	.9568
Average			**.2016**	**.5083**	**.1575**	**.1291**	**.0175**	**.9976**	**.9904**	**.6822**

Notes: RE, WLS denotes the random-effects and unrestricted weighted least squares meta-analysis averages, respectively, and PET-PEESE is the meta-regression publication bias corrected estimate.

Table 2: Bias, power and level of alternative meta-methods with no reporting selection

Design			Average		Bias			Power/Type I Error		
d	**m**	**σ_h**	**Bias**	**I^2**	***RE***	***WLS***	*PET-PEESE*	***RE***	***WLS***	***PET***
0	10	0	-.0010	.1044	-.0002	-.0000	-.0087	.0334	.0468	.0507
0	10	6.25	.0001	.1530	.0005	.0004	-.0104	.0560	.0610	.0607
0	10	12.5	-.0005	.2840	-.0005	-.0008	-.0166	.0839	.0799	.0713
0	10	25	.0003	.5931	.0003	-.0002	-.0253	.1038	.1063	.0969
0	10	50	.0009	.8541	.0012	.0021	-.0413	.1148	.1213	.0983
0	20	0	.0004	.0824	.0000	-.0000	-.0082	.0354	.0459	.0522
0	20	6.25	-.0002	.1388	.0000	.0000	-.0082	.0554	.0587	.0534
0	20	12.5	-.0002	.3120	.0004	.0004	-.0097	.0775	.0816	.0745
0	20	25	-.0009	.6514	-.0006	-.0004	-.0173	.0787	.1105	.0965
0	20	50	-.0001	.8803	.0000	.0013	-.0255	.0768	.1152	.0967
0	40	0	-.0002	.0608	.0000	.0000	-.0047	.0428	.0541	.0519
0	40	6.25	.0005	.1283	-.0001	-.0001	-.0071	.0501	.0572	.0609
0	40	12.5	-.0004	.3334	-.0007	-.0008	-.0091	.0621	.0785	.0807
0	40	25	-.0004	.6797	-.0006	-.0008	-.0137	.0638	.1017	.0906
0	40	50	.0004	.8917	.0006	.0009	-.0188	.0670	.1166	.0924
0	80	0	.0003	.0428	.0005	.0005	-.0031	.0422	.0527	.0482
0	80	6.25	-.0000	.1188	-.0001	-.0001	-.0046	.0533	.0602	.0587
0	80	12.5	-.0006	.3519	-.0004	-.0002	-.0051	.0569	.0823	.0740
0	80	25	.0003	.6909	.0005	.0006	-.0079	.0574	.1056	.0922
0	80	50	.0003	.8957	.0003	.0001	-.0155	.0590	.1149	.0935
Average type I error rate (size)								**.0635**	**.0825**	**.0747**
0.2	10	0	.0029	.1042	.0001	.0000	-.0254	.9771	.9789	.4301
0.2	10	6.25	.0019	.1489	-.0005	-.0006	-.0275	.9617	.9572	.3907
0.2	10	12.5	.0025	.2887	-.0001	-.0006	-.0335	.8909	.8711	.3359
0.2	10	25	-.0001	.5954	-.0026	-.0048	-.0517	.6808	.6316	.2321
0.2	10	50	.0030	.8537	-.0006	-.0095	-.0858	.3871	.3529	.1555
0.2	20	0	.0019	.0811	-.0005	-.0005	-.0139	.9997	.9999	.7048
0.2	20	6.25	.0027	.1402	.0000	-.0002	-.0165	.9990	.9990	.6403
0.2	20	12.5	.0015	.3105	-.0010	-.0015	-.0249	.9922	.9906	.5242
0.2	20	25	.0036	.6516	.0006	-.0023	-.0427	.8800	.8460	.3240
0.2	20	50	.0027	.8790	-.0002	-.0103	-.0774	.5337	.4905	.1799
0.2	40	0	.0021	.0615	-.0002	-.0002	-.0045	1.0000	1.0000	.9325
0.2	40	6.25	.0024	.1290	.0001	-.0001	-.0061	1.0000	1.0000	.8935
0.2	40	12.5	.0025	.3354	.0000	-.0005	-.0117	.9999	1.0000	.7600
0.2	40	25	.0027	.6779	.0004	-.0016	-.0280	.9893	.9825	.5007
0.2	40	50	.0022	.8903	-.0006	-.0106	-.0673	.7613	.6994	.2317
0.2	80	0	.0022	.0440	-.0003	-.0004	-.0026	1.0000	1.0000	.9978
0.2	80	6.25	.0023	.1200	-.0003	-.0004	-.0028	1.0000	1.0000	.9931
0.2	80	12.5	.0018	.3499	-.0006	-.0012	-.0051	1.0000	1.0000	.9521
0.2	80	25	.0023	.6914	-.0004	-.0031	-.0199	.9999	.9997	.7069
0.2	80	50	.0009	.8953	-.0022	-.0130	-.0620	.9485	.9020	.3281
0.5	10	0	.0066	.1076	.0000	-.0003	-.0057	1.0000	1.0000	.9760
0.5	10	6.25	.0047	.1495	.0002	-.0003	-.0051	1.0000	1.0000	.9549
0.5	10	12.5	.0058	.2843	-.0010	-.0026	-.0178	1.0000	1.0000	.8721
0.5	10	25	.0067	.5843	-.0003	-.0054	-.0515	.9986	.9938	.6137
0.5	10	50	.0041	.8511	-.0038	-.0250	-.1354	.9101	.8123	.3072
0.5	20	0	.0059	.0819	-.0002	-.0003	-.0048	1.0000	1.0000	1.0000
0.5	20	6.25	.0056	.1337	-.0002	-.0007	-.0052	1.0000	1.0000	.9996
0.5	20	12.5	.0049	.3074	-.0015	-.0030	-.0092	1.0000	1.0000	.9923
0.5	20	25	.0050	.6454	-.0014	-.0069	-.0293	1.0000	1.0000	.8457
0.5	20	50	.0048	.8772	-.0028	-.0251	-.1168	.9925	.9680	.4332
0.5	40	0	.0044	.0630	-.0011	-.0013	-.0057	1.0000	1.0000	1.0000
0.5	40	6.25	.0047	.1258	-.0010	-.0013	-.0060	1.0000	1.0000	1.0000
0.5	40	12.5	.0051	.3275	-.0008	-.0023	-.0081	1.0000	1.0000	1.0000
0.5	40	25	.0060	.6728	-.0009	-.0078	-.0202	1.0000	1.0000	.9786
0.5	40	50	.0062	.8886	-.0014	-.0267	-.0960	1.0000	.9995	.6185
0.5	80	0	.0053	.0433	-.0006	-.0007	-.0056	1.0000	1.0000	1.0000
0.5	80	6.25	.0049	.1161	-.0009	-.0012	-.0063	1.0000	1.0000	1.0000
0.5	80	12.5	.0053	.3438	-.0010	-.0026	-.0089	1.0000	1.0000	1.0000
0.5	80	25	.0054	.6858	-.0013	-.0079	-.0188	1.0000	1.0000	.9999
0.5	80	50	.0057	.8932	-.0019	-.0273	-.0685	1.0000	1.0000	.8500
Average			**-.0003**	**.5083**	**-.0005**	**-.0035**	**-.0249**	**.9476**	**.9369**	**.7164**

Notes: RE, WLS denotes the random-effects and unrestricted weighted least squares meta-analysis averages, respectively, and PET-PEESE is the meta-regression publication bias corrected estimate.

Table 3: Bias, power and level: 50% reporting selection for larger sample sizes

Design			Average		Bias			Power/Type I Error		
d	**m**	**σ_h**	**Bias**	**I^2**	***RE***	***WLS***	*PET-PEESE*	***RE***	***WLS***	***PET***
0	10	0	.1659	.5296	.1273	.1065	.0011	.7948	.4688	.0001
0	10	6.25	.1718	.5694	.1364	.1136	.0109	.7750	.4899	.0007
0	10	12.5	.1836	.6601	.1554	.1287	.0310	.7286	.4813	.0123
0	10	25	.2188	.8116	.2021	.1692	.0743	.6499	.4511	.0532
0	10	50	.3066	.9283	.2992	.2587	.1481	.5623	.4016	.0841
0	20	0	.1663	.5309	.1276	.1063	.0001	.9957	.9458	.0001
0	20	6.25	.1710	.5838	.1364	.1132	.0114	.9885	.8990	.0017
0	20	12.5	.1837	.6864	.1561	.1291	.0343	.9711	.8273	.0201
0	20	25	.2191	.8405	.2034	.1693	.0789	.9085	.7181	.0784
0	20	50	.3071	.9428	.2999	.2564	.1524	.8065	.6187	.1168
0	40	0	.1656	.5336	.1276	.1063	.0014	1.0000	1.0000	.0000
0	40	6.25	.1711	.5903	.1367	.1132	.0121	1.0000	.9992	.0031
0	40	12.5	.1839	.7002	.1568	.1292	.0372	1.0000	.9925	.0313
0	40	25	.2186	.8527	.2033	.1688	.0851	.9979	.9480	.1202
0	40	50	.3062	.9485	.2995	.2569	.1653	.9807	.8813	.1713
0	80	0	.1665	.5314	.1281	.1066	.0006	1.0000	1.0000	.0007
0	80	6.25	.1713	.5929	.1370	.1134	.0134	1.0000	1.0000	.0032
0	80	12.5	.1837	.7047	.1568	.1288	.0406	1.0000	1.0000	.0575
0	80	25	.2192	.8578	.2042	.1699	.0971	1.0000	.9994	.2025
0	80	50	.3065	.9512	.2999	.2566	.1788	.9998	.9934	.2704
Average type I error rate (size)								**.9080**	**.8058**	**.0614**
0.2	10	0	.0884	.1815	.0404	.0338	-.0288	1.0000	1.0000	.5095
0.2	10	6.25	.0942	.2787	.0510	.0410	-.0264	1.0000	1.0000	.4647
0.2	10	12.5	.1072	.4741	.0737	.0574	-.0155	1.0000	.9959	.4189
0.2	10	25	.1458	.7504	.1259	.0975	.0084	.9888	.9278	.3338
0.2	10	50	.2366	.9190	.2271	.1854	.0690	.9036	.7576	.2549
0.2	20	0	.0882	.1615	.0393	.0336	-.0100	1.0000	1.0000	.8610
0.2	20	6.25	.0942	.2841	.0513	.0416	-.0051	1.0000	1.0000	.7956
0.2	20	12.5	.1073	.5200	.0748	.0573	.0028	1.0000	1.0000	.6677
0.2	20	25	.1466	.7971	.1285	.0985	.0268	1.0000	.9980	.4997
0.2	20	50	.2354	.9355	.2269	.1825	.0792	.9930	.9426	.3567
0.2	40	0	.0881	.1514	.0389	.0337	-.0043	1.0000	1.0000	.9921
0.2	40	6.25	.0937	.2934	.0507	.0410	.0036	1.0000	1.0000	.9775
0.2	40	12.5	.1072	.5468	.0754	.0573	.0180	1.0000	1.0000	.9092
0.2	40	25	.1464	.8153	.1291	.0985	.0475	1.0000	1.0000	.7347
0.2	40	50	.2376	.9428	.2297	.1841	.1019	1.0000	.9985	.5213
0.2	80	0	.0883	.1446	.0390	.0341	-.0036	1.0000	1.0000	1.0000
0.2	80	6.25	.0941	.3045	.0511	.0411	.0042	1.0000	1.0000	.9998
0.2	80	12.5	.1079	.5630	.0764	.0575	.0224	1.0000	1.0000	.9952
0.2	80	25	.1462	.8241	.1295	.0990	.0633	1.0000	1.0000	.9438
0.2	80	50	.2371	.9457	.2293	.1828	.1264	1.0000	1.0000	.7590
0.5	10	0	.0294	.0935	.0070	.0055	-.0108	1.0000	1.0000	1.0000
0.5	10	6.25	.0307	.1983	.0087	.0054	-.0118	1.0000	1.0000	.9990
0.5	10	12.5	.0375	.4344	.0169	.0079	-.0141	1.0000	1.0000	.9676
0.5	10	25	.0656	.7427	.0487	.0273	-.0220	1.0000	.9994	.7410
0.5	10	50	.1499	.9138	.1385	.0929	-.0208	.9959	.9548	.4470
0.5	20	0	.0283	.0698	.0060	.0050	-.0110	1.0000	1.0000	1.0000
0.5	20	6.25	.0308	.2041	.0085	.0055	-.0120	1.0000	1.0000	1.0000
0.5	20	12.5	.0367	.4967	.0166	.0074	-.0127	1.0000	1.0000	.9999
0.5	20	25	.0652	.7933	.0499	.0263	-.0052	1.0000	1.0000	.9430
0.5	20	50	.1500	.9323	.1401	.0905	.0043	1.0000	.9994	.6506
0.5	40	0	.0290	.0500	.0062	.0056	-.0107	1.0000	1.0000	1.0000
0.5	40	6.25	.0309	.2039	.0082	.0055	-.0122	1.0000	1.0000	1.0000
0.5	40	12.5	.0372	.5298	.0172	.0077	-.0127	1.0000	1.0000	1.0000
0.5	40	25	.0649	.8117	.0503	.0257	-.0012	1.0000	1.0000	.9983
0.5	40	50	.1484	.9394	.1395	.0896	.0343	1.0000	1.0000	.8637
0.5	80	0	.0288	.0325	.0056	.0052	-.0112	1.0000	1.0000	1.0000
0.5	80	6.25	.0305	.2095	.0079	.0052	-.0124	1.0000	1.0000	1.0000
0.5	80	12.5	.0370	.5490	.0170	.0074	-.0131	1.0000	1.0000	1.0000
0.5	80	25	.0650	.8200	.0508	.0261	-.0007	1.0000	1.0000	1.0000
0.5	80	50	.1495	.9425	.1407	.0890	.0468	1.0000	1.0000	.9842
Average			**.1354**	**.5858**	**.1111**	**.0883**	**.0257**	**.9970**	**.9893**	**.8147**

Notes: RE, WLS denotes the random-effects and unrestricted weighted least squares meta-analysis averages, respectively, and PET-PEESE is the meta-regression publication bias corrected estimate.

Table 4: Bias, power and level: 50% reporting selection and smaller sample sizes

Design			Average		Bias			Power/Type I Error		
d	**m**	**σ_h**	**Bias**	**I^2**	***RE***	***WLS***	*PET-PEESE*	***RE***	***WLS***	***PET***
0	10	0	.3760	.4698	.3371	.3180	-.0701	.8008	.5812	.0002
0	10	6.25	.3778	.4827	.3396	.3200	-.0670	.7899	.5603	0.0000
0	10	12.5	.3837	.5017	.3469	.3260	-.0610	.7767	.5564	0.0000
0	10	25	.4047	.5794	.3726	.3475	-.0368	.7291	.5157	.0004
0	10	50	.4725	.7282	.4473	.4102	-.0049	.6517	.4438	.0024
0	20	0	.3754	.4760	.3365	.3172	-.0739	.9939	.9756	0.0000
0	20	6.25	.3778	.4837	.3397	.3200	-.0678	.9936	.9712	.0003
0	20	12.5	.3828	.5108	.3462	.3249	-.0642	.9895	.9592	.0002
0	20	25	.4052	.5956	.3728	.3470	-.0427	.9737	.9205	.0005
0	20	50	.4752	.7490	.4508	.4119	-.0126	.9254	.8235	.0025
0	40	0	.3750	.4766	.3367	.3174	-.0699	1.0000	1.0000	.0003
0	40	6.25	.3780	.4891	.3398	.3197	-.0717	1.0000	1.0000	.0002
0	40	12.5	.3835	.5174	.3469	.3255	-.0636	1.0000	1.0000	.0004
0	40	25	.4058	.6031	.3738	.3478	-.0407	1.0000	.9999	.0011
0	40	50	.4722	.7598	.4483	.4089	-.0134	.9989	.9930	.0039
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 (size)								**.9312**	**.8650**	**.0010**
0.2	10	0	.2914	.2746	.2445	.2332	-.1509	.9980	.9896	.0054
0.2	10	6.25	.2939	.2915	.2474	.2354	-.1514	.9977	.9879	.0053
0.2	10	12.5	.3016	.3271	.2566	.2425	-.1487	.9967	.9852	.0058
0.2	10	25	.3257	.4472	.2863	.2658	-.1319	.9871	.9522	.0117
0.2	10	50	.3967	.6693	.3665	.3271	-.1368	.9387	.8399	.0185
0.2	20	0	.2917	.2696	.2441	.2333	-.1547	1.0000	1.0000	.0059
0.2	20	6.25	.2952	.2874	.2479	.2364	-.1531	1.0000	1.0000	.0076
0.2	20	12.5	.3022	.3351	.2572	.2436	-.1464	1.0000	1.0000	.0100
0.2	20	25	.3254	.4716	.2859	.2646	-.1405	1.0000	.9998	.0176
0.2	20	50	.3984	.7060	.3684	.3275	-.1441	.9988	.9928	.0208
0.2	40	0	.2923	.2683	.2443	.2336	-.1552	1.0000	1.0000	.0104
0.2	40	6.25	.2941	.2893	.2468	.2352	-.1560	1.0000	1.0000	.0114
0.2	40	12.5	.3014	.3396	.2561	.2424	-.1480	1.0000	1.0000	.0145
0.2	40	25	.3253	.4874	.2860	.2644	-.1400	1.0000	1.0000	.0205
0.2	40	50	.3958	.7205	.3666	.3248	-.1485	1.0000	1.0000	.0211
0.2	80	0	.2922	.2741	.2440	.2333	-.1539	1.0000	1.0000	.0141
0.2	80	6.25	.2947	.2924	.2471	.2355	-.1535	1.0000	1.0000	.0197
0.2	80	12.5	.3014	.3483	.2562	.2422	-.1473	1.0000	1.0000	.0204
0.2	80	25	.3252	.4978	.2866	.2649	-.1310	1.0000	1.0000	.0273
0.2	80	50	.3968	.7265	.3678	.3254	-.1481	1.0000	1.0000	.0192
0.5	10	0	.1814	.1061	.1294	.1247	-.2355	1.0000	1.0000	.0927
0.5	10	6.25	.1838	.1215	.1319	.1265	-.2378	1.0000	1.0000	.0863
0.5	10	12.5	.1919	.1562	.1412	.1340	-.2374	1.0000	1.0000	.0916
0.5	10	25	.2183	.2948	.1722	.1574	-.2376	1.0000	.9997	.0766
0.5	10	50	.2920	.5953	.2532	.2116	-.3074	.9960	.9790	.0473
0.5	20	0	.1820	.0794	.1277	.1242	-.2202	1.0000	1.0000	.1729
0.5	20	6.25	.1856	.0930	.1317	.1278	-.2175	1.0000	1.0000	.1739
0.5	20	12.5	.1925	.1374	.1399	.1341	-.2206	1.0000	1.0000	.1716
0.5	20	25	.2173	.3129	.1708	.1563	-.2281	1.0000	1.0000	.1277
0.5	20	50	.2950	.6448	.2573	.2141	-.3123	1.0000	1.0000	.0636
0.5	40	0	.1814	.0558	.1257	.1235	-.1763	1.0000	1.0000	.3429
0.5	40	6.25	.1849	.0694	.1301	.1272	-.1738	1.0000	1.0000	.3442
0.5	40	12.5	.1926	.1232	.1386	.1335	-.1839	1.0000	1.0000	.3018
0.5	40	25	.2188	.3263	.1710	.1561	-.2081	1.0000	1.0000	.2142
0.5	40	50	.2944	.6671	.2579	.2137	-.3044	1.0000	1.0000	.0800
0.5	80	0	.1821	.0374	.1255	.1240	-.1107	1.0000	1.0000	.6159
0.5	80	6.25	.1845	.0526	.1283	.1262	-.1121	1.0000	1.0000	.6013
0.5	80	12.5	.1918	.1068	.1374	.1330	-.1169	1.0000	1.0000	.5560
0.5	80	25	.2172	.3349	.1697	.1546	-.1524	1.0000	1.0000	.3977
0.5	80	50	.2942	.6766	.2578	.2128	-.2889	1.0000	1.0000	.1201
Average			**.3131**	**.4100**	**.2714**	**.2518**	**-.1374**	**.9978**	**.9932**	**.1241**

Notes: RE, WLS denotes the random-effects and unrestricted weighted least squares meta-analysis averages, respectively, and PET-PEESE is the meta-regression publication bias corrected estimate.

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 (Table 2) and others in which half of the reported results have undergone a process of selection to be statistically significant and positive (Tables 1, 3 and 4). 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. The simulations results reported in Table 1 for the 50% selective reporting case correspond quite closely to what the Open Science Collaboration (2015) and Richard and Bond's (2003) broad surveys find. Table 1 reveals that 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

Tables 1-4 report the average biases of random-effects (*RE*), the unrestricted weighted least squares (*WLS*) and the conditional meta-regression estimator *PET-PEESE*. 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: $\mu_d=0$). When the mean true effect is zero (i.e., $\mu_d =0$), these proportions represent the observed frequency of a type I error (aka 'size'). About one-third of the way down Tables 1-4, the average 'sizes' are displayed in the last three columns. When the true effect is not zero (*i.e.*, $\mu_d =.2$ or $\mu_d =.5$), these proportions represent the power of these alternative estimators to identify a nonzero overall effect. At the bottom of Tables 1-4, the average powers are displayed along with the average biases and average I^2.

The simulations revealed in Table 1 assume that the distribution of sample sizes is n = {15, 35, 50, 100, 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 (.2941) when there is no true effect, $\mu_d=0$. Although random-effects (*RE*) reduces this bias somewhat (.1575, overall), *RE* can give the appearance of a small effect (.2446) when there is none ($\mu_d=0$). Worse still, *RE* makes a type I

error nearly 92% of the time (.9198). Thus, conventional random-effects meta-analysis does not 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 (82.5%, on average). The known relationship between *WLS* and fixed-effect (*FE*) implies that *FE* will always have worse type I error rates than *WLS* when there is any heterogeneity and is thus not reported.

Only the precision-effect test (*PET*) has acceptable type I error rates (3.5% on average), which is less than the nominal 5% level used by all of these simulations. Likewise, the related *PET-PEESE* conditional meta-regression estimator successfully reduces average bias to practical insignificance (.0175). Also, the average of absolute bias of *PET-PEESE* remains practically insignificant—.0382. But *PET* and *PET-PEESE* too has their limitations—see the 'Discussion and Comments' below.

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 size, so only its statistical power is relevant. For a small effect, $\mu_d=.2$, *PET*'s power reaches 50% if there are 40 or more estimates. However, when there is a medium-size effect, $\mu_d=.5$, *PET*'s power is almost always greater than 80%. The only exceptions to this positive evaluation of *PET* and *PET-PEESE* for these typical social/personality psychology conditions (Table 1) occur when there is very high heterogeneity. See 'Discussion and Comments' below for the meaning of these limitations and how they might be mitigated.

Simulations reported in Table 2 calculate the same statistics for the exact same design parameters as those that generate Table 1's 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 to 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 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, as before, *PET*'s power can be rather low for small meta-regression samples 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, 1997; Stanley, 2008).

To explore other weaknesses of these meta-analysis methods, 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 a different distribution 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 (*e.g.,* Stead et al., 2008; Stanley and Doucouliagos, 2014). 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 there is no overall true effect (μ_d =0). Average selective reporting bias =.2093, *RE'*s average bias is .1847, and *WLS* has an average bias of .1550. 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 Discussion and Comments below.

The simulations displayed in Table 4 assume yet another sample size distribution, n={10, 18, 25, 33, 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 and 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, .1241 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.

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, the precision-effect test (*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.*, μ_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 SEs). 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, σ_h =50, PET's size becomes inflated, larger than the nominal 5% level. This type I error inflation actually worsens as the meta-analysis sample increases. Although a serious problem, this type I error inflation is minor compared with very high type I error inflation rates that are typical of conventional meta-analysis: random-effects (*RE*) and weighted least squares (*WLS*)—recall Table 1. 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, σ_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 σ_h =50, the typical

variation of true effects from their mean true effect is $\pm$0.5d. This implies that nearly 16% of the time the *true* effect is actually negative when the mean true effect, μ_d, is positive and medium-size (μ_d =0.5). Heterogeneity means that there is no single 'true' effect, but rather 'true' effects vary from study to study by the equivalent of d=$\pm$.5 for σ_h =50. Thus, 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 zero. 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 psychological effect is.

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? We recommend that no meta-analysis method be used if I^2 is greater than 80%. Because tests of heterogeneity are widely known to have low power and to be statistically unreliable, formal hypothesis testing of I^2 or its related sample variance, MSE from equation (4), is unlikely to be useful in practice. Thus, we recommend this 80% cutoff only as an application 'rule of thumb.' When applied to these simulation results, *PET-PEESE* would not be calculated for many of the instances where heterogeneity is at its highest level, σ_h =50. As a result, most of the worrisome cases for *PET-PEESE* and *PET* would be eliminated, and the average power/type I error for *PET* improves—.7257 and .0135 for average power and size, respectively, for the simulations reported in Table 1. However, as discussed above, when observed heterogeneity is higher than 80%, the very meaning of social/personality psychological phenomenon is questionable. With a typical true effect of μ_d=.2 and a very high level of heterogeneity (σ_h =50), the *true effect* will have the opposite sign as μ_d over one-third of the time (.3446).

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 powers) found among the primary studies in the social/personality psychological research literature. For typical sample sizes found in social/personality psychology (Fraley and 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, unable to identify a genuine effect should it exist. In this worst-case scenario, the average power is only .1241, but the type I error rate is practically zero, .001—see Table 4. When reviewers observe that all the sample sizes in a research literature are small, *PET*'s statistical properties would improve notable if a one-tail test with alpha of 10% were used rather than the conventional two-tail test at 5%. Nonetheless, great caution should be used in interpreting any meta-analysis, regardless of the methods used when all studies are underpowered, because the research record contains little genuine information.

It is important to put *PET-PEESE's* limitations in context. First, in all these cases where the use of *PET-PEESE* is compromised, conventional meta-analysis (*RE* and *FE*) is much worse. In all three cases: small meta-analysis samples, high heterogeneity and research literatures comprised of only small-sample studies, *RE* and *WLS* are much worse than *PET-PEESE*. Thus, the limitations identified by our simulations are not challenges for *PET-PEESE* alone but apply to all meta-analysis methods.

Furthermore, meta-analysis's limitations may alternatively be regarded as inadequacies of the research record. If all studies in an area of social/personality psychology research are greatly underpowered, this can only be seen as weakness of that area of research. For over 30 years, psychologists have been acutely aware of the critical importance of statistical power (Cohen, 1988; Fraley and 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). Meta-analysis can effectively increase statistical power by combining several underpowered primary results only if they are known to be unbiased. With selective reporting bias, some adequately-powered studies are required to distinguish the genuine signal 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 under-researched, then there will insufficient research knowledge to be confident about the phenomenon in question. Lastly is the issue of very high levels of heterogeneity. The source of such a confused effect is not meta-analysis, but rather some combination of the

social/personality psychological phenomenon and the research methods used to study it. In some cases, social/personality psychological effects may vary greatly by socio-economic status, age, gender, culture, or the passage of time. Or, the instruments used to measure social/personality psychological effects may have low reliability and biases, causing the appearance of heterogeneity in reported outcomes. Before meta-analysis can reliably reduce ubiquitous selective reporting biases, the research record must contain some adequately powered studies.

Conclusion

We investigate the statistical properties and limitations of the *PET-PEESE* approach to identifying a genuine effect in the presence of selective reporting bias. Our simulations reveal that these meta-analysis methods are valid for the typical social/personality psychological area of research, but they do have important limitations. First, very large heterogeneity (I^2>80%) can reduce power and raise the probability of a type I error. Second, their reliability and statistical power depends on the distribution of sample sizes found in the research record in question. If all studies are small, *PET-PEESE* is almost powerless to identify a genuine empirical effect. Third, recent or sparse areas of research which have only a few 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.

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,* 419-63.

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.

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 edn Hillsdale: Erlbaum.

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(10): e109019. doi:10.1371/journal.pone.0109019.

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-Analysis in Social Research*, Beverly Hills: Sage.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Birt, A., Brand, R., & Cannon, T. (2016). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science.* [Epub ahead of print].

Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Orlando: 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). (2016). The power of bias in economics research. *The Economic Journal*, forthcoming, also as *SWP*, Economics Series 2016/2. http://www.deakin.edu.au/__data/assets/pdf_file/0007/477763/2016_1.pdf Deakin University, Australia. [Accessed on 21 July 2016].

Maxwell, S.E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies, *Psychological Methods, 9*, 147-63.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716–aac4716. doi:10.1126/science.aac4716

Popper, K. (1959). *The Logic of Scientific Discovery.* New York: Basic Books.

Richard, F.D. & Bond, C.F. (2003). One hundred years of social psychology quantitatively described, *Review of General Psychology, 7*, 331–63.

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: 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*,103-27.

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.

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*, 2116-27.

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

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

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*, 30-4.