Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015

Michael Inzlicht[1,2]

Will M. Gervais[3]

Elliott T. Berkman[4,5]

[1]Department of Psychology, University of Toronto

[2]Rotman School of Management, University of Toronto

[3]Department of Psychology, University of Kentucky

[4]Department of Psychology, University of Oregon

[5]Center for Translational Neuroscience, University of Oregon

Cite as:

Inzlicht, M., Gervais, W., & Berkman, E. (2015). Bias-Correction Techniques Alone Cannot Determine Whether Ego Depletion is Different from Zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015. *Available at SSRN: http://ssrn.com/abstract=2659409 or http://dx.doi.org/10.2139/ssrn.2659409*

Abstract

Carter, Kofler, Forster, & McCullough (2015) conducted a bias-corrected meta-analysis of the so-called ego depletion effect to determine its real size and robustness. Their efforts have raised awareness of how badly meta-analyses can mislead when the articles that go into them are products of publication bias. Despite our genuine enthusiasm for their work, we worry that in their zeal to correct the record of publication bias, they have drawn too heavily on largely untested statistical techniques that can be insensitive and sometimes misleading. We tested a set of bias-correction techniques, including those favored by Carter and colleagues, by simulating 40,000 meta-analyses in a range of situations that approximate what is found in the ego depletion literature, most notably the presence of heterogeneous effects filtered by publication bias. Our simulations revealed that not one of the bias-correction techniques revealed itself superior in all conditions, with corrections performing adequately in some situations but inadequately in others. Such a result implies that meta-analysts ought to present a range of possible effect sizes and to consider them all as being possible. The problem with the ego depletion literature is that the bias-corrected estimates for the overall effect do not converge, with estimates ranging from g=0 to g=0.24 to g=0.26. Despite our admiration for this program of meta-research, we suggest that bias-corrected meta-analyses cannot yet resolve whether the overall ego depletion is different from zero or not.

Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015

Psychology owes a debt of thanks to Evan Carter and Michael McCullough. They and their colleagues (Carter, Kofler, Forster, & Mccullough, 2015; Carter & McCullough, 2014) have courageously raised awareness of how publication bias can inflate—if not altogether warp—effect size estimates for an entire literature. Because of them, the "there are 100 studies demonstrating X" is no longer a viable defense against charges of non-replicability and non-robustness. We now know that meta-analyses can be misleading when the articles that go into them are products of the multifarious forms of bias that plague our field. And for this, the field ought to be grateful.

Despite our admiration and overall support for their recent program of meta-research, we worry that in their zeal to correct the record on the state of the ego depletion effect, they have drawn too heavily on new and largely untested statistical techniques that can themselves be biased and insensitive, and as a result misleading.

Careful, crucial, and complete

To quickly recap, Carter and colleagues (2015) conducted a meta-analysis of the so-called ego depletion effect (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Baumeister, Vohs, & Tice, 2007) to determine its real size and robustness. Ego depletion describes the phenomenon whereby self-control performance deteriorates over time from repeated exertion (Baumeister et al., 2007); it is when self-control tends to fail when it follows previous acts of effortful self-control, as if some central resource has been depleted (Muraven & Baumeister, 2000; but see, Inzlicht, Schmeichel, & Macrae, 2014). According to proponents of the resource model of self-control, self-control relies on some central yet limited resource that gets consumed by all manner of control until the resource is depleted (i.e., ego depletion) with subsequent control slacking as a result (Baumeister et al., 1998; Muraven & Baumeister, 2000).

Understanding the size and robustness of the ego depletion effect is of some import, given how foundational the phenomenon has become to our understanding of control and effort (Westbrook & Braver, 2015), how widely it has spread to other disciplines (e.g., Vohs, Baumeister, Joiner, & Rudd, 2000), and how broadly it is recognized as a fact by the lay public (Baumeister & Tierney, 2011). This public acceptance was accompanied by apparent scientific support, as a meta-analysis of nearly 200 ego depletion studies revealed a healthy effect size of d= 0.62 (Hagger, Wood, Stiff, & Chatzisarantis, 2010).

Carter and colleagues' (2015) meta-analysis is an improvement over this past attempt (Hagger et al., 2010) for a number of reasons. First, they only included studies that tested the core ego depletion effect—the waning of self-control over time when it follows previous acts of control—excluding studies that tried to build on this central premise, but which did not directly test it. Second, they only included studies that unambiguously manipulated and measured self-control, which again allows for a direct test of the core ego depletion effect. Third, they scoured the literature to identify and include a significant number of non-published studies, thus attenuating some of the serious problems associated with the file drawer (Rosenthal, 1979). Fourth, and most critical, they addressed the inflationary effects of publication bias and other small study effects by using several new techniques to identify and correct for them. With this last step, they follow on their earlier bias-corrected meta-analysis, which suggested that the depletion effect was either small (d=0.25 [0.18, 0.32]) or, perhaps, not meaningfully different from zero (d= -0.10 [-0.23, 0. 02]) (Carter & McCullough, 2014).

Carter and colleagues’ (2015) meta-analysis poses a serious blow to the resource model. Even ignoring their extraordinary attempts to correct the literature for publication bias, their analyses reveal that the resource model is in need of serious re-thinking. Despite the resource model suggesting that all forms of control draw upon a common central resource, the current analyses suggest that this is often not the case. Carter and colleagues (2015) found that in two, or possibly, in three of the eight self-control domains the basic ego depletion effect did not hold, even when relying on standard meta-analytic techniques. This suggests that self-control might not be so central after all, with perhaps some behaviors more likely to wane over time than others. While we (e.g., Inzlicht & Berkman, 2015) and others (e.g., Evans, Boggero, & Segerstrom, 2015; Kool & Botvinick, 2014; Kurzban, Duckworth, Kable, & Myers, 2013) have proposed alternatives to the resource model, all of these alternatives have assumed that control and effort are central, an assumption that is clearly in need of revision.

In presenting their crucial work, Carter and colleagues (2015) were very careful, very thoughtful, and very thorough. They analyzed their data in multiple complementary ways; they applied many independent tests; and, they generally wrote about their results using measured language. We are grateful their paper was published for many reasons, not the least of which is that they raised awareness of how badly publication bias and small study effects can lead a field astray. That is, even if we are not persuaded by all of their arguments, their analysis, as well as that of countless others (e.g., Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011), makes clear that publication bias and the use of questionable research practices can lead a field to believe an effect is real and robust when it is neither of those things.

Is it warranted to conclude that there is no evidence that ego depletion is real?

Despite being legitimately admiring of their work and grateful for the many things we have learned, we ask if their most provocative claim, that there is "very little evidence that the depletion effect is a real phenomenon," (p. 796) is premature (see also, Hagger & Chatzisarantis, 2014). Our critique is that their conclusion that the overall depletion effect is indistinguishable from zero is based on new and largely untested bias-correction techniques.

Carter and colleagues (2015) actually find a significant overall depletion effect (*g*=.43 [0.34, 0.52]) when relying on a standard random-effects meta-analysis, despite it consisting of more than 40% unpublished studies and containing over 50% statistically non-significant results. Further, their use of the widely used (though widely criticized; e.g., Moreno, Sutton, Turner, et al., 2009; Simonsohn, Nelson, & Simmons, 2014) trim-and-fill procedure imputed 25% extra studies, yet it too returned a significant depletion effect (*g*=.24 [0.13, 0.34]). Notwithstanding, Carter and colleagues (2015) chose to make the claim that depletion is not distinguishable from zero based on new bias-correction techniques that, while used sparingly in the field of economics (e.g., Stanley, 2008), have not been tested or validated in the field of psychology. More critical, these statistical techniques are already known to perform poorly under the very conditions that characterize the field of ego depletion (if not all of social psychology), namely heterogeneous effects (Moreno, Sutton, Ades, et al., 2009; Reed, Florax, & Poot, 2015); although, to be fair, we note that nearly all meta-analytic techniques perform poorly under these conditions.

We think bias-correction techniques are necessary to help ascertain the size and robustness of effects in the face of publication bias and small study effects (e.g., Simonsohn et al., 2014; van Assen, van Aert, & Wicherts, 2015). However, we also believe that such techniques should be properly vetted and tested before concluding that an entire field has produced empirical results that are not meaningfully different from zero. The results of our own tests, based on 40,000 simulated meta-analyses that cover a range of conditions that one might find in social psychology indicate that these meta-regression techniques are wanting: they too often suggest no effect when an effect is in fact present (false negative); they are unable to reliably discriminate between real and non-real effects; and they return imprecise results that sometimes contain the real effect only because their suggested confidence intervals are unacceptably wide. For these reasons, these new techniques might be especially ill-equipped to arbitrate between real and non-real effects, which is precisely what Carter and colleagues (2015) have attempted to do.

Putting PET and PEESE to the test

Carter and colleagues (2015) perform a series of tests to determine whether ego depletion “ought to be convincing, even to a skeptical audience” (p. 797). We suggest that the specific tests Carter and colleagues employed might themselves be found lacking by such a skeptical audience. As previously mentioned, no previously published literature has evaluated the tests they employed in the context of social psychology, and the authors provide no evidence that these techniques work in the context in which they employed them.

While Carter and colleagues (2015) conducted many analyses, the ones they drew upon to conclude that there is scant evidence that ego depletion is a real phenomenon are meta-regression techniques known as the Precision Effect Test (PET) and the Precision Effect Estimation with Standard Error (PEESE). These techniques, which rely on weighted least squares regression, estimate meta-analytic effect sizes by taking into account the association between effect size and sample size (or standard error, to be precise). The logic here is that the size of an effect should be independent of sample size (e.g., the true effect of gender on height should not vary as a function of whether one samples 40, 80, or 200 men and women). However, because sample sizes in psychology are too small (Fraley & Vazire, 2014) and because journals tend not to publish null results, the effect sizes found in our journals tend to be inflated, with smaller studies seeming to produce larger effects. The association between effect size and sample size, in other words, is taken as a proxy for publication bias. While this is not an uncontroversial assumption (Borenstein, Hedges, Higgins, & Rothstein, 2009), there is now evidence that large (and presumably high-powered) studies do not show this association (Kühberger, Fritz, & Scherndl, 2014) providing some justification for the association between effect size and sample size as a metric of bias. PET and PEESE simply account and correct for this association, and when applied to the ego depletion literature, they return meta-analytic estimates that are not different from zero.

Before concluding that there is little evidence that ego depletion is real, we should probably also know something about the performance of the techniques upon which these conclusion hinge. After all, as new techniques developed in the field of economics, they were tested under conditions that might not be common in psychology. For example, the developers of these techniques (Stanley & Doucouliagos, 2013; Stanley, 2008) tested them in a limited range of situations and only tested them with a true effect size of r=0.3 (or d=0.6), which would be quite large for social psychology, placing it around the 75th percentile of effect sizes in our field (Richard, Bond, & Stokes-Zoota, 2003). We agree with Reed and colleagues (2015), who suggested that little effort was made to "ensure that the simulated samples 'look like' the kinds of samples used in actual meta-analysis studies" (p. 3). The tests of these techniques, in other words, were not done with parameters that resemble the kinds of studies found in social psychology. What is more, when tested across a fuller range of possible situations, these techniques were found to be wanting. Their inaccuracy was particularly apparent in the presence of heterogeneous effects and publication bias (Reed et al., 2015), the very conditions that Carter and colleagues (2015) confirm are very much present in the ego depletion literature. What is more troubling is that these same comparisons suggested that estimators that correct for publication bias produce estimates that are in some conditions more biased and less precise than estimators that do not correct for publication bias (Reed et al., 2015). As PET and PEESE are new and largely untested across the conditions that are typical in social psychology, it would be imprudent to assume that their novelty implies that they are more accurate or rigorous than available alternatives.

To further compare the performance of PET and PEESE, we ran a series of simulations examining how the methods performed when estimating the magnitude of true effects of various sizes, including a nil effect, in the presence of both publication bias and heterogeneity. All data and R code related to this manuscript, including the R code for these simulations, can be found at: https://osf.io/fcts8/. We aimed to make our simulations as realistic as possible, trying to capture what the current state of the social psychology literature looks like. Further, we modeled our simulated data after the data analyzed by Carter and colleagues (2015). To this end, all simulations were based on the same moderate degree of heterogeneity ($\tau = 0.4$) observed by Carter and colleagues, and the same degree of publication bias observed by Carter and colleagues. As with Carter and colleague’s data, we simulated conditions such that 42% of studies were subject to publication bias. In other words, publication bias was modeled such that 42% of the distribution of effects was significant ($p<0.05$) and directionally supportive, whereas the other 58% of the distribution consisted of the unfiltered effects, be they significant, directionally supportive, or not[1]. Modeled this way, the rate of significance in our simulations roughly matched the rate in Carter and colleagues' (2015) analysis.

We simulated four different true population effect sizes, g=0, g=0.16, g=0.36, and g=0.62, to correspond with a nil effect, the modal effect in psychology (which also correspond closely to a recent effect size estimate of the depletion effect derived from a recent paper that was free of publication bias; Tuk, Zhang, & Sweldens, 2015), the median effect in social psychology, and Hagger and colleagues' (2010) estimate of the ego depletion literature, respectively (Richard et al., 2003). For sample sizes, we sampled from a distribution closely matching that revealed by a recent survey of the social psychological literature (Fraley & Vazire, 2014), yielding sample sizes averaging per-condition n = 51 (but ranging between, on average, 11 and 170). These values approximate those identified by Carter and colleagues (2015). Finally, each simulated meta-analysis consists of 110 studies (k = 110), which is based on k = 116 for Carter and colleagues' (2015) analysis.

In additional to presenting the results of the uncorrected random effects meta-analysis (which acts as the baseline), each simulated meta-analysis was corrected using PET and PEESE, with 10,000 meta-analyses per effect size, yielding 40,000 simulated meta-analyses in total. In addition to these meta-regression techniques, we also corrected these meta-analyses using the so-called Trim and Fill and Top10 methods. Trim and Fill (Duval & Tweedie, 2000) is the most popular method for examining and correcting for publication bias and involves examining and correcting for funnel plot asymmetry by imputing assumed to be missing studies. The Top10 method (Stanley, Jarrell, & Doucouliagos, 2010)—which was meant to be a heuristic correction that is easy to implement—simply estimates effects size from the top 10% of studies in terms of precision (standard error) and discards the other 90%. Finally, as a second reference point, we

[1] We thank Evan Carter and Michael McCullough for this suggestion about how to model publication bias.

added a random meta-analytic estimate, which presents the least accurate (i.e., most random) correction we could imagine. This estimator picked a random effect size estimate along the distribution of social psychology effect sizes (Richard, et al., 2007), bounded between 0 and .36; confidence intervals varied uniformly in width between .2 and .8.

We evaluated our meta-analytic corrections on a number of outcome measures, including bias (estimated effect minus true effect), precision (mean square error and width of confidence intervals), the percentage of times the estimate was deemed to be zero (which should be high with nil effects, but low with real effects), and 95% coverage (the proportion of time the true value fell within the 95% confidence interval of the estimator), which by definition should be 95% of the time. In addition, we included outcome of signal detection analysis to help us determine sensitivity (the ability to differentiate a real effect from a nil; d') and bias (the extent to which one response is more probable than another).

Inspection of the results is quite deflating (Table 1; Figures 1a-1d). None of the estimates is particularly good, meaning that in the presence of heterogeneity and publication bias—the very conditions present in the ego depletion literature if not all of social psychology—none of the estimates capture reality very well, and all were biased, imprecise, and insensitive. This suggests that meta-analyses, at least when the current crop of corrections are used, should themselves be treated with skepticism.

Despite this, some estimates were better than others. The uncorrected random effects estimate fared very poorly. It routinely gave estimates that were far larger than the true value, it was insensitive to nil results, and its 95% confidence intervals covered far less variability than it was supposed to. It is no wonder, then, that Evan Carter and Michael McCullough (Carter et al., 2015; Carter & McCullough, 2014) sought to correct the recorded literature. Our results also show that our random correction was, well, random. It was unable to differentiate between true and nil effects and generally lacked precision. The only exception to this was when the true effect was zero, in which case the random correction outperformed all the others. That is, picking a correction at random provided estimates that were least biased, had the best coverage, and were the most precise. This surprised us. It suggests that when a true effect is zero and there is heterogeneity and publication bias in the literature, we are better off guessing at the size of an effect than in trusting and correcting the published literature. The problem, of course, is that we never know when an effect is real or not a priori.

A second surprise was that Trim and Fill performed better than we expected. While it consistently returned meta-analytic estimates that were too large, its bias was relatively low and it was best at discriminating between real and nil effects, at least when the true effect was small (g=.16; see Figure 1b). This latter finding is critical because one unbiased estimate of the depletion effect (Tuk et al., 2015) suggests the effect might be small in size, implying Trim and Fill might provide the best (or, perhaps, least bad) correction. In contrast, its 95% confidence interval too rarely covered the true effect and when an effect was nil, it rarely suggested nil. This is a major shortcoming. Thus, while Trim and Fill performed better than expected, especially when effects were small, this is not a strong endorsement. Its performance makes it little wonder that this method is so widely criticized (e.g., Simonsohn et al., 2014).

While the results for PET and PEESE were generally better, they were still far from good. PET is especially problematic. It performs well when there is a nil effect, but worse than randomly guessing. The far bigger problem is that it returns nil far too often, even when effect sizes are large. For example, when the true effect is g=0.62, PET returns a nil effect 12% of the time; it returns a nil 33% of the time when the true effect is g=0.32. Most critically, when there is a true effect of about the same magnitude as one unbiased estimate of the ego depletion literature (Tuk et al., 2015; g=0.16), PET returns a nil 53% of the time. It thus routinely underestimates true effect sizes, and is thus biased toward Type II error. For this reason, we cannot recommend PET be used under any circumstance. While it provides the best estimate when the true effect is zero (other than randomly guessing, that is), because zero is returned so often, it makes for a very insensitive tool. For this reason, estimates that combine PET and PEESE by applying PEESE conditional on PET not returning a zero (Carter & McCullough, 2014) are also insensitive and should be avoided.

While none of the estimates was particularly good, Top10, PEESE, and Trim and Fill performed adequately *under certain—but by no means all—conditions*. While not as accurate as PET (or randomly guessing), Top10 performed adequately when the true effect was zero (see Figure 1a), with bias and 95% coverage being acceptable. PEESE, here, misses the mark badly. PEESE returned a value that is consistent with zero only 17% of the time even when the simulated true effect size was, in fact, zero. Its 95% coverage of the true effect was also poor. Only Trim and Fill (and the uncorrected estimate, of course) performed worse than PEESE when the effect was truly zero. When the effect was large (see Figure 1d), Top10 was the most sensitive, with the best coverage, and acceptable bias, although it was not particularly precise. PEESE also performed well in terms of sensitivity, bias, coverage, and precision. When there is no true effect or when the effect is large, Top10 performed best. PEESE was good when effects were large, but not when it was nil. But, again, because meta-analysts have no idea of the size of the true effect (i.e., this is why they are conducting the meta-analysis to begin with), knowing which technique to use is fraught with difficulty. This is a recurring theme.

Things were even less clear with small and moderate effects (see Figures 1b and 1c). As mentioned above, with small effects (g=.16), Trim and Fill was the most sensitive at differentiating real from nil results, with relatively precise and unbiased estimates; however, its 95% coverage was poor. PEESE had only acceptable coverage and it was somewhat precise, although it did return inappropriate nil results 3% of the time. With moderate effects (g=.36), in contrast, PEESE was quite good, returning results that were sensitive, precise, relatively unbiased, and with good coverage. Here, Trim and Fill was also good, being somewhat sensitive, precise, and relatively unbiased, and with decent coverage. PEESE is clearly preferable with moderate effects, but Trim and Fill (and to a lesser extent Top10) is acceptable, at least in comparison to the other poorly-performing techniques to which it was compared.

We found our simulations eye opening. None of the corrections performed well in all or even most conditions. This is a real problem because it suggests that meta-analysts need to know something about the true state of their data in order to decide which correction to apply. This is difficult, of course, because meta-analysts don’t know this a priori. This general problem aside, we draw a few conclusions. First, we do not think PET (or anything conditional on PET) should be used to correct for bias. It over-corrects, and too often returns zero when true effects exist, sometimes even large effects. This is important when evaluating Carter and McCullough’s (2014) original bias-corrected meta-analysis, which relied on PET (but not PEESE) to conclude that their “results do not support the claim that the depletion effect is meaningfully different from zero” (p. 7). Second, PEESE, Top10, and (to our surprise) Trim and Fill might be decent, but not excellent, all-purpose corrections. PEESE was good for medium and large effects, acceptable for small effects, but woeful for nils. Top10 was good for nils and large effects, decent for medium effects, but ineffective for small effects. Trim and Fill was good for small and large effects, middling for medium effects, and atrocious for nils.

Based on our simulations, PET is not well-suited for estimating effect sizes under realistic conditions of social psychology research. While PEESE is considerably better, it too left a lot to desire. Top10 and Trim and Fill were also better all-around than PET, but had noted weaknesses as well. Our findings are consistent with Reed and colleagues (2015) who tested a number of meta analytic estimators (including PET and PEESE) under a range of conditions and found that no one estimator was generally superior. In some cases they found that the meta-analytic estimators “that do not explicitly correct for publication bias perform as well, if not better, than those that do” (Reed et al., 2015; p. 19). While the bias correction techniques in our own simulations always returned estimates that were superior to the uncorrected estimates, it is notable that we are not the only ones who find these new meta-analytic techniques wanting. Even though we tested these new meta-analytic estimators in a range of conditions that are found in social psychology—and especially in the conditions analyzed by Carter and colleagues—testing even more conditions (e.g., ranges of heterogeneity and publication bias) was outside of the scope of the current commentary. It is thus possible that the estimators would perform better or worse in still different conditions. Nevertheless, we draw the same conclusions as Reed and colleagues (2015), finding no one estimator being superior to others. Thus, despite confident reports that these meta-regression techniques can reliably detect and estimate meta-analytic effects in the presence of publication bias (Carter et al., 2015), there is lack of consensus among statisticians on this matter, who have stated that “a general conclusion remains elusive for now” (Reed et al., 2015, p. 4).

What is the real size of the ego depletion effect?

Given the overall poor performance of all of the bias correction techniques, what conclusions should we draw about the size of the ego depletion effect? Some of the techniques were acceptable under some conditions and not others, but not one of them showed general superiority; so, what conclusions should we draw? We follow Reed and colleagues (2015), who drew similar conclusions about the use of PET and PEESE, and think the safest route is to present a variety of meta-analytic estimators, not just one or two, and to consider the range of possible effects these techniques suggest. Carter and colleagues (2015) do an admirable job of presenting several techniques and examining their results from many angles; however, they rely almost exclusively on PET and PEESE—which we have found to be lacking under certain key conditions—to conclude here and elsewhere that their analyses do not support the claim that the depletion effect is different from zero (Carter & McCullough, 2014).

Instead, we suggest it is more prudent to present a range of possible effect sizes and to consider them all as being plausible to differing degrees. Our own analysis suggests that, at least under some conditions, PEESE, Trim and Fill, and Top10 might perform adequately (if not particularly well); in contrast, we cannot recommend PET because it performs poorly in almost all conditions. Carter and colleagues' (2015) analysis thus offers three possible effect sizes for the overall ego depletion effect: (1) the estimate derived from PEESE (g = 0 [-0.14, 0.15]); (2) the estimate derived from Trim and Fill (g = 0.24 [0.13, 0.34]); and (3) the estimate derived from Top10, which we calculated from Carter and colleagues' own data (g = 0.26 [.07, .44]).

As we can see, the estimates of the true effect fluctuate, and do not converge on a reasonable range of estimates. This is most unsatisfying for those, like us, who want to know the real size of the ego depletion effect. That said, we should note that the estimate favored by Carter and colleagues (2015) is discrepant from the other two estimates: While Trim and Fill and Top10 return effect sizes that are non-zero and positive, PEESE returned a correction that was not distinguishable from zero. The problem, for us, is that these three estimates do not converge, precluding us from saying anything meaningful about the size of the ego depletion effect; it might be meaningfully different from zero, or it might not.

This is not the first time that multiple meta-analytic bias-correction techniques contradict one another. Just recently, multiple meta-analytic corrections were applied to the religious priming literature, with PET-PEESE suggesting an effect that was not different from zero, but with Bayesian bias correction suggesting an effect that was, in fact, different from zero (van Elk et al., 2015). The lack of convergence when using multiple bias-correction techniques is problematic, and we agree with the authors of the religious priming bias-corrected meta-analysis that these results "demonstrate that meta-analytic techniques alone may not be sufficiently robust to firmly establish the presence or absence of an effect (van Elk et al., 2015, p.1).

What is really needed is a high-powered and good-faith attempt to replicate the basic effect with large samples to estimate the true size of the ego depletion effect.

Registered Replication Report. Fortunately, this sort of multi-lab pre-registered replication effort has just been conducted, and to our surprise it too revealed an effect size that was no different than zero (Hagger et al., in press). One might be tempted to look at this report as a vindication of the PET-PEESE methods advocated by Carter and colleagues (2015). And on the face of it, this seems like strong validation of these new techniques. We believe such a conclusion is premature, as even faulty methods can make correct predictions from time to time.

The registered replication, involving 24 different labs, (Hagger et al., in press) uncovered a meta-analytic effect that was no different than zero (d=0.02 [-0.09, 0.13]). While the details of the replication are beyond the scope of this commentary, we note that this multi-lab effort focused on a response interference task that shares many features with the Stroop task (something called the multi-source interference task; Bush & Shin, 2006), and in fact used this task as the dependent variable. The use of a Stroop-like task is relevant here because Carter and colleagues (2015) provided meta-analytic estimates for that subset of the ego depletion literature that used the Stroop as a dependent variable, and we think it is instructive to compare these estimates and see how they compare to the results of the multi-lab replication. Even though the uncorrected meta-analytic effect for the Stroop task was significant (g = 0.24 [0.07, 0.41]), all three of what we suggest are the best (or least bad) bias-correction estimators suggest something else altogether: PEESE (g = -.07 [-0.24, 0.11]), Trim and Fil (g = 0.11 [-0.07, 0.29]), and Top10 (g = 0.06 [-0.02, 0.14]) converge to suggest that the effect of ego depletion on the Stroop task is not meaningfully different from zero.

While we urge caution in over-interpreting meta-analytic estimates based on such a small number of studies, in this case 16, here is a case where all three bias-correction estimates speak with the same voice. The registered replication (Hagger et al., in press) is thus consistent not just with PEESE, but also with Trim and Fill and Top10. One interpretation of this is that bias-correction techniques ought to be taken seriously when multiple corrections converge on the same range of estimates. However, when there is a lack of convergence among corrections, as is the case with the overall ego depletion literature, the way forward is less clear. How do we know what to conclude when some corrections indicate an effect is different from zero but when others indicate otherwise? In such a case, we echo previous claims that meta-analytic techniques alone are not sufficient to determine the presence or absence of an effect (van Elk et al., 2015).

Despite our belief that registered replications are the best way to resolve the issue, we also agree with Carter and colleagues (2015), when they wrote that: “If replication efforts focus only on a single combination of manipulation and outcome tasks, [as was the case with Hagger and colleagues (in press)], results from such efforts can only answer the question of whether the depletion effect exists when measured with those tasks” (p. 813). That is, while the multi-lab replication strongly suggests that ego depletion does not impact Stroop-like tests—thereby vindicating PEESE, but also Trim and Fill and Top10—it says less about the robustness of the overall effect than we’d like. These reservations notwithstanding, we maintain that this multi-lab replication is the best piece of evidence (so far) about the state of the ego depletion literature, and it should act as a wakeup call to ego depletion researchers (including ourselves) to design more robust studies or risk abandoning the phenomenon altogether.

Conclusion

We close by once again expressing our admiration for Evan Carter, Michael McCullough, and their colleagues (Carter & McCullough, 2014; Carter et al., 2015) for engaging with the ego depletion literature in such a comprehensive manner. Their detailed analyses make clear that meta-analyses are only as good as the studies that go into them. And when studies are filtered through the lens of publication bias and the questionable research practices that we now know are commonplace (Simmons et al., 2011), great efforts need to be taken to correct the record. There work also makes clear that the resource model is in need of revision, as two or three domains were not found to be depletable, even when using standard meta-analytic techniques. Our admiration, notwithstanding, our own simulations and analyses, as well as those of others (Reed et al., 2015; van Elk et al., 2015) suggest that bias-correction techniques can give divergent effect size estimates. Some meta-analytic techniques perform well in some conditions, but poorly in others, with the problem that meta-analysts have no idea what condition they are operating in a priori. For these reasons, bias-corrected meta-analysis alone cannot resolve whether an effect is present or not.

References

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*, 1252–1265. doi:10.1037/0022-3514.74.5.1252

Baumeister, R. F., & Tierney, J. (2011). *Willpower: Rediscovering the Greatest Human Strength* (p. 304). New York, NY: Penguin Press HC. Retrieved from https://sivers.org/book/Willpower

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The Strength Model of Self-Control. *Current Directions in Psychological Science*, *16*(6), 351–355. doi:10.1111/j.1467-8721.2007.00534.x

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. New York, NY: John Wiley & Sons.

Bush, G., & Shin, L. M. (2006). The Multi-Source Interference Task: an fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nature Protocols*, *1*(1), 308–313. doi:10.1038/nprot.2006.48

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.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*(July), 1–11. doi:10.3389/fpsyg.2014.00823

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

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. doi:10.1111/j.0006-341x.2000.00455.x

Evans, D. R., Boggero, I. a., & Segerstrom, S. C. (in press). The Nature of Self-Regulatory Fatigue and "Ego Depletion": Lessons From Physical Fatigue. *Personality and Social Psychology Review*. doi:10.1177/1088868315597841

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

Hagger, M. S., & Chatzisarantis, N. L. D. (2014). It is premature to regard the ego-depletion effect as "Too Incredible". *Frontiers in Psychology*, *5*, 298. doi:10.3389/fpsyg.2014.00298

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Birt, A., Brand, R., … Cannon, T. (in press). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological Bulletin*, *136*(4), 495–525. doi:10.1037/a0019486

Inzlicht, M., & Berkman, E. (2015). Six Questions for the Resource Model of Control (and Some Answers). *Social and Personality Psychology Compass*, 1–14. doi:10.1111/spc3.12200

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, *18*(3), 127–33. doi:10.1016/j.tics.2013.12.009

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology. General*, *143*(1), 131–41. doi:10.1037/a0031048

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PloS One*, *9*(9), e105825. doi:10.1371/journal.pone.0105825

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and Brain Sciences*, *36*(6), 661–79. doi:10.1017/S0140525X12003196

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. doi:10.1186/1471-2288-9-2

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. *BMJ (Clinical Research Ed.)*, *339*, b2981. doi:10.1136/bmj.b2981

Muraven, M., & Baumeister, R. F. (2000). Self-Regulation and Depletion of Limited Resources : Does Self-Control Resemble a Muscle ?, *126*(2), 247–259.

Reed, W. R., Florax, R. J. G. M., & Poot, J. (2015). A Monte Carlo Analysis of Alternative Meta-Analysis Estimators in the Presence of Publication Bias. *Economics Discussion Papers*, *2015-9*. Retrieved from http://www.economics-ejournal.org/economics/discussionpapers/2015-9

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. doi:10.1037/1089-2680.7.4.331

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–66. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*, *9*(6), 666–681. doi:10.1177/1745691614553988

Stanley, T. D. (2008). Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection. *Oxford Bulletin of Economics and Statistics*, *70*, 103–127. doi:10.1111/j.1468-0084.2007.00487.x

Stanley, T. D., & Doucouliagos, H. (2013). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*, 60–78. doi:10.1002/jrsm.1095

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*(1), 70–77. doi:10.1198/tast.2009.08205

Tuk, M. A., Zhang, K., & Sweldens, S. (2015). The Propagation of Self-Control: Self-Control in One Domain Simultaneously Improves Self-Control in Other Domains. *Journal of Experimental Psychology: General*. doi:10.1037/xge0000065

Van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies. *Psychological Methods*, *20*, 293–309. doi:http://dx.doi.org/10.1037/met0000025

Van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: a skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, 1365. doi:10.3389/fpsyg.2015.01365

Vohs, K. D., Baumeister, R. F., Joiner, T. E., & Rudd, M. D. (2000). Escaping the self consumes regulatory resources: A self-regulatory model of suicide. In *Suicide science: Expanding the boundaries.* (p. 33). Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2001-16049-004&site=ehost-live&scope=site

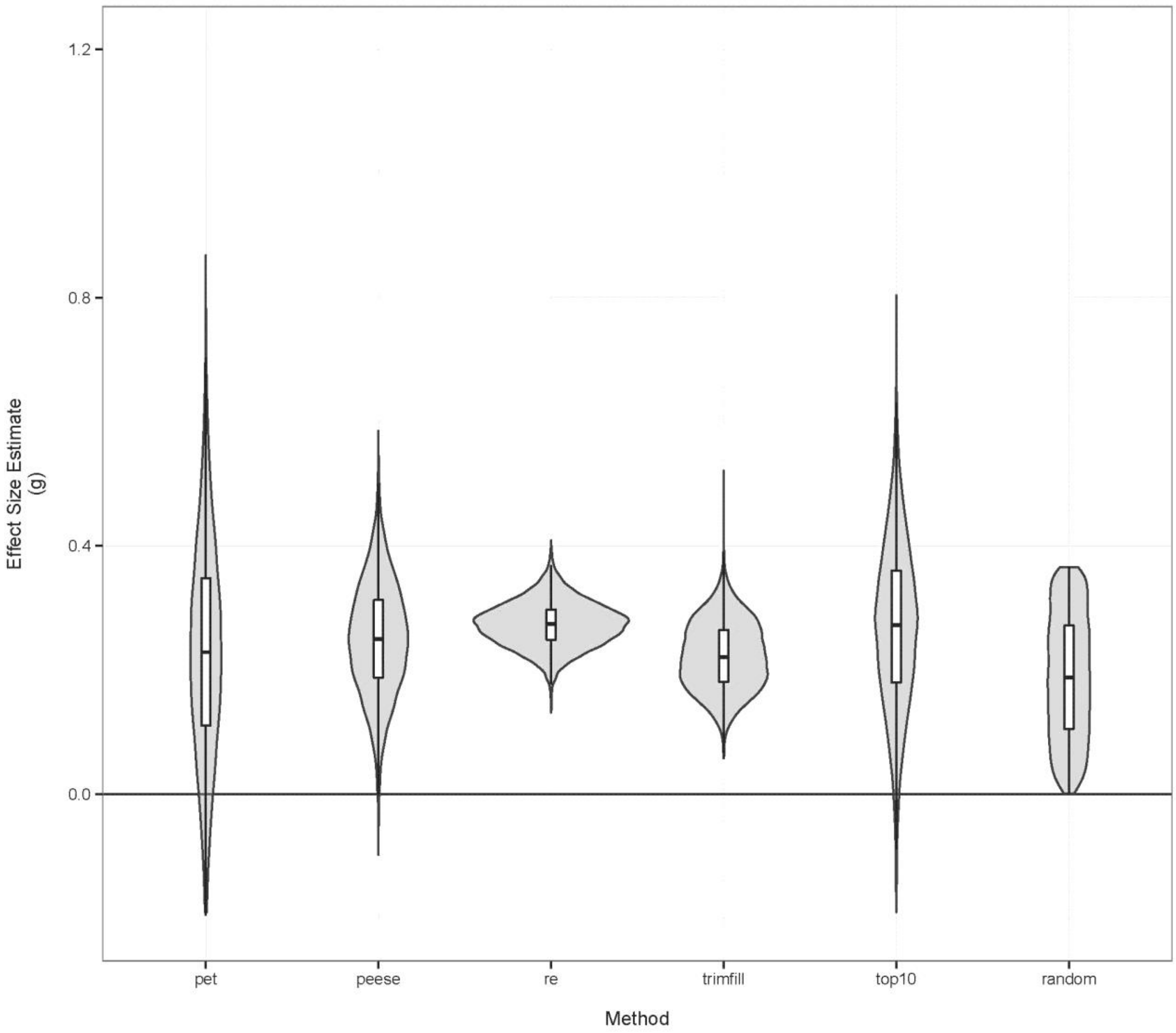
Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*.

Table 1: Results of 40,000 simulated meta-analyses after being corrected by various bias-correction techniques

	M estimate	**CI width**	**Bias**	**Coverage**	**% zero**	**MSE**	**Sensitivity (d')**	**Bias (-C)**
g = 0								
PET	.23	.60	.23	.65	.65	.08	--	--
PEESE	.25	.31	.25	.17	.17	.07	--	--
RE	.27	.18	.22	.00	.00	.08	--	--
TrimFill	.22	.19	.22	.01	.01	.05	--	--
Top10	.27	.57	.27	.53	.53	.09	--	--
Random	.19	.50	.19	.67	.67	.05	--	--
g = .16								
PET	.27	.57	.11	.83	.53	.04	.31	-.23
PEESE	.32	.29	.16	.42	.03	.03	.93	1.42
RE	.38	.17	.22	.00	.00	.02	.00	3.89
TrimFill	.31	.18	.15	.16	.00	.02	1.56	3.11
Top10	.36	.53	.20	.65	.45	.06	.20	.03
Random	.19	.50	.03	.94	.67	.01	.00	-.44
g = .36								
PET	.35	.54	-.01	.87	.33	.03	.83	.03
PEESE	.43	.28	.07	.76	.00	.01	2.94	2.42
RE	.52	.16	.16	.02	.00	.03	.00	3.89
TrimFill	.43	.17	.09	.66	.00	.01	1.56	3.11
Top10	.49	.51	.13	.79	.07	.04	1.55	.70
Random	.18	.50	-.18	.70	.68	.04	-.03	-.45
g = .62								
PET	.47	.52	-.15	.70	.12	.06	1.56	.39
PEESE	.59	.27	-.03	.81	.00	.01	2.94	2.42
RE	.71	.16	.09	.37	.00	.01	.00	3.89
TrimFill	.62	.17	0	.67	.00	.01	1.56	3.11
Top10	.68	.51	.06	.84	.00	.03	3.97	1.91
Random	.18	.50	-.44	.10	.69	.20	-.06	-.47

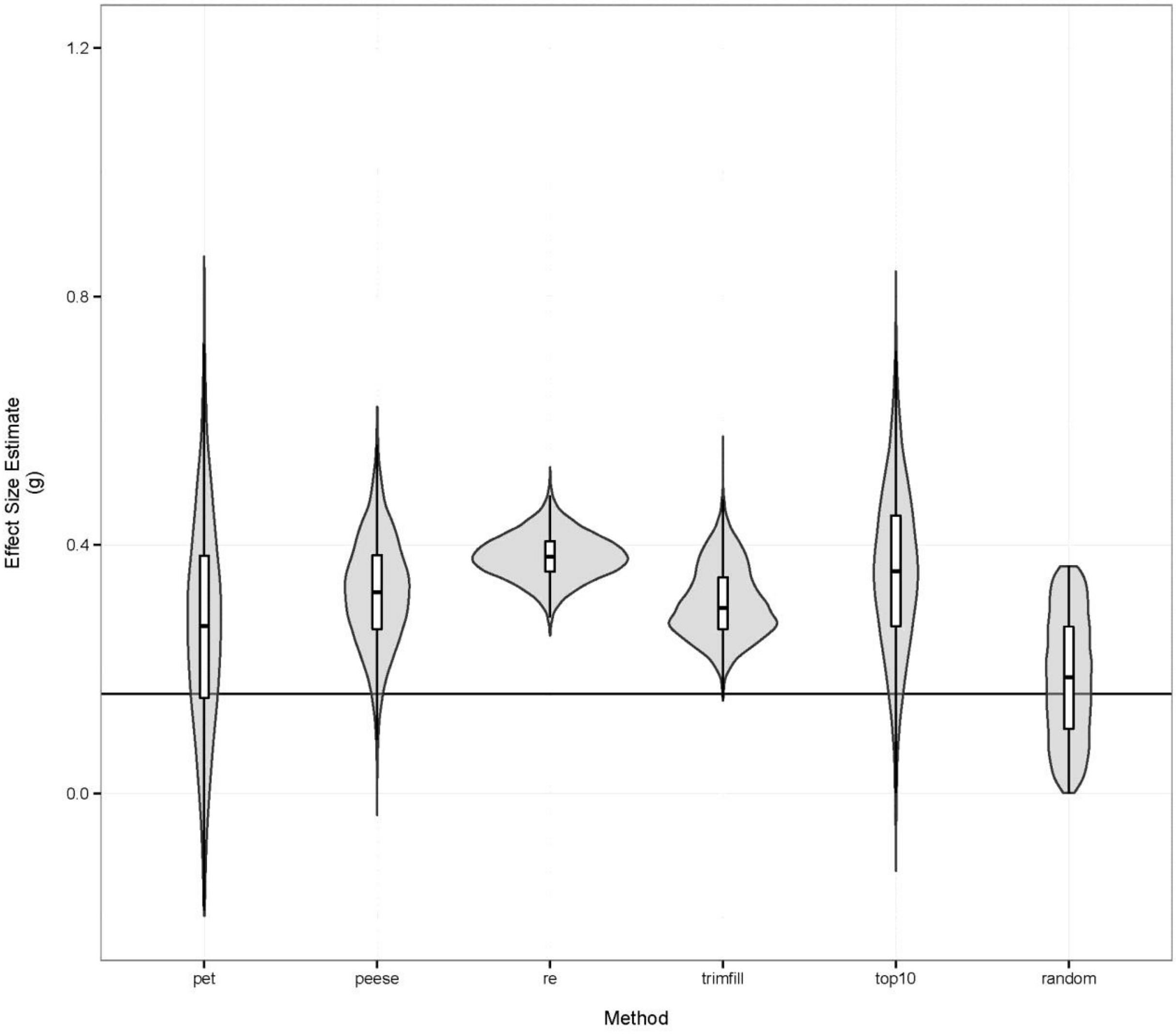
Note: PET = Precision Effect Test. PEESE =Precision Effect Estimation with Standard Error test. RE = Random Effect meta analysis. M Estimate = Mean estimate of the meta-analytic effect. CI Width = width (in percent) of the confidence interval around the mean. Bias = Bias in meta-analytic estimate. Coverage = Proportion of time (in %) the true value fell within the 95% confidence interval of the estimator. % zero = Percentage of times the estimated mean was zero. MSE = mean squared error of the meta-analytic estimate, a measure of precision. Sensitivity (d') = the sensitivity of the test to distinguish a real effect from a nil effect, using signal detection analysis. Bias (-C) = the extent that one response is more probable than another, using signal detection analysis.

Figure 1a. Violin plots depicting the effectiveness of various meta-analytic techniques to estimate the size of an effect, when the true effect (solid horizontal line) is nil (g=0).



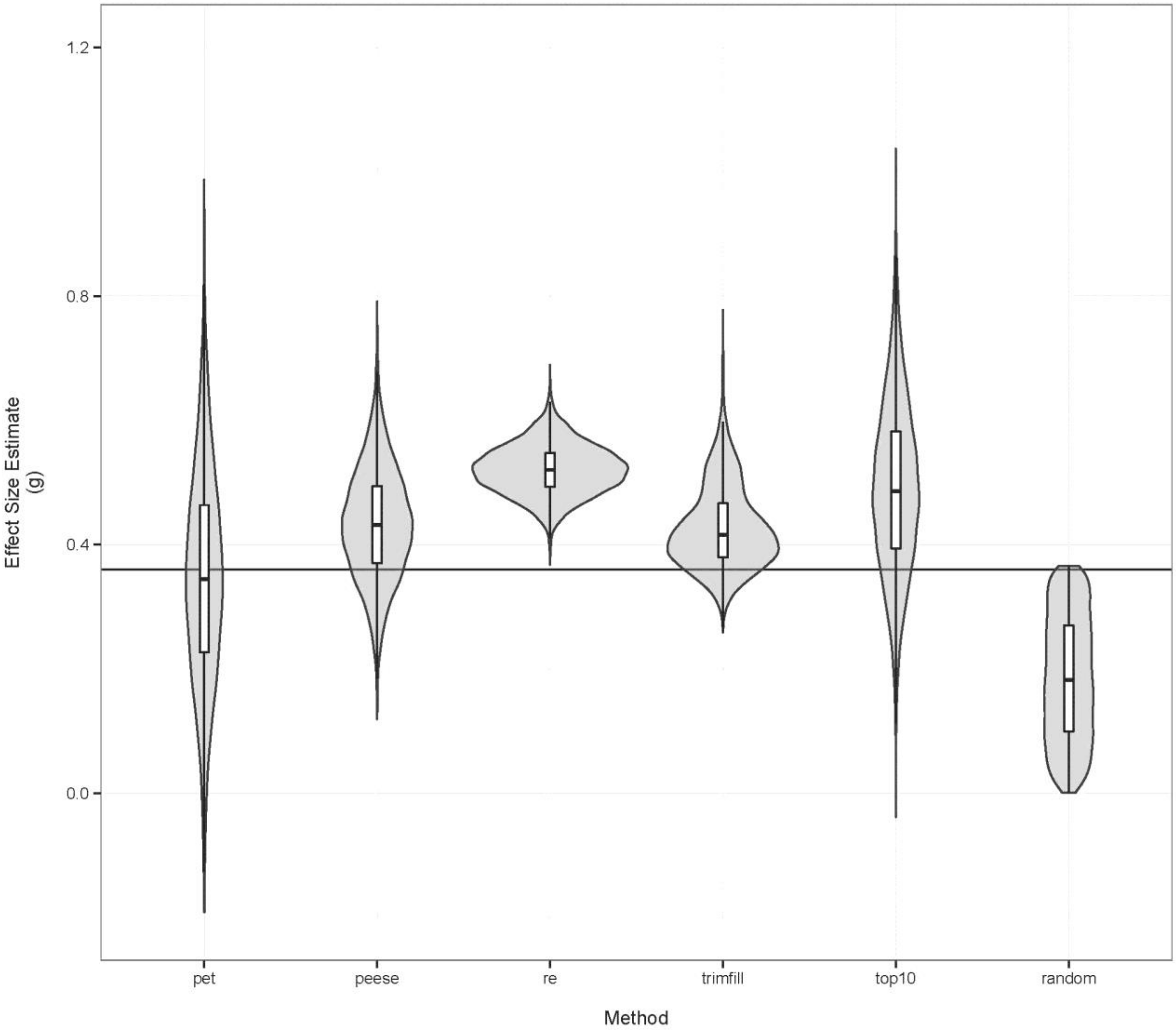
Note: The width of each plot corresponds to the frequency of estimates at that level. PET = Precision Effect Test. PEESE =Precision Effect Estimation with Standard Error test. RE = Random Effect meta-analysis. Trimfill = Trim and Fill. Top10 = Top 10 estimator. Random = Random correction.

Figure 1b. Violin plots depicting the effectiveness of various meta-analytic techniques to estimate the size of an effect, when the true effect (solid horizontal line) is small (g=.16).



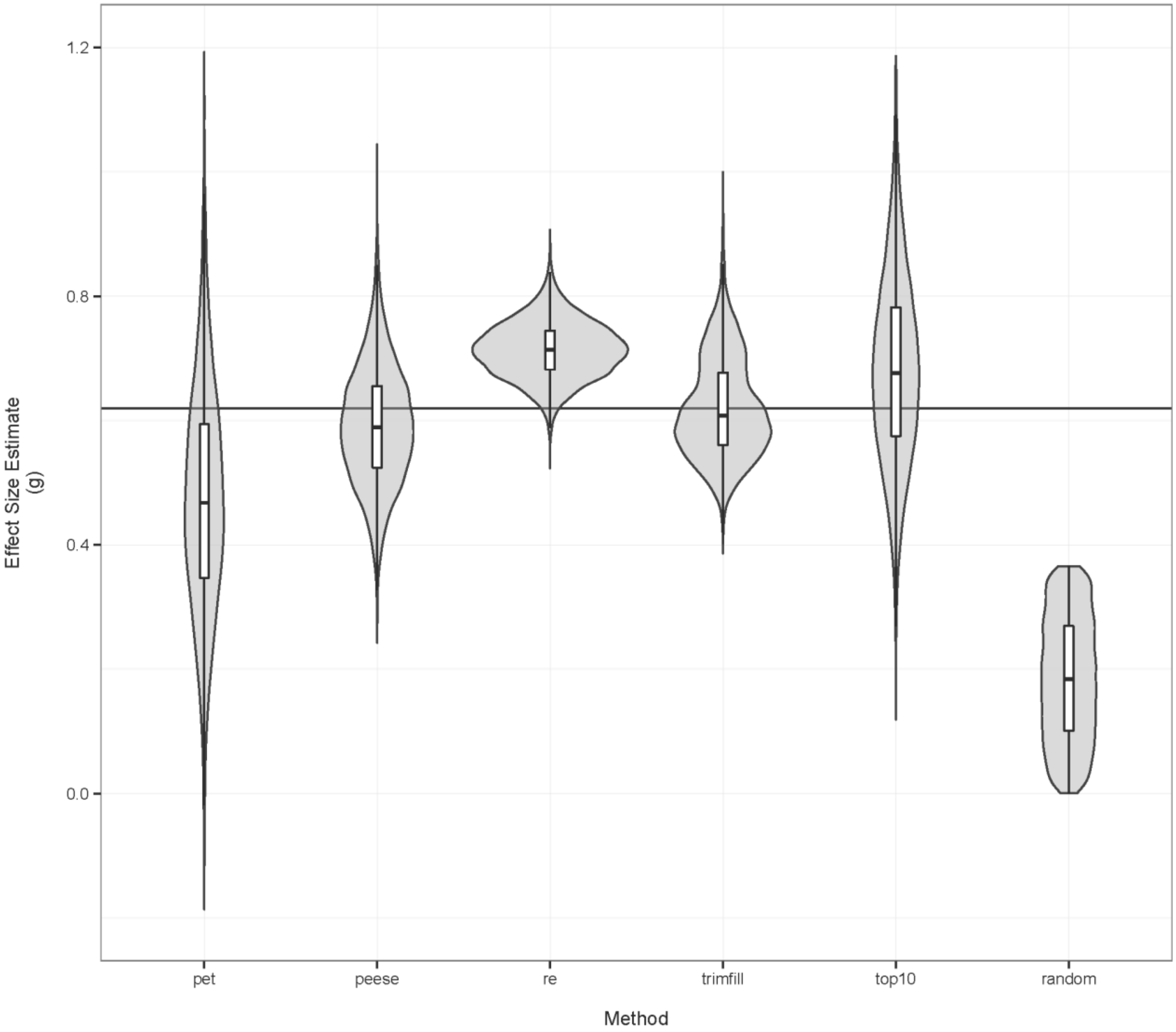
Note: The width of each plot corresponds to the frequency of estimates at that level. PET = Precision Effect Test. PEESE =Precision Effect Estimation with Standard Error test. RE = Random Effect meta-analysis. Trimfill = Trim and Fill. Top10 = Top 10 estimator. Random = Random correction.

Figure 1c. Violin plots depicting the effectiveness of various meta-analytic techniques to estimate the size of an effect, when the true effect (solid horizontal line) is moderate (g=.36).



Note: The width of each plot corresponds to the frequency of estimates at that level. PET = Precision Effect Test. PEESE =Precision Effect Estimation with Standard Error test. RE = Random Effect meta-analysis. Trimfill = Trim and Fill. Top10 = Top 10 estimator. Random = Random correction.

Figure 1d. Violin plots depicting the effectiveness of various meta-analytic techniques to estimate the size of an effect, when the true effect (solid horizontal line) is large (g=.62).


Effect Size Estimate (g)
pet
peese
re
trimfill
top10
random
Method