In Press, *Advances in Methods and Practices in Psychological Science* (*AMPPS*)

A simple, principled approach to combining evidence from meta-analysis and high-quality replications

Evan C. Carter[1,2], Michael E. McCullough[3]*

[1]Department of Neuroscience, University of Minnesota, 321 Church Street SE, Minneapolis, MN 55455

[2]United States Army Research Laboratory, Aberdeen Proving Ground, Aberdeen, MD 21005

[3]Department of Psychology, University of Miami, 1320 S. Dixie Highway, Coral Gables, FL 33146

*Correspondence to: Evan C. Carter (evan.c.carter@gmail.com).

Abstract

Recent discussions of the influence of publication bias and questionable research practices on psychological science have increased researchers' interest in both bias-correcting meta-analytic techniques and preregistered replication. Both approaches have their strengths: For example, meta-analyses can quantitatively characterize the full body of work being done in the field of interest and preregistered replications can be immune to bias. Both approaches also have clear weaknesses: Decisions about which meta-analytic estimates to interpret tend to be controversial and replications can be discounted for failing to fulfill key boundary conditions that are specified by theory to produce an effect. Using the experimental literature on the ego depletion phenomenon as a case study, we illustrate a principled approach to combining information from meta-analysis with subsequently conducted bias-free replications. This approach (1) compels researchers to explicate their beliefs in meta-analytic conclusions (and also, when controversy arises, to defend their warrants for embracing those beliefs), (2) encourages consideration of practical significance, and (3) facilitates the process of planning replications by specifying the sample sizes necessary to have a reasonable chance of changing the minds of other researchers.

Key Words: meta-analysis; replication; Bayesian estimation; publication bias; ego depletion.

A simple, principled approach to combining evidence from meta-analysis and high-quality replications

Valid inference in empirical psychology is threatened by data that are biased by questionable research practices (QRPs; Simmons, Nelson, & Simonsohn, 2011) and publication bias (Rothstein, Sutton, & Bornstein, 2006). To cope with these threats, researchers often use meta-analytic tools that can adjust for their distortionary effects. However, many bias-correcting estimators exist, and they often yield different conclusions. There is currently no consensus as to which is most appropriate in any given circumstance (e.g., Moreno et al., 2009; Reed et al., 2015; McShane et al., 2016; van Aert et al., 2016; Stanley, 2017; Carter, Schonbrodt, Hilgard, & Gervais, 2017).

Some have argued instead that inference should be based primarily on preregistered replications (van Elk, et al., 2015), which are uncorrupted by QRPs or publication bias *ipso facto*. Any individual preregistered replication, however, requires a specific operationalization of a general claim and as a result, replication efforts cannot easily reflect the potentially important methodological heterogeneity in how studies within any given area are executed. Moreover, discounting all non-preregistered experiments (possibly numbering in the hundreds) from consideration during scientific inference almost surely underestimates their potential evidential value: Just because a literature is biased does not mean that it must yield hopelessly misleading conclusions (van Elk, et al., 2015). How can scientists optimally combine what they believe based on previous (possibly biased) findings with new and less biased findings from preregistered replications? Here, we describe a simple, principled approach to answering this question and illustrate its application to research on the topic of ego depletion. Note that

this method is not designed to necessarily improve statistical estimation or hypothesis testing—instead, we intend it as a tool to make discourse more principled and thereby facilitate our field's ability to accumulate and apply knowledge. We provide R scripts and a tutorial as supplemental material so researchers can apply our method to their own areas of interest.

Our approach should be thought of as a computational process model (Lee, in press; Lee & Wagenmakers, 2013) of how real-world scientists should change their beliefs about hypothesized states of the world as new data come to light (Okasha, 2016). Broadly, this model involves using *M* meta-analytic estimates of an effect to specify *M* prior beliefs that researchers could hold regarding the effect size of a hypothesized phenomenon. These beliefs are then updated based on the results from subsequent high-quality replications using Bayesian meta-analysis (Heck, Gronau, & Wagenmakers, 2017). The result is a set of *M posterior beliefs*, which indicate how initial beliefs formed from existing meta-analytic work have changed in the face of new evidence (i.e., the replications).

Our model assumes that researchers' beliefs can be described as probability distributions over candidate values for the true effect, θ. When a value of θ is given high probability, this indicates that it is relatively more plausible—or credible—to the researcher. Additionally, our model holds that researchers change their minds in a mathematically correct fashion—that is, on the basis of Bayes' rule, which specifies that distributions of prior belief, $p(\theta)$, be updated in the face of data, D, to form distributions of posterior belief, $p(\theta \mid D)$. By reflecting on his or her own beliefs, a researcher can study how his or her mind would change in the face of new data, and therefore, use this

model as a means of performing Bayesian inference about the true magnitude of a given phenomenon. However, as mentioned, our approach is not so much a statistical advance as a conceptual and discursive one inasmuch as it encourages researchers to explicate and justify their beliefs and then to update them rationally (i.e., according to Bayes' rule). It additionally encourages researchers to examine the practical significance of what they believe—after seeing new data from high-quality replications—to be plausible values of the effect of interest. Beyond these applications, our model also allows us to study the beliefs of hypothetical researchers, and therefore, to (1) understand the impact of a particular set of replication results and (2) provide guidance for planning preregistered replications by answering the question, "how much data do we need to collect to change someone else's mind?" Notably, although our approach is not designed to provide statistical advantages, the aforementioned conceptual advantages are accompanied by the typical advantages to inference offered by a Bayesian approach—a topic beyond our current scope, but thoroughly covered by others (e.g., Lee & Wagenmakers, 2013; Kruschke, 2013; Kruschke & Liddell, 2017; Wagenmakers et al., 2017; Lee, in press).

Case study: Ego depletion

The method we describe is general and can applied to any area of research in which a meta-analysis has been conducted and high-quality replication data have been collected. For illustrative purposes, here we apply the method to data on the ego depletion effect, which are often marshaled to support the inference that an act of self-control impairs people's subsequent self-control (a state known as ego depletion; Baumeister, Bratslavsky, Muraven, & Tice, 1998).

Hagger et al. (2010) published the first meta-analysis on this effect and concluded that the depletion effect was, on average, medium or large in magnitude and robust across various experimental methods. Because of our own failure to replicate the depletion effect in a large sample (Carter & McCullough, 2013), we re-analyzed Hagger et al.'s (2010) data set with an eye toward detecting and correcting for publication bias and other small-study effects (Carter & McCullough, 2014). Our analysis indicated that controlling for such effects potentially reduced the depletion effect to the point where it was indistinguishable from zero.

Importantly, one could argue that the Hagger et al. (2010) data set includes inappropriate tests of the depletion effect. We found four reasons that this might be the case: (1) Hagger et al. (2010) included an experiment if its authors claimed it tested depletion, rather than creating an a priori definition of the depletion effect; (2) some experiments' putative measures of self-control were so non-specific as to provide support for the depletion effect regardless of the outcome; (3) Hagger et al. (2010) included many experiments that started with the premise that the depletion effect was real and then sought to use the effect to test whether a given task required self-control; and (4) Hagger et al. (2010) included only published experiments (Carter, Kofler, Forster, & McCullough, 2015). If it is true that the Hagger et al. (2010) data set includes inappropriate tests of the depletion effect, then the validity of Hagger et al.'s (2010) conclusions, as well as the conclusions we drew based on our re-analysis of Hagger et al.'s data (Carter & McCullough, 2014), are undermined. Therefore, we collected a data set following new inclusion criteria with the aim of providing a test of the depletion effect that could convince even a skeptical audience. Based on the resulting data set of 116

effect estimates, we again concluded that there was little convincing evidence that the true magnitude of the depletion effect differed meaningfully from zero on average (Carter et al., 2015).

Shortly thereafter, a large-scale preregistered replication report corroborated our conclusions with data from 23 separate replication attempts (Hagger et al., 2016), as did one additional preregistered replication (Lurquin et al., 2016), and four (see below) experiments that Tuk, Zhang, and Sweldens (2015) claimed to reflect their laboratory's entire "file drawer" of behavioral tests of the depletion effect (Tuk, Zhang, & Sweldens, 2015). Compared to the data examined in Hagger et al.'s (2010) meta-analysis and our later meta-analytic effort (Carter et al., 2015), one can be far more certain that these most recent 28 results permit higher-quality inference by virtue of their relatively low likelihood of bias due to QRPs or publication bias. (Hereafter, we refer to these data as the HQ28.) But how should researchers integrate the previous meta-analytic data with these new findings? Should the new findings represent the final word on the depletion effect? To answer such questions, we apply and analyze our model of how meta-analyses and replication data combine to inform what researchers believe.

Disclosure

All analyses were conducted in R (R Core Team, 2016). Data, scripts, and a tutorial are available online as supplementary material. We declare that we have no conflicts of interest with respect to the authorship or the publication of this article. ECC and MEM developed the ideas for this project, ran the analyses, and wrote the manuscript.

Method

Combining evidence from meta-analysis and high-quality replications

The first step in our approach is to define a set of *M* prior beliefs based on previous meta-analytic estimates. The published literature (Carter & McCullough, 2014; Carter et al., 2015) includes eight relevant results from four estimators (i.e., random-effects meta-analysis, trim-and-fill, PET, PEESE) applied to both the Hagger et al. (2010) data set and our updated data set (*M* = 8 = 2 data sets * 4 estimators). Priors take the form of probability distributions over the average true value of the depletion effect, μ, so a prior based on the random-effects meta-analysis estimate (d = 0.43, se = 0.05) from our data set (Carter et al., 2015) would be translated to a prior of the form $p(\mu) = N(0.43, 0.05)$. In principle, it is possible to add a new estimate by applying any version of any estimator to any relevant data set; however, doing so requires justifying such choices. We focus only on this set of *M* = 8 estimates, not because we think these estimates are unequivocally the most valid (see, e.g., Moreno et al., 2009; Reed et al., 2015; McShane et al., 2016; van Aert et al., 2016; Stanley, 2017; Carter et al., 2017), but because of the possibility that real world researchers may have formed their beliefs about the depletion effect on the basis of these previously published and relatively well-circulated estimates. Said slightly differently, the *M* prior beliefs need not be designed to represent best practices in meta-analytic estimation—indeed, we think the use of estimators like random-effects meta-analysis or trim-and-fill is naïve in this case; however, there are certainly researchers who would disagree with us on this point, and the purpose of our approach is to be able to explore the implications of such a position.

The second step in our approach is to define a set of replication data that one believes to be of appropriate quality. In other words, just as our approach allows researchers freedom to define their prior beliefs in the form of *M* meta-analytic

estimates, it also allows researchers to define the replication studies they find to be most convincing[1]. In our view, the HQ28 we have described above are of sufficient quality, but it would have easily been possible to make a different choice. For example, Tuk et al. (2015) in fact reported a total of nine experiments. Here, we include only those experiments that made use of behavioral tasks to manipulate and measure self-control—an inclusion criterion we also used in our own meta-analysis (Carter et al., 2015). Applying this inclusion criterion resulted in the omission of data from the five of Tuk et al.'s nine experiments that measured participants' forecasts of their own self-control in responses to hypothetical vignettes.

In the following, we briefly describe each of the four estimators we applied to the Hagger et al. (2010) and Carter et al. (2015) data sets and compare the associated estimates to the HQ28. We then detail how we combine information from these estimators with the HQ28 through Bayesian model averaged meta-analysis (Scheibehenne, Gronau, Jamil, & Wagenmakers, 2017).

Random-effects meta-analysis. This method is based on the assumption that the i^{th} study estimates a specific true effect, δ_i, which represents an observation drawn from a distribution of true effects centered on μ (Cooper et al., 2009) with a standard deviation of τ. By modeling a distribution of true effects, random-effects meta-analysis allows for the possibility that different studies will provide different results due to the influence of moderator variables—that is, factors like the location of the study, the population sampled, or the study-specific operationalization may all modify the *true* value of effect of interest. This approach provides an estimate of μ—the average of the

distribution of true effects, and importantly, it makes no correction for the possibility of bias.

Trim-and-fill. Based on the logic that a scatter plot of effect size against some proxy for sample size—often called a funnel plot—will have a predictable form of symmetry in the absence of publication bias (Duval & Tweedie, 2000), the trim-and-fill method first removes observations from the funnel plot until a criterion for symmetry is met, and then fills the “trimmed” observations back into the plot along with new imputed observations of the opposite sign, thereby taking an informed guess at the nature of the data that were potentially removed due to publication bias. Other meta-analytic methods, such as random-effects meta-analysis, can then be used to summarize this new expanded data set.

Precision effect test (PET). This approach assumes that the influence of small-study effects (for example, publication bias) is revealed by the relationship between the study-level effect size estimates and their standard errors. The intercept of a weighted-least squares meta-regression where effect sizes are predicted by their standard errors provides an estimate of the effect that controls for small-study effects—that is, an effect size estimate net the potential influence of publication bias (Stanley & Doucouliagos, 2014).

Precision effect estimate with standard error (PEESE). This method is conceptually and methodologically similar to PET, but it models the influence of small-study effects as a non-linear effect: Rather than regressing effect sizes on their standard errors, one regresses effect sizes on their variances (i.e., standard error squared).

Figure 1 compares the eight meta-analytic estimates to the HQ28. One might argue that the area of overlap of the 95% confidence intervals from the HQ28 (blue vertical lines in Figure 1) accords best with our results from the PEESE estimator; however, we cannot know which estimator is the best for these particular data sets because the performance of any depends very much on the unknown processes that generated the data (Moreno et al., 2009; Reed et al., 2015; Stanley, 2017; Carter et al., 2017). Moreover, there is no objective truth regarding which meta-analytic data set most appropriately represents tests of the depletion effect: Cunningham and Baumeister (2016), for example, recently speculated that our inclusion criteria (Carter et al., 2015)—namely, our decision to exclude experiments that did not measure self-control via an objective measure of cognitive performance or some other behavioral measure, and our decision to actively seek out unpublished experiments—systematically removed evidence for the depletion effect by removing published data from expert researchers and including unpublished work by inexperienced researchers, respectively. Therefore, instead of choosing a single estimate from a single data set—and thereby making many untestable assumptions—our approach allows us to study how these beliefs change in the face of the HQ28.

To model how prior beliefs change in the face of the HQ28, we will implement the Bayesian model averaged meta-analytic approach proposed by Scheibehenne, et al., 2017. In this approach, two Bayesian meta-analytic models are applied to the data. The first is the random-effects model as described above. The second is the fixed-effect model, which differs from the random-effects model in that it assumes there is only one true effect, μ, and each study-specific estimate differs from

this value only because of sampling error (i.e., it does not allow for the influence of moderators). Both models are realistic when considering a meta-analysis of high-quality replications: On one hand, one might assume that the fixed-effect model is the most appropriate for a set of exact replication estimates, such as those from Hagger et al.'s (2016) registered replication report. On the other hand, there are good reasons to think that individual studies may measure different true effects, even when they are carefully conducted exact replications (Klein et al., 2014; Hagger et al., 2016), and therefore, one should prefer the random-effects model. The choice between these two models is further complicated by other issues (e.g., Rice, Higgins, & Lumley, 2017), and for these reasons, Scheibehenne et al. (2017) have proposed a Bayesian model averaging approach that creates a combined estimate that is influenced by both models in proportion to how likely they are given the observed data[2].

Applying both the random-effects and fixed-effect models requires first describing prior belief distributions over the parameters in the models. For both the random-effects and fixed-effect models we set $p(\mu)$ as a normal distribution with mean and standard deviation set equal to a given point estimate and standard error from the estimates shown in Figure 1. For example, $p(\mu)$ = N(0.00, 0.07) when based on the PEESE estimate (d = 0.00, se = 0.07) from our data set (Carter et al., 2015). The random-effects model includes a second parameter, τ, for which we set a prior p(τ) as following a half-Student-t distribution with scale and degrees of freedom set to 1. The prior over τ, therefore, allows for positive values only and is right-skewed, much as appears to be the case in psychology in general (van Erp, Verhagen, Grasman, & Wagenmakers, 2017).

Our goal is to derive a posterior distribution, $p(\theta|D)$, where θ represents an arbitrary set of parameters of interest (e.g., μ and τ) and D represents the data. From Bayes' rule, we know that

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)},$$

where $p(\theta)$ is the prior based on previous meta-analytic results, $p(D \mid \theta)$ is the likelihood (the standard normal likelihood in this case), and $p(D)$ is the marginal likelihood of the data. Instead of deriving the posterior analytically, which can be problematic in many cases, we rely on recent advances in Markov chain Monte Carlo techniques to approximate the posterior through Gibbs Sampling as implemented with JAGS software (Plummer, 2015) through R (R Core Team, 2016) with the r2jags package (Su & Yajima, 2015).

In the case of Bayesian model averaged meta-analysis, $p(\theta \mid D)$ is the weighted sum of the posteriors from each of the L models considered, where the weights are given as the posterior model probabilities, $p(M \mid D)$ (i.e., the relative plausibility of a model M given the data):

$$p(\theta|D) = \sum_{i=1}^{L} p(\theta|M_i, D)p(M_i|D),$$

where

$$p(M_i|D) = \frac{p(D|M_i)p(M_i)}{\sum_{j=1}^{L} p(D|M_j)p(M_j)}.$$

In the above, $p(D|M_i)$ and $p(M_i)$ are the marginal likelihood and the prior probability of the i^{th} model, respectively. In our application, $L = 2$ (the random-effects and fixed-effect meta-analysis models with priors as described above), and both models are taken to be equally likely *a priori.* We derived the marginal likelihoods using bridge sampling

techniques (Meng & Wong, 1996; Gronau, et al., 2017) as implemented with the bridgesampling package in R (Gronau & Singmann, 2017). Furthermore, the Bayesian model-averaged marginal posterior—p(μ|Data)—is of primary interest, as it represents how, in the face of the HQ28 and for a given set of prior beliefs, one ought to distribute belief over a range of candidate values for the average true magnitude of the depletion effect, μ.

Assessing practical significance

Bayesian inference, as we have described it, lends itself easily to assessing the practical significance of an effect through a focus on effect size estimates. Practical significance can be operationalized in this context by setting a Region of Practical Equivalence (ROPE) in the posterior wherein the true effect is considered practically equivalent to zero. Furthermore, our approach allows researchers to assess how *any* idiosyncratic definition of a ROPE compares to the most credible estimates of the effect of interest as represented by the posterior distributions. To us, a true effect in the ROPE of $-0.15 < \mu < 0.15$ should be considered practically non-significant for the depletion effect. Furthermore, the limited strength model clearly involves a directional hypothesis, so we set our ROPE as $\mu < 0.15$.

Just as setting prior beliefs is subjective, so too is defining a ROPE, and just as with setting prior beliefs, it is necessary to explain one's reasoning behind a given ROPE. We came to our ROPE of $\mu < 0.15$ in part on the basis of the estimated real world influence of $\mu = 0.15$: We translated $\mu = 0.15$ from a standardized mean difference into more directly interpretable metrics. For example, in their third experiment, Vohs and Heatherton (2000) tested whether depletion increased ice cream consumption. Based

on the pooled standard deviation in that experiment, $\mu = 0.15$ implies depleted participants consume 15g more ice cream—something like one to two more bites. Similarly, Muraven, Collins, and Neinhaus (2002) studied whether depletion increased beer consumption. There, an effect of $\mu = 0.15$ corresponds to 38.4ml additional beer consumption—about 2.6 tablespoons. If the depletion effect were concretely related to more devastating outcomes such as death or disease instead of consumption of ice cream and beer, this analysis would need to account for that by modifying the definition of practical significance; however, we judge that effects such as 15g more ice cream or 38.4ml of additional beer following a depleting episode would ultimately be of little consequence in the real world.

An additional reason to consider $\mu = 0.15$ as practically non-significant is that achieving 80% power for such an effect in a simple two-group comparison requires 699 participants *per experimental condition*—about five times more participants than in the largest experiment *in total* we included in our previous meta-analysis (Carter et al., 2015). Importantly, research on the depletion effect frequently takes the form of testing whether some moderator reduces the depletion effect, such as glucose ingestion (Gailliot et al., 2007), or subjects' belief in the nature of willpower (Job, Dweck, & Walton, 2010). For such attenuation interaction hypotheses, achieving 80% power requires *twice* the number of subjects *per experimental condition* than the simple two-group design does (Simonsohn, 2011). Therefore, if the magnitude of the depletion effect is $\mu = 0.15$, the total sample size required to test whether drinking lemonade replenishes self-control following a depletion manipulation, for example, at 80% power is $N = 699 * 4 * 2 = 5592$. This sample size, which is for a single study, is fully 85% of

the *total* number of subjects in our meta-analytic sample (Carter et al., 2015). Moreover, $N = 5592$ is about *92 times* the sample size in one of the original, iconic tests of this form (Experiment 7 in Gailliot et al., 2007). We believe such sample sizes are prohibitively large: If the true magnitude of the depletion effect were $\mu = 0.15$, researchers would only very rarely have the resources to detect the phenomenon with an acceptable level of statistical power, a fact that should raise doubts about the practicality of studying the depletion effect.

Planning replication efforts that can change partisans' minds

Researchers can also use the approach we describe here to estimate the amount of data necessary to have a reasonable chance of shifting hypothetical skeptics' or proponents' prior beliefs. To do so, one would simulate registered replication reports where k teams of simulated researchers each produce an experiment with the target sample size of n subjects per cell in a standard two-group design. By applying the Bayesian model-averaged meta-analytic approach described above to these simulated data, one can track the proportion of times that a given goal is met (Kruschke, 2015). This value is akin to statistical power, but instead of the probability of rejecting a false null hypothesis, this quantity represents the probability that the simulated researcher either correctly finds only practically significant values of the effect to be plausible (i.e., the posterior remains outside of the ROPE) or correctly finds practically non-significant values to be plausible (i.e., the posterior overlaps with the ROPE).

Imagine that the beliefs of proponents of ego depletion can be modeled as we have done thus far: $p(\mu) = N(0.43, 0.05)$ based on the random-effects estimate from our data set (Carter et al., 2015). Similarly, the beliefs of skeptics are correctly specified

as $p(\mu)$ = N(0.00, 0.07) based on our PEESE estimate. Imagine further that both skeptics and proponents can agree on a ROPE of $\mu < 0.15$ as the range in which the depletion effect could be called practically non-significant and that both proponents and skeptics are willing to have their minds changed by the data from the replication effort being planned (a crucial assumption that cannot necessarily be taken for granted). For two possible realities of the true magnitude of the depletion effect in which it is either μ = 0.43 or μ = 0, one can then map the probability of correctly changing either a proponent's or a skeptic's mind. Here, we define changing a simulated researcher's mind as when values in the inner 95% of posterior distribution (what is sometimes called the highest density or highest credibility interval) is contained within the ROPE. In other words, a proponent or skeptic is now willing to consider practically non-significant values to be relatively plausible. We conducted a simulation of this situation across a range of the number of research teams conducting replication (k = {6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36}) and the number of subjects in each replication study (n = {20, 50, 80, 110, 140, 170, 200, 230, 260, 320}). Each unique combination of k and n was simulated 100 times and we assessed the number of times out of 100 that (1) a proponent—$p(\mu)$ = N(0.43, 0.05)—changed his or her mind when μ = 0 and (2) a skeptic—$p(\mu)$ = N(0.00, 0.07)—changed his or her mind when μ = 0.43.

Results

The eight sets of prior and posterior distributions are displayed in Figure 2. As is evident from the figure, one's prior beliefs about the magnitude of the depletion effect play a powerful role in how those beliefs are changed in the face of the HQ28. For example, if one believed that Hagger et al.'s (2010) data were the most valid and that no

correction for publication bias was necessary—and one therefore preferred the estimate d = 0.67 (se = 0.03)—the 28 recent high-quality replications averaging roughly d = 0 come nowhere near dissuading one from believing in the reality of the ego depletion phenomenon (see the distributions in the top panel drawn with dashed lines). This conclusion is clear from the fact that a value of zero (or any values in the ROPE) is far from the 95% HDI. Moreover, examining this 95% HDI, one can see that a hypothetical researcher who holds these prior beliefs would find a value of d = 0.65 to be the most credible estimate of the true magnitude of the depletion effect, with the range of the most credible estimates spanning d = 0.59 to d = 0.71. A similar conclusion is reached if one bases one's prior beliefs on the trim-and-fill estimate from the Hagger et al. (2010) data or the random-effects estimate from our data set (Carter et al., 2015). Alternatively, if one's prior beliefs accord with any of the five other prior distributions we specified, then the 28 subsequent replications compel one to believe that the true magnitude of the depletion effect is very likely either μ = 0 or practically equivalent to μ = 0 (as we have defined it). This is indicated by the fact that the 95% HDIs of these posterior distributions contain values from the ROPE (Figure 2).

Figure 3 indicates the amount of data required to change the minds of either proponents or skeptics when the beliefs they hold are incorrect. For example, proponents could aim for 24 replications with target sample sizes of 110 per cell (total N = 2,640) to reach 80% power for convincing skeptics—defined as holding prior beliefs such that $p(\mu)$ = N(0.00, 0.07)—when the true magnitude of the effect is μ = 0.43. Alternatively, if proponents who believed the true magnitude of the effect is μ = 0.43 wished to be able to change their own minds, Figure 3 indicates that they should aim to

collect, for example, 15 replications with target samples sizes of 170 per cell (total N = 2,550). As a comparison, Hagger et al.'s (2016) RRR included data from 23 labs with about 47 subjects per cell (total N = 2,141), which, although impressive, falls below our estimate of the sample size required to change the mind of a proponent.

We examined the sensitivity of our results to the definition of prior belief (not shown, but code, data, and figures are included in the supplement). Specifically, there is reason to think that our definition of a proponent's beliefs is too skeptical: For example, Cunningham and Baumeister (2016) clearly rejected the inclusion criteria used to collect the Carter et al. (2015) data set. If, instead, a proponent's beliefs are modeled as based on the random effects analysis of the Hagger et al (2010) data—$p(\mu) = N(0.67, 0.03)$—it becomes essentially impossible to change proponents' minds when they are incorrect. We simulated this possibility (Figure S1) and found that having 36 research teams run experiments with target sample sizes of n = 320 per group (total N = 23,040) resulted in 0% chance of a proponent correctly accepting the null hypothesis. This result likely stems from the fact that the distribution of prior beliefs associated with the random-effects estimate based on the Hagger et al. (2010) data is relatively precise (i.e., has a small standard deviation; see Figure 2). A prior distribution with higher precision represents more certain beliefs in that credibility is more concentrated over a smaller range of candidate values. Significantly shifting such a distribution will require much more contrary data than we simulated. This result, therefore, represents how difficult it can be to change strongly held beliefs.

Discussion

For psychology's foreseeable future, scientific inference will likely be stuck between controversial meta-analyses and preregistered replications that can realize only some of the conditions under which a phenomenon has been hypothesized to occur. The approach outlined here makes the best of this tricky situation by modeling how researchers combine their prior beliefs (which they could form on the basis of meta-analytic conclusions about a possibly biased literature) with higher-quality evidence from subsequent preregistered replications. Our approach forces both the proponents and critics of an idea to make their initial beliefs explicit and to defend the methods upon which they acquired those beliefs. Then it shows them how their beliefs should change when they are confronted by new data. It also provides guidance about the amount of new data that would need to be collected to change those beliefs when a replication effort is being planned.

The case of ego depletion illustrates how difficult it can be to resolve the difference between a skeptic's view of a phenomenon and a proponent's view. As a first step, such a resolution requires that researchers clearly specify their beliefs as well as the conditions under which they are willing to change those beliefs. Our meta-analysis (Carter et al., 2015) was designed to provide a test that would be stringent enough to convince a skeptical audience, so it is not surprising that it seemed overly strict to proponents of the effect and may have done little to change their minds (e.g., Cunningham & Baumeister, 2016). However, if one were willing to entertain some skepticism—for example, by forming one's prior beliefs based on estimates from the trim-and-fill, PET, or PEESE as applied to the to the Carter et al. (2015) data set—then

the HQ28 replication data would lead one to conclude that the depletion effect very plausibly $\mu \leq 0.15$, which we argue is practically non-significant. Of course, if a researcher believed that the HQ28 includes inappropriate replications, then he or she would be justified in changing his or her beliefs only after a new massive replication effort was produced. Thus, the way forward for ego depletion research, as for all research generally where inferences must be made on the basis of meta-analytic data and preregistered replications, will involve carefully considering what it will take to change the minds of researchers on both sides of this issue, and then applying adequate effort to do so.

Our approach highlights that rational people can remain unconvinced by dozens of preregistered experiments that question the reality of a highly studied phenomenon, although maintaining such a belief could (depending on which meta-analytic data set and estimator it depends) require them to defend rather optimistic assumptions about the amount of bias in the literature. Along similar lines, our approach can highlight when researchers' beliefs regarding an effect are so strong that it is practically impossible to collect enough data to change their minds. We are not sure if the debate about the depletion effect is at such an impasse, but it might be: If proponents' prior beliefs can be described in terms of the original Hagger et al. (2010) estimate—$p(\mu) = N(0.67, 0.03)$—but the true effect is actually described by our bias-corrected estimate (i.e., $\mu = 0.00$), then a replication effort on the order of 36 teams each collecting approximately 320 people per group (total $N = 23{,}040$) will have *no chance* of correctly changing such confident proponents' minds (Figure S1). This is an interesting problem facing psychological science that, to our knowledge, has received

little discussion. Indeed, it may be that becoming more explicit about what we each believe and how strongly we believe it—and what those beliefs imply about how a rational person should react to subsequent empirical evidence—could be extremely beneficial as we move forward from science's current crisis in confidence.

Footnotes

[1]Reader's might find this degree of freedom disconcerting. One should note, however, that Bayesian methods are not necessarily more flexible than frequentist methods. For example, when defining a maximum likelihood estimate or a *p*-value, one can pick from a variety of assumptions about the form of the likelihood distribution, the familywise error level, factors that determine the distribution of test statistics (e.g., how the data were sampled), and of course, the design of the data collection effort. A critical advantage of the Bayesian approach, however, is that these assumptions tend to be made more explicitly than in the frequentist approach (Wagenmakers et al., 2017). This can temper the abuse of flexibility in that unjustified and extreme choices (e.g., overly favorable prior distributions) will be easy to spot and dismiss.

[2]Any meta-analytic model, not just the fixed-effect and random-effects models, could be used in Bayesian model averaging for this second step. For our purposes, the only requirement is that the model be appropriate for combining high-quality replications. For example, this could include any meta-analytic model that does not explicitly correct for publication bias (e.g., Henmi & Copas, 2010; Baker & Jackson, 2013; Stanley & Doucouliagos, 2015; Schmid, 2017).

References

Baker, R. D., & Jackson, D. (2013). Meta-analysis inside and outside particle physics: two traditions that should converge? Research Synthesis Methods, 4(2), 109-124.

Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science*, *11*(4), 574-575.

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*(5), 1252.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97-111.

Carter, E. C., & McCullough, M. E. (2013). After a pair of self-control-intensive tasks, sucrose swishing improves subsequent working memory performance. *BMC psychology*, *1*(1), 1.

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*, 823.

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*(4), 796.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2017, September 19). Correcting for bias in psychology: A comparison of meta-analytic methods. Retrieved from osf.io/rf3ys

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). The handbook of research synthesis and meta-analysis. New York, NY: Russell Sage Foundation.

Cunningham, M. R., & Baumeister, R. F. (2016). How to Make Nothing Out of Something: Analyses of the Impact of Study Sampling and Statistical Interpretation in Misleading Meta-Analytic Conclusions. *Frontiers in Psychology*, *7*, 1639.

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.

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., ... & Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: willpower is more than a metaphor. *Journal of personality and social psychology*, *92*(2), 325.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... & Steingroever, H. (2017). A tutorial on bridge sampling. arXiv preprint arXiv:1703.05984.

Gronau, Q. F., & Singmann, H. (2017). Bridge Sampling for Marginal Likelihoods and Bayes Factors. R package version 0.2-2.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., & Zwienenberg, M. (2016). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 2.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological bulletin*, *136*(4), 495.

Heck, D. W., Gronau, Q. F., & Wagenmakers, E.-J. (2017). metaBMA: Bayesian model averaging for random and fixed effects meta-analysis. Retrieved from https://github.com/danheck/metaBMA. doi:10.5281/zenodo.835494

Henmi, M., & Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. Statistics in medicine, 29(29), 2969-2983.

Job, V., Dweck, C. S., & Walton, G. M. (2010). Ego Depletion—Is It All in Your Head? Implicit Theories About Willpower Affect Self-Regulation. *Psychological Science*, *21*(11), 1686-1693.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. Social psychology.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General, 142(2), 573.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd ed.)*. San Diego, CA: Academic Press.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review, 1-29.

Lee, M.D. (in press). Bayesian methods in cognitive modeling. Submitted for The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition.

Lee, M.D., & Wagenmakers, E.-J. (2013). Bayesian Cognitive Modeling: A Practical Course. Cambridge University Press.

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., von Bastian, C. C., Carruth, N. P., & Miyake, A. (2016). No Evidence of the Ego-Depletion Effect

across Task Characteristics and Individual Differences: A Pre-Registered Study. *PloS one*, *11*(2), e0147770.

Meng, X. L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica, 831-860.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. Perspectives on Psychological Science, 11(5), 730-749.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. BMC medical research methodology, 9(1), 2.

Muraven, M., Collins, R. L., & Neinhaus, K. (2002). Self-control and alcohol restraint: an initial application of the self-control strength model.*Psychology of Addictive Behaviors*, *16*(2), 113.

Okasha, S. (2016). Philosophy of science: A very short introduction (2nd ed.). Oxford, UK: Oxford University Press.

Plummer, M. (2015). JAGS Version 4.0. 0 user manual. See https://sourceforge. net/projects/mcmc-jags/files/Manuals/4. x.

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Reed, W. R., Florax, R. J., & Poot, J. (2015). *A Monte Carlo analysis of alternative meta-analysis estimators in the presence of publication bias* (No. 2015-9). Economics Discussion Papers.

Rice, K., Higgins, J., & Lumley, T. (2017). A re-evaluation of fixed effect (s) meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society).

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.

Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Manuscript submitted for publication.

Schmid, C. H. (2017). Heterogeneity: multiplicative, additive or both?. Research synthesis methods, 8(1), 119-120.

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*, 0956797611417632.

Simonsohn, U. (2011, March). No-Way Interactions. Retrieved from http://datacolada.org/17

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. Social Psychological and Personality Science, 1948550617693062.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60-78.

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: weighted least squares meta-analysis. Statistics in medicine, 34(13), 2116-2127.

Su, Y. S., & Yajima, M. (2015). R2jags: A Package for Running jags from R. R package version 0.5-7.

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*, *144*(3), 639.

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. Perspectives on Psychological Science, 11(5), 713-729.Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. Proceedings of the National Academy of Sciences, 201521897.

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*.

Van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017, August 18). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in Psychological Bulletin From 1990-2013. Retrieved from osf.io/4xnm3

Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *Psychological science*, *11*(3), 249-254.

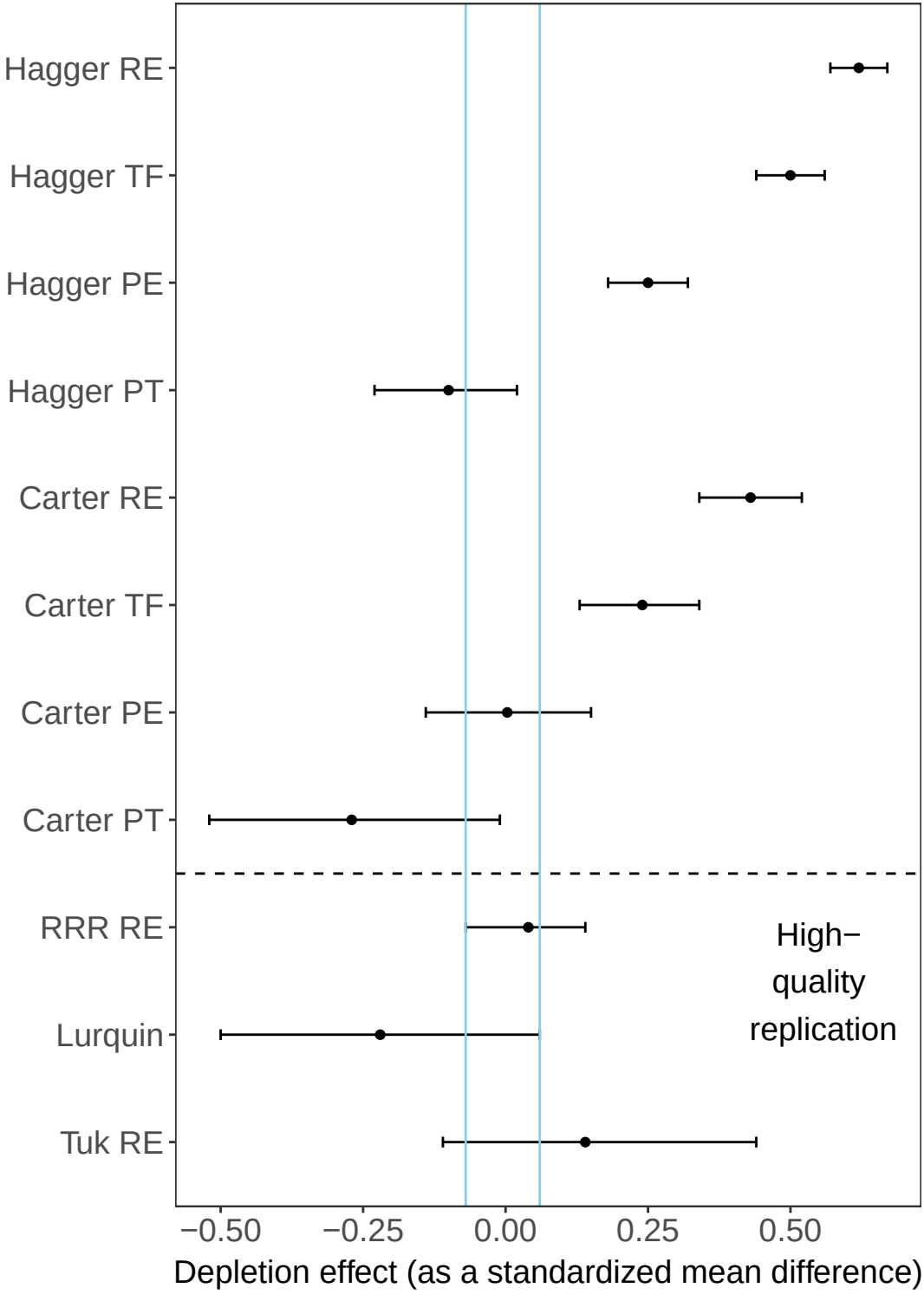
Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 1-23.

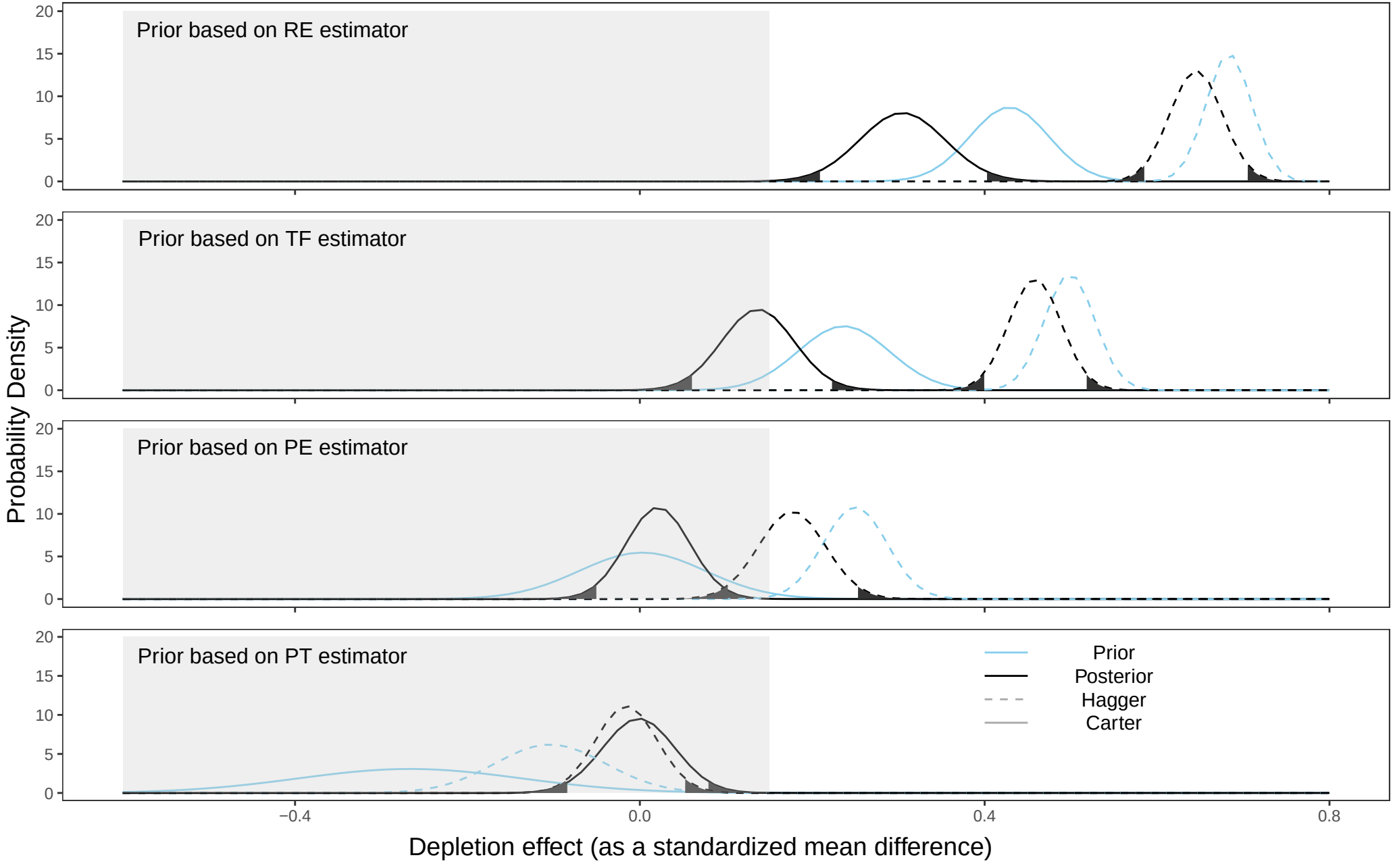
Figure Captions

Figure 1. Meta-analytic estimates of the depletion effect based on Hagger et al.'s (2010) and Carter et al.'s (2015) data (above the dashed line) and recent, high-quality replications (below the dashed line). Dots represent meta-analytic point estimates; whiskers represent the 95% confidence intervals. Vertical blue lines represent the region of overlap for the 95% confidence intervals for the high-quality replication data. RE = random-effects meta-analysis; TF = the trim-and-fill; PE = PEESE (precision effect estimate with standard error); PT = PET (precision effect test).

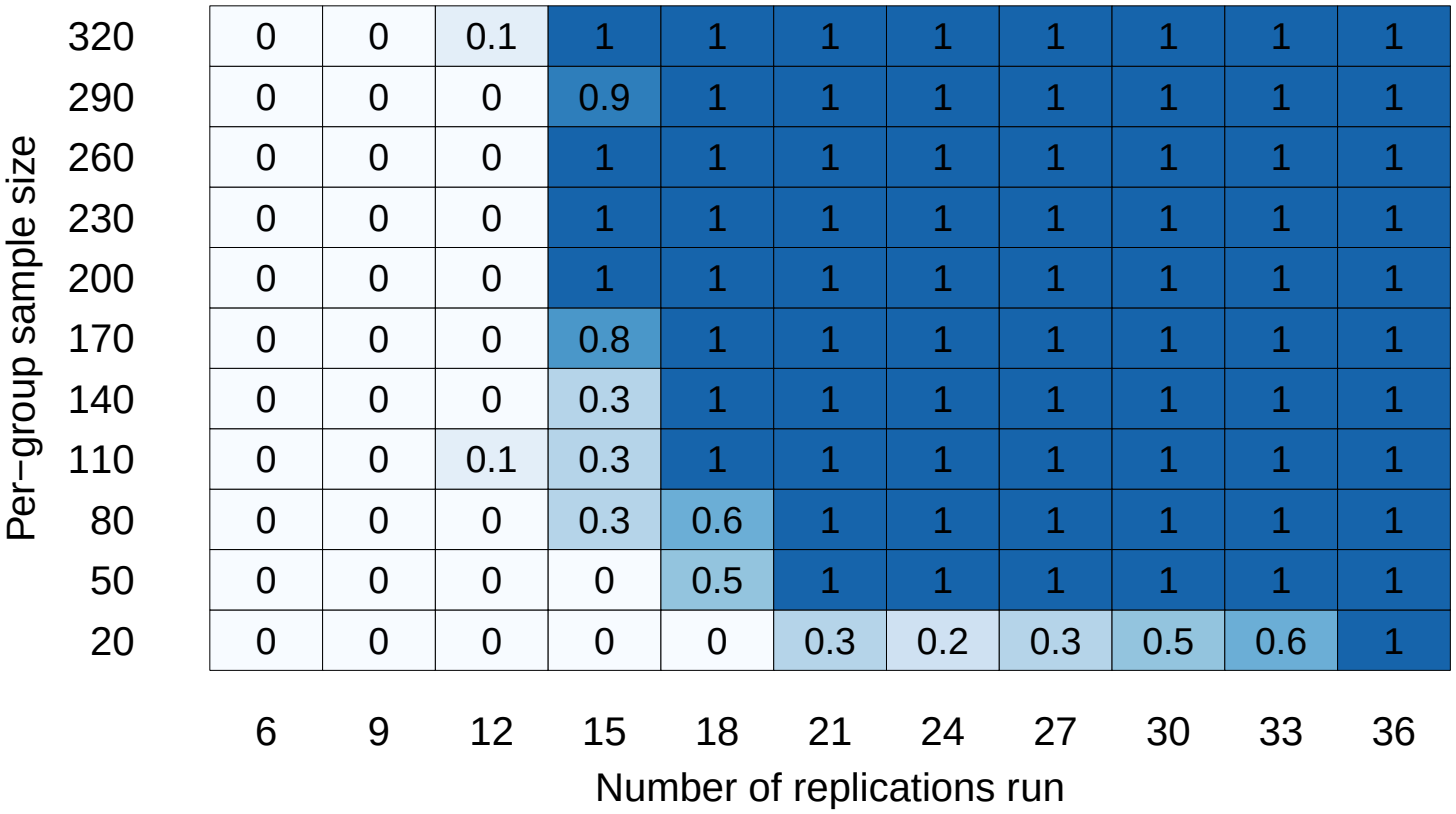
Figure 2. Posterior distributions are based on a Bayesian model averaged meta-analysis applied to the HQ28 data where distributions of prior belief are specified from previous meta-analytic estimates from either the Hagger et al. (2010) or Carter et al. (2015) data. Unshaded areas of the posterior distributions represent the 95% Highest Density Interval (HDI). RE = random-effects meta-analysis; TF = the trim-and-fill; PE = PEESE (precision effect estimate with standard error); PT = PET (precision effect test).

Figure 3. (A) The probability of convincing a proponent that he or she is wrong—that is, observing overlap between the 95% HDI and the ROPE = $\mu < 0.15$ given that $\mu = 0.00$ and that the proponent's prior belief is represented by $\mu \sim N(0.43, 0.05)$. **(B)** The probability of convincing a skeptic that he or she is wrong—that is, observing overlap between the 95% HDI and the ROPE = $\mu < 0.15$ given that $\mu = 0.43$ and that the skeptic's prior belief is represented by $\mu \sim N(0.00, 0.07)$. This probability is given as a function of *k*—the number of replications conducted, and *n*—the per-group sample size in each simulated replication.


Hagger RE
Hagger TF
Hagger PE
Hagger PT
Carter RE
Carter TF
Carter PE
Carter PT
RRR RE
Lurquin
Tuk RE
High−
quality
replication
−0.50
−0.25
0.00
0.25
0.50
Depletion effect (as a standardized mean difference)


Prior based on RE estimator
Prior based on TF estimator
Prior based on PE estimator
Prior based on PT estimator
Probability Density
Depletion effect (as a standardized mean difference)
Prior
Posterior
Hagger
Carter
0
5
10
15
20
−0.4
0.0
0.4
0.8

(A) Probability of convincing a proponent given μ = 0 & ROPE: μ< 0.15



(B) Probability of convincing a skeptic given μ = 0.43 & ROPE: μ< 0.15