

A Series of Meta-Analytic Tests of the Depletion Effect: Self-Control Does Not Seem to Rely on a Limited Resource

Evan C. Carter

University of Miami and University of Minnesota

Lilly M. Kofler

University of Miami and University of Chicago

Daniel E. Forster and Michael E. McCullough

University of Miami

Failures of self-control are thought to underlie various important behaviors (e.g., addiction, violence, obesity, poor academic achievement). The modern conceptualization of self-control failure has been heavily influenced by the idea that self-control functions as if it relied upon a limited physiological or cognitive resource. This view of self-control has inspired hundreds of experiments designed to test the prediction that acts of self-control are more likely to fail when they follow previous acts of self-control (the *depletion effect*). Here, we evaluated the empirical evidence for this effect with a series of focused, meta-analytic tests that address the limitations in prior appraisals of the evidence. We find very little evidence that the depletion effect is a real phenomenon, at least when assessed with the methods most frequently used in the laboratory. Our results strongly challenge the idea that self-control functions as if it relies on a limited psychological or physical resource.

Keywords: ego depletion, self-control, self-regulation, meta-analysis, publication bias

Supplemental materials: <http://dx.doi.org/10.1037/xge0000083.supp>

From subcellular processes to trophic interactions, every adaptive biological process depends on the management of energy. The *limited strength model* of self-control (Baumeister, Bratslavsky, Muraven, & Tice, 1998) is a scientifically and popularly acclaimed application of this truth to the study of brain and behavior. The primary assumption of the limited strength model is that self-control (the process by which responses to thoughts and emotions are directed to serve higher order goals) relies on a limited physiological or cognitive resource, and thus, fails as the resource is depleted. This assumption leads to the prediction that acts of self-control following previous acts of self-control will be less successful (the *depletion effect*; Baumeister et al., 1998). Scientists' interest in the depletion effect has resulted in over 200 published experiments (Hagger, Wood, Stiff, & Chatzisarantis, 2010), many of which appear to support the conclusion that behaviors like racism (Muraven, 2008), violence (Stucke &

Baumeister, 2006), risk taking (Freeman & Muraven, 2010), and addiction (Christiansen, Cole, & Field, 2012), are caused by depleted self-control.

Based on a 2010 meta-analysis of 198 published experiments, Hagger, Wood, Stiff, and Chatzisarantis (2010) concluded that the depletion effect is real, robust to experimental context, and, in terms of a standardized mean difference (i.e., Cohen's *d*), of medium-to-large magnitude: $d = 0.62$ (95% CI [0.57, 0.67]). Here, we present a series of meta-analyses that (a) test the depletion effect with data from both published and unpublished experiments, (b) are based on improved inclusion criteria, and (c) employ cutting-edge statistical techniques. Our results revealed signals of *small-study effects* (when larger samples produce smaller effect size estimates; Egger, Davey Smith, Schneider, & Minder, 1997), which can indicate a distortion of meta-analytic estimates due to *publication bias* (when counterfactual, statistically nonsignificant results are less likely to be included in a meta-analysis; Franco, Malhotra, & Simonovits, 2014). After accounting for small-study effects, we found only scant evidence that the depletion effect is distinguishable from zero.

Therefore, despite hundreds of apparently supportive tests, the available meta-analytic evidence does not allow one to conclude that the depletion effect—as commonly operationalized—is a real behavioral phenomenon.

Concerns About the Appropriateness of Previous Meta-Analytic Efforts

To test the depletion effect, researchers typically use the *sequential task paradigm* (Baumeister et al., 1998), during which participants complete at least two tasks that are thought to require

This article was published Online First June 15, 2015.

Evan C. Carter, Department of Psychology, University of Miami and Department of Ecology, Evolution and Behavior, University of Minnesota; Lilly M. Kofler, Department of Psychology, University of Miami and Division of the Social Sciences, University of Chicago; Daniel E. Forster and Michael E. McCullough, Department of Psychology, University of Miami.

This work was funded by a grant from the John Templeton Foundation and from a fellowship awarded to Evan C. Carter by the National Science Foundation.

Correspondence concerning this article should be addressed to Evan C. Carter or Michael E. McCullough, Department of Psychology, University of Miami, 5665 Ponce De Leon Boulevard, Coral Gables, FL 33124-0751. E-mail: evan.c.carter@gmail.com or mikem@miami.edu

self-control. We refer to these two tasks as the manipulation task and the outcome task, respectively. Participants in the experimental condition complete a version of the manipulation task that ostensibly requires more self-control relative to the version completed by control participants. Following the manipulation task and any intermediate tasks (e.g., questionnaires), all participants complete the outcome task. The depletion effect is quantified as the mean difference in performance between the two groups on the outcome task. Hagger et al. (2010) meta-analyzed 198 experiments that used the sequential task paradigm; however, the goal of meta-analysis is to provide inferences about the underlying effect(s) tested by the universe of experiments for which the meta-analytic data set is an appropriate sample (Cooper, Hedges, & Valentine, 2009), and we find four reasons for concern that Hagger et al.'s (2010) meta-analytic sample is less-than-ideal for drawing inferences about the depletion effect.

First, Hagger et al. (2010) included all experiments using the sequential task paradigm, regardless of whether the manipulation or outcome tasks could be considered valid operationalizations of self-control. The range of tasks that have been used in the sequential task paradigm either to manipulate or to measure self-control is extremely broad—from throwing darts (Englert & Bertrams, 2012) to self-reported likelihood of cheating on a hypothetical romantic partner (Gailliot & Baumeister, 2007)—and the validity of such tasks as manipulations or measures of self-control is generally unknown. To the extent that tasks do not measure or manipulate self-control, experiments that use them cannot be said to test the depletion effect.

Second, Hagger et al.'s (2010) inclusion criteria resulted in the inclusion of experiments that used measures of self-control that were so weakly linked to theory that, regardless of results, findings could be interpreted as support for the depletion effect. In one study, for example, higher donations to charity were treated as evidence for depletion (Janssen, Fennis, Pruyn, & Vohs, 2008), whereas in another, fewer hours volunteered to help a stranger in need were treated as evidence for depletion (DeWall, Baumeister, Gailliot, & Maner, 2008). When experiments use measures of self-control that are divorced from theory to the extent that they do not allow for the falsification of the depletion effect, they cannot logically be considered tests of the depletion effect.

Third, Hagger et al. (2010) included what might be thought of as “extension experiments”—experiments that began from the premise that the depletion effect was real and then sought to examine the extent to which the depletion effect explained other phenomena (e.g., anxiety about death; Gailliot, Schmeichel, & Baumeister, 2006). Extension experiments of this sort are, by definition, not tests of the depletion effect.

Finally, Hagger et al. (2010) included only published experiments, which means that, in the presence of publication bias, their estimates are based on a particular form of unrepresentative sampling that profoundly exaggerates the estimate of a conjectured effect. Previously, we reanalyzed Hagger et al.'s (2010) data to assess and correct for small-study effects (Carter & McCullough, 2013a; Carter & McCullough, 2014). We found compelling evidence that small-study effects biased their results, that these small-study effects were very likely due in part to publication bias, and that the resulting bias was extreme enough that the appearance of a robust, medium-to-large magnitude effect might have been spurious. Hagger and Chatzisarantis (2014) have since independently

verified our statistical conclusions, although they disagreed with our interpretation. For example, they remained skeptical that the degree of publication bias operating on the literature is strong enough to have inflated a truly null effect to the extent that they originally reported (Hagger et al., 2010). We return to this issue in the Discussion section.

The first three concerns listed above about the appropriateness of the Hagger et al. (2010) data set for making inferences about the depletion effect apply to any conclusions based on that data set, so that it is entirely possible that our previous conclusion—that the depletion effect has been severely overestimated due to small-study effects (Carter & McCullough, 2014)—is also invalid. It may be, for example, that small-study effects (e.g., publication bias, the undisclosed use of researcher degrees of freedom) only affected those experiments that did not truly manipulate and measure self-control, and therefore, our previous attempts at correcting for such influences masked true evidence for the depletion effect. Therefore, just as it is necessary to reassess Hagger et al.'s conclusion that the depletion effect is a real and robust phenomenon, it is also necessary to reassess our conclusion that the apparent effect simply reflects bias.

The Current Study

Given the concerns listed here, any conclusion based on the Hagger et al.'s (2010) data set would likely be unconvincing to a skeptical audience. Therefore, our goal for the current study was to address these problems, and thereby provide the most appropriate meta-analytic tests of the depletion effect possible. In other words, one might usefully consider our approach to differ from that of Hagger et al.'s (2010) in that it derives from a skeptical perspective of the limited strength model. The benefit of such an approach is that any inferences about the depletion effect (i.e., either that it is real and robust or that it is merely a statistical artifact) ought to be convincing, even to a skeptical audience.

To address the first three concerns described above, we included here only experiments that involved *both* frequently used manipulation tasks *and* frequently used outcome tasks (see Method section). This approach follows the logic that researchers tend to select tasks that seem to be the most valid operationalizations of self-control and that provide the most interpretable results. Additionally, because extension experiments include tasks that allow one to test whether the depletion effect applies to other constructs, such as death-related anxiety, our approach excludes these experiments. To minimize the fourth concern—overestimation of mean effect sizes due to publication bias—we searched for, retrieved, and included results from as many unpublished experiments as possible. Moreover, we applied both classic and more recently developed statistical techniques to assess and correct meta-analytic estimates for the influence of small-study effects such as publication bias (Duval & Tweedie, 2000; Ioannidis & Trikalinos, 2007; Stanley & Doucouliagos, 2014).

Rather than perform a single meta-analysis, we grouped effect size estimates based on the outcome tasks from which they were derived and performed a series of meta-analyses on these more methodologically homogeneous data sets. This strategy limited the methodological variability across the experiments being synthesized, a factor that might contribute to statistical heterogeneity in meta-analysis (Cooper et al., 2009), and enabled us to avoid the

assumption that all outcome tasks are commensurate measures of self-control (i.e., another means of addressing the first three concerns described above).

Method

Inclusion Criteria

For inclusion in our analyses, an effect size had to have resulted from a true experimental test with a behavioral outcome task. We omitted tests of the depletion effect that were correlational or quasi-experimental, as well as those that tested whether judgments, ratings, or responses to hypothetical situations or requests were affected by previous exertion of self-control. For experiments in which an individual difference variable was thought to moderate the depletion effect, only the main effect for the depletion manipulation was included. For experiments that included an experimental manipulation as a moderator (e.g., administration of glucose to half of the participants; DeWall et al., 2008), we followed Hagger et al. (2010) in only including the effect size derived from the level of the moderator not thought to attenuate the depletion effect. This approach to moderation by individual differences and by experimental manipulation was favored because our goal was to obtain estimates of the depletion effect from samples that were most comparable with the samples used in other experiments where no moderator was considered. These inclusion criteria were set prior to data collection and data analysis.

Additionally, only effect sizes derived from instances of the sequential task paradigm in which both the manipulation task and the outcome task were frequently used tasks were included. Following Hagger et al. (2010), we defined frequently used tasks as those that have been used in *at least 10* independent tests of the depletion effect (the 10 instances must have been either all as a manipulation or all as an outcome). This criterion was also chosen based on sample size recommendations for the statistical techniques we planned to apply (Sterne et al., 2011).

Data Collection

We conducted an exhaustive literature search using the following strategies: Searching online databases (i.e., EBSCO, ISI Web of Science, and Proquest), searching online lists of conference abstracts (i.e., lists for annual conferences for the Association for Psychological Science [APS] and the Society for Personality and Social Psychology [SPSP]), personal communications with experts in the field, and issuing several calls for unpublished data through the listserv of the Society for Personality and Social Psychology (SPSP). Overall, we individually contacted over 200 researchers with requests for unpublished data. Additionally, all studies that were included in Hagger et al. (2010) were examined for inclusion here.

The following online databases were searched: ISI Web of Science, EBSCO (including MEDLINE, PsycINFO, PsychARTICLES, PsychEXTRAS, ERIC), and ProQuest (including American Periodicals, Ethnic NewsWatch, FRANCIS, GenderWatch, PAIS, PILOTS, ProQuest Dissertations & Theses: History, ProQuest Dissertations & Theses: Social Sciences, ProQuest Research Library: Social Sciences, ProQuest Social Science Journals, ProQuest Sociology, Social Services Abstracts, Sociological Abstracts). Publication type was set to

articles, proceedings papers, reviews, and meeting abstracts for ISI Web of Science; periodicals, reviews, reports, and dissertations for EBSCO; and conference papers and proceedings, dissertations and theses, reports, and scholarly journals for ProQuest. Each search was limited to results in English that used human subjects and that were dated from 1998 to 2012.

Exact search terms were as follows (an asterisk indicates a truncated search word, which includes all versions of the word in the search; e.g., “deplet*” includes the words deplete, depletion, depleted, and depletes in the search): For ISI Web of Science, the full search term was (“Self Regulat*” or “Self Control” or “Impulse” or “Ego”) AND (“Resource” or “Deplet*” or “Perform*”). For EBSCO, the full search term was (“Self Regulat*” or “Self Control” or “Impulse” or “Ego”) AND (“Resource” or “Deplet*” or “Perform*”). And for ProQuest, the full search term was EXACT (“Self Control” or “self regulat*”) OR (“impuls*” or “ego”) AND (deplete* or resource* or perform*) AND CAU(Baumeister R). In the search term for ProQuest, the code CAU(Baumeister R) specifies that the search only return hits that cite an author with the last name Baumeister and first initial R. This option was only available for ProQuest, but reduced the total returned hits by several thousand.

Each search returned the following number of hits: 3,851 for ISI Web of Science, 7,889 for EBSCO, and 853 for ProQuest. These abstracts were then examined for general relevance. This resulted in 177 abstracts for ISI Web of Science, 132 abstracts for EBSCO, and 54 for ProQuest. With duplicates removed, this resulted in a combined total of 269 abstracts for which obtained full-text articles.¹ From these articles, there were 328 independent experiments within 141 articles that made use of the sequential task paradigm.

Conference programs for the annual meetings of Society for Personality and Social Psychology (SPSP) and Association for Psychological Science (APS) were obtained for each year between 2003 (the earliest available year) and 2011. Using the find function, the search term “deplet” returned 31 poster and symposium presentation abstracts from the APS Convention Programs and 149 from the SPSP Meeting Programs. The authors for each of these posters or presentations were sent an email request for information about methods, statistics, and any other unpublished data.

In December 2012, a second wave of data collection was conducted to keep the data set updated. This second wave was conducted in exactly the same way as the first, except that databases were searched from 2011 onward. From the online databases, each search returned the following number of hits: 1,209 for ISI Web of Science, 694 for EBSCO, and 72 for ProQuest. These abstracts were then examined for general relevance. This resulted in 90 abstracts for ISI Web of Science, 87 abstracts for EBSCO, and 14

¹ Eighteen of the abstracts deemed irrelevant were done so during a second phase of examination that was prompted by those abstracts being both unavailable through the University’s holdings and unobtainable through interlibrary loan. These abstracts were judged to very likely be irrelevant or unnecessary (e.g., data that we had located in another form, such as preliminary data that later led to a article that we had already located). Unfortunately, we did not keep detailed records of these abstracts, and so cannot report their exact nature or contact the original authors for more information. In principle, it is possible that some of these abstracts would have led to additional data; however, as mentioned, we judged this to be very unlikely.

for ProQuest. Removing duplicates yielded a total of 138 abstracts. From this list of abstracts, we obtained 133 full-text articles. From these articles, there were 83 independent experiments within 47 articles that made use of the sequential task paradigm.

At this time, we also searched conference programs for the annual meetings of Society for Personality and Social Psychology (SPSP) and Association for Psychological Science (APS) were searched for the years 2012–2013. Using the find function, the search term “deplet” returned 16 poster and symposium presentation abstracts from the APS Convention Programs and 54 from the SPSP Meeting Programs. The authors for each of these posters or presentations were sent the same email request for information.

In total, after adding experiments that were emailed to us to the set of experiments located via searching online databases, our search resulted in 620 individual instances of the sequential task paradigm. Each of these was then grouped by the type of manipulation task and the type of outcome task used. Following this grouping procedure, the data set was organized by manipulation task in ascending order of the number of times each task was used. Ten categories of manipulation task emerged as frequently occurring (i.e., appearing 10 or more times in the data set). These 10 categories comprised a total of 359 experiments. The 359 experiments that used a frequently used manipulation task were then organized by the type of outcome task used. The result was eight classes of outcome tasks that included 10 or more experiments. In total, our literature search produced 157 experiments that contained both frequently used manipulation tasks and frequently used outcome tasks. The categories of tasks are described in detail below.

Of the 157 experiments, 41 were excluded for analyses for one of three possible reasons: First, 15 experiments did not contain enough information to code (all authors had been contacted about the missing information, but at the time that analyses were conducted, no reply had been received). Second, 19 experiments included experiment-level moderators that did not have appropriate controls, and thus, no clear test of the depletion effect was available. And third, in seven additional experiments, the manipulation task was not used to manipulate use of self-control, but rather, as a means of inducing ego depletion in all the participants in the sample. Thus, the final sample was composed of 116 independent instances of the sequential task paradigm, two of which used both impossible anagrams and Stroop as outcome tasks, and could therefore be broken down into two dependent (i.e., derived from the same sample) effect sizes.

Compared with Hagger et al. (2010), our data collection efforts occurred more recently, included unpublished experiments, and used substantially different inclusion criteria. Therefore, it is unsurprising that our sample overlapped minimally with the set of experiments analyzed by Hagger et al. (2010): Only 28 of our 116 experiments (24.14%) were included in Hagger et al.’s (2010) data set. Moreover, of our 116 experiments, 48 (41.38%) were unpublished (in contrast to zero of the 198 in Hagger et al.’s (2010) sample), and 59 (50.86%) yielded statistically nonsignificant effects, (in comparison with 47 of the 198, or 23.74%, in Hagger et al.’s, 2010 data set).

Frequently Used Manipulation and Outcome Tasks

Manipulation tasks. The following 10 frequently used manipulation tasks were identified as described above (k is the num-

ber of effect sizes that made use of the corresponding manipulation task). (a) *Attention essay* ($k = 10$): Participants are asked to write about a topic (e.g., a recent vacation). Participants in the experimental condition are told that they cannot use some set of commonly occurring letters, usually a and n , while writing. Participants in the control condition are told that they cannot use uncommon letters, for example, q and z . (b) *Attention video* ($k = 19$): Participants watch a silent video during which stimuli occasionally appear. Participants in the control condition are given no instructions other than to watch the video, whereas participants in the experimental condition are told to ignore the stimuli when they appear. The video is usually of a woman being interviewed while words are displayed in the bottom right corner. (c) *Crossing out letters* ($k = 20$): Participants are given sheets of paper with printed text. For the first page, participants are asked to cross out certain letters following certain rules. On the following page, participants in the experimental condition are given a different, more complex set of rules. Participants in the control condition continue on with the same rule. (d) *Emotion video* ($k = 16$): Participants are shown an emotionally evocative video (e.g., a video of animals being harmed). Participants in the experimental condition are given instructions to regulate their emotions in some way (e.g., to either suppress or exaggerate them), whereas participants in the control condition are told to watch the video as they would any other video. (e) *Food temptation* ($k = 8$): Participants in the experimental condition are told to resist the temptation to eat some type of food, usually a dessert. For example, participants are shown a plate of chocolates and a plate of radishes. Participants in the experimental condition are told to only eat radishes and keep from eating chocolates, whereas participants in the control condition are told to eat chocolate. Participants are commonly told that they are taking part in a taste test. (f) *Math* ($k = 1$): Participants in the experimental condition are given more difficult math problems to complete (e.g., 3-digit multiplication) than participants in the control condition (e.g., single-digit addition). (g) *Stroop* ($k = 9$): Participants are shown color words (e.g., the word yellow) printed in colored ink (e.g., blue) and told to name the color of the ink. Generally, participants in the experimental condition are shown all incongruent trials (i.e., when the color of the ink does not match the color to which the word refers), whereas participants in the control condition are shown all congruent trials. (h) *Social exclusion* ($k = 4$): Participants are led to feel socially excluded. For example, while completing an ostensible task as a group, participants in the exclusion condition are told that no other participants wanted to work them. Participants in the control condition are typically included. (i) *Thought suppression* ($k = 17$): Participants are asked to refrain from thinking about a certain topic. The most common version of this task is also known as “the white bear” paradigm because participants in the experimental condition are told that they can think about anything they want, except for a white bear. In contrast, participants in the control condition are told to think about whatever they want. (j) *Transcription* ($k = 6$): Participants are given a sheet of text and told to transcribe it. Participants in the experimental condition are told to transcribe the text without using certain keys, such as the space bar. Participants in the control condition are not given any additional instructions. (k) *Working memory* ($k = 7$): Participants in the experimental condition perform a task that is high in working memory load (e.g., remembering information while performing another task), whereas

participants in the control condition perform a task that is relatively low in working memory load. Note that five effect sizes were derived from experiments in which participants completed pairs of manipulation tasks from the above list (see Table 1).

Outcome tasks. The following eight frequently used outcome tasks were identified. (a) *Food consumed* ($k = 14$): The amount of food (e.g., ice cream) that participants consume in the laboratory is measured. Higher amounts of food are thought to be indicative of lower levels of self-control. (b) *Hand grip* ($k = 13$): Participants hold the arms of a hand grip closed for as long as possible (or hold a dynamometer at some percentage of their maximum grip strength). The length of time that participants are able to persist at this painful task is considered to indicate levels of self-control, where shorter times mean lower levels of self-control. (c) *Impossible anagrams* ($k = 20$): Participants are given a set of anagrams to solve, some of which are designed to be impossible to solve. Persistence at this impossible task is thought to measure self-control, with less time spent (or lower numbers of attempts) indicating worse self-control. (d) *Impossible puzzles* ($k = 16$): Participants are asked to solve puzzles (e.g., tracing geometric shapes printed on paper without going back over previous lines). Unbeknownst to participants, the puzzles are unsolvable. As with impossible anagrams, persistence (either as time or as number of attempts) at this impossible task is used to index self-control. (e) *Possible anagrams* ($k = 12$): Participants are given a large set of anagrams and told to solve as many as possible. Lower numbers of solved anagrams are considered to be indicative of lower self-control. (f) *Standardized tests* ($k = 13$): Participants are given problems from some standardized test, typically the graduate record exam (GRE). The number of problems solved, the number of problems attempted, and the proportion of problems correct out of problems attempted are all used as indexes of self-control (with worse performance being interpreted as lower self-control). (g) *Stroop* ($k = 14$): As described above, participants must identify the ink color of color words. Self-control is measured as the number of correct trials, as well as reaction time (RT) on trials (with slower RT meaning less self-control). (h) *Working memory* ($k = 11$): Participants perform some tasks designed to measure working memory. For example, the operation span task, in which participants must remember words or letters while solving simple math problems. Worse working memory performance (as indicated in a variety of ways, e.g., fewer words recalled overall) is thought to indicate lower self-control. Note that two experiments used more than one of the above outcome tasks and that these effect sizes were calculated separately for the primary analyses (see Table 1).

Effect Size Coding

We quantified the depletion effect as bias-corrected standardized group mean differences (i.e., Hedge's g). A single effect size estimate was taken from each of the 116 experiments that met our inclusion criteria, except for two estimates each which were taken from two experiments that included two frequently used outcome tasks (see Table 1). Hedge's g can be derived from any experiment that provides information about samples sizes, means, and sample standard deviations for the two groups. It is also possible to calculate g from test statistics or p values when means and standard deviations are unavailable. We calculated g based on means and standard deviations when they were available and from other

metrics when means and standard deviations were not available (Cooper et al., 2009). When none of the necessary information was available, we contacted the original authors. When authors report only information from analyses that are more complex than simple comparisons of means from two groups (i.e., paired-sample t tests, repeated measures analysis of variance, and analysis of covariance), additional information is needed to calculate g , such as the correlation between pre- and posttest scores or the correlation between the outcome and the covariate. When this information was available, g was calculated, and when it was not, the authors were contacted or, in some cases, an estimate was made (e.g., if a replication was available in which the necessary information was given, that information was used to estimate the missing information in the experiment for which it was missing).

In the case where authors only reported the overall sample size, it was assumed that sample sizes were equal across groups (if the total sample size was odd, the remainder was placed in the experimental group).

If multiple effect size estimates were available from one outcome measure, a composite of the estimates was calculated. For example, there is no a priori reason to prefer RT to accuracy for the Stroop task as a measure of self-control, and because both measures should reflect depletion, an aggregate of the two was computed using the method described by Gleser and Olkin (1994). This method assumes that the two outcomes are correlated at the level of $r = .50$ by default. When the true correlation between the multiple outcomes was not available, the default was used; however, if analogous experiments contained information about the correlation of interest, these values were used instead.

All procedures for coding effect sizes were set prior to data analysis.

Coding Experiment Attributes

Each experiment was coded for the following attributes, and in the case of significant statistical heterogeneity, these codes were used as meta-analytic moderators (see below): (a) publication status, (b) source laboratory, (c) the number of manipulation tasks, and (d) the number of outcome tasks. For publication status, experiments that were published in peer-reviewed journals, in press, under review, or being sent in for review were coded as one, whereas all other experiments were coded as zero. For source laboratory, experiments were coded as one if one of the authors was associated with the Baumeister and Tice laboratory at Florida State University or a laboratory of a student from the Baumeister and Tice laboratory (this procedure was adopted from Hagger et al., 2010): If any of the authors, or a committee member on a dissertation or master's thesis, was Roy Baumeister, Diane Tice, Kathleen Vohs, C. Nathan DeWall, Mark Muraven, Brandon Schmeichel, or Matthew Gailliot, the experiment was coded as 1, whereas all other experiments were coded as 0. For the number of tasks used in an experiment, if more than one manipulation or outcome task was used, the experiment was coded as a 1; otherwise, it was coded as 0.

Decisions about which experimental attributes to code and how to code them were made prior to data collection and analysis. Additionally, prior to data collection and analysis, we made the decision that we would only attempt to explore statistical heterogeneity (i.e., variability between effect size estimates) using the

Table 1
Coded Characteristics of Experiments Across Data Sets

| Outcome | Author(s) | Exp | Yr | IV | Mult. IV | Mult. DV | Pub | Lab | <i>g</i> | <i>v</i> | <i>n1, n2</i> | |
|---------------------|--------------------|-------------|------|-------|----------|----------|-----|------|----------|----------|---------------|--------|
| Food consumed | BaumeisterD | 2 | 2005 | SE | 0 | 0 | 1 | 1 | 0.88 | 0.12 | 19, 19 | |
| | ChristiansenC | 0 | 2012 | EV | 0 | 0 | 1 | 0 | 0.66 | 0.05 | 40, 40 | |
| | DewitteB | 2a | 2009 | FT | 0 | 0 | 1 | 0 | -0.54 | 0.06 | 35, 41 | |
| | DingemansM | 0 | 2009 | EV | 0 | 0 | 1 | 0 | -0.02 | 0.06 | 33, 33 | |
| | FrieseH | 2 | 2008 | EV | 0 | 0 | 1 | 0 | 0.34 | 0.06 | 33, 33 | |
| | FrieseH | 3 | 2008 | EV | 0 | 0 | 1 | 0 | 0.3 | 0.09 | 25, 21 | |
| | HofmannR | 0 | 2007 | EV | 0 | 0 | 1 | 0 | -0.11 | 0.08 | 26, 24 | |
| | ImhoffS | 1 | 2013 | S | 0 | 0 | 1 | 0 | 0.69 | 0.03 | 69, 68 | |
| | LattimoreM | 0 | 2004 | S | 0 | 0 | 1 | 0 | -0.54 | 0.07 | 29, 30 | |
| | MuravenC | 0 | 2002 | TS | 0 | 0 | 1 | 1 | 0.42 | 0.07 | 29, 29 | |
| | OatenW | 1 | 2008 | SE | 0 | 0 | 1 | 0 | 2.66 | 0.09 | 37, 36 | |
| | StillmanT | 3 | 2009 | AV | 0 | 0 | 1 | 1 | 0.09 | 0.06 | 33, 33 | |
| | VohsB | 5 | 2013 | AE | 0 | 1 | 1 | 1 | 0.72 | 0.13 | 15, 15 | |
| | VohsH | 3 | 2000 | EV | 0 | 0 | 1 | 1 | 0.73 | 0.11 | 18, 18 | |
| | Hand grip | BrayM | 0 | 2008 | S | 0 | 0 | 1 | 0 | 0.46 | 0.08 | 26, 23 |
| | | BrayM | 0 | 2011 | S | 0 | 0 | 1 | 0 | 0.18 | 0.07 | 33, 28 |
| EganH | | 2 | 2012 | CL | 0 | 0 | 1 | 0 | 1.18 | 0.12 | 21, 20 | |
| Litvin | | 0 | 2012 | TS | 0 | 0 | 1 | 0 | 0.16 | 0.03 | 54, 108 | |
| MartijnT | | 1 | 2002 | EV | 0 | 0 | 1 | 0 | 0.7 | 0.12 | 17, 16 | |
| MoldenD | | 2 | 2012 | CL | 0 | 0 | 1 | 0 | 1.05 | 0.19 | 11, 11 | |
| MuravenT | | 1 | 1998 | EV | 0 | 0 | 1 | 1 | 0.67 | 0.08 | 40, 20 | |
| MurtaghT | | 1 | 2004 | S | 0 | 0 | 1 | 0 | 0.07 | 0.06 | 42, 27 | |
| Neale-Lorello | | 1 | 2009 | CL | 0 | 0 | 0 | 0 | 0.22 | 0.07 | 30, 29 | |
| SeeleyG | | 1 | 2003 | TS | 0 | 0 | 1 | 0 | 0.31 | 0.06 | 37, 36 | |
| SeeleyG | | 2 | 2003 | TS | 0 | 0 | 1 | 0 | 0.79 | 0.08 | 28, 27 | |
| TylerB | | 2 | 2009 | CL | 0 | 0 | 1 | 0 | 1.17 | 0.08 | 30, 30 | |
| TylerB | | 3 | 2009 | TS | 0 | 0 | 1 | 0 | 1.12 | 0.12 | 20, 20 | |
| Impossible anagrams | | BarberR | 1 | 2011 | EV | 0 | 1 | 0 | 0 | 0.32 | 0.11 | 18, 18 |
| | | BarberR | 2 | 2011 | AE | 0 | 0 | 0 | 0 | -0.01 | 0.06 | 24, 24 |
| | | BarberR | 3 | 2011 | AE | 0 | 0 | 0 | 0 | 0.14 | 0.06 | 24, 24 |
| | DarowskiH | 2 | 2010 | M | 0 | 0 | 0 | 0 | 0.53 | 0.10 | 16, 13 | |
| | DvorakS | 0 | 2009 | EV | 0 | 0 | 1 | 0 | 0.88 | 0.02 | 90, 90 | |
| | EganH | 1 | 2012 | TS | 0 | 0 | 1 | 0 | 0.72 | 0.12 | 17, 16 | |
| | Gohar | 2 | 2011 | T | 0 | 0 | 0 | 0 | 0.16 | 0.13 | 14, 14 | |
| | Gohar | 3 | 2011 | T | 0 | 0 | 0 | 0 | 0.79 | 0.15 | 16, 12 | |
| | Holmqvist | 1 | 2008 | AV | 0 | 0 | 0 | 0 | -0.02 | 0.05 | 33, 29 | |
| | Holmqvist | 2 | 2008 | WM/AV | 1 | 0 | 0 | 0 | 0.1 | 0.05 | 51, 15 | |
| | Holmqvist | 3 | 2008 | WM/AV | 1 | 0 | 0 | 0 | -0.13 | 0.03 | 74, 27 | |
| | MuravenS | 4 | 2005 | T | 0 | 1 | 1 | 1 | 0.95 | 0.08 | 57, 19 | |
| | MuravenT | 2 | 1998 | TS | 0 | 0 | 1 | 1 | 0.92 | 0.1 | 17, 34 | |
| | Myers | 2 | 2010 | AV | 0 | 0 | 0 | 0 | 0 | 0.09 | 25, 21 | |
| | Ruci | 2 | 2003 | S | 0 | 1 | 0 | 0 | 0.58 | 0.06 | 30, 37 | |
| | Impossible puzzles | ScherschelM | 1 | 2011 | AE | 0 | 0 | 0 | 0 | 0.41 | 0.06 | 35, 33 |
| ScherschelM | | 2 | 2011 | T | 0 | 0 | 0 | 0 | 0.25 | 0.07 | 24, 31 | |
| SegerstromN | | 0 | 2007 | FT | 0 | 0 | 1 | 0 | 0.22 | 0.05 | 41, 42 | |
| Smith | | 1 | 2002 | TS | 0 | 0 | 0 | 1 | 1.77 | 0.19 | 14, 14 | |
| Smith | | p1 | 2002 | TS | 0 | 0 | 0 | 1 | 1.47 | 0.22 | 10, 12 | |
| Wan | | 6 | 2007 | TS | 0 | 0 | 0 | 0 | 1.25 | 0.17 | 14, 13 | |
| BaumeisterB | | 1 | 1998 | FT | 0 | 0 | 1 | 1 | 1.31 | 0.05 | 25, 44 | |
| BaumeisterD | | 3 | 2005 | SE | 0 | 0 | 1 | 1 | 1.25 | 0.22 | 10, 10 | |
| GeeraertC | | 1 | 2013 | TS/FT | 1 | 0 | 0 | 0 | 1.02 | 0.1 | 15, 15 | |
| GeeraertC | | 2 | 2013 | FT | 0 | 0 | 0 | 0 | 0.64 | 0.1 | 15, 15 | |
| GeeraertY | | 1b | 2007 | FT | 0 | 0 | 1 | 0 | 0.52 | 0.07 | 24, 20 | |
| KlaphakeS | | 2.1 | 2011 | AE | 0 | 0 | 0 | 1 | -0.07 | 0.08 | 20, 20 | |
| MuravenS | | 1 | 2003 | TS | 0 | 0 | 1 | 1 | 0.57 | 0.1 | 22, 21 | |
| SatoH | | 0 | 2010 | CL | 0 | 0 | 1 | 1 | 0.14 | 0.02 | 86, 109 | |
| VohsH | | 2 | 2000 | FT | 0 | 0 | 1 | 1 | 0.8 | 0.14 | 14, 14 | |
| WallaceB | | 0 | 2002 | S | 0 | 0 | 1 | 1 | 1.09 | 0.19 | 11, 11 | |
| Wan | 1 | 2007 | CL | 0 | 0 | 0 | 0 | 1.25 | 0.18 | 13, 12 | | |
| Wan | 2 | 2007 | CL | 0 | 0 | 0 | 0 | 1.25 | 0.17 | 14, 13 | | |
| Wan | 3 | 2007 | CL | 0 | 0 | 0 | 0 | 0.9 | 0.09 | 24, 24 | | |
| Wan | 4 | 2007 | CL | 0 | 0 | 0 | 0 | 0.96 | 0.12 | 39, 38 | | |
| Wan | 7 | 2007 | CL | 0 | 0 | 0 | 0 | 0.84 | 0.16 | 13, 13 | | |
| Wan | 8 | 2007 | CL | 0 | 0 | 0 | 0 | 1.18 | 0.15 | 15, 14 | | |

(table continues)

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1 (continued)

| Outcome | Author(s) | Exp | Yr | IV | Mult. IV | Mult. DV | Pub | Lab | <i>g</i> | <i>v</i> | <i>n1, n2</i> | |
|-------------------|--------------------|-----------|-------|------|----------|----------|-----|------|----------|----------|---------------|--------|
| Possible anagrams | BaumeisterB | 3 | 1998 | EV | 0 | 0 | 1 | 1 | 0.74 | 0.13 | 15, 15 | |
| | BoucherK | 2 | 2012 | TS | 0 | 0 | 1 | 0 | 1.01 | 0.19 | 11, 11 | |
| | ClarksonH | 1 | 2010 | CL | 0 | 0 | 1 | 0 | 0.76 | 0.12 | 16, 16 | |
| | ConverseD | 1 | 2009 | WM | 1 | 0 | 1 | 0 | -0.46 | 0.05 | 38, 37 | |
| | ConverseD | 2 | 2009 | CL/S | 1 | 0 | 1 | 0 | -0.71 | 0.11 | 20, 20 | |
| | DamanM | 3 | 2013 | CL | 0 | 0 | 0 | 1 | -0.01 | 0.04 | 54, 53 | |
| | DewitteB | 2b | 2009 | FT | 0 | 0 | 1 | 0 | 0.36 | 0.05 | 38, 38 | |
| | MasicampoR | 5 | 2011 | TS | 0 | 0 | 1 | 1 | 0.75 | 0.08 | 27, 27 | |
| | MoldenD | 1 | 2012 | CL | 0 | 0 | 1 | 0 | 0.31 | 0.05 | 43, 42 | |
| | MurtaghT | 2 | 2004 | TS | 0 | 0 | 1 | 0 | -0.08 | 0.06 | 26, 50 | |
| | UzielL | 3 | 2012 | T | 1 | 0 | 1 | 1 | 0.59 | 0.1 | 20, 23 | |
| | vanDellenM | 2 | 2012 | AV | 0 | 0 | 1 | 0 | 0.19 | 0.06 | 56, 22 | |
| | Standardized tests | ConverseD | 3a | 2009 | CL | 0 | 0 | 1 | 0 | 0.54 | 0.03 | 15, 15 |
| | | ConverseD | 3b | 2009 | CL/S | 1 | 0 | 1 | 0 | -0.30 | 0.13 | 15, 15 |
| KlaphakeS | | 1.1a | 2012 | AE | 0 | 0 | 0 | 1 | -0.32 | 0.12 | 10, 10 | |
| KlaphakeS | | 1.1b | 2012 | WM | 0 | 0 | 0 | 1 | -0.14 | 0.08 | 11, 10 | |
| KlaphakeS | | 1.2a | 2012 | AE | 0 | 0 | 0 | 1 | 0.44 | 0.11 | 19, 20 | |
| KlaphakeS | | 1.2b | 2012 | WM | 0 | 0 | 0 | 1 | 0.04 | 0.08 | 20, 26 | |
| KlaphakeS | | 1.3a | 2012 | AE | 0 | 0 | 0 | 1 | 0.43 | 0.04 | 20, 20 | |
| KlaphakeS | | 1.3b | 2012 | WM | 0 | 0 | 0 | 1 | 0.02 | 0.06 | 20, 20 | |
| KlaphakeS | | 1.4a | 2012 | AE | 0 | 0 | 0 | 1 | 0.42 | 0.08 | 13, 14 | |
| KlaphakeS | | 1.4b | 2012 | WM | 0 | 0 | 0 | 1 | 0.52 | 0.05 | 14, 13 | |
| PondD | | 3 | 2011 | AV | 0 | 0 | 0 | 1 | 0.35 | 0.07 | 65, 63 | |
| SchmeichelV | | 1 | 2003 | AV | 0 | 0 | 1 | 1 | 1.29 | 0.06 | 12, 12 | |
| SchmeichelV | | 3 | 2003 | AV | 0 | 1 | 1 | 1 | 0.54 | 0.09 | 18, 18 | |
| Stroop | | BarberR | 1 | 2011 | EV | 0 | 1 | 0 | 0 | 0.11 | 0.08 | 18, 18 |
| | BoucherK | 1 | 2012 | CL | 0 | 0 | 1 | 0 | 0.69 | 0.15 | 14, 13 | |
| | Cesario | 0 | 2011 | AV | 0 | 0 | 0 | 0 | 0.25 | 0.05 | 31, 30 | |
| | Davisson | 1 | 2009 | CL | 0 | 0 | 0 | 0 | 0.26 | 0.04 | 37, 40 | |
| | DeWallB | 3 | 2008b | SE | 0 | 0 | 1 | 1 | 1.07 | 0.15 | 14, 14 | |
| | Friese | 2 | 2012 | TS | 0 | 0 | 0 | 0 | 0.2 | 0.07 | 29, 32 | |
| | Friese | 3 | 2012 | EV | 0 | 0 | 0 | 0 | 0.27 | 0.05 | 41, 38 | |
| | GailliotB | 7 | 2007c | AV | 0 | 0 | 1 | 1 | 0.62 | 0.09 | 16, 15 | |
| | HedgcockV | 1 | 2012 | AV | 0 | 1 | 1 | 1 | -0.04 | 0.05 | 30, 30 | |
| | InzlichtG | 0 | 2007 | EV | 0 | 0 | 1 | 1 | 0.64 | 0.12 | 15, 18 | |
| | MuravenS | 4 | 2005 | T | 0 | 1 | 1 | 1 | 0.69 | 0.06 | 38, 38 | |
| | Myers | 1 | 2010 | T | 0 | 0 | 0 | 0 | -0.09 | 0.06 | 24, 26 | |
| | PondD | 1 | 2011 | AV | 0 | 0 | 0 | 1 | 0.02 | 0.03 | 56, 60 | |
| | XuH | 0 | 2012 | EV | 0 | 0 | 1 | 0 | 0.65 | 0.07 | 24, 23 | |
| Working memory | YostM | 1 | 2009 | AV | 0 | 0 | 0 | 0 | 0.10 | 0.01 | 129, 122 | |
| | YostM | 1 | 2013 | AV | 0 | 0 | 0 | 0 | -0.14 | 0.02 | 45, 45 | |
| | CarterM | 1a | 2013 | CL | 1 | 0 | 1 | 0 | -0.09 | 0.03 | 71, 71 | |
| | DarowskiH | p2 | 2011 | AE | 0 | 0 | 0 | 0 | -0.22 | 0.13 | 14, 14 | |
| | DarowskiH | p3 | 2011 | AE | 0 | 0 | 0 | 0 | -0.05 | 0.07 | 23, 44 | |
| | HealeyH | 1 | 2011 | AV | 0 | 0 | 1 | 0 | 1.28 | 0.12 | 19, 19 | |
| | HealeyH | 2 | 2011 | AV | 0 | 0 | 1 | 0 | 0.11 | 0.08 | 25, 25 | |
| | HealeyH | 3 | 2011 | AV | 0 | 0 | 1 | 0 | 0.7 | 0.11 | 19, 18 | |
| | HealeyH | 4 | 2011 | AV | 0 | 0 | 1 | 0 | -0.13 | 0.08 | 27, 22 | |
| | KlaphakeS | 2.2 | 2012 | TS | 0 | 1 | 0 | 1 | 0.16 | 0.09 | 21, 21 | |
| | Schmeichel | 1a | 2007 | AV | 0 | 0 | 1 | 1 | 0.45 | 0.05 | 41, 38 | |
| | Schmeichel | 1b | 2007 | AV | 0 | 0 | 1 | 1 | 0.50 | 0.06 | 31, 31 | |
| | Schmeichel | 3 | 2005 | EV | 0 | 0 | 0 | 1 | 0.62 | 0.08 | 22, 22 | |
| | Schmeichel | 2 | 2007 | AE | 0 | 0 | 1 | 1 | 0.51 | 0.05 | 32, 29 | |
| Schmeichel | 4 | 2007 | EV | 0 | 0 | 1 | 1 | 0.53 | 0.06 | 32, 33 | | |

Note. Author names = the last name of the first author and the first letter of the last name of the second author. Exp = the number given to the experiment in the original article (0 = only one experiment was conducted in the original article; the addition of a letter or a decimal indicates subsamples). Yr = the year the experiment was published or, when that information was unavailable, the year we retrieved the data. IV = the outcome task; AE = attention essay; AV = attention video; CL = crossing out letters; EV = emotion video; FT = food temptation; M = Math; S = Stroop; SE = social exclusion; T = transcription; TS = thought suppression; WM = working memory. Pub = 1 when the experiment was under review, in press, or being sent to or published in a peer-review journal. Lab = 1 when one of the authors was associated with the Baumeister and Tice lab. Mult. IV and Mult. DV = 1 when more than one task was used as a manipulation or outcome task, respectively. *g* = the adjusted standardized mean difference and *v* is its associated variance.

four experimental attributes described above. In principle, one might conceive of many other possible experimental attributes that moderate the depletion effect (e.g., the number of impossible anagrams given to participants, the amount of time participants spent in an experimental session, the ratio of incongruent to congruent trials presented during the Stroop task). We chose to

limit ourselves only to the factors listed above because they are interpretable and theoretically important for all of the outcome tasks we observed, and we committed ourselves to only examining these factors because examinations of meta-analytic moderators are notoriously sensitive to Type I error (Thompson & Higgins, 2002). Furthermore, as mentioned above, the goal of this study was

to provide a critical test of the depletion effect, and therefore, producing and testing an exhaustive list of the experimental characteristics that might modulate the depletion effect was deemed unnecessary. Put differently, for the depletion effect to be considered robust and consistent with predictions of the limited strength model, we would need to find evidence for it that did not depend on moderating factors beyond the four listed above.

Table 1 displays the coded experiment-level characteristics for each experiment organized by the category of outcome task. Note that, because two experiments included multiple outcome tasks that were both frequently used tasks (impossible anagrams and Stroop), these experiments were analyzed in two samples: Experiment 4 from MuravenS 2006 and Experiment 1 from BarberR 2011. See the supplementary materials for the citations to the experiments included our data sets.

Reliability

The second and third authors made all of the coding decisions regarding the number of manipulation tasks and the number of outcome tasks used. The first, second, and third authors independently made each coding decision for the other variables discussed above. As recommended (Cooper et al., 2009), interrater agreement for nominal data, such as categorization of the number of manipulation tasks used, was calculated as Cohen’s κ coefficient, and interrater agreement for continuous data was calculated using the intraclass correlation coefficient (ICC). Reliabilities are presented in Table 2. The weighted reliability for each coded variable was satisfactory. After reliability was calculated, all disagreements were resolved with a discussion between coders before a final code was given.

Analyses

Analyses were conducted using R (version 2.15; R Core Team, 2014). All of the data and scripts for our analyses are available as

supplementary materials. Except where noted, decisions on which analyses to conduct were made prior to data collection and analysis.

We applied random/mixed-effects meta-analysis models to our eight data sets (Cooper et al., 2009); between-study variance (τ^2) was estimated using restricted maximum-likelihood estimation, and all models were calculated using the Knapp and Hartung correction (Viechtbauer, 2010), which has been shown to be an improvement over standard methods, particularly when meta-analytic samples are small (Int’Hout, Ioannidis, & Borm, 2014). When we observed evidence for statistical heterogeneity among effect sizes (i.e., variation due to a source other than sampling error), mixed-effects models including experiment attributes as meta-analytic moderators were conducted (if, that is, these variables divided the sample up into subgroups containing more than one experiment—in other words, if only one experiment in a sample was unpublished, then publication status was not used as a moderator). When the overall F test for moderators was not statistically significant, it was concluded that none of the attributes moderated the overall effect.

In addition to the random/mixed-effects meta-analysis models, we also applied a method for assessing excess statistical significance (the Test for Excess Significance, or TES; Ioannidis & Trikalinos, 2007). TES evaluates whether there is an excess of statistically significant results in a set of tests of an effect by comparing the observed number of statistically significant findings to the expected number based on estimates of statistical power. This method makes no assumptions about the actual cause of the discrepancies between observed and expected findings, which may be due to publication bias, undisclosed use of researcher degrees of freedom in the primary literature, fabrication of data, or randomness.

There are two methods for calculating power for TES (Schimmack, 2012): First, it is possible to obtain an estimate of the true effect size using information from the entire set of studies (as in a meta-analysis, for example), and to then calculate power for each study given that estimate. One can also use the upper and lower limits of the 95% confidence interval around the meta-analytic estimate to calculate power for each study. Second, it is possible to use the effect size estimates from each study to calculate post hoc power for each study independently. In both cases, the power estimates can be averaged to give the expected proportion of statistically significant results in a set of studies, E . The observed number of studies with statistically significant results, O , is then compared with E using a binomial test where the null hypothesis is that O was produced by a binomial distribution where the probability of observing a statistically significant result is E . The significance of this binomial test reflects the probability that there is an excess of significant studies (i.e., smaller p values suggest lower probability that there is *not* an excess of statistical significance).

A significance threshold for TES of $p < .10$ has been suggested (Ioannidis & Trikalinos, 2007); however, we use TES less as a hard line for making dichotomous decisions about the existence of bias, and more as a descriptive statistic (Schimmack, 2012). We conducted TES using power calculated from the estimated true effect from random-effects meta-analysis models, from the upper and lower limit of the 95% confidence intervals surrounding these

Table 2
Interrater Reliability for Coded Variables Across all Data Sets

| Variable | Measure | Reliability by rater pairs | | Number coded (k) | | Weighted reliability |
|--------------------|----------|----------------------------|-------|----------------------|-----|----------------------|
| | | R1&R3 | R2&R3 | R1 | R2 | |
| Publication status | κ | 1.00 | 0.87 | 100 | 24 | 0.97 |
| Source lab | κ | 0.92 | 0.47 | 100 | 24 | 0.83 |
| IV category | κ | . | 0.91 | 0 | 116 | 0.91 |
| DV category | κ | . | 0.97 | 0 | 116 | 0.97 |
| IV count | ICC | . | 0.94 | 0 | 116 | 0.94 |
| DV count | ICC | . | 1.00 | 0 | 116 | 1.00 |
| $n1$ | ICC | 0.92 | 0.99 | 100 | 17 | 0.93 |
| $n2$ | ICC | 0.93 | 1.00 | 100 | 17 | 0.94 |
| d | ICC | 0.98 | 0.91 | 145 | 29 | 0.97 |
| v | ICC | 0.96 | 0.99 | 145 | 29 | 0.97 |

Note. Raters: R1 is rater 1 (first author), R2 is rater 2 (second author), R3 is rater 3 (third author). κ = Cohen’s κ ; ICC = the intraclass correlation coefficient. $n1$ and $n2$ are the sample sizes for the experimental and control groups, respectively. R3 coded all observations, whereas R1 and R2 coded only some. Weighted reliability = the reliability scores between the pairs of raters weighted by the number of effect sizes coded by those pairs. The numbers of effect sizes coded do not necessarily match the final number because many effect sizes in the analyzed data sets are aggregations of several effect sizes, each of which was coded separately.

estimates, and from the effect size estimates provided by each experiment.

In addition to TES, we applied the popular trim and fill method (Duval & Tweedie, 2000), which is an iterative nonparametric test that estimates the number and magnitude of effect sizes missing due to publication bias and then corrects the estimate of the overall effect size. Recently, this method has been criticized for undercorrecting for publication bias, as well as being based on overly specific assumptions: It is specifically a model of how missingness in a meta-analytic data set can lead to overestimation of the true underlying effect (Moreno et al., 2009a).

We also analyzed our data sets using two meta-regression methods for correcting for the influence of small-study effects (Stanley & Doucouliagos, 2014). The most concerning cause of small-study effects is publication bias, which produces a monotonic relationship between effect size estimates and the standard errors of those estimates. Traditionally, small-study effects are visually examined with a funnel plot (Figures 1 and 2) and quantified as the slope coefficient—the so-called funnel plot asymmetry test (FAT)—for a weighted least squares (WLS) regression model in which effect size estimates are regressed on the standard error of those estimates, weighted by the inverse of the variances. Recently, the

intercept of this same WLS regression model has been shown to provide an accurate estimate of the underlying effect that is uninfluenced by publication bias and other small-study effects (Precision Effect Test or PET; Stanley & Doucouliagos, 2014). The intercept from a similar model that uses variances instead of standard errors as the predictor has also been shown to be extremely useful as an estimate of the underlying effect that is robust to small-study effects (Precision Effect Estimation with Standard Error or PEESE; Stanley & Doucouliagos, 2014). Although small-study effects are not necessarily due to publication bias, it is still useful to correct for their influence, particularly when the theory inspiring the observed studies cannot reasonably account for the presence of such effects, as is the case for the depletion effect as predicted by the limited strength model (see below).

Two points should be considered when using PET and PEESE. First, it may seem that these estimators should not be applied when FAT is not statistically significant—the logic being that, with a nonsignificant test for small-study effects, the influence of such effects need not be controlled. Unfortunately, FAT has been shown to have particularly low statistical power, and Stanley and Doucouliagos (2014) have recommended that PET and PEESE be applied regardless of the statistical significance of FAT. Second, it

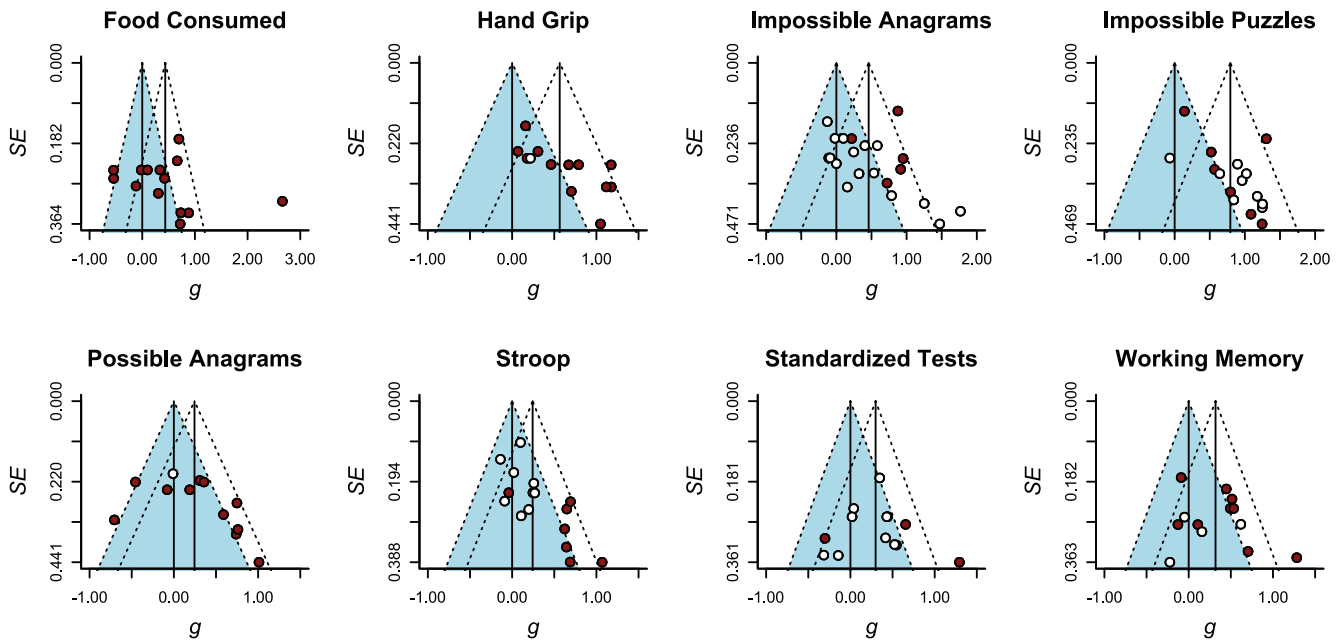


Figure 1. Contour-enhanced funnel plots for the eight data sets. The experiments in each of the eight data sets are displayed in funnel plots: standard error (*SE*) on the (inverted) vertical axis and the standardized mean difference (*g*) on the horizontal axis. Published experiments are plotted as color-filled circles, unpublished experiments as white-filled. Experiments plotted outside of the shaded contour are statistically significant ($p < .05$). The unshaded triangular area is centered on the random-effects meta-analysis estimate of the depletion effect, and in the absence of statistical heterogeneity, 95% of experiments should fall within this area. In the absence of small-study effects (and other forms of heterogeneity), one expects the largest experiments (lowest *SE*) to center on the true effect and the smaller experiments to be symmetrically distributed to the left and right as *SE* increases. Small-study effects disrupt this expected symmetry, creating instead a monotonic relationship between *SE* and *g*. Confidence that the cause of small-study effects is related to bias increases if experiments approximately track the rightmost side of the shaded contour, which indicates that results tend to be just statistically significant enough (i.e., at levels closest to $p = .05$). Bias is also indicated by a trend for published experiments to be significant and unpublished experiments to be nonsignificant (i.e., color-filled circles outside of the shaded area and white-filled circles within it). See the online article for the color version of this figure.

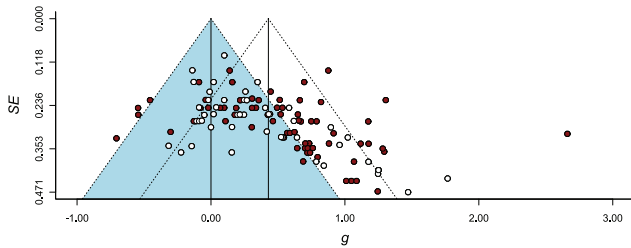


Figure 2. Contour-enhanced funnel plot for the combined data set. *SE* = Standard error of the effect size estimate; *g* = bias-corrected standardized mean difference between performance on the outcome task by the control and experimental groups. The unfilled triangular contour is centered on the random-effects meta-analysis estimate of the depletion effect. White-filled circles are unpublished effect size estimates, color-filled are published. Effect sizes overlapping the shaded contour are not statistically significant. See the online article for the color version of this figure.

has been shown that, in the presence of an underlying effect that is truly nonzero, PET provides an underestimate (i.e., overcorrects for the influence of small-study effects). In contrast, when the underlying effect is zero, PEESE provides an overestimate (i.e., undercorrects for small-study effects). As a result, Stanley and Doucouliagos (2014) have suggested using PET and PEESE together as a conditional estimator (PET-PEESE) where the estimate of the underlying effect is given by PET when PET is not statistically significant and PEESE otherwise. There is promising support for this method, although it is relatively new (Stanley & Doucouliagos, 2014). Here, we report both estimates.

Concerns have been raised in the literature about the application of each of the above techniques (TES, the trim and fill, FAT, PET, and PEESE) in the presence of moderate to extreme statistical heterogeneity (e.g., $I^2 > 50\%$). The basis for these concerns is that statistical heterogeneity suggests that multiple true underlying effects are being measured by studies in the meta-analytic sample, and thus, the assumption upon which the above techniques are based is violated (Ioannidis & Trikalinos, 2007; Terrin, Schmid, Lau, & Olkin, 2003). Substantial statistical heterogeneity seems to be particularly problematic for TES and the trim and fill—which are both methods designed specifically for assessing missingness due to bias (Ioannidis & Trikalinos, 2007; Schimmack, 2012; Terrin et al., 2003); however, statistical heterogeneity is less of a concern for the more general WLS models, PET and PEESE (Stanley & Doucouliagos, 2014). If one conceptualizes small-study effects as a special case of statistical heterogeneity, then the logical next step is to explain this heterogeneity by examining possible meta-analytic moderators—that is, by fitting a meta-regression model to statistically control for the effect of some variable (such as small-study effects) on the underlying effect of interest (Rücker, Schwarzer, Carpenter, Binder, & Schumacher, 2011; Stanley & Doucouliagos, 2014). Thus, from this perspective, heterogeneity is not a problem for the application of meta-regression methods such as PET and PEESE: Indeed, between-study heterogeneity is the statistical condition that these models are designed to explain.

Nevertheless, simulation studies do suggest that PET and PEESE become relatively more inaccurate in the face of extreme statistical heterogeneity (e.g., Stanley & Doucouliagos, 2014).

Therefore, it is important to keep the amount of statistical heterogeneity in mind when interpreting results for these methods (as is the case for interpreting any parameter estimate from a meta-analytic model applied to heterogeneous data). However, these meta-regression methods have outperformed other methods in terms of reducing the inflation of effect size estimation due to publication bias, and although the resulting estimates are not perfectly accurate, they tend to be the most accurate (Moreno et al., 2009a; Rücker et al., 2011; Stanley & Doucouliagos, 2014).

In addition to the a priori analyses described above, we also conducted two sets of post hoc analyses. First, it is possible that any small-study effects observed in our data sets can be explained if the degree to which a manipulation task depletes self-control is somehow related to how many participants are typically included in experiments that make use of it. This explanation for small-study effects would be consistent with the limited strength model if manipulation task potency was negatively correlated with sample size (i.e., participants in smaller experiments tend to experience more depletion, and thus, such experiments produce larger estimates of the depletion effect). To investigate the possibility that manipulation task and sample size were related, we modeled sample size as a function of manipulation task using a generalized linear model.

Second, we decided to apply the methods described above to a data set comprising all of the effect sizes from our eight separate data sets. The primary motivation for this post hoc analysis was to provide estimates based on the highest degree of precision and statistical power we could achieve; however, it is important to keep in mind that estimates from such a combined model represent an average over several potentially distinct effects, and interpreting such estimates effectively assumes that the eight outcome tasks we identified were comparably valid measures of self-control.

Results

As described above, eight samples were created by dividing the effect size estimates we collected into groups on the basis of outcome task category. The seven statistical techniques described above (Random-effects meta-analysis, mixed-effects meta-analysis, TES, the trim and fill, and estimators based on WLS meta-regression, FAT, PET, and PEESE) were applied to each sample.

Random-Effects Meta-Analysis

Overall, the estimates of the depletion effect based on the random-effects meta-analysis models tended to be statistically significant: All estimates were significant except for those for the food consumption and possible anagrams data sets (see Table 3). The statistically significant estimates ranged from $g = 0.24$ (95% CI = [0.07, 0.41]) for Stroop to $g = 0.79$ (95% CI = [0.56, 1.02]) for impossible puzzles. There was considerable variation among our estimates of the depletion effect based on different outcome tasks—note, for example, that the confidence intervals just described for standardized tests and for impossible puzzles do not overlap. These results strongly suggest that the magnitude of the depletion effect is highly dependent on the outcome task used to operationalize it, or some other factor that covaries with the outcome task that researchers select for their experiments.

Table 3
Parameter Estimates for the Random-Effects Models

| Sample | <i>k</i> | <i>g</i> | <i>Q</i> | τ^2 | <i>I</i> ² |
|---------------------|----------|-----------------------|------------|-------------------|-----------------------|
| Food consumption | 14 | 0.44* [-0.01, 0.89] | 96.75**** | 0.52 [0.24, 1.51] | 88.54 [77.96, 95.75] |
| Hand grip | 13 | 0.56**** [0.31, 0.81] | 26.85**** | 0.09 [0.01, 0.39] | 55.73 [12.31, 83.94] |
| Impossible anagrams | 21 | 0.46**** [0.23, 0.69] | 64.66**** | 0.15 [0.06, 0.48] | 68.39 [46.37, 87.49] |
| Impossible puzzles | 16 | 0.79**** [0.56, 1.02] | 40.71**** | 0.12 [0.02, 0.31] | 57.16 [19.24, 77.58] |
| Possible anagrams | 12 | 0.24 [-0.07, 0.56] | 33.13**** | 0.16 [0.04, 0.69] | 69.26 [37.58, 90.80] |
| Standardized tests | 13 | 0.30** [0.05, 0.54] | 22.04** | 0.06 [0.00, 0.42] | 43.32 [0.00, 84.74] |
| Stroop | 16 | 0.24*** [0.07, 0.41] | 27.87** | 0.04 [0.00, 0.21] | 46.81 [1.41, 82.60] |
| Working memory | 13 | 0.32** [0.08, 0.56] | 25.58** | 0.07 [0.01, 0.39] | 51.69 [8.79, 85.73] |
| Combined | 116 | 0.43**** [0.34, 0.52] | 375.76**** | 0.16 [0.12, 0.26] | 71.55 [64.54, 80.15] |

Note. Numbers within brackets are the upper and lower limits of 95% confidence intervals. *k* = number of experiments; *g* = estimate of the average underlying effect; *Q* = Cochran's *Q* statistic for statistical heterogeneity, τ^2 is the estimate of the between-study variance, and *I*² is a metric generally interpreted to indicate the percentage of variance due to sources other than sampling error.

* *p* < .10. ** *p* < .05. *** *p* < .01. **** *p* < .001.

In addition to the fact that most of the random-effects meta-analytic estimates of the depletion effect were statistically significant, it is noteworthy that all the data sets showed statistically significant heterogeneity at a *p* < .05 level (see Table 2). When statistical heterogeneity was quantified in terms of the popular *I*² statistic (Cooper et al., 2009)—typically interpreted as the percent of total variation in effect sizes due to between-study variation rather than sampling error—estimates of heterogeneity ranged from approximately 43.32% (for standardized tests) to 88.54% (for food consumption; see Table 3). Using the recommended overlapping descriptors (Higgins, 2008), five of our data sets fell within the range described as moderate heterogeneity (i.e., 30% < *I*² < 60%) and five fell within the range described as “substantial” heterogeneity (i.e., 50% < *I*² < 90%) (estimates from two data sets were in the overlap of the two ranges). Importantly, such thresholds are only rough guides, and it should be noted that the confidence intervals surrounding estimates of both *I*² and of τ^2 for our data sets are quite wide: For example, for the three data sets with the smallest amounts of heterogeneity (standardized tests, Stroop, and working memory), the lower limits of the 95% confidence intervals for *I*² were less than 10%, but the upper limits were all greater than 80% (i.e., “substantial” heterogeneity). Thus, our measures of statistical heterogeneity can only be interpreted as suggesting that it is likely that at least some heterogeneity (but possibly a great deal of it) exists across our data sets.

Mixed-Effects Meta-Analysis

We attempted to explain the heterogeneity in our data sets using mixed-effects meta-analysis models including terms for experiment-level characteristics. We used this approach only for data sets in which the coded experiment characteristics divided data sets up into subgroups of more than one experiment (see Table 1), so we were able to apply mixed-effects models to all but the hand grip data set.

For only the impossible anagrams and Stroop data sets was the *F* test for the inclusion of moderators statistically significant (see Table 4). And despite this significant test for the impossible anagrams data set, the test for residual heterogeneity for that data set remained statistically significant, suggesting that the inclusion of meta-analytic moderators (i.e., predictors) did not fully explain

the observed heterogeneity. For this model, only the regression coefficients for the intercept (*b* = 0.33, *p* = .02) and for Source Lab (*b* = 0.77, *p* = .01) were statistically significant (see Table 4), which implies that the average effect size of the depletion effect as measured with impossible anagrams as the outcome task is medium in magnitude, but that it should be expected to become three times larger (*b* = 0.33 + 0.77 = 1.10) when observed in experiments conducted by experimenters affiliated with the Baumeister and Tice laboratory.

For the Stroop data set, only the coefficient for publication status was statistically significant (*b* = 0.58, *p* = .004), whereas the intercept for this model was not statistically significant (*b* = 0.10, *p* = .14), suggesting that the average effect size derived from unpublished tests of the depletion effect using the Stroop task as an outcome is indistinguishable from zero. However, when taken from published tests of the depletion effect, the average effect size can be expected to become nearly seven times larger. These results are consistent with publication bias as an explanation for the small-study effects observed in this data set (see below). It is also noteworthy that accounting for publication status in this data set apparently fully explains the initially observed “moderate” degree of heterogeneity (compare *p* values for *Q* in Table 1 to *p* values for *Q_e* in Table 4).

Similar—although not consistently statistically significant—patterns were seen in the standardized tests and working memory data sets. For both of these data sets, the test for residual heterogeneity was nonsignificant, as were the terms for the intercepts, whereas the terms for publication status and source lab were positive and associated with small *p* values (.35 ≥ *p* ≥ .03). Of course, for both of these data sets the *F* test for the inclusion of moderators was nonsignificant (*p* = .08 for standardized tests; *p* = .24 for working memory), so no firm conclusions should be drawn from these findings.

Notably, the evidence for an effect of source lab was less consistent than the evidence for an effect of publication status across all of our data sets. In every case where publication status was included in a model, the effect was positive, whereas in two of the seven cases in which source lab was included, the effect was negative. Furthermore, in the mixed-effects model applied to the combined data set (see below), only the effect for publication status was nearly statistically significant (*p* = .06). Thus, we think

Table 4
Mixed-Effects Models

| Sample | Parameter estimates | | | Test of moderators | | Residual heterogeneity | |
|---------------------|---------------------|----------|----------|--------------------|----------|------------------------|----------|
| | Moderators | <i>b</i> | <i>p</i> | <i>F</i> | <i>p</i> | <i>Q_e</i> | <i>p</i> |
| Food consumption | Intercept | 0.37 | .19 | 0.17 | .69 | 96.03 | <.001 |
| | Source lab | 0.19 | .69 | | | | |
| Impossible anagrams | Intercept | 0.34 | <.001 | 3.99 | .02 | 31.02 | .01 |
| | Publication | 0.13 | .55 | | | | |
| | Source lab | 0.76 | .01 | | | | |
| | Multiple IV | -0.37 | .17 | | | | |
| | Multiple DV | -0.01 | .97 | | | | |
| Impossible puzzles | Intercept | 0.90 | <.001 | 0.94 | .42 | 33.68 | .001 |
| | Publication | 0.17 | .59 | | | | |
| | Source lab | -0.39 | .22 | | | | |
| Possible anagrams | Intercept | 0.27 | .14 | 2.9 | .11 | 21.32 | .01 |
| | Source lab | 0.33 | .24 | | | | |
| | Multiple IV | -0.58 | .07 | | | | |
| Standardized tests | Intercept | -0.57 | .16 | 3.31 | .08 | 13.46 | .20 |
| | Source lab | 0.80 | .05 | | | | |
| | Publication | 0.67 | .03 | | | | |
| Stroop | Intercept | 0.10 | .14 | 4.92 | .02 | 12.73 | .39 |
| | Source lab | -0.07 | .66 | | | | |
| | Publication | 0.58 | .004 | | | | |
| | Multiple DV | -0.23 | .22 | | | | |
| Working memory | Intercept | -0.01 | .22 | 1.67 | .24 | 17.76 | .06 |
| | Source lab | 0.32 | .20 | | | | |
| | Publication | 0.23 | .23 | | | | |
| Combined | Intercept | 0.33 | <.001 | 3.53 | .01 | 321.96 | <.001 |
| | Source lab | 0.11 | .24 | | | | |
| | Publication | 0.18 | .06 | | | | |
| | Multiple IV | -0.47 | .01 | | | | |
| | Multiple DV | -0.20 | .18 | | | | |

Note. Up to four coded experiment-level characteristics could have been used as predictors (moderators) in the mixed-effects models. Residual heterogeneity is given as *Q_e* (compare with the *Q* statistic in Table 1).

there is little reason to conclude that results produced by experimenters who are affiliated with the Baumeister and Tice laboratory are stronger than are results produced by other experimenters, whereas it seems as though results that were ultimately published tended to be more in favor of the limited strength model than those that were not published.

Test for Excess Statistical Significance (TES)

TES was applied to our data sets to assess for excess statistical significance (see Table 5). As mentioned, TES is based on the statistical power of a set of studies, so we calculated statistical power from four separate estimates of the depletion effect for each data set: (a) the effect size estimates from the individual experiments, (b) the effect size estimates from the random-effects meta-analysis models, and (c) the lower and (d) the upper limits of the 95% confidence intervals surrounding the random-effects meta-analysis estimates.

For every data set, when TES was calculated based on the limits of the confidence intervals for the random-effects estimate, it was either always small enough for concern (*ps* ≤ .03, for the lower limits) or always large enough to indicate a lack of bias (*ps* > .64 for the upper limits). When TES was instead calculated using the random-effects meta-analysis effect size estimate, it was small enough to potentially suggest bias in only four of our eight data sets (food consumption, impossible puzzles, possible anagrams,

and working memory). Given this range of results, and the fact that some have questioned the validity of this measure (Simonsohn, 2012), it is difficult to draw any firm conclusions from the TES results. What does seem clear, however, is that the statistical power for the experiments examining the depletion effect is chronically low, almost regardless of which estimate one prefers for the true effect (see Table 5).

The Trim and Fill

The trim and fill method was used to impute data for experiments that might have been missing as a function of effect size and sample size (e.g., due to publication bias), and then to reestimate the random-effects meta-analysis estimate based on the imputed data set. For four of our eight data sets (food consumption, impossible anagrams, standardized tests, and working memory), the trim and fill did not impute any missing studies, and thus, the estimates of the overall effect were not adjusted (see Table 6). For the four remaining data sets, the trim and fill estimated between one (possible anagrams) and five (impossible puzzles and Stroop) missing studies and reduced the random-effects meta-analysis estimates of the overall effect by 17% (possible anagrams), 26% (impossible puzzles), 36% (hand grip), and 46% (Stroop). Additionally, the overall effect size for the Stroop data set following application of the trim and

Table 5
Test for Excessive Significance (TES) and Average Power for Each Data Set

| Sample | Measure | Effect size estimate used for power calculation | | | |
|---------------------|------------|---|-------|-------|-------|
| | | Individual | RE | RE LL | RE UL |
| Food consumption | Avg. power | .48 | .39 | .05 | .89 |
| | TES | .35 | .13 | <.001 | .99 |
| Hand grip | Avg. power | .51 | .53 | .22 | .80 |
| | TES | .52 | .59 | .01 | .99 |
| Impossible anagrams | Avg. power | .42 | .36 | .13 | .62 |
| | TES | .55 | .32 | <.001 | .98 |
| Impossible puzzles | Avg. power | .63 | .64 | .41 | .81 |
| | TES | .11 | .11 | .001 | .64 |
| Possible anagrams | Avg. power | .40 | .15 | .06 | .53 |
| | TES | .34 | .005 | <.001 | .70 |
| Standardized tests | Avg. power | .27 | .15 | .05 | .36 |
| | TES | .72 | .30 | .03 | .89 |
| Stroop | Avg. power | .29 | .17 | .06 | .37 |
| | TES | .71 | .28 | .01 | .90 |
| Working memory | Avg. power | .35 | .22 | .06 | .52 |
| | TES | .13 | .01 | <.001 | .56 |
| Combined | Avg. power | .42 | .33 | .23 | .45 |
| | TES | .10 | <.001 | <.001 | .25 |

Note. Avg. power is the mean statistical power for the given sample and effect size estimate. Effect size estimates are taken from the individual experiments (Individual), the estimate from the random-effects meta-analysis model (RE), and the lower (LL) and upper (UL) limits of the 95% confidence interval on random-effects estimate.

fill procedure became statistically nonsignificant, $g = 0.11$, 95% CI [-0.07, 0.29].

Weighted-Least Squares Meta-Regression Models (Funnel Plot Asymmetry Test [FAT], Precision Effect Test [PET], and Precision Effect Estimate With Standard Error [PEESE])

We applied two WLS meta-regression models to test for small-study effects (using FAT) and to correct for their influence (using PET and PEESE) in our data sets (Tables 7). For FAT, three of our eight data sets—hand grip, impossible puzzles, and Stroop—showed coefficients that were positive and significant at the recommended threshold of $p < .10$ (Table 7; Egger et al., 1997). For the four other data sets—food consumption, impossible anagrams, possible anagrams, and working memory—coefficients for FAT were not statistically significant, but were positive and ranged from $b = 1.91$ to $b = 3.50$, consistent with the presence of small-study effects. For the remaining data set—standardized tests—the coefficient for FAT was positive, but nonsignificant and small, $b = 0.12$, $p = .95$. Moreover, examination of the funnel plot (see Figure 1) indicated that the fact that this coefficient was positive was likely due only to a single observation that represented both largest effect size estimate and the smallest sample size in the standardized tests data set. Upon removing this observation, the coefficient for FAT became negative: $b = -1.32$, 95% CI [-4.90, 2.25]. Thus, it seems unlikely that theoretically meaningful small-study effects were present in the standardized tests data set, but, given the problems with FAT as a definitive test for the small-study effects (Egger et al., 1997; Stanley & Doucouli-

gos, 2014; Sterne et al., 2011), a case can at least be made for their presence in the remaining seven data sets.

The results from PET and PEESE are clearer than those for FAT. For all data sets, estimates of the depletion effect based on PET and PEESE were statistically nonsignificant (the first two columns of Table 7), which suggests that the apparent evidence observed when data sets were analyzed using random-effects meta-analysis was likely due to small-study effects. However, there are two important points to note about these results.

First, for the standardized tests data set, after removing the one seemingly outlying observation, the PET estimate ($b = 0.60$) and PEESE estimate ($b = 0.46$) of the depletion effect were both larger than the random-effects meta-analysis estimate ($g = 0.30$). In other words, the application of PET and PEESE to this data set actually provided increased estimates of the depletion effect (although these estimates were nonsignificant because WLS meta-regression models produce wider confidence intervals compared to random-effects meta-analysis). Given the comparison between PET, PEESE, and the random-effects meta-analysis estimates, the negative coefficient for FAT, and the fact that the contour-enhanced funnel plot for this data set is entirely inconsistent with small-study effects, particularly publication bias, suggest to us that the depletion effect, when derived from experiments using standardized tests as an outcome variable, is of medium magnitude and distinguishable from zero (i.e., $g = 0.30$), consistent with the limited strength model.

A second important detail about the PET and PEESE estimates of the depletion effect is that, in some cases, the estimates provided by PET are not only negative, but can be described as large effect sizes (hand grip: $b = -0.76$ and possible anagrams: $b = -0.71$). In contrast, all estimates from PEESE (with the exception of that for standardized tests) are small in magnitude and close to zero: The estimates range from $b = -0.23$ (possible anagrams) to $b = 0.22$ (impossible puzzles). Therefore, it seems extreme to claim that the depletion effect is actually strongly negative (e.g., $b < -0.70$, as indicated by the estimates for hand grip and possible anagrams) and that small-study effects have resulted in both the published and unpublished literature almost exclusively showing effect sizes around 0 or higher (see Figure 1). Instead, a more reasonable interpretation seems to be that the PEESE estimates are more accurate than the PET estimates.

Indeed, Stanley and Doucouliagos (2014) have shown that, when the true effect is nonzero, PEESE tends to provide a more

Table 6
Estimates of the Depletion Effect Based on the Trim and Fill

| Sample | g | p | $+k$ |
|---------------------|--------------------|-------|------|
| Food consumption | 0.44 [-0.01, 0.89] | .06 | 0 |
| Hand grip | 0.36 [0.09, 0.63] | .008 | 4 |
| Impossible anagrams | 0.46 [0.23, 0.69] | <.001 | 0 |
| Impossible puzzles | 0.62 [0.39, 0.84] | <.001 | 5 |
| Possible anagrams | 0.20 [-0.07, 0.47] | .15 | 1 |
| Standardized tests | 0.30 [0.05, 0.54] | .02 | 0 |
| Stroop | 0.11 [-0.07, 0.29] | .23 | 5 |
| Working memory | 0.32 [0.08, 0.56] | .01 | 0 |
| Combined | 0.24 [0.13, 0.34] | <.001 | 29 |

Note. g = the (adjusted) estimate of the overall true effect after experiments have been imputed (the p -value corresponds to this estimate). $+k$ = the number of experiments imputed by the trim and fill.

Table 7
Parameter Estimates for PET, PEESE, and FAT

| Sample | PET | | PEESE | | FAT | |
|---------------------|---------|----------------|-------|---------------|----------|----------------|
| Food consumption | -0.21 | (-2.35, 1.93) | -0.01 | (-1.13, 1.11) | 2.38 | (-5.94, 10.69) |
| Hand grip | -0.76 | (-1.55, 0.04) | -0.11 | (-0.54, 0.32) | 4.76*** | (1.81, 7.71) |
| Impossible anagrams | 0.04 | (-0.66, 0.74) | 0.15 | (-0.22, 0.53) | 1.51 | (-1.18, 4.20) |
| Impossible puzzles | -0.16 | (-0.76, 0.43) | 0.22 | (-0.16, 0.60) | 3.02*** | (0.98, 5.06) |
| Possible anagrams | -0.71 | (-1.93, 0.51) | -0.23 | (-0.83, 0.38) | 3.50 | (-1.18, 8.17) |
| Standardized tests | 0.27 | (-0.85, 1.38) | 0.27 | (-0.37, 0.90) | 0.12 | (-3.96, 4.20) |
| Stroop | -0.27* | (-0.58, 0.04) | -0.07 | (-0.24, 0.11) | 2.35*** | (0.84, 3.86) |
| Working memory | -0.15 | (-1.20, 0.99) | 0.09 | (-0.47, 0.65) | 1.79 | (-2.39, 5.97) |
| Combined | -0.27** | (-0.52, -0.01) | 0.003 | (-0.14, 0.15) | 2.54**** | (1.52, 3.55) |

Note. PET and PEESE are estimates of the underlying effect that are robust to small-study effects.
* $p < .10$. ** $p < .05$. *** $p < .01$. **** $p < .001$.

accurate estimate: This is because PET overcorrects for small-study effects in such cases. Thus, it may be that the depletion effect, although essentially zero on average, is indeed nonzero in specific cases, but small in magnitude and both positive (as appears to be the case for impossible anagrams, impossible puzzles, and working memory)—as predicted by the limited strength model—and negative—contrary to the limited strength model (as appears to be the case for food consumption, hand grip, possible anagrams, and Stroop). If this is true, than PEESE ought to provide the most accurate estimates.

Post Hoc Analyses

We conducted two sets of analyses post hoc. The first, which applied the statistical methods described above to a combined data set including all of the effect sizes across our eight data sets, was conducted to produce estimates with the highest possible statistical power and precision. The second, which represented an examination of sample size as a function of manipulation task, was designed to rule out a potential limited-strength-model-consistent explanation for small-study effects. Two experiments produced effect sizes that fell into two data sets (i.e., MuravenS, 2005, Experiment 4, and BarberR, 2011, Experiment 1, both from the impossible anagrams and Stroop data sets), so for our post hoc analyses, we aggregated the pairs of estimates produced by these two experiments.

The estimate of the depletion effect from the random effects model applied to this data set was $g = 0.43$, 95% CI [0.34, 0.52], which was considerably smaller than (and nonoverlapping with) the overall mean effect size for the depletion effect that Hagger et al. (2010) estimated using random effects meta-analysis, $d = 0.62$, 95% CI [0.57, 0.67]. This difference can be attributed to the differences in inclusion criteria between the two meta-analytic data sets (e.g., our inclusion of unpublished data, our exclusion of experiments that we deemed inappropriate tests of the depletion effect). As expected from our primary analyses, the estimated effect for our combined data set was qualified by clear signs of statistical heterogeneity, $Q = 375.76$, $p < .001$; $I^2 = 71.55\%$. The mixed-effects model applied to this data set revealed a statistically significant intercept, $b = 0.33$, $p < .001$, in addition to a nearly significant positive effect for publication status, $b = 0.18$, $p = .06$, and a negative effect for the presence of multiple manipulation tasks, $b = -0.47$, $p = .001$, which suggests that, contrary to the

limited strength model, the depletion effect reverses (i.e., subsequent acts of self-control are less likely to fail) in the presence of multiple “depleting” tasks (see Table 4). Application of the trim and fill reduced the overall effect by 44%, $g = 0.24$, 95% CI [0.13, 0.34], based on the addition of 29 effect sizes (see Table 6).

FAT was statistically significant and in the direction consistent with small-study effects (see Table 7), and the overall effect for our combined data set when estimated using PET was negative and statistically significant, $b = -0.27$, 95% CI [-0.52, -0.01]. The estimate from PEESE was essentially zero and not statistically significant, $b = 0.003$, 95% CI [-0.14, 0.15] (note that applying PET-PEESE in this case would result in favoring the PEESE estimate). Moreover, our post hoc analysis of the combined data set highlights an important point: Sample sizes across our data sets were chronically small (see Figure 3)—specifically, the minimum total sample size in the combined data set was $N = 20$, the 25% quantile was $N = 31.5$, the median was $N = 48$, the 75% quantile was $N = 67.5$, and the maximum sample size was $N = 251$. Average statistical power was also quite low (ranging from 23% to 45%; Table 5), and the results from TES suggest that, given these levels of statistical power, it is very unlikely that as many statis-

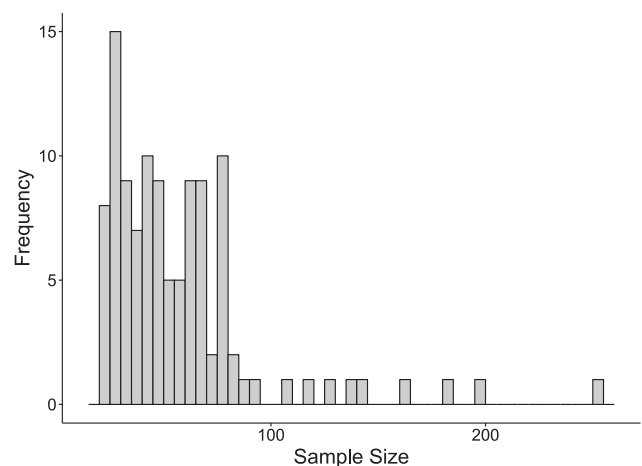


Figure 3. Histogram of sample sizes. Only independent effect sizes derived using a single manipulation task are shown (i.e., 111 of the 116 independent effect size estimates).

tically significant findings as were observed in this data set were generated without the influence of some form of bias (see Table 5).

To help rule out a limited-strength-model-consistent explanation for small-study effects (i.e., that the degree to which manipulation tasks deplete self-control is somehow related to the number of participants that are exposed to it), we modeled sample size as a function of manipulation task. For this analysis, we used the combined data set, but we also removed any effect size that had been produced from an experiment that used more than one manipulation task. Therefore, this model was based on 111 observations, rather than 116.

Importantly, sample size is a count variable (i.e., nonnegative integer), and therefore we did not expect it to follow a normal distribution (and it did not; Figure 3). Poisson regression models, which are the standard for such data (Gardner, Mulvey, & Shaw, 1995), assume that the variance and the mean of the outcome are equal, but in our case, the average sample size was 57.04 and the variance was 1,374.96. Not surprisingly then, a test for overdispersion (Zeileis, Kleiber, & Jackman, 2008) rejected the null hypothesis that the mean and the variance were equal: estimated dispersion parameter = 20.42, $z = 3.51$, $p < .001$. As a result, standard Poisson regression was inappropriate for these data. Instead, we used a quasi-Poisson regression model, which we preferred to a negative binomial model because it does not require any assumptions about the underlying probability distribution (Gardner et al., 1995).

In this model, every term was statistically nonsignificant except for the intercept, $b = 3.76$, $p < .001$ (which indicates, unsurprisingly, that the average sample size was nonzero) and the coefficient for the attention video manipulation task, $b = 0.51$, $p = .04$, which suggests that experiments that use the attention video manipulation also tend to have larger-than-average sample sizes (Table 8). Critically, the significant result for the attention video manipulation seemed to be due to a single extreme observation, YostM, 2009, Experiment 1, which included the largest sample size ($N = 251$) across all of our data sets (to visualize the extremeness of this observation, compare it to the median of the attention video group, and compare this median to those of the other groups in Figure 4). With this one data point removed, the coefficient for the attention video task decreased by about 30%

Table 8

Quasi-Poisson Regression Model Predicting Sample Size as a Function of Manipulation Task

| | <i>b</i> | <i>SE</i> | <i>p</i> |
|----------------------|----------|-----------|----------|
| Intercept | 3.76 | 0.21 | <.001 |
| Attention video | 0.51 | 0.25 | .04 |
| Crossing out letters | 0.36 | 0.25 | .16 |
| Emotion video | 0.32 | 0.26 | .22 |
| Food temptation | 0.30 | 0.32 | .35 |
| Math | -0.39 | 0.91 | .67 |
| Stroop | 0.43 | 0.30 | .16 |
| Social exclusion | -0.08 | 0.43 | .86 |
| Transcription | 0.08 | 0.35 | .82 |
| Thought suppression | 0.21 | 0.27 | .44 |
| Working memory | -0.03 | 0.39 | .94 |

Note. Only independent effect sizes derived using a single manipulation task are shown (i.e., 111 of the 116 for the full sample). There is only one observation that uses the Math manipulation task.

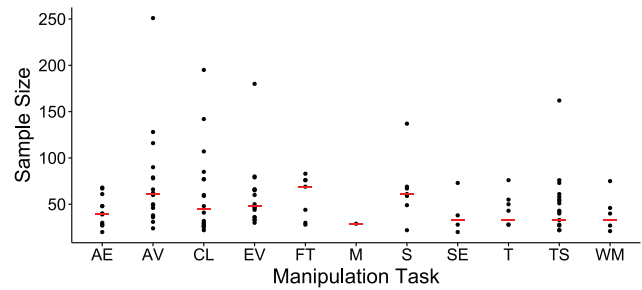


Figure 4. Sample size as a function of manipulation task type. Horizontal bars indicate median sample sizes for each category. AE = attention essay; AV = attention video; CL = crossing out letters; EV = emotion video; FT = food temptation; M = Math; S = Stroop; SE = social exclusion; T = transcription; TS = thought suppression; WM = working memory. Note that only one observation exists for the Math manipulation task. Only independent effect sizes derived using a single manipulation task are shown (i.e., 111 of the 116 for the full sample). See the online article for the color version of this figure.

to $b = 0.36$, $p = .12$. Moreover, removing this effect size from the analysis of the Stroop and combined data sets left the results of the analyses of those data sets essentially unchanged. Thus, given this context, as well as the fact that this analysis was exploratory, we believe that the best interpretation of the data is that sample sizes did not differ as a function of manipulation task, and that it would therefore be inaccurate to claim that the small-study effects we observed were due to more potent manipulations tending to be used in conjunction with smaller sample sizes.

Summary

In all but two samples (food consumption and possible anagrams), the estimates of the average overall effect from the random-effects meta-analysis models were statistically significant (see Table 1). However, in all samples the average overall effect was also qualified by moderate to substantial between-study heterogeneity (see Table 3), suggesting that the experiments in these samples were not all measuring the same effect, so that any single summary estimate was in fact an average across measures of multiple effects. Although our mixed-effects meta-analysis models were mostly unsuccessful at explaining this heterogeneity, there did appear to be some evidence that publication status explained some degree of heterogeneity (i.e., published experiments tended to have produced higher effect size estimates on average than unpublished ones).

In five of the samples (hand grip, impossible anagrams, impossible puzzles, Stroop, and working memory), the average overall effect appeared to be dependent on the presence of small-study effects (see Table 7). In four of these five samples (hand grip, impossible anagrams, Stroop, and working memory), the small-study effects in question were plausibly due to publication bias (given, for example, the placement of published effect sizes as compared to unpublished effect sizes on the contour-enhanced funnel plots; Figure 1), whereas the cause was less obvious in the impossible puzzles data set. Regardless, in each of the samples where evidence for an influence of small-study effects was found, controlling for this influence through the application of PET and

PEESE reduced the estimate of the overall average effect to nonsignificance.

Upon removing the single most extreme observation from the standardized tests data set, the estimates of the depletion effect from PET and PEESE were nonsignificant *but* larger than the random-effects estimate, and the coefficient for FAT was negative. As such, we believe that the coefficients for the WLS models for this data set, in conjunction with inspection of the funnel plot in Figure 1, imply that small-study effects are an unlikely explanation for the positive and statistically significant random-effects estimate for the standardized tests data set. Therefore, we find that the existing evidence provides support for the claim that previous acts of self-control do impair subsequent performance on standardized tests.

The application of TES and the trim and fill methods provided some evidence for the presence of bias. Results from TES were primarily ambiguous because of the large degree of between-study heterogeneity in the samples and the resulting wide confidence intervals around the estimate of the average overall effect from the random-effects meta-analysis models; however, the sample sizes in the experiments meta-analyzed were typically small (see Figure 3), and the resulting average statistical power for estimates from the combined data set was chronically low (see Table 5), mirroring what we found previously in the Hagger et al. (2010) dataset. Application of the trim and fill method reduced the estimate of the overall effect in four of our eight data sets, and in the case of the Stroop data set, the reduction left the overall effect statistically nonsignificant.

Results from the combined data set were consistent with our primary analyses: The depletion effect was statistically significant and positive when estimated using the random-effects meta-analysis model, the trim and fill estimate was reduced, and estimates from PET and PEESE were indistinguishable from zero. TES and FAT clearly suggested the presence of bias, but the contour-enhanced funnel plot (see Figure 2) does not necessarily suggest a pure case of publication bias (although the mixed-effects model indicated that published experiments tended to produce

larger effect size estimates than unpublished ones). Interestingly, there seemed to be a relationship between effect size and sample size for unpublished effect sizes only (see Figure 2), raising the possibility that small-study effects, such as the application of undisclosed researcher degrees of freedom, might be influence the reporting of some unpublished results, such as those included in dissertations or theses. Regardless, the overall pattern of results from the combined data set is in general agreement with that from the primary analysis: The depletion effect is not robust to context, and estimates that account for small-study effects, regardless of what those effects may be, suggest that the depletion effect is, on average, indistinguishable from zero.

Discussion

Assuming that the frequently used manipulation and outcome tasks that we identified are valid operationalizations of self-control, full support of the limited strength model requires that the depletion effect be distinguishable from zero in all eight of our data sets, as well as our combined data set (Baumeister et al., 1998). However, in only the standardized tests data set did we find convincing evidence that the depletion effect was different from zero; otherwise, random effects meta-analytic estimates were either not statistically significant (i.e., food consumption: $g = 0.44$, 95% CI = [-0.01, 0.89]; possible anagrams: $g = 0.24$, 95% CI = [-0.07, 0.56]) or were qualified by the apparent presence of small-study effects to the extent that correcting for such effects resulted in the overall effect not being distinguishable from zero (see Table 1; see Table 9 for a summary of how our results do or do not support the existence of the depletion effect as proposed by the limited strength model).

Hagger et al. (2010) described the results of their meta-analysis as "... demonstrating that the ego-depletion effect exists, its associated confidence intervals do not include trivial values, and it is generalizable across spheres of self-control" (p. 515). Our results contradict each of those claims, and it appears that self-control functions as predicted by limited strength model only when

Table 9

Interpreting Our Results in Terms of Evidence for the Depletion Effect as Laid out in the Limited Strength Model of Self-Control

| Key questions | Sample | | | | | | | | | Basis for the conclusion | |
|---|--------|----|----|----|----|----|---|----|---|--------------------------|-------|
| | FC | HG | IA | IP | PA | ST | S | WM | C | Test | Table |
| Q1: Is the average overall depletion effect statistically significant? | N | Y | Y | Y | N | Y | Y | Y | Y | RE model | 3 |
| Q2: If the depletion effect is moderated by an experiment-level characteristic, is the effect consistent with the limited strength model? | . | . | Y | . | . | . | N | . | N | ME model | 4 |
| Q3: After imputing experiments that are potentially missing due to publication bias, is the overall average depletion effect significant? | . | Y | Y | Y | N | . | N | . | Y | Trim & fill | 6 |
| Q4: Is the overall average depletion effect still significant after correcting for small-study effects? | N | N | N | N | N | . | N | N | N | PET & PEESE | 7 |
| Q5: Does the evidence support the existence of the depletion effect as proposed in the limited strength model (i.e., the answers to Q1 through Q4 are never "N")? | N | N | N | N | N | Y | N | N | N | | |

Note. FC = food consumption sample; HG = handgrip sample; IA = impossible anagram sample; IP = impossible puzzles sample; PA = possible anagram sample; ST = standardized tests sample; S = Stroop sample; WM = working memory sample; C = The combined data set. "Y" = yes; "N" = no; "." = not applicable. Random-effects model = the RE model; Mixed-effects model = the ME model.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the outcome task is performance on standardized tests. Notably, without applying any corrections for small-study effects, the estimate of the depletion effect derived from the standardized tests data set was less than half the size of the overall estimate provided by Hagger et al. (2010), and the lower limit of the 95% CI was nearly zero ($g = 0.05$).

Publication bias seems to explain the small-study effects in four of our data sets: The funnel plots (see Figure 1) for hand grip, impossible anagrams, Stroop, and working memory show the publication-bias-consistent pattern in which published results possess effect sizes that tend to exceed the threshold for statistical significance compared to effect sizes for unpublished results. This pattern is less apparent in the data sets for impossible puzzles, suggesting that publication bias does not fully explain the small-study effects for in this data set. Nevertheless, the fact that controlling for small-study effects in these data sets reduced the depletion effect to nonsignificance should inform our confidence in the depletion effect: To maintain belief in the depletion effect in the face of these results, one would have to argue that smaller experiments were somehow more effective at depleting self-control. Because experiments in each of the eight data sets used the same outcome tasks and a limited set of manipulation tasks, such a state of affairs seems unlikely. Moreover, we found that sample size was likely unrelated to the type of manipulation task used (see Table 8), placing in doubt the possibility that smaller experiments involved more effective manipulation tasks.

We favor an interpretation of our findings that depends on the validity of the WLS meta-regression estimators PET and PEESE, but because such methods are relatively infrequently used in psychology (but not, for example, economics, Costa-Font et al., 2011; Doucouliagos & Stanley, 2009; Havranek, 2010; or medicine, Hemingway et al., 2010; Moreno et al., 2009a, 2009b; Nüesch et al., 2010), it seems likely that some readers will not find those results completely convincing. Even ignoring the regression-based estimates, however, our findings still present critical problems for the limited strength model of self-control. First, the food consumption and possible anagrams data sets did not produce statistically significant overall effects as estimated by standard random-effects meta-analysis models, and the estimate of the overall effect for the Stroop data set was reduced to nonsignificance by the commonly used trim and fill method.

Based on the limited strength model, one would clearly predict that each of our samples would have shown significant overall effects, and the fact that three do not, suggests either that (a) only some behaviors can be “depleted” or that (b) these tasks do not in fact measure self-control. Accepting the former interpretation would necessitate a complete revision of the limited strength model in that it would mean that only certain behaviors, rather than any act of self-control, show the depletion effect. Accepting the latter interpretation has two implications: First, that any instance of the sequential task paradigm in which such a behavior was used as a manipulation would not inform the evidence for (or against) the limited strength model. For example, several experiments in our data sets used the Stroop task to manipulate subsequent performance on other commonly used tasks, such as impossible puzzles, but if the Stroop task does not require self-control, then these other experiments cannot be said to have tested the depletion effect. Second, any extension work that takes the validity of such tasks as an initial assumption are meaningless in terms of determining

whether self-control was necessary for the behavior in question. For example, DeWall, Baumeister, Stillman, and Gailliot (2007) tested whether aggression in the face of provocation could be manipulated via the Stroop task, and Gailliot et al. (2007) reported that the depletion effect, as measured by Stroop performance, was obviated by having participants consume a drink sweetened with sugar rather than an artificial sweetener. If the Stroop task does not require self-control, then the work by DeWall et al. (2007) and Gailliot et al. (2007) tells us nothing about the proposals that self-control is related to aggression or that blood glucose levels are related to the depletion effect.

Using standard mixed-effects meta-analysis techniques, we also found support for the notion that self-control actually improves if more than a single manipulation task is completed: This effect was nearly statistically significant for the impossible anagrams and possible anagrams data sets, and statistically significant for the combined data set. These findings are entirely inconsistent with the limited strength model, and have been better accounted for by theories such as learned industriousness (see, for example, discussions by Carter & McCullough, 2013b and Converse & DeShon, 2009). It is worth noting that at least one additional experiment, which was not included here due to its too-recent completion date, has been published in support of the pattern that completing more manipulation tasks (i.e., exercising greater self-control) results in improvement in subsequent self-control performance (Xiao, Dang, Mao, & Liljedahl, 2014).

It is also worth noting that, by applying the widely used trim and fill method, we found that the majority of our data sets do not allow one to rule out effect sizes of a magnitude that one might consider trivial. Specifically, in six of our eight data sets, and in our combined data set, the lower limit of the 95% confidence interval surrounding the estimate from the trim and fill method was less than $g = 0.15$ (see Table 6), an effect size magnitude that is small enough to warrant serious reconsideration of whether the depletion effect can even be productively studied in the laboratory because achieving 80% power to observe an effect of that size in a two-group design would require 699 subjects *per group*. Therefore, whether or not one chooses to be confident in more recent estimators like PET and PEESE, there are ample reasons that our findings present serious problems for both the generality and usefulness of the limited strength model.

Hagger and Chatzisarantis (2014) previously expressed skepticism that publication bias could be operating strongly enough to suppress the number of results necessary to inflate a null average effect into an effect of medium magnitude—such as the estimate of $d = 0.62$ for the overall average depletion effect reported by Hagger et al. (2010). Importantly, simply censoring results from meta-analytic samples is not the only way in which a meta-analytic effects can be inflated. First, as we have argued previously (Carter & McCullough, 2014), it is possible that the undisclosed use of researcher degrees of freedom turned unfavorable results into favorable ones, thereby skewing the estimate of the overall average depletion effect without requiring that any results be lost to “the file drawer.” Second, the fact that the present study generally found estimates of the depletion effect that were smaller than the estimate reported by Hagger et al. (2010) raises an additional point against the view espoused by Hagger and Chatzisarantis (2014): Compare, for example, Hagger et al.’s (2010) estimate of $d = 0.62$ with our estimates of the depletion effect for the average published data set (given either as the intercept or the sum of

the intercept and the coefficient for publication status; see Table 4) from the possible anagrams ($b = 0.27$), standardized tests ($b = 0.10$), and working memory ($b = 0.22$) data sets. This pattern suggests the possibility that the original estimate of $d = 0.62$ may have been inflated by influences other than publication bias—namely, the inclusion of estimates of effects that here we have argued are not appropriate tests of the depletion effect (e.g., estimates of effects derived from extension experiments, estimates of effects that can be claimed as supporting the depletion effect regardless of the direction of the effect). Therefore, even if Hagger and Chatzisarantis (2014) were correct about the influence of publication bias in the literature on the limited strength model, it is still plausible that Hagger et al.'s (2010) estimate of the depletion effect represented an overestimate of an effect that was small or zero on average.

Readers may wonder whether Hagger and Chatzisarantis's (2014) intuition that an extensive number of unresponsive or contrary findings likely do not exist is validated by the number of unpublished experiments we were able to collect and include here. The collection and inclusion of unpublished data, although critical, is an imperfect process (e.g., many "failed" experiments are never written up, and the lack of any record often makes retrieval of previously collected data extremely difficult and time-consuming for researchers; Franco et al., 2014). Thus, retrieving unpublished data cannot entirely replace the statistical approaches we employed (e.g., PET and PESE), and it is extremely unlikely that the unpublished data we collected is the entire body of unpublished work on the depletion effect.

Arguably, a set of preregistered replication efforts would help to settle the issue of whether the depletion effect is real. Recently, Alex Holcombe and Martin Hagger have proposed just such a registered replication effort,² and although we believe this project to be worthwhile, any attempt to replicate the depletion effect should consider two points: First, our results indicate that replicating the depletion effect would require large samples. We believe that the only convincing evidence for the depletion effect in our data sets was found for the standardized tests data set (standardized tests, $g = 0.30$); but, if one is unconvinced by the WLS estimators, and would instead prefer to look to more standard meta-analytic techniques, then likely the best estimate of the overall depletion effect is given by the trim and fill method as applied to our combined data set ($g = 0.24$). In either case, many subjects per group (e.g., 273 per group for the estimate based on the combined data sets) in a between-subjects design are required to achieve 80% power to detect the depletion effect. Our trim-and-fill-derived estimate is in stark contrast to Hagger et al.'s (2010) estimate of $d = 0.62$ —the estimate on which the current replication project bases its suggested sample size of at least 84 per group. Second, the limited strength model holds that any act of self-control should result in decreases in performance on any task that also requires self-control. As a result, for the limited strength model to be supported, the depletion effect would need to be successfully replicated using multiple combinations of manipulation and outcome tasks (presumably those self-control tasks that are thought to be the most valid). If, instead, replication efforts focus only on a single combination of manipulation and outcome tasks, as is the case for the currently proposed replication, results from such efforts can only answer the question of whether the depletion effect exists *when measured with those tasks*. This finding would be necessary to support the limited strength model, but not sufficient. Therefore, for large-scale replication efforts to support the notion that self-control functions as if it relies on a limited resource—rather than simply the idea that a given

pair of manipulation and outcome tasks show the depletion effect—a suite of experiments making use of a variety of tasks will be required.

There are two important points related to this issue of generalizability. First, the claims we make about the depletion effect only apply to the depletion effect *as it is typically measured in the laboratory*—that is, as it is measured by experiments for which our samples can be considered representative (e.g., instances of the sequential task paradigm involving only the most frequently used manipulation and outcome tasks). Our conclusions should not be taken to necessarily apply to every instance of the sequential task paradigm; however, given that our results are based on the types of experiments that we argue to be the core of research on the depletion effect, we believe our findings are sufficient to raise serious concerns about the entire body of evidence thought to support the limited strength model. Second, our conclusions and those of Hagger et al. (2010), as well as those of the authors of the work that has been meta-analyzed here or previously, define the depletion effect in terms of performance on the sequential task paradigm. It may be that self-control does truly "deplete" when it is measured in different ways (e.g., in terms of performance on a single task as a function of time-on-task, as in the literature on cognitive fatigue; Ackerman, 2011). Examining such alternative operationalizations of self-control failure that might plausibly be linked to resource depletion—or at least a process that resembles resource depletion—would be very useful for determining the nature of self-control failure; however, to avoid the same problems we have identified in the literature on the sequential task paradigm (e.g., low statistical power, publication bias), researchers would do well to collect the largest samples possible, preregister their experiments, and make their data, regardless of results, easily accessible.

Conclusion

We designed our tests to provide a critical examination of the depletion effect, one for which even the most skeptical reader would have needed to revise his or her beliefs had the findings supported the limited strength model. However, our results were inconsistent with the predictions of the limited strength model (see Table 9). For example, it seems that previous acts of self-control reduce performance on subsequent standardized tests, but the lack of support for the notion that this effect also applies to more classic self-control tasks, such as the Stroop task, strongly suggests that self-control in general does not decrease as a function of previous use. Given the overall picture provided by our analyses, we conclude that the meta-analytic evidence does not support the proposition (and popular belief) that self-control functions as if it relies on a limited resource, at least when measured as it typically is in the laboratory. We encourage scientists and nonscientists alike to seriously consider other theories of when and why self-control might fail.

² <https://osf.io/jymhe/>

References

- Ackerman, P. L. (2011). *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications*. Washington, DC: American Psychological Association. <http://dx.doi.org/10.1037/12343-000>
- Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Person-*

- ality and Social Psychology, 74, 1252–1265. <http://dx.doi.org/10.1037/0022-3514.74.5.1252>
- Carter, E. C., & McCullough, M. E. (2013a). Is ego depletion too incredible? Evidence for the overestimation of the depletion effect. *Behavioral and Brain Sciences*, 36, 683–684. <http://dx.doi.org/10.1017/S0140525X13000952>
- Carter, E. C., & McCullough, M. E. (2013b). After a pair of self-control-intensive tasks, sucrose swishing improves subsequent working memory performance. *BMC psychology*, 1, 22.
- 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. <http://dx.doi.org/10.3389/fpsyg.2014.00823>
- Christiansen, P., Cole, J. C., & Field, M. (2012). Ego depletion increases ad-lib alcohol consumption: Investigating cognitive mediators and moderators. *Experimental and Clinical Psychopharmacology*, 20, 118–128. <http://dx.doi.org/10.1037/a0026623>
- Converse, P. D., & Deshon, R. P. (2009). A tale of two tasks: Reversing the self-regulatory resource depletion effect. *Journal of Applied Psychology*, 94, 1318–1324. <http://dx.doi.org/10.1037/a0014604>
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.
- Costa-Font, J., Gammill, M., & Rubert, G. (2011). Biases in the healthcare luxury good hypothesis: A meta-regression analysis. *Journal of the Royal Statistical Society, A*, 174, 95–107.
- DeWall, C. N., Baumeister, R. F., Gailliot, M. T., & Maner, J. K. (2008). Depletion makes the heart grow less helpful: Helping as a function of self-regulatory energy and genetic relatedness. *Personality and Social Psychology Bulletin*, 34, 1653–1662. <http://dx.doi.org/10.1177/0146167208323981>
- DeWall, C. N., Baumeister, R. F., Stillman, T. F., & Gailliot, M. T. (2007). Violence restrained: Effects of self-regulation and its depletion on aggression. *Journal of Experimental Social Psychology*, 43, 62–76. <http://dx.doi.org/10.1016/j.jesp.2005.12.005>
- Doucouliaqos, C. H., & Stanley, T. D. (2009). Publication selection bias in minimum wage research? A meta-regression analysis. *British Journal of Industrial Relations*, 47, 406–429. <http://dx.doi.org/10.1111/j.1467-8543.2009.00723.x>
- Duval, S., & Tweedie, R. (2000). A Nonparametric “Trim and Fill” Method of Accounting for Publication Bias in Meta-Analysis. *Journal of the American Statistical Association*, 95, 89–98.
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634. <http://dx.doi.org/10.1136/bmj.315.7109.629>
- Englert, C., & Bertrams, A. (2012). Anxiety, ego depletion, and sports performance. *Journal of Sport & Exercise Psychology*, 34, 580–599.
- Freeman, N., & Muraven, M. (2010). Self-Control Depletion Leads to Increased Risk Taking. *Social Psychological & Personality Science*, 1, 175–181.
- Gailliot, M. T., & Baumeister, R. F. (2007). Self-regulation and sexual restraint: Dispositionally and temporarily poor self-regulatory abilities contribute to failures at restraining sexual behavior. *Personality and Social Psychology Bulletin*, 33, 173–186. <http://dx.doi.org/10.1177/0146167206293472>
- 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, 325–336. <http://dx.doi.org/10.1037/0022-3514.92.2.325>
- Gailliot, M. T., Schmeichel, B. J., & Baumeister, R. F. (2006). Self-regulatory processes defend against the threat of death: Effects of self-control depletion and trait self-control on thoughts and fears of dying. *Journal of Personality and Social Psychology*, 91, 49–62. <http://dx.doi.org/10.1037/0022-3514.91.1.49>
- Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118, 392–404. <http://dx.doi.org/10.1037/0033-2909.118.3.392>
- Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York, NY: Russell Sage Foundation.
- 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. <http://dx.doi.org/10.3389/fpsyg.2014.00298>
- 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, 495–525. <http://dx.doi.org/10.1037/a0019486>
- Havraneck, T. (2010). Rose effect and the Euro: Is the magic gone? *Review of World Economics*, 146, 241–261. <http://dx.doi.org/10.1007/s10290-010-0050-1>
- Hemingway, H., Philipson, P., Chen, R., Fitzpatrick, N. K., Damant, J., Shipley, M., . . . Hingorani, A. D. (2010). Evaluating the quality of research into a single prognostic biomarker: A systematic review and meta-analysis of 83 studies of C-reactive protein in stable coronary artery disease. *PLoS Medicine*, 7, e1000286. <http://dx.doi.org/10.1371/journal.pmed.1000286>
- Higgins, J. P. (Ed.). (2008). *Cochrane handbook for systematic reviews of interventions* (Vol. 5). Chichester, UK: Wiley-Blackwell.
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 1255484.
- IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14, 25. <http://dx.doi.org/10.1186/1471-2288-14-25>
- Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4, 245–253. <http://dx.doi.org/10.1177/1740774507079441>
- Janssen, L., Fennis, B. M., Pruyn, A. T. H., & Vohs, K. D. (2008). The path of least resistance: Regulatory resource depletion and the effectiveness of social influence techniques. *Journal of Business Research*, 61, 1041–1045. <http://dx.doi.org/10.1016/j.jbusres.2007.09.013>
- Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009a). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, 9, 2. <http://dx.doi.org/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. (2009b). Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *British Medical Journal*, 339, b2981.
- Muraven, M. (2008). Prejudice as self-control failure. *Journal of Applied Social Psychology*, 38, 314–333. <http://dx.doi.org/10.1111/j.1559-1816.2007.00307.x>
- Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W., Tschannen, B., Altman, D. G., . . . Jüni, P. (2010). Small study effects in meta-analyses of osteoarthritis trials: Meta-epidemiological study. *British Medical Journal*, 341.
- R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rücker, G., Schwarzer, G., Carpenter, J. R., Binder, H., & Schumacher, M. (2011). Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics*, 12, 122–142. <http://dx.doi.org/10.1093/biostatistics/kxq046>

- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566. <http://dx.doi.org/10.1037/a0029487>
- Simonsohn, U. (2012). It does not follow evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspectives on Psychological Science, 7*, 597–599. <http://dx.doi.org/10.1177/1745691612463399>
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*, 60–78. <http://dx.doi.org/10.1002/jrsm.1095>
- Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., . . . Higgins, J. P. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343*, d4002.
- Stucke, T. S., & Baumeister, R. F. (2006). Ego depletion and aggressive behavior: Is the inhibition of aggression a limited resource? *European Journal of Social Psychology, 36*, 1–13. <http://dx.doi.org/10.1002/ejsp.285>
- Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*, 2113–2126. <http://dx.doi.org/10.1002/sim.1461>
- Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*, 1559–1573. <http://dx.doi.org/10.1002/sim.1187>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.
- Xiao, S., Dang, J., Mao, L., & Liljedahl, S. (2014). When more depletion offsets the ego depletion effect. *Social Psychology, 45*, 421–425. <http://dx.doi.org/10.1027/1864-9335/a000197>
- Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software, 27*, 1–27.

Received February 22, 2015

Revision received April 23, 2015

Accepted April 24, 2015 ■

Correction to Kreuzbauer, King, and Basu (2015)

In the article “The Mind in the Object—Psychological Valuation of Materialized Human Expression” by Robert Kreuzbauer, Dan King, and Shankha Basu (*Journal of Experimental Psychology: General*, Advance online publication. June 15, 2015. <http://dx.doi.org/10.1037/xge0000080>), the labels on the X-axis of Figure 1 “Remove Variance” and “Preserve Variance” should be switched.

<http://dx.doi.org/10.1037/a0039581>